Missing data in randomised controlled trials — a practical guide

James R. Carpenter

&

Michael G. Kenward

November 21, 2007

Competing interests: None

Word count (main text): 54,158

Funder: NHS (NCCRM) Grant no. RM04/JH17/MK

EXECUTIVE SUMMARY

Randomised Controlled Trials (RCTs) are well established as the preferred method for evaluating interventions. Unlike studies based on observational data, the randomisation of patients to interventions means that a direct causal link can be made between an intervention and its effect.

In order to measure the effect of a treatment we need, at least, a measure of each patient's response at the end of the trial. More commonly, a series of responses will be measured at baseline and throughout follow-up. Inevitably, it is not possible to collect all the intended data on each individual. Unfortunately, as one might expect, simply analysing the data that were collected without any further reflection generally leads to misleading conclusions. Specifically, when the data are incomplete the causal link between intervention and response is broken.

We refer to data that we intended to collect, but for one reason or another were unable to, as *missing data*. Anyone with practical experience of trials knows that missing data are ubiquitous. Nevertheless, a recent survey of 383 parallel group trials showed that 69% did not report how attrition was handled.

This monograph reviews the issues of raised by missing data in clinical trials, and describes and illustrates a principled approach to analyses in such settings. It is divided into three parts.

Part I gives a non-technical overview of the issues raised by missing data. We propose a systematic approach to handling missing data in clinical trials, and discuss the implications of this for design, 'intention to treat' and 'per-protocol' analyses. This leads to a critique of the current Committee for Proprietary Medicinal Products guidelines for missing data, together with many of the ad-hoc statistical methods often used by statisticians for the analysis of trials with missing data. We argue that analyses should be principled, that is, follow well-defined and accepted statistical arguments, using models and assumptions that are transparent, and hence open to criticism and debate.

When data are missing any attempt to draw conclusions from a statistical analysis rests on untestable assumptions concerning the relationship between the unobserved data and the reasons for them being missing (the missing value mechanism). In this way, missing data introduce ambiguity into the analysis beyond conventional sampling imprecision and the assumptions behind any such analyses form a crucial part of the argument behind any conclusions drawn. We argue that primary analyses should rest on a central assumption about this relationship, the so-called missing at random assumption. Broadly, this is the most general assumption that allows valid analyses to be made independently of the missing value mechanism.

Part II shows how primary analyses in a range of settings can be carried out under the so-called missing at random assumption. This assumption has a central role in underpinning the most important classes of primary analysis, such as those based on likelihood. However, as its validity cannot be assessed from the data under analysis, Part III outlines practical methods for assessing the sensitivity of conclusions drawn from the analyses in part II to the missing at random assumption. We compare and contrast the two main approaches to this in the literature, again giving examples and code.

In summary:

- From the design stage onwards, our principled approach to handling missing data should be adopted, and
- This monograph outlines how this principled approach can be practically, and directly, applied to the majority of trials with longitudinal follow-up.

ABSTRACT

Missing data in clinical trials — a practical guide

James R. Carpenter* and Michael G. Kenward

Medical Statistics Unit, London School of Hygiene & Tropical Medicine, UK

*Corresponding author

Objective: Missing data are ubiquitous in clinical trials, yet recent research suggests many statisticians and investigators appear uncertain how to handle them. The objective of this monograph is to set out a principled approach for handling missing data in clinical trials, and provide examples and code to facilitate its adoption.

Data sources: An asthma trial from GlaxoSmithKline, a asthma trial from AstraZeneca, and a dental pain trial from GlaxoSmithKline.

Methods: Part I gives a non-technical review of how missing data are typically handled in the analysis of clinical trials, and outlines the issues raised by missing data. When faced with missing data, we show no analysis can avoid making additional untestable assumptions. This leads to a proposal for a principled, systematic approach for handling missing data in clinical trials, which in turn informs a critique of current Committee of Proprietary Medicinal Products guidelines for missing data, together with many of the ad-hoc statistical methods currently employed.

Part II shows how primary analyses in a range of settings can be carried out under the so-called missing at random assumption. This key assumption has a central role in underpinning the most important classes of primary analysis, such as those based on likelihood. However its validity cannot be assessed from the data under analysis, so in Part III two main approaches are developed and illustrated, for the assessment of the sensitivity of the primary analyses to this assumption.

Examples: Throughout, examples are used to illustrate the arguments and analyses. Code for the analyses (mostly in SAS) is given in Appendix C. The end of each example is indicated with a '□'.

Results: The literature review revealed missing data are often ignored, or poorly handled in the analysis. Current guidelines, and frequently used ad-hoc statistical methods, are shown to be flawed. A principled, yet practical, alternative approach is developed, which examples show leads to inferences with greater validity. SAS code is given to facilitate its direct application.

Conclusions: From the design stage onwards, a principled approach to handling missing data should be adopted. Such an approach follows well-defined and accepted statistical arguments, using models and assumptions that are transparent, and hence open to criticism and debate. This monograph outlines how this principled approach can be practically, and directly, applied to the majority of trials with longitudinal follow-up.

ACKNOWLEDGEMENTS

We would like to thank GlaxoSmithKline and AstraZeneca for permission to use their data.

We gratefully acknowledge funding from the NHS (CCRM).

James Carpenter would like to thank colleagues at the Department for Medical Biometry and Statistics, University Hospital, Freiburg, for their hospitality and encouragement during his sabbatical visit, where the first draft of this monograph was completed.

We are grateful for stimulating conversations with a many colleagues. In particular, Prof. James H. Roger, who also contributed a SAS macro, and Prof. Geert Molenberghs. We thank Stephen Evans and three anonymous referees for their helpful comments on the first draft, which have led to a greatly improved manuscript. The remaining errors are ours.

We invite comments and corrections. Please email these to james.carpenter@lshtm.ac.uk. This book has a homepage at www.missingdata.org.uk. In due course we intend to publish here a supplementary chapter on time to event outcomes.

James Carpenter & Mike Kenward
London School of Hygiene & Tropical Medicine, Spring 2007

CONTRIBUTIONS OF THE AUTHORS

The monograph was planned jointly.

James Carpenter did the majority of the writing and computation.

Mike Kenward reviewed draft Chapters and wrote some additional material.

PRINCIPAL ABBREVIATIONS

COPD	Chronic Obstructive Pulmonary Disease
CPMP	Committee for Proprietary Medicinal Products
GEE	Generalised Estimating Equation
GLM	Generalised Linear Model
GLMM	Generalised Linear Mixed Model
ICH	International Council on Harmonisation
LOCF	Last Observation Carried Forward
MAR	Missing At Random
MCAR	Missing Completely At Random
MI	Multiple Imputation
MNAR	Missing Not At Random
PM	Pattern Mixture
PA	Population Averaged
SS	Subject Specific

Table of Contents

List of Tables

List of Figures

Part I

Chapter 1

Missing data: principles

1.1 Introduction

Randomised clinical trials are now well established as the key tool for establishing the efficacy of new medical interventions. Their widespread use, both as part of the formal drug licensing process and more generally, has thrown up many statistical issues relating to their scope, design, analysis and reporting. In 1998, these resulted in the International Council on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) issuing a guideline on statistical methodology, known as ICH E9 (ICH E9 Expert Working Group (1999); see also www.ich.org).

Through this guideline, the ICH seeks to achieve greater international harmonisation of the statistical aspects of clinical trials in order to (i) make more economical use of the resources (human, animal and material) involved, and (ii) to reduce unnecessary delay in the development and availability of products internationally. At the same time safeguards and standards to protect public health should be maintained. The guideline quickly became a key document for medical statisticians, setting the statistical standards for clinical trials (Lewis, 1999). By setting down principles, not procedures, it has ensured its continuing influence on statistical methods for clinical trials.

The discussion on data analysis contains the following section (5.3) on missing data:

> Missing values represent a potential source of bias in a clinical trial. Hence, every effort should be undertaken to fulfil all the requirements of the protocol concerning the collection and management of data. In reality, however, there will almost always be some missing data. A trial may be regarded as valid, none the less, provided the methods of dealing with missing values are sensible, and particularly if those methods are predefined in the protocol. Definition of methods may be refined by updating this aspect in the statistical analysis plan during the blind review. Unfortunately, no universally applicable methods of handling missing values can be recommended. An investigation should be made concerning the sensitivity of the results of analysis to the method of handling missing values, especially if the number of missing values is substantial.

This guideline recognises that there will almost always be some missing data. Further, the CONSORT guidelines for reporting clinical trials (Moher *et al.*, 2001) state that the number of patients with missing data should be reported by treatment arm. Nevertheless, Chan and Altman (2005) estimate that 65% of studies in PubMed journals do not report the handling of missing data. Further, a recent review by Wood *et al.* (2004) identified serious weaknesses in both description of missing data and in the statistical methodology adopted. Together, these findings suggest that trialists are unsure how to handle missing data, in both analysis and reporting, and therefore reluctant to discuss it.

This is perhaps surprising, as there is now a large literature on missing data, not only in clinical trials but more widely (for example, see the bibliography at www.missingdata.org.uk). It suggests a gap remains between clinicians and statisticians focused on running and analysing trials, and those involved in more developmental research. The aim of this monograph is to address this gap. We write with busy trials statisticians in mind, but also intend the early parts, which deal with principles, will be accessible to a much wider range of those involved in trials, from clinicians to data managers. This is because issues raised by missing data are wider than the technical details of statistical analyses.

The aim of this Chapter is to give an accessible outline of the key principles that should be kept in mind when considering missing data in trials. We seek to flesh out the early part of the ICH E9 guideline, to equip readers to constructively discuss the issues raised by missing data. We hope our examples will enable readers to relate the concepts to their own work.

We begin by discussing what we mean by 'missing data' and the issues it is likely to raise. We then describe what we mean by sensible analysis with missing data, and discuss how this relates to the ICH E9 statement that, despite missing data, 'a trial may be regarded as valid'. As part of this, we discuss statistical jargon which unfortunately often masks ideas that bear directly on this.

We conclude by discussing the implications of missing data on the trial design, and on two common analyses. In ICH E9 terms these are (i) the 'full data analysis' including every patient who is randomised, to estimate the effect of intending to give patients a particular intervention, and (ii) the 'per-protocol' analysis, which includes only that group of patients who comply[1] with the intervention, to estimate its actual effect.

1.2 What do we mean by missing data

In a trial context, missing data are data we intended to collect, but for one reason or another did not. By this, we do not mean 'counter-factual' data, e.g. the response a patient *might* have given *if* they were randomised to the active drug instead of the placebo.

Consider a typical clinical trial, where patients are observed at the start (baseline), randomly allocated their treatment, and then followed up at a number of visits. Data could then be missing at baseline, and at one or more of the follow-up visits.

At baseline or follow-up visits, specific readings could be missing, perhaps because a machine has broken down. Or it may be that all the data from a particular baseline or follow-up visit are missing because a patient was not present. The patient may return for future follow-up visits, or may withdraw.

We distinguish three broad occurrences:

1. If a patient missed a follow-up visit but attended at least one subsequent follow-up visit, we refer to the resulting missing data as *interim missing data*.

[1] Strictly, who comply with the intervention to the minimum extent required by the protocol.

2. If a patient is no longer seen after a certain follow-up visit, we say the data is missing due to *withdrawal*.

   (a) In some trials, when a patient stops complying with the intervention they will be withdrawn and subsequent follow-up data will be missing.

   (b) In others, follow-up will continue, at least for an initial period after compliance stops.

When we talk about 'missing data', the ideas relate to all these settings. However, the details of the statistical analyses vary. To keep the language simple below, *we use the term* withdrawal *to refer to situation (2a) above*, and give further details when we are considering situations such as (2b). In the literature, the terms *dropout*, *attrition* and *loss to follow-up* are also used, typically fairly synonymously with withdrawal.

We begin by focusing on withdrawal, which in our experience is the source of most missing data and most directly affects the interpretation of the trial results.

1.3 Trial validity and sensible analyses

The ICH E9 guideline says that despite missing data, a trial may still be *valid*, provided the statistical methods used are *sensible*. On reflection, it becomes apparent that the terms 'valid' and 'sensible' should mean the same whether or not there are any missing data[2]. We now consider this further.

We refer to a group of patients with a particular disease as a *patient population*. Suppose this population may benefit from a new intervention. A trial is set up, and patients recruited from this population are randomly allocated to either the new or the best existing intervention.

Suppose the trial found that average survival was one year longer with the new intervention. Broadly speaking, if we can infer that using the new intervention in the patient population will improve survival by 1 year, we say the trial is *valid*. For this to be the case, as described in the ICH guidelines, a substantial number of conditions need to be satisfied. From our viewpoint, the statistical analyses must be appropriate, so that

1. any variation between the intervention effect estimated from patients in the trial and that in the population is random. In other words it is not systematically biased in one direction;

2. as we include more and more patients from the population in our trial, the variation between the intervention effect estimated from patients in the trial and that in the population gets smaller and smaller. In other words, as the size of the trial increases, the estimated intervention effect homes in on the true value in the population. Such intervention estimates are called *consistent* in statistical jargon, and

[2]The term 'external validity' is often used in this context. For the results of a trial to have external validity, the analysis must be sensible. Additionally, however, external validity may require certain conditions on the representativeness of the patient population and those recruited into the trial.

3. our estimate of the extent of variability between the trial intervention effect and the true effect in the population (in statistical terms, the *standard error*) is accurate.

If all these conditions hold, we follow the ICH E9 guideline and call an analysis *sensible*[3].

We assume that the trial was validly designed and run. Then, if data are missing, drawing valid conclusions depends on sensible statistical analyses. Such analyses may well be different from complete data analyses and will usually require additional assumptions. They may be more difficult. Further, as data are missing, we are in effect missing information which we would otherwise use to estimate the effect of intervention. Thus, conclusions will be less precise. They can nevertheless be valid, in the sense described above.

1.4 How much should we bother about missing data?

This is almost the first question asked by trialists when faced with missing data. The sub-text is usually 'given these missing data, is the originally planned full-data analysis acceptable'? Although it would be nice to think a universal cut-and-dried answer could be given, the variety of trial designs, and occurrences of missing data, make this unrealistic. Consider the following examples.

EXAMPLE 1.1 *Trial with binary outcome*

Imagine a trial where patients with *known* outcomes respond as shown in Table 1.1. The odds ratio in favour of treatment B is 2.41 (95% CI 1.34–4.32), so the data support the hypothesis that B is preferable.

Treatment	A	B
Good outcome	50	70
Poor outcome	50	29

Table 1.1: Hypothetical trial: number of patients with good/poor outcomes in treatment groups A and B

Now suppose there are 2 further patients (one receiving treatment A and the other B) whose outcomes are missing. Whether these outcomes are 'Good' or 'Poor' will not change the conclusions.

Conversely, if there are 30 further patients with missing outcomes then, depending on both the treatment to which these 30 were allocated and their unknown outcomes, combining these with the data in Table 1.1 could lead to very different conclusions. □

[3] We also want our estimates of confidence intervals and p-values to have the correct properties. Thus, a 95% confidence interval, estimated from trial data, should include the true intervention effect in the population from which the patients are drawn in 95% of trials. Likewise, if a p-value is < 0.05, the chance of the observed trial intervention effect occurring by coincidence if there is no intervention effect in the patient population is less than 5%. However, these usually follow if the above conditions hold.

EXAMPLE 1.2 *Asthma trial*

Busse *et al.* (1998) report the results of randomising 473 patients with chronic asthma to either a placebo or 200, 400, 800 or 1600 mcg of budesonide daily. The two outcome variables of particular interest were forced expiratory volume in 1 second (FEV_1) and peak expiratory flow (PEF). FEV_1 represents the maximum volume of air, in litres, an individual can exhale in one second. It was recorded at clinic visits at baseline, 2, 4, 8 and 12 weeks after randomisation. PEF represents the maximum rate an individual can exhale air, in litres per second. It was recorded by individuals twice daily at home.

Treatment efficacy was assessed using FEV_1 and PEF. Table 1.2 shows the number of patients randomised to each treatment group, and how many remained in the trial at each scheduled visit. Amongst patients who completed, in the placebo arm the average FEV_1 was 2.072 litres,

Treatment group:	Number Randomised	Number at week 2	Number at week 4	Number at week 8	Number at week 12
Placebo	92	82	57	42	34
200 mcg	91	91	81	75	68
400 mcg	93	92	91	86	80
800 mcg	99	97	94	91	84
1600 mcg	98	97	94	90	88

Table 1.2: Number of patients randomised to each treatment group, and number remaining in the trial at each scheduled clinic visit

while in the 1600 mcg dose arm it was 2.324 litres. Comparing these two arms, the baseline adjusted estimate of treatment difference is 0.377 litres (s.e. 0.0974) which is highly significant ($p = 0.0002$).

Is the intention to treat patients with 1600 mcg beneficial? It depends on what happened to those who withdrew. If we assume patients who withdrew in the placebo arm would, had they continued, all have had FEV_1 lower than their last one, while those who withdrew in the 1600 mcg arm would all have had higher FEV_1 than their last one, then treatment is beneficial. More plausibly, patients' missing data could be closely related to their responses prior to withdrawal. In this case one cannot be confident the treatment is beneficial without a more detailed analysis.

Suppose we are instead interested in the 'per-protocol' treatment effect, that is the effect of treatment had all patients complied with the protocol throughout the trial. Should we only include in the analysis those who did not withdraw, or is this likely to be over optimistic? □

Both examples illustrate the same points. The number, or proportion, of missing observations alone is not sufficient to indicate whether missing data are an issue or not. Rather their impact is determined by

1. the question;
2. the information in the observed data, and
3. the reason for the missing data.

The question usually focuses on estimating the effect of intending to give patients a particular intervention, or estimating the 'per-protocol' (loosely, actual) effect of an intervention. As the examples illustrate, the information in the observed data depends not only on the question, but also crucially on the reason for the missing data. As one would expect, this point turns out to be at the heart of statistical analyses for partially observed data.

All who have been faced with missing data know that the uncomfortable truth is that, while we may have some knowledge about why data are missing we do not usually know for certain. So missing data bring a fundamental ambiguity into the analysis of an RCT. This ambiguity is different from the imprecise estimate of intervention effects due to sampling variation. We are in control of how many patients enter the trial and we randomly allocate them to intervention. We are not usually in control of when and why data are missing.

The above discussion also shows that one cannot make universal recommendations on how to proceed based on the proportion of missing data. The error induced by a given proportion of missing data depends critically on the context. For example, if an event (e.g. death or a serious side effect) is rare, missing data on very few patients can markedly alter estimated event rates. Also, if the proportion of patients withdrawing varies by intervention arm, estimated intervention effects are more likely to be affected than if patients withdraw independently of intervention. Missing data also cause errors in estimation of the standard error. Often, patients who remain are too similar, resulting in an overly precise estimate of the intervention effect. Thus, missing data on even relatively few patients may alter the conclusions.

In summary, errors arise when the intended full data analyses are carried out and interpreted as if there were no missing data. It is not possible to give any general rule relating the proportion of missing data to the size of these errors. Instead of adopting ad-hoc rules for various situations, a systematic approach is needed. The following example emphasises this.

EXAMPLE 1.3 *Stent vs Angioplasty trial*

Savage *et al.* (1997) randomised 220 patients, whose coronary bypass graft had become obstructed, to either balloon angioplasty or stent insertion. Table 1.3 summarises the results; restenosis (re-obstruction of the artery) is the poor outcome. Among those whose outcome is known, the odds ratio in favour of stent insertion is 0.69 (95% CI 0.37–1.28). However, the outcome is unknown for 54 patients.

The ambiguity introduced by the patients whose outcome is unknown is illustrated in Table 1.4. This shows the results that would be seen if

1. in each treatment group, unknown outcomes had the same proportion of a good results as the known outcomes;
2. all unknown outcomes were poor;
3. all unknown outcomes were good, and
4. in the stent group, unknown outcomes were 30% more likely to be good than known outcomes, whereas in the angioplasty group, unknown outcomes had the same proportion of good results as known outcomes.

		Stent	Angioplasty
Restenosis	No	54	43
	Yes	32	37
	Unknown	24	30
Total randomised		110	110

Table 1.3: Results of RCT comparing angioplasty with inserting a stent among patients whose coronary bypass graft has become obstructed. Restinosis is a poor outcome

	Assumption 1		Assumption 2		Assumption 3		Assumption 4	
	Stent	A'plasty	Stent	A'plasty	Stent	A'plasty	Stent	A'plasty
Good	69	59	54	43	78	73	74	59
Poor	41	51	56	67	32	37	36	51
Total	110	110	110	110	110	110	110	110
Odds ratio: 0.69 95% CI: 0.40–1.18			Odds ratio: 0.67 95% CI: 0.39–1.14		Odds ratio: 0.81 95% CI: 0.46–1.43		Odds ratio: 0.56 95% CI: 0.32–0.97	

Table 1.4: Trial results under assumptions 1–4 above

Table 1.4. underlines how the missing outcomes have limited the the extent to which the trial can inform clinical practice. In fact, depending on the actual outcomes of the unobserved patients the results could decisively favour either treatment.

This example also shows how, with missing data, extra assumptions about the reasons for the missing data underpin all analyses. We might feel the reason we have lost track of the patients is down to chance and has nothing to do with the outcome. Thus, in each treatment arm we might follow assumption (1) above. In this case, the estimated treatment effect remains unchanged. However, the effect of our assumption is that we obtain a narrower confidence interval.

On the other hand, if we feel pessimistic, we might assume the reason we have lost track of patients is because they have died. Thus we treat all unknown outcomes as poor (assumption (2)). Perhaps unexpectedly, the odds ratio in favour of stent is now 0.67, less than that using the observed data alone. This illustrates how assumptions about missing data may have unexpected effects. The opposite 'optimistic' assumption (3) — that all patients with unknown outcomes are better and so have not bothered to return to hospital — reduces the evidence in favour

of stent. Then again, experience might lead us to believe that the reason a patient's outcome is unknown is more likely to be good for those patients who have had a stent inserted. Assumption (4) is an example of this. Under this assumption the data are consistent with preferring stent.

In the light of the wide range of conclusions it is possible to draw from this trial under various assumptions, it may be tempting to conclude that trials with non-trivial degrees of missing data must be discarded. However, although some information is irretrievably lost, we can salvage something. The success of the salvage operation depends on (i) the extent to which we can identify a set of plausible reasons, or mechanisms for the data being missing and (ii) the degree to which conclusions are robust to these different reasons/mechanisms. For example, suppose data could be missing for reasons A, B and C. If, under assumptions A, B and C in turn, sensible analyses always show a significant treatment effect, then we can be confident of the treatment efficacy, despite the missing data.

As we shall see below, often the data themselves indicate why information is missing. Thoughtful design maximises the chance of this. Though such information is never definitive, it can nevertheless be very useful. In other cases, there may be a degree of consensus amongst investigators or other experts about why data are missing, which will allow conclusions to be drawn. Ideally, both sources of information are present.

The main focus of this book is on using information in the trial data, although we discuss the use of expert opinion in §6.4. However, as this example illustrates, in order to arrive at useful conclusions a more systematic approach needs to be adopted. □

1.5 Towards a systematic approach

We propose that a systematic approach begins with considering the reason, or mechanism, which caused the data to be missing. As this plays a central role in our discussion, we refer to it more succinctly as the *missingness mechanism*. We may think of the missingness mechanism as a second stage of sampling. It samples from the data we intended to collect leaving us with the data we actually observe. Now, if we do not know how individuals came to be included in a study, or selected for intervention, we cannot draw definite conclusions from the study. Similarly, as discussed above, unless we know the 'missingness mechanism', we generally cannot draw definitive conclusions.

However, in discussion with investigators and/or regulators, and by examining the observed data, we can often come up with one or more likely missingness mechanisms. In an asthma study, for example, it may be those with additional complications at baseline are more likely to withdraw.

Then, it turns out that there are two broad approaches for incorporating into the analysis the necessary extra assumptions that must be made when data are missing. We outline these below, and illustrate them by considering a trial with no missing data up to and including the penultimate follow-up visit, but some missing data at the final follow-up visit. We suppose interest focuses on the estimated intervention effect at the final visit.

The first approach focuses on the details of the missingness mechanism. Specifically, after taking account of all the information about the missingness mechanism in the observed data, it considers how the missingness mechanism depends on the unseen data. This then informs the

probability distribution of the missing data, and thus the analysis. The focus on the mechanism by which the data become missing (or alternatively are selected for observation) leads to the term *selection modelling* for this approach. Taking the example from the previous paragraph, we would first consider how the reason for missing the final visit depended on previous visits and baseline data. Then we would consider how, in addition to this, the missingness mechanism might depend on the unseen measurement. This then affects the probability distribution of the missing data, and thus the estimated intervention effect at the final visit.

The second approach focuses on the possible distribution of the missing data given the observed data. In the example, this means focusing on whether the distribution of patients' unseen observations at the final visit, given their observations at previous visits and baseline, is different from that seen among the patients who have no missing data. In other words the focus is on whether the 'pattern' of the data is the same in patients who do, and do not, have missing data. To estimate the intervention effect at the end of the trial, we have to make an assumption about how the patterns differ in the two groups of patients. This leads to an estimated intervention effect amongst those who do, and do not have, missing data, which has to be averaged, or mixed, to arrive at the overall estimate of the effect of intervention. Hence the name for this is a *pattern mixture* approach. Example 1.4 illustrates this in a simple setting.

Although both approaches appear different, we can actually go from one to the other, although this is usually not straightforward (Molenberghs *et al.*, 2003). Whichever approach we adopt, we need to make assumptions about either (i) the missingness mechanism, or (ii) how the distribution, or pattern, of missing data differs between patients we actually observed and those we intended to observe, but did not. Note that (i) implies things about (ii) and vice versa. We term these assumptions the *missing data model*.

If we adopt a missing data model, we can then determine a sensible analysis and draw conclusions. These conclusions will be correct if our adopted missing data model is correct. However, if it is not correct, the conclusions will generally be wrong. We can then adopt another missing data model, and re-analyse the data. In fact, we can repeat this process as often as we wish.

A more systematic, and informative, approach is as follows. Either before the trial is conducted, or during a blind review of the data, the trialists meet and discuss various missing data models that may be appropriate. Ideally, there will be agreement on the relative plausibility of these missing data models. Then, under each missing data model, the statistician can plan a sensible analysis. After the blinding is broken, these analyses are performed. The results reflect the range of conclusions that are consistent with the observed data and the assumed models.

Taking these conclusions and the relative plausibility of the missing data models together, the trial can be interpreted as follows:

1. Under the most plausible missing data model, a, we conclude A.
2. Under a range of similar missing data models, b, c, d, we conclude B, C, D.
3. Under slightly different missing data models, e, f, g, which cannot be ruled out, we conclude E, F, G.

In line with ICH E9, a valid interpretation of the trial, which explores the sensitivity of the conclusions to the missing data models, presents all these analyses. Hopefully, and quite often

in our experience, the conclusions will not be too sensitive to the more plausible missing data models. A *valid* interpretation of the trial would then be to act on the basis of the common themes running through conclusions A–D, possibly in a way that minimises the risk to patients if E–G turn out to be correct.

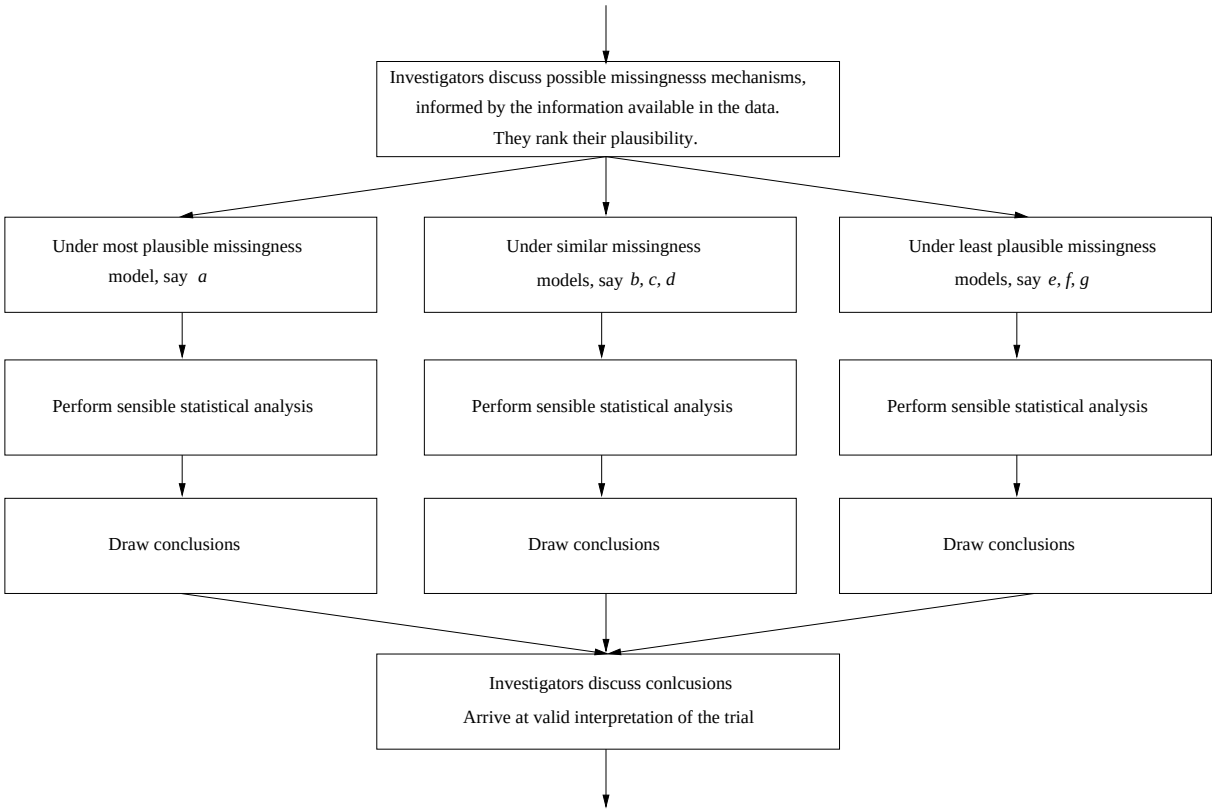

Figure 1.1: A systematic approach for analysing a trial with missing data

Figure 1.1 shows this approach. As discussed above, although the observed data usually cannot definitively identify the missing data model, they can often provide useful guidance about what is and is not plausible, in the given trial context. Thus, careful analysis of the data should play an important role in formulating missing data models. At the design stage, data from similar previous studies may be used. Data from blind reviews may also provide useful information. We consider design issues and interactions with the regulatory authorities further at the end of this Chapter.

As the missing data model is only ever a *working* proposition under which the analysis is performed, we regard considering the effect of several missing data models on the conclusions of the analysis as an essential part of the analysis process. Following ICH E9, we call this *sensitivity analysis*.

This approach is fundamentally different from common practice where the analyst regards missing data as a 'problem' and casts around for a 'solution', usually a computationally simple procedure. Once the data have been analysed using this procedure the problem is regarded as having been 'solved'. Such an approach is contrary to ICH E9, and may well lead to misleading conclusions.

Statisticians and programmers will notice we have deliberately avoided discussing what is computationally feasible. This is because we believe that the principles of the analysis should be

laid down before turning to computational and methodological issues. Although at first sight our approach might appear much harder work for the statistician, often analyses under slightly different missingness mechanisms are quite similar from one trial to another, allowing programs to be reused with relatively minor changes.

Our proposed approach may give sharper conclusions than the practice of replacing each missing observation with either a best or worst value, seeking that combination which gives the smallest estimated intervention effect. We do not believe this makes it misleading, though. Rather, best/worst values, which often represent extremely implausible, degenerate[4] probability distributions for the missing data, are more likely to mislead.

When analysing a trial without missing data, we do not cycle through a series of analyses presenting only the one giving the strongest or weakest estimated intervention effect. Instead, at the design stage we plan the analysis which will hopefully give the most sensible estimate of the intervention effect. After the trial, we report the results of this analysis.

In the same way, when anticipating missing data (preferably at the design stage but possibly at a blind review of the data) we believe it is sensible to discuss possible missingness mechanisms and missing data models. These should take into account any information present in the data. Such information very often shows that the best/worst value approach (two paragraphs above) gives rise to extremely implausible results for individual patients. Once missing data mechanisms and models have been identified, sensible analyses can be planned to give estimates of the intervention effect, and valid conclusions drawn.

1.6 Missing data mechanisms

We now discuss possible missingness mechanisms in more detail. In terms of their implication for the analysis, we shall see they fall into three broad classes. This is encouraging, as it suggests that the approach outlined in Figure 1.1 is practical. In the statistical literature, these classes are given names like 'missing completely at random', 'missing at random' 'missing not at random' and '(un)ignorable'. These are supposed to be succinct summaries of the analysis implications of a missingness mechanism belonging to one of these classes, and were introduced by Rubin (1976). We discuss these classes in a trials context.

1.6.1 Missing completely at random

Suppose the missingness mechanism is unrelated to any inference we wish to draw about the intervention effect. For example, some observations may be missing because of equipment failure in the clinic, or because a member of staff was ill, or because a patient was unable to attend for some reason not related to his/her illness or its intervention (e.g. his/her child was unwell).

Such events are as likely to occur for one patient as for another, whatever their disease severity or intervention. Thus the average effect of intervention will be the same among those who do, and do not, have missing data. This means that estimating the effect of intervention from those

[4] A probability distribution which says a single particular value is certain to occur is termed *degenerate*. With missing data, all we can estimate is the distribution of the missing data given the observed data, under certain assumptions. Imputing a single, worst/best value, usually therefore implicitly assumes a very implausible degenerate distribution for the missing data given the observed data.

who do not have missing data will give a sensible estimate of intervention effect. It is as if, after randomising the patients to intervention, we further randomly decide who to observe.

Data that are missing for reasons unrelated to any inference we wish to draw about the intervention are called *missing completely at random* or *MCAR*. As we argued above, with data MCAR, analysing only those with fully observed data gives sensible results. Of course, the results are inevitably less precise than if the full data had been observed.

EXAMPLE 1.2 *Asthma trial (ctd)*

Table 1.5 shows the mean and standard error at week 2 of the 91 patients randomised to 200 mcg of budesonide. It also shows the mean and standard error in 5 situations when 10 of these 91 observations are MCAR. In each of these 5 cases, the data consists of 81 observations drawn randomly from the full set of 91. In all cases, the mean is close to 2.170 litres, the value in the full data. This illustrates that if data are MCAR, analysing only the observed data gives sensible results. However, notice that in all cases, the estimated variability of our intervention estimate (the standard error of the mean) is larger than that for the full data. This illustrates that information is lost if data are MCAR.

Lastly, the right hand column shows the mean and standard error when the 10 largest observations are omitted. Here the bias is obvious, and the standard error decreases. This shows what happens when, despite missing data, only the planned full data analysis is carried out. This implicitly assumes missing data are MCAR. If they are not, results may be misleading. □

	Full data	10 obs MCAR					Missing 10
	91 obs	case 1	case 2	case 3	case 4	case 5	largest obs
mean (litres)	2.170	2.196	2.167	2.154	2.160	2.137	2.002
standard error:	0.078	0.081	0.085	0.085	0.083	0.081	0.066

Table 1.5: Illustration of the effect of data missing completely at random. Data from week 2 follow-up of the 200 mcg arm of the asthma trial. As in Example 1.2, the outcome is patient FEV_1

1.6.2 Is MCAR likely in practice?

We have seen that when data are MCAR, we can set the details of the missingness mechanism to one side and analyse the observed data. That is to say, a sensible analysis simply includes only those patients who have complete data on the variables needed for that analysis. All we have lost is some information.[5] We therefore need to consider whether MCAR is likely in practice, and how we might detect it. Recall that the definition of MCAR data is that the missingness mechanism is unrelated to anything we wish to infer from the data. Assuming that reasonably

[5] As discussed in later chapters, depending on whether covariates or outcomes are MCAR, it may be possible to use partial observations and recover some of this information.

careful follow up arrangements are in place, it follows that the proportion of patients with data MCAR is likely to be small. Further, this proportion will not vary with any of the observed covariates (e.g. intervention group, sex, age, illness severity).

Unfortunately, these points are only *consistent* with MCAR, they are not sufficient to show that data are definitely MCAR. For example, there could be a variable related to intervention, associated with the chance of a patient withdrawing, which was not measured. In the asthma study, withdrawal of hayfever sufferers might depend on the local pollen count. Withdrawal may be unassociated with any of the data recorded. Thus, if local pollen count is not recorded, data may appear MCAR. But they are not MCAR. Rather, as high local pollen count exacerbates a patient's asthma, we are left with data from patients who either do not have hayfever, or whose hayfever is better controlled. This is a non-random sample of those enrolled in the study.

An extreme example of this would be if withdrawal depended on a sudden, unpredicted, change in the response, e.g. a sudden deterioration in FEV_1. Again, looking at the observed data, patients may appear to be MCAR, but in fact patients who withdraw are systematically different from those who do not — just in an unobserved way.

The last two paragraphs underline how the extra assumptions required for an analysis when data are missing cannot be verified from the data. In spite of what our observed data may suggest, we can never be sure that data are MCAR. Nevertheless, the observed data can rule out MCAR. We can investigate whether there is any relationship between observed data and the occurrence of missing data. If there is, data are not MCAR. We can investigate this more formally. For an example in a longitudinal context, see Diggle (1989).

EXAMPLE 1.2 *Asthma trial (ctd)*

Are patients MCAR in the asthma trial?[6] From Table 1.2 the chance of a patient staying in the trial to the end clearly depends on treatment; those in the placebo or lowest dose arm are much less likely to complete. A chi-square test confirms this, $p < 0.001$. Patients are clearly not MCAR. Of course, patient withdrawal may also depend on other factors besides treatment. □

1.6.3 Missing at random

In practice trial data are rarely MCAR. Usually there is an association between the chance of patient withdrawal and observations — typically intervention, baseline and (in longitudinal follow-up) measurements prior to withdrawal. In this case, it is not sensible to include in the analysis only those with complete data.

For example, suppose that worse health at baseline is associated both with increased risk of withdrawal and poor response to intervention. Analysing data from the patients who remain to the end of the trial will thus give an over optimistic view of the intervention effect. However, if we can identify those variables which are associated with an increased risk of withdrawal, we can carry out a sensible analysis. We illustrate this key idea with the asthma trial.

EXAMPLE 1.2 *Asthma trial (ctd)*

In the placebo arm, only 35 out of 92 patients completed the trial. The average FEV_1 of the 92 patients at the start of the trial was 2.050 litres. Suppose we are interested in the FEV_1 at 12

[6] This is a minor abuse of our definitions: more strictly, is the reason for patients' withdrawal (and hence their unseen responses) in the 'MCAR' class — i.e. independent of anything we wish to make inferences about.

weeks, and whether there is any evidence of a 'placebo effect' whereby patients taking a drug improve, even though their drug contains no active ingredients.

If we believe the patients are MCAR at 12 weeks, then a valid estimate of the average FEV_1 at 12 weeks is obtained from the 35 who complete. Their average is 2.072 litres, suggesting no placebo effect.

However, we need to check if MCAR is plausible. So we need to look to see if any variables in the data are associated with withdrawal. As one suspects that patients with worse asthma initially are more likely to withdraw, an obvious place to look is baseline FEV_1. Suppose we classify baseline FEV_1 as 'low' if it is below 2.015 litres and 'high' otherwise. Table 1.6 shows how patients with low FEV_1 at baseline are much more likely to have withdrawn by 12 weeks. An analysis that assumes MCAR is not sensible.

		Baseline FEV_1	
		low	high
At 12 weeks	present	15	20
	absent	31	26

Table 1.6: Proportion of placebo patients who have withdrawn by 12 weeks, by baseline FEV_1

However, suppose the 15 patients with low baseline FEV_1 who we see at 12 weeks are drawn randomly from the 46 patients with low baseline. In other words, *within the group of patients with low baseline FEV_1*, patients are MCAR. Then, for those patients with low baseline, a sensible estimate of the average 12 week FEV_1 is given by averaging the 12 week values for the 15 patients who we see. This is 1.861 litres.

Likewise, suppose the 20 patients with high baseline FEV_1 who we see at 12 weeks are drawn randomly from the 46 patients with high baseline. Arguing in the same way, for those patients with high baseline, a sensible estimate of the average 12 week FEV_1 is given by averaging the 12 week FEV_1 for these 20 patients who we see. This is 2.230 litres.

Figure 1.2 shows this graphically. At 12 weeks, we assume that we observe a random selection of the patients in the 'low' and 'high' group. A sensible estimate of FEV_1 at 12 weeks *within* these groups is therefore the average of the observed values in these groups.

The overall average 12 week FEV_1 can thus be sensibly estimated by averaging the estimates from the 'low' and 'high' groups, allowing for there being 46 patients in the low group and 46 in the high group. This is

$$\frac{46 \times 1.861 + 46 \times 2.230}{92} = 2.046 \text{ litres.}$$

Comparing this with the average baseline FEV_1 of 2.050 litres, confirms that there is no evidence of a 'placebo effect'. □

The key steps in the above example are:

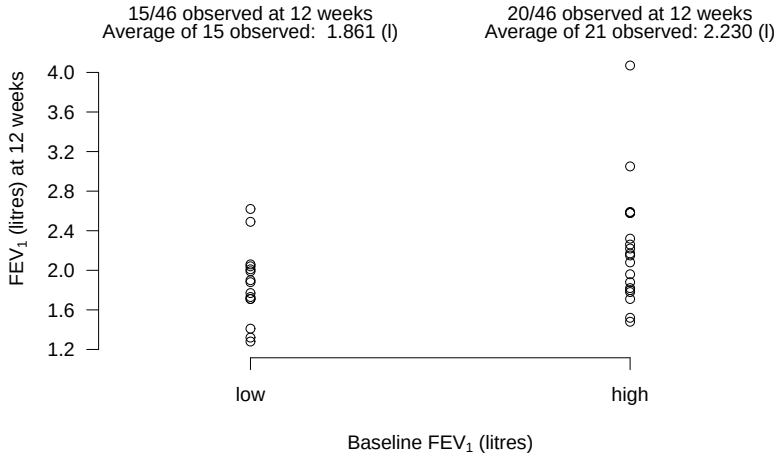

Figure 1.2: Graphical illustration: within the two groups defined by 'low' and 'high' baseline FEV_1, we assume that we observe a random selection of patients at 12 weeks

1. identify a fully observed variable whose values predict the occurrence of missing data;
2. within groups defined by this variable, assume data are MCAR;
3. within these groups, sensible estimates can thus be obtained from the observed data, and
4. to obtain a sensible estimates overall, average the estimates from the groups in (3), allowing for there being different numbers of patients in the different groups.

When we can find fully observed variables which define groups within which the data are MCAR, we say data are *missing at random (MAR)*. In the asthma trial, within the two baseline FEV_1 groups, response at 12 weeks is MCAR. The expression 'MAR' is thus a convenient shorthand. Instead of saying 'assume there exists a variable, say baseline FEV_1, and that among those with the same baseline FEV_1, the FEV_1 at 12 weeks is MCAR' we can just write 'assume FEV_1 at 12 weeks is MAR'.

The asthma example also enables us to introduce the term *conditional*, common in missing data literature. Instead of saying 'among those with the same baseline FEV_1, the FEV_1 at 12 weeks is MCAR' we say *conditional* on baseline FEV_1, the FEV_1 at 12 weeks is MCAR.

Before going on, we underline an aspect of MAR data implicit in what has gone before. In the asthma study, we wanted an overall average 12 week FEV_1, not an average in the 'low' and 'high' baseline group. We call this overall average *marginal*. If no patients withdrew, the marginal average is simply the average of the values for the 92 patients.

With MAR data, averaging the values of the 35 patients who do not withdraw is wrong. We can no longer go directly to the marginal average. Instead we have to calculate averages *conditional* on baseline, and then take a further step to estimate the marginal average.

We repeat the previous paragraph more abstractly. If values of a variable Y are MAR, then statistics calculated using only observed values of Y, (e.g. mean, standard deviation, confidence intervals, p-values) are wrong. As values are MAR we cannot go directly to marginal statistics. Suppose that conditional on the fully observed variable X, Y is MCAR. Then we have to calculate statistics conditional on X, and then take a further step to estimate the marginal statistics.

Generally, variables that we condition on do not have to take on discrete values (as baseline FEV_1 was forced to in the example above). When we are interested in averages, and have quantitative variables, we can condition using *regression*. The *regression* of Y on X estimates the average of Y conditional on X. To do this it estimates the numbers α and β so that

$$\text{average value of } Y = \alpha + \beta \times X. \tag{1.1}$$

For example, if $\alpha = 1$ and $\beta = 2$, then the average value of Y conditional on $X = 5$ is $1 + 2 \times 5 = 11$.

We now use the same idea from the asthma example again. Recall that Y is partially observed, X fully observed, and conditional on X, Y is MCAR. We estimate α and β by fitting the regression to the subset of individuals on whom both Y and X are observed. Once we have α and β we can use (1.1) to get the conditional average of Y for *each* value of X for which Y is missing. Then we average these conditional values with the observed Y values to give the marginal average of Y.

EXAMPLE 1.2 *Asthma trial (ctd)*

Above, we split baseline FEV_1 into two groups to introduce the ideas. However this is not necessary. Suppose 12 week FEV_1 is MAR; specifically that conditional on baseline FEV_1, 12 week FEV_1 is MCAR. Then we can estimate the average 12 week FEV_1 conditional on a value of baseline FEV_1 by fitting the regression model

$$\text{Average 12 week FEV}_1 = \alpha + \beta \times \text{baseline FEV}_1$$

to the 35 patients on whom we observe both. Doing this, we find that

$$\text{Average 12 week FEV}_1 = 0.923 + 0.535 \times \text{baseline FEV}_1 \tag{1.2}$$

Thus, conditional on a baseline FEV_1 of 2.0 litres the average 12 week FEV_1 is $0.923 + 0.535 \times 2 = 1.993$ litres

The average 12 week FEV_1 is obtained by (i) calculating the conditional 12 week FEV_1 for each of the 57 patients with missing 12 week FEV_1, and (ii) averaging these 57 values and the 35 observed values. Numerically, this gives:

$$\begin{aligned}
&\frac{1}{92}\{(0.923 + 0.535 \times \text{baseline FEV}_1 \text{ of 1st patient with missing 12 week FEV}_1) \\
&\quad + (0.923 + 0.535 \times \text{baseline FEV}_1 \text{ of 2nd patient with missing 12 week FEV}_1) + \cdots \\
&\quad + (0.923 + 0.535 \times \text{baseline FEV}_1 \text{ of 57th patient with missing 12 week FEV}_1) \\
&\quad + 12 \text{ week FEV}_1 \text{ of 1st patient with observed 12 week FEV}_1 + \cdots \\
&\quad + 12 \text{ week FEV}_1 \text{ of 35th patient with observed 12 week FEV}_1\} \\
&= 2.109 \text{ litres.}
\end{aligned}$$

So, assuming 12 week FEV_1 is MCAR given baseline FEV_1, a sensible estimate of average 12 week FEV_1 is 2.019 litres, down from the baseline mean of 2.050 litres. This is more accurate than the estimate obtained before, when we lost information by splitting the quantitative variable, baseline FEV_1 into two groups, 'high' and 'low'. □

Notice that if we say a partially observed variable is MAR, then that means that we have fully observed variables, conditional on which the partially observed variable is MCAR. In other words, conditional on these fully observed variables, *the reason for the missing values does not depend on the unobserved values themselves*. This is the aspect of MAR that is usually mentioned in quick descriptions in the literature. We emphasise that this is a conditional statement. If data are MAR, the reason for the missing values will often depend on the unseen values. However, *conditional on other fully observed values this association will be broken.*

EXAMPLE 1.2 *Asthma study (ctd)*

Consider again the placebo arm of the asthma study. It is plausible that the probability of the 12 week FEV_1 being missing is higher the lower the value of that (unseen) observation. Assuming MAR though, conditional on a patient's baseline FEV_1, the probability of them being missing at 12 weeks no longer depends on their FEV_1 at 12 weeks. □

In summary, if we have a fully observed variable whose values affect the chance of seeing missing data, those missing data are not MCAR. But, if conditional on this fully observed variable, we *assume* the chance of seeing the partially observed variable does not depend on its values, the data are MAR. The important word is *assume*. Usually we do not know whether MAR is actually true or not.

The final implication of MAR is that the statistical distribution of potentially missing data is the same (conditionally) for all patients who share the same *observed* data, *whether or not they withdraw*. This is implicit in the example above, where in the placebo group we:

1. estimated the conditional distribution of week 12 FEV_1 given baseline FEV_1 from the 35 patients on whom both was observed;

2. assumed this distribution was identical in the 57 patients whose week 12 FEV_1 was not observed, and

3. then used this distribution to estimate the mean week 12 FEV_1 for these 57 patients.

Or, more generally in a longitudinal study design, under MAR subjects who withdraw share the same conditional statistical behaviour in their (unobserved) future, given their observed past, as those who do not withdraw. It is this property that allows principled methods of analysis, like those based on likelihood, to make the appropriate adjustments for withdrawal under MAR.

EXAMPLE 1.2 *Asthma study (ctd)*

In the previous example, we fitted the regression relating average 12 week FEV_1 to baseline FEV_1, equation (1.2). Implicit in this is that, amongst the 35 patients with both 12 week and

baseline observations, the conditional distribution of a patient's 12 week FEV_1 given their baseline FEV_1 is x, say, is normal, with estimated mean $0.923 + 0.535x$, and estimated variance 0.213.

The MAR assumption means that the distribution of 12 week FEV_1 given baseline FEV_1 for the 57 patients with missing 12 week FEV_1 *is the same*, and that its parameters are consistently[7] estimated by values given above, which are calculated using the data from the 35 patients on whom both are observed. □

This facet of MAR means that (assuming data are MAR) joint modelling of complete and partially observed response data, conditional on fully observed data, is a natural way to approach the analysis. This motivates the approach we develop in Chapter 3.

1.6.4 Missing not at random

If data are neither MCAR nor MAR, then they are *missing not at random* (MNAR). Such data are also often referred to as *not missing at random* (NMAR) or *informatively missing* (IM) or *non-ignorable*. This means that, even given the information about the missingness mechanism in the fully observed data, the reason for an observation being missing still depends on the unseen value of that observation.

EXAMPLE 1.2 *Asthma trial (ctd)*

Patients are MNAR from the placebo arm at 12 weeks if conditional on their baseline FEV_1, the chance of their being present at 12 weeks still depends on their FEV_1 at 12 weeks. □

Estimating effects when data are MNAR is much more difficult. This is because we now need to either (i) describe the statistical relationship between the chance of seeing a variable and its (unseen) value or (ii) describe how the distribution of the data differs among patients with missing observations. Clearly the observed data can tell us nothing definitive about either of these. As mentioned above, when data are missing there can be no *definitive* analysis. It is important therefore to remember that the move from an MAR to MNAR analysis does not necessarily bring us nearer the "truth". One is not searching for the "correct" MNAR model; it will never be identifiable. Rather, such models are a vehicle for expressing in a formal, and ideally transparent way, possible departures from the MAR assumptions which can be used to underpin a principled sensitivity analysis. The choice of MNAR model(s) may well be informed by expert opinion, or other information from the substantive setting. While these can be used to help delineate meaningful directions of departure for the assessment of sensitivity, they cannot be substitutes for the sensitivity analysis itself.

Usually in clinical trials data are MNAR, at least to some degree. However, that does not mean that methods valid under MAR are of little use; quite the contrary. First, it quite often happens that after accounting for the information about the missingness mechanism in the observed data, there is relatively little information remaining in the unseen data (Rubin *et al.*, 1995). There may be additional measurements that predict withdrawal and can be used in the analysis to reduce still further the dependence of missingness on the unseen data. In this case, the MAR model may well give quite accurate answers. Second, because there is a sense in which the

[7] Informally, *consistent* means that as the data set gets larger, the parameter estimate 'homes in' on the true value.

MAR assumptions are the weakest required to justify an analysis that ignores the missing value mechanism, they provide a very sensible starting point for sensitivity analysis. All MNAR models represent departures in one way or another from this crucial set of assumptions.

EXAMPLE 1.4 *Asthma study*

In §1.6.3 we assumed 12 week FEV_1 was MCAR conditional on baseline FEV_1. Now suppose that even conditional on baseline the chance of seeing FEV_1 at 12 weeks depends on the value at 12 weeks — *i.e.* 12 week FEV_1 is MNAR.

We have to make further assumptions in order get an estimate of the average FEV_1 at 12 weeks. Taking the *pattern mixture* approach (see p. 11), suppose that the average FEV_1 of those who withdraw at 12 weeks is 10% less than MAR would predict. Assuming this is correct, we can obtain sensible estimates for the 12 week FEV_1. We need to (i) assume MAR and calculate the expected 12 week FEV_1 for the 57 patients who are missing at 12 weeks; (ii) reduce this by 10% and (iii) average these values with the 35 observed 12 week FEV_1's.

Under our MAR model above, the average 12 week FEV_1 for the 57 patients who are missing at 12 weeks is

$$\begin{aligned}
&\frac{1}{57}\{(0.923+0.535\times\text{baseline FEV}_1\text{ of first missing patient})\\
&\quad+(0.923+0.535\times\text{baseline FEV}_1\text{ of 2nd missing patient})+\cdots\\
&\quad+(0.923+0.535\times\text{baseline FEV}_1\text{ of 57th missing patient})\}\\
&=0.923+0.535\times\text{average baseline FEV}_1\text{ of 57 missing patients}\\
&=0.923+0.535\times 1.988=1.987\text{ litres.}
\end{aligned}$$

However, as these patients dropped out, under our model we now reduce this by 10% on average to $0.9\times 1.987=1.788$ litres. Finally we combine this figure with the data from the 35 patients who completed the trial; their average 12 week FEV_1 is 2.072 litres. Our estimate of 12 week FEV_1 is thus

$$\frac{1}{92}(57\times 1.788+35\times 2.072)=1.896\text{ litres.}$$

As expected given our assumptions, our MNAR estimate of 12 week FEV_1 is below that obtained from our MAR analysis. Both analyses are sensible if their respective assumptions are true. However, the conclusion of the MAR analysis, that FEV_1 remains unchanged in the placebo arm, is sensitive to the assumed MAR mechanism. The more we assume patients who withdraw have a worse FEV_1, the lower the average FEV_1 at 12 weeks. □

1.7 Some other terms that may confuse

As mentioned above, the word 'ignorable' is often used in connection with missing data, although strictly it refers to the probability model (likelihood) for the data. It summarises whether a joint model for the observed data and the missingness mechanism is necessary for a valid analysis (as it is under MNAR), or whether the model for the missingness mechanism can be ignored. More loosely, 'ignorable missing data' means that the missing data mechanism is either MCAR or MAR. It does *not* mean that that missing data can be ignored, and sensible marginal results obtained from just analysing the subset of patients with no missing data.

Another term that sometimes occurs is *covariate dependent missing completely at random*. This just denotes data that are MAR, or MCAR given covariates. However, in a trials context

the 'covariates' referred to are measured at baseline. The term is thus used to distinguish (i) MAR data where the reason for withdrawal depends on baseline data alone from (ii) MAR data where the reason for withdrawal depends on baseline data and post-randomisation data prior to withdrawal. Our modelling approach in part II encompasses both these situations naturally within the same framework, so we do not consider this distinction further.

1.8 Implications

Hopefully, the importance of clarifying possible missingness mechanisms is clear. Once this is done, each mechanism can be classed as either MCAR, MAR or MNAR, sensible analyses performed, and conclusions drawn. It is important to know whether these conclusions are sensitive to the missingness mechanisms as this implies how precise the implications of the trial are.

Further, our discussion has explained why the CONSORT guidelines say authors 'should report the number of patients with missing data by treatment arm: imbalance is likely to cause bias when the outcome of interest is associated with the reason for patient withdrawal.' If patient withdrawal is truly MCAR conditional on intervention *alone*, most statistical analyses are sensible, as they condition on intervention to estimate its effect. But readers should be aware of such imbalance, to alert them to the fact that withdrawal may be MNAR, and enable them to consider the direction and extent of any resulting biases. Imbalance in patient withdrawal by intervention arm is not itself a problem, but is a possible indicator of other problems.

1.8.1 Design

Good design is the key to minimising the ambiguity introduced by the inevitable missing data. At the design stage, the statistician should stress the ambiguity that missing data can cause, and convey to the investigators the extent to which the trial conclusions can become ambiguous even with a fairly small number of missing observations. Then protocol modifications that can reduce the chance of missing data occurring can be considered. As part of this, plausible missingness mechanisms should be considered.

The more confident we are that data are MAR, the less ambiguity is introduced into the analysis. Thus, it may be that slight modifications to the proposed observations, or slight changes to the information collected prior to patient withdrawal, can make the missingness mechanism MAR. Obviously, much can be learned from missingness mechanisms in previous trials in similar areas. The measurements that are likely to prove useful will depend on the postulated missingness mechanisms. However, measurements that correspond exactly with withdrawal are of little help, as they cannot be used as part of an MAR analysis. Thus recording 'patient withdrawal due to intervention failure' for all patients who withdraw is not helpful. Instead, prior to withdrawal, we need variables recorded on all patients (e.g. at baseline or during follow-up), some of whom subsequently withdraw. Sometimes, the withdrawal process can be triggered by values of these variables, forcing data to be MAR.

EXAMPLE 1.2 *Asthma study (ctd)*

In the asthma study, we could require patients whose FEV_1 falls 10% below their baseline to withdraw. Their subsequent data would be missing, but given their observed data, the missingness mechanism would by definition not depend on their missing data. Thus data would be MAR.

As in this set-up we know the missingness mechanism, the ambiguity introduced by the missing data is reduced, although we still have to make a distributional assumption about data we have not seen. This approach is described in the context of hypertension trials by Murray and Findlay (1988). □

For the same reasons that complete data analyses are pre-specified, it is very helpful to pre-specify analyses for the most plausible missingness mechanisms in consultation with the regulators. Figure 1.1 provides a possible structure for doing this.

1.8.2 Missing data and per-protocol analyses

Broadly speaking, a per-protocol analysis seeks to estimate the intervention effect that would be seen if all patients undertook the intervention as per the protocol. Thus in a per-protocol analysis, when a patient withdraws we want to estimate the distribution of their response(s) had they continued to adhere to the protocol. Assuming the data are MAR, sensible estimates of intervention effect address precisely this hypothesis. To see this, consider the MAR asthma example again (see p. 15). Given 'low' or 'high' baseline FEV_1, we assume the chance of observing 12 week FEV_1 is random. In other words, within each baseline group, the distribution of 12 week FEV_1 values is the same for observed and unobserved patients. This assumes that *the unobserved patients continued with treatment per-protocol, as the observed patients did.* Under MAR, the estimated treatment effect is thus the per-protocol treatment effect.

This applies quite generally, and there is an analogy with the rationale for trials. Patients enrolled in a trial are representative of a wider patient population. The trial provides information about that wider population because we assume the distribution of patients' response to intervention is representative of patients not included in the trial. Similarly, the MAR assumption says that if a group of patients have similar observations until some withdraw, the distribution of response to the intervention for the whole group is represented by those who complete. Thus, the hypothesis of a per-protocol analysis can be directly addressed if the missing data are MAR and the observed data are from those who adhered to the protocol.

Nevertheless, we should still carry out sensitivity analyses to the MAR assumption. For example, this assumption is likely to be

(a) implausible if patients' missing data are due to some unobserved deterioration, but

(b) much more plausible if patients' missing data are simply due to loss to follow-up.

Sensitivity analyses (through either a selection or pattern mixture approach) should use any information bearing on (a) and (b) above to frame possible departures from MAR. For per-protocol analyses, the question turns on how the distribution of the unseen response data of patients who withdrew differs from those who did not, under the assumption that the former continued with the intervention.[8] There are clear links here with the literature on randomisation based estimates of causal intervention effects. See for example, White *et al.* (1999); Peduzzi *et al.* (1993); White *et al.* (2003) and references therein.

[8] to the minimum extent required by the protocol.

1.8.3 Missing data and intention to treat (ITT) analyses

With no missing data, 'Intention To Treat' (ITT) analyses include every patient who was randomised, regardless of adherence to the protocol, to estimate the effect of intending to give an intervention. Thus if we continue to follow up patients after they cease to adhere to the intervention protocol, whatever intervention or treatment they then receive, we have the data we need for a ITT analysis. By contrast, the per-protocol analysis would regard a patient's data as effectively missing from the time they ceased to adhere to the protocol.

However, if the post-protocol adherence data are missing, then—because of the different hypotheses underlying per-protocol and ITT analyses—there must be a difference in the way missing data are handled. In other words there must be a difference between the conditional distribution of the missing data given the observed when addressing the ITT and the per-protocol hypotheses. Moreover, as we argued above, a MAR analysis directly addresses the per-protocol hypotheses[9]. Thus the ITT interpretation cannot be directly adopted when outcome data are missing (Hollis and Campbell, 1999), a fact that appears to remain quite widely misunderstood (Wood *et al.*, 2004).

To address the ITT hypothesis when data are missing, we need to consider plausible models for the distribution of unseen responses when patients cease to comply with the protocol. Such distributions should, as usual, condition on the observed responses. Assume for now that patient responses are observed if, and only if, they are complying with the protocol. Then an ITT analysis implicitly assumes a MNAR mechanism, as the pattern of responses is different for those who do and do not continue with intervention per-protocol. Mathematically, this can be shown to be equivalent to the reason for withdrawal depending on the unseen responses, even after taking into account the information in the observed data. We have already noted that, under MNAR, a range of missingness mechanisms, all equally consistent with the data, can operate. Thus, under MNAR there can be no definitive estimate of intervention effect. So, if there are missing values, there can no longer be an unequivocal ITT analysis (Brown, 2003).

However, if we are careful at the design stage, we may be able to ensure the collection of valuable information to inform ITT analyses. For instance, consider an asthma study comparing placebo and two active drugs: the standard treatment and a new treatment. Suppose that, following withdrawal, we note the treatment patients switch to. Suppose further that most of them go onto the standard treatment. From the trial, we have information about how patients assigned to the standard treatment respond over time. We can use this information to inform estimates of the distribution of responses among patients who discontinue their randomised treatment and switch to the standard treatment. In this way we can arrive at an ITT treatment estimate. This approach is described in some detail by Little and Yau (1996). We can also use expert opinion to inform a model for the change in the pattern of patient response following withdrawal, and hence an ITT analysis. In Chapter 6 we describe possible approaches, and consider these issues further. As usual, there is a key role for sensitivity analysis in exploring the robustness of conclusions to modelling assumptions.

[9] If missing data — due to patients who withdraw — are MAR and observed data are from patients who adhere to the protocol.

1.8.4 Composite hypotheses

When patients withdraw, Shih and Quan (1997) advocate an ITT analysis based on a joint test of (i) whether patients who complete, benefit from the intervention and (ii) whether those who withdraw have excessive adverse effects. As an illustration, consider a trial to reduce blood pressure. Suppose some patients withdraw from the trial, and subsequently some of these die of a heart attack. This approach divides patients into two groups (i) those who complete and (ii) those who do not. Amongst group (i) we compare the effect of the intervention on blood pressure, as originally intended. Amongst group (ii) we compare the effect of the intervention (or intending to give the intervention) on the risk of death.

This approach is attractive as it hard to conceive of a patient's ITT blood pressure after they have died. However, it entails a unified definition of 'intervention benefit' across groups (i) and (ii) so a null hypothesis can be specified and tested. This means explicitly weighing the benefits of, for example, blood pressure reductions and heart attacks. When faced with multiple interventions, factorial designs, and/or unexpected adverse events, we anticipate such definitions of 'intervention benefit' are increasingly intricate, difficult to define in advance, difficult to communicate to non-statisticians and difficult to defend.

It is impossible to give completely general guidance, as, clearly, a lot depends on the interplay between the treatment and the adverse event. However, we feel that usually greater clarity emerges by keeping the analysis of the primary response separate from the adverse event rate. Nevertheless, the relevance of the ITT hypothesis becomes questionable.

Returning to the blood pressure reduction trial illustration, heart attack is not wholly attributable to blood pressure. If the predominant reason for withdrawal is a heart attack, a per-protocol analysis together with a comparison of the heart attack rate may provide a clear basis to interpret the trial. Here the per-protocol analysis estimates the effect of treatment if no heart attack occurs.

When the heart attack occurs after withdrawal, there is no substitute for post-protocol adherence follow-up data. With this, we can make progress towards evaluating the ITT hypothesis. Again, the ITT analysis estimates the effect of treatment if no adverse event occurs.

On the other hand, if heart attack might be directly linked to the treatment, by whatever mechanism, the primary focus switches to the end point of heart attack and composite hypotheses are of secondary interest.

In summary, without ruling out this approach, we believe that a trial can usually be more clearly interpreted by separating, rather than combining, the response of interest and adverse events. Composite hypotheses sidestep, rather than address, the ITT hypothesis. We therefore do not consider composite hypotheses further here.

1.9 A critique of CPMP guidelines

The Committee for Proprietary Medicinal Products (CPMP) (2001) adopted some 'points to consider on missing data' (henceforth referred to as the CPMP-PTC), which aim to put flesh on the ICH E9 bones. We now review these in the light of the principles discussed in this Chapter.

First, this Chapter reinforces the important points made in CPMP-PTC that it is necessary

(i) to take care, in both design and implementation, to try to minimise the number of missing observations;

(ii) to consider how to cope with missing data when drawing up the analysis plan;

(iii) where possible, to agree in advance the nature and scope of sensitivity analysis, and

(iv) to look closely at the data, especially the proportion of missing data by time of withdrawal and treatment arm.

The CPMP-PTC is also surely right to make the point that there can be no universal analysis when data are missing. However, unfortunately, it overlooks the fact that there are general principles to follow in the analysis of missing data, and these principles can and should inform specific analyses. It is these principles that we have aimed to set out in this Chapter. Without clear principles to inform the document, the motivation of various statements is unclear, and the application of the guidelines to specific problems unnecessarily difficult.

Perhaps this goes some way to explain a key misunderstanding on the effect of a relationship between treatment and missing outcome data. On page 4, at the top, the guidelines say

> "mixed effects models ... assume that there is no relationship between treatment and missing outcome, and this generally cannot be assumed"

This is incorrect. Mixed models are a form of multivariate regression models; these do not rely on there being no association between missing outcome and treatment. Instead, the MAR principle described above applies directly: if outcome data are MCAR given treatment alone (*i.e.* MAR), then including treatment in the model (*i.e.* conditioning on it) gives a valid treatment estimate. Almost all primary analyses include treatment in this way.

In addition, the above statement implicitly contradicts page 2 of the guidelines, where it states:

> "In principle missing values will not be expected to lead to bias if they are only related to the treatment..."

However, the principles described in this Chapter show this is only true if data are MAR. We can never know if this is the case; indeed often, as discussed above, it will not be. As it stands, therefore, this statement is of very limited usefulness for informing the analysis of a trial with missing data.

A further point of concern is the discussion of the need to impute missing data to perform an analysis. Yet, we have seen that under MAR this is not necessary. In Chapter 6 we also see that this is not necessary for MNAR analyses. Of course, one way to do MAR and MNAR analyses is via imputation, but that is a separate issue.

Considerable space is also given to the discussion of Last Observation Carried Forward (see Chapter 2), and imputing the best or worst values for missing data. Again, this is unhelpful. The principled approach set out here clearly shows that if data are missing, extra assumptions are

needed to estimate the *probability distribution* of patients' missing data given their observed data. Only under very implausible scenarios will this probability distribution be focused entirely on a single value. Away from the case of binary data, additional, yet more implausible assumptions, are needed if this value is the best or worst possible. Even the definition of best or worst values is problematic in many settings, and triggers extensive, unilluminating, debate.

To impute single values of any kind is to lose sight of the principles underpinning sensible analyses with missing data. From another viewpoint, in the analysis of observed data we always allow for the fact measurements are made with error; that is why we use regression methods. To assume that, if data are missing, we can impute the missing value without error (*i.e.* impute a single value) is consequently illogical.

A similar concern runs through the discussion of sensitivity analysis. Here the document focuses on exploring sensitivity to methods, rather than assumptions. Unfortunately, a large number of methods (e.g. mixed models, multiple imputation, mean score imputation, EM algorithm, hot-deck imputation, ...) all usually rely on the same assumption: that data are MAR. Therefore, following the guidelines, sensitivity analysis using a variety of methods could lead to a misleadingly optimistic view of the robustness of the conclusions.

Instead, sensitivity analysis needs to vary the assumptions, then use appropriate techniques to estimate the effect of intervention under these alternative assumptions. The aim is to establish how robust such intervention estimates are to such alternatives.

Thus, while these guidelines attempt to answer some very important questions, and put necessary flesh on the bones provided by the paragraph in the ICH E9 guideline on missing data, they fall well short of their goal. Indeed, to the extent that the ICH E9 guideline calls for a principled approach to the analyses of trials with missing data, it conflicts with the CPMP points to consider. A substantial revision of the latter is overdue.

1.10 Inferential approach

Although Bayesian approaches are gaining ground in the analysis of clinical trials (Spiegelhalter *et al.*, 2003), a frequentist approach still predominates. Thus this monograph mainly adopts a frequentist approach, especially when analysing data under the assumption of MAR. However, all the mixed models we describe when data are MAR (for both continuous and discrete data) can be fitted using Bayesian methods (e.g. `winBUGS`). In this case, all the missing observations are treated as unknown parameters and their posterior distribution sampled from. However, fitting such models in `winBUGS` is both more complex and has more pitfalls than using maximum likelihood methods in `SAS`. Moreover, it does not get around the issue of population averaged versus subject specific parameter estimates for discrete data. Also, the distribution of individual missing observations is not usually of much interest[10], rather the focus is on treatment parameters. Thus, unless we wish to incorporate prior beliefs about certain parameters, a likelihood analysis is in many ways preferable to a Bayesian one. Note though that multiple imputation is essentially a Bayesian method which approximates frequentist inference under certain conditions. Thus, where analysts have prior beliefs about certain parameters, under MAR this can often be most naturally handled using multiple imputation.

[10] In any case, they can be easily approximated from a likelihood analysis.

When analysing data under the MNAR mechanism, it is often useful to bring in expert opinion, for example about likely differences in response between patients who complete and those who do not. This essentially requires a Bayesian approach, and the methods in Chapter 6 reflect this.

Whether Bayesian or frequentist, the methods we discuss in Chapters 3–6 below are centred around the likelihood. Another approach is centred around weighting by the inverse probability of withdrawal. We have considered this elsewhere (Carpenter *et al.*, 2006) and found that simple inverse probability weighting is usually quite inefficient relative to likelihood methods. Methods to improve efficiency are promising, but not yet sufficiently developed to cope with more than a few special situations. We refer to them from time to time, but do not describe them in detail.

1.11 Summary

This Chapter has sought to flesh out the ICH E9 guideline relating to missing data and discuss its implications for trial design, analysis and regulatory guidelines. We have seen that there can be no universal statistical method when data are missing, but there are universal principles which apply to every situation. We have further seen that with missing data we require extra, unavoidable, assumptions to inform the analysis. Such assumptions take the form of specifying the mechanism by which the data become missing, and/or the differences in the distribution of the data between patients who do, and do not, complete.

As when there are no missing data, the aim remains obtaining valid inferences about the intervention effect. With missing data, we must additionally show that inferences about the intervention are robust to different assumptions about the reason for missing data, or patient withdrawal. Such sensitivity analyses inevitably entail additional work at both the design and analysis stage. However, the cost of this additional work is marginal to the overall cost of the trial, and the benefits — especially if missing data and appropriate analysis are considered from the design stage onwards — substantial. A least we know what we have to assume to infer an intervention effect; more often we will be able to infer an intervention effect robust to the missing data. In short, even with missing data, a principled approach leads to valid inferences about treatment effects. It should be adopted as a matter of course.

Chapter 2

A critique of common approaches to missing data

2.1 Introduction

In this Chapter we consider some commonly used approaches to handling missing data in clinical trials. These methods share computational simplicity, but this comes at a price: the resulting analyses and conclusions are often not sensible, or only sensible in particular circumstances or under an extremely restrictive missingness mechanism. However, we saw in Chapter 1 that we can never know the missingness mechanism. In practice, therefore, we advocate using methods that give sensible inferences when the missingness mechanism belongs to the broad class of MAR mechanisms (which includes MCAR mechanisms), before going on to explore the sensitivity of inferences to possible MNAR mechanisms.

EXAMPLE 2.1 *Isolde trial*

Throughout this Chapter, we illustrate the various methods with data from the *Isolde* trial (Burge *et al.*, 2000). 751 patients with chronic obstructive pulmonary disease (COPD) were randomised to receive either 50 mg/day of fluticasone propionate (FP) or an identical placebo. Patients were followed up for 3 years, and their FEV_1 (litres) was recorded every 3 months, although here we only use the 6 monthly measures. Interest focuses on how patients respond to treatment over time; especially in any treatment by time interaction. Patients with COPD are also liable to suffer acute exacerbations, and the number occurring between follow-up visits was also recorded.

Visit	Number of patients attending visit in	
	Placebo arm	FP arm
Baseline	376	374
6 months	298	288
12 months	269	241
18 months	246	222
24 months	235	194
30 months	216	174
36 months	168	141

Table 2.1: *Isolde* trial: Number of patients attending follow-up visits, by treatment group

As Table 2.1 indicates, only 45% of FP patients completed, compared with 38% of placebo

patients. Of these, many had interim missing values. To identify key predictors of withdrawal, we carried out a logistic regression of the probability of completion on the available baseline variables together with the post-randomisation exacerbation rate and rate of change in FEV_1. Table 2.2 shows the results after excluding variables with p-values > 0.06. The effect of age, sex and BMI are all in line with expectations from other studies. After adjusting for these, it is clear that the response to treatment is a key predictor of patient withdrawal. In particular, discussions with the trialists suggested that high exacerbation rates were probably acting as a direct trigger for withdrawal. □

Variable	Odds ratio	(95% CI)	p-value
Exacerbation rate (no./year)	1.51	(1.35, 1.69)	< 0.001
BMI (kg/m^2)	0.95	(0.92, 0.99)	0.025
FEV_1 slope (ml/year)	0.98	(0.96, 0.99)	0.003
Age (years)	1.03	(1.01, 1.06)	0.011
Sex (Male vs Female)	1.51	(0.99, 2.32)	0.057

Table 2.2: *Isolde* trial: Adjusted odds ratios for withdrawal

2.2 Complete cases

Consider a clinical trial where some patients withdraw, so their response is missing. A *complete case* analysis excludes these patients, including only patients who did not withdraw. If the missingness mechanism is MCAR, a complete case analysis is sensible, although it may well not use all the available information in the data. However, if the missingness mechanism is not MCAR, complete case analysis is not sensible.

EXAMPLE 2.1 *Isolde trial (ctd)*

Suppose we wish to estimate the effect of treatment at 3 years. Table 2.3 shows the results of a t-test to estimate this, using data from complete cases only.

There does not appear to be a treatment effect. However, of the 705 patients randomised, only 309 (44%) completed and are included in this analysis. Further, Table 2.2 shows the data from the patients who withdrew is very unlikely MCAR. The unseen data could substantially change the conclusions. A complete case analysis is therefore not sensible. □

Like *Isolde*, many trials follow up patients longitudinally, obtaining several measurements over the course of the trial. Suppose the trial has finished, and that we wish to estimate the treatment effect half way through the follow up. A complete case analysis would only include data from patients who went on to complete the trial. An *available case* analysis (sometimes called an *all observed data* analysis) includes data from all patients who have not withdrawn from the trial at the half way point. If the missingness mechanism is MCAR, both the complete case analysis and the available case analysis will give sensible answers. In this case, the available case analysis

Group	No. of patients	Mean FEV_1 (litres)	SD
Active treatment (FP)	168	1.33	0.46
Placebo	141	1.30	0.49

$t = 0.48$, 307 degrees of freedom, $p = 0.63$
95% CI for difference $(-0.08, 0.13)$ litres

Table 2.3: *Isolde* trial, complete case analysis: t-test of treatment effect 3 years after randomisation

would be preferable, as it will include more patients and so give more precise estimates of treatment effects. However, if the missingness mechanism is not MCAR, then neither method is sensible. Only if data were MCAR in the early part of the trial and not MCAR later would observed case analysis be sensible but complete case analysis not sensible. This is very unlikely in practice.

2.3 Last observation carried forward

Suppose a trial has longitudinal follow up, and that patients withdraw over the course of the follow up. After a patient withdraws, their subsequent responses are missing. Suppose that, for each patient who withdraws, we set their missing responses equal to their last observed response. This is called *Last Observation Carried Forward* (LOCF). If some patients withdraw before the first follow-up visit, then their baseline observation can be carried forward. Using LOCF gives a data set with no missing values, to which the analysis method intended for the fully observed data can be directly applied. We say the missing values have been *imputed using LOCF*. We refer to the assumption that a missing patient's responses are equal to their last observed response as the *LOCF assumption*.

EXAMPLE 2.1 *Isolde (ctd)*

Table 2.4 shows follow-up data from 4 patients. The first completed the trial. The subsequent 3 have had their missing data imputed using LOCF (values shown in italics).

To illustrate the use of LOCF, we impute the missing responses for every patient following their withdrawal, apart from the 134 who withdrew before the first follow-up visit. Figure 2.1 shows the mean FEV_1 at each follow-up visit, by treatment group, using (i) all available data at each follow-up visit and (ii) LOCF to impute the missing data. The LOCF imputed means are similar for the FP arm, but markedly lower for the placebo arm. The exception is the last visit, where LOCF gives a higher mean for the FP arm. Table 2.5 shows a t-test for treatment effect using the LOCF imputed data. In contrast to Table 2.3, the estimated treatment effect is now significant at the 5% level.

However, this raises a number of questions. Under which missing data models is LOCF sen-

Patient	Years of follow-up	FEV_1 (litres) at follow-up visit:					
		6 months	1 year	1.5 years	2 years	2.5 years	3 years
1	3	1.3	1.2	1.0	1.0	1.0	1.1
2	0.5	0.7	*0.7*	*0.7*	*0.7*	*0.7*	*0.7*
3	1	1.7	1.5	*1.5*	*1.5*	*1.5*	*1.5*
4	1.5	0.9	1.0	1.2	*1.2*	*1.2*	*1.2*

Table 2.4: *Isolde* trial: After withdrawal, patients have had their missing data imputed using LOCF (imputed values shown in *italics*)

sible, and are these plausible here? We need to be confident about the answers to these before concluding a treatment effect actually exists. □

Group	No. of patients	Mean FEV_1 (litres)	SD
Active treatment (FP)	316	1.35	0.47
Placebo	301	1.26	0.48

$t = 2.28$, 615 degrees of freedom, $p = 0.02$
95% CI for difference $(0.01, 0.16)$ litres

Table 2.5: *Isolde* study, LOCF imputed data: t-test of treatment effect 3 years after randomisation

LOCF is a popular method for handling missing data. The above example illustrates its simplicity, and it can be argued that much of its popularity is due to this. We now consider whether it is a sensible method.

Two principles emerged in Chapter 1. First, when a patient withdraws, we can rarely hope to recover their missing values. Second, suppose we assume the withdrawn patient's missing data are MAR. Then, suppose we can find a group of patients whose members, prior to the patient withdrawing, share similar responses to the patient who withdrew. Then, at least under the per-protocol hypothesis (§1.8.2), the subsequent responses of this group give an estimate of the likely distribution of the withdrawn patient's missing responses (e.g. Figure 2.2, left panel).

LOCF generally goes against both these principles. It imputes a single value for each missing response. The subsequent analysis gives these imputed responses the same status as actual observed responses. This is unsatisfactory, as a single value is being used as an estimate of a *distribution*. This can only be generally correct in the extremely implausible event that the dis-

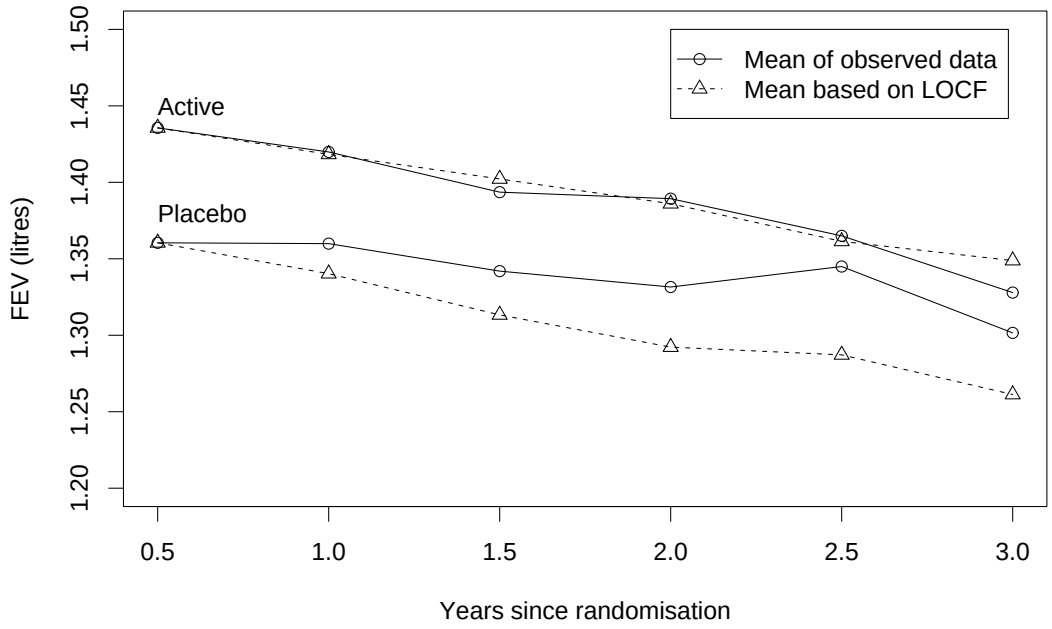

Figure 2.1: *Isolde* trial: mean FEV_1 (litres) at each follow-up visit, by treatment arm. Solid line, means calculated using all available data at each visit. Broken line, means calculated after imputing missing data using LOCF. Note that 134 patients with no readings after baseline are omitted

tribution is degenerate[1]. Such a degenerate distribution will never be implied by the multivariate normal distribution, or any standard distributions.

At best, estimating a distribution by a single value potentially underestimates its variance[2]. This explains why LOCF analyses for the per-protocol hypothesis may underestimate the information lost due to missing data, resulting in standard errors that are too small and confidence intervals that are too narrow.[3] Further, under the per-protocol hypothesis, suppose the withdrawn patient's data are approximately MAR. Then the group of patients who complete, but who share similar characteristics and responses to this patient prior to withdrawal, will usually give a better estimate of the distribution of the missing values than the last response before the patient withdrew. Yet LOCF ignores this information. Thus LOCF is likely to give biased imputations for the missing data leading in turn to biased estimates of treatment effect.

On the other hand, if we focus on the ITT analysis, and believe the distribution around the marginal (*i.e.* treatment group) mean stays the same for patients who withdraw, we should perform a 'principled LOCF' and 'carry forward' this distribution, not the last observation. The

[1] A probability distribution which says a single particular value is certain to occur is termed *degenerate*. With missing data, all we can estimate is the distribution of the missing data given the observed data, under certain assumptions. Imputing a single, worst/best value, usually therefore implicitly assumes a very implausible degenerate distribution for the missing data given the observed data.

[2] As response variability usually increases over time.

[3] Some have described hypotheses where the LOCF analysis has the correct size (Shao and Zhong, 2003), but these are *not* the per-protocol or ITT hypotheses (Carpenter *et al.*, 2004).

one exception is if we are prepared to accept that for each patient who drops out, *before* their last observation, their condition had stabilised — so that the distribution of their responses does not change *at all* for the remainder of the study[4]. Under this strong assumption, the patient's last observation is a genuine observation from their stable response distribution, and could have equally been seen just before withdrawal as at the end of the study. We can therefore use this last observation as the patient's response in the cross sectional analysis of treatment effect at the end of the trial follow-up. However, this corresponds to a very counter-intuitive missingness mechanism. Indeed it is hard to think of why patients would withdraw *after* they had stabilised unless either the protocol were very demanding or they had no expectation that their condition would change whether they were in or out of the trial. We reiterate that in most settings this is very implausible. Patients often change intervention when they withdraw from a trial (the desire to do this may well trigger withdrawal) in the hope of getting a better response.

It is sometimes suggested that, where the focus is on estimated treatment differences at the end of the follow-up, because ITT analyses 'need' a response from each randomised patient, LOCF is appropriate. We disagree. As discussed in §1.8.3, when patients withdraw, and almost certainly change their intervention regime, an ITT analysis needs to estimate the distribution of their unseen response at the end of the trial *under this new regime*. It is highly implausible that this distribution is adequately represented by their last observation.

Defenders of LOCF sometimes argue that it leads to *conservative* estimates of treatment effects. However, it is easy to show that this cannot be true in general (Molenberghs *et al.*, 2004). Rather, the direction of the bias depends on the (unknown) true treatment effect and the missing value mechanism. In general, LOCF is biased even when a complete case analysis is sensible (Molenberghs *et al.*, 2004). If investigators or regulators have strong prior beliefs about the relationship between missing and observed responses, the correct way to allow for these is through a sensitivity analysis, examples of which which we discuss in Chapter 6.

EXAMPLE 2.2 *LOCF is not sensible when data are MCAR*

Consider a hypothetical study where we have a placebo and an active treatment group, both with 100 patients. At the first post-randomisation visit, both groups have a mean FEV_1 of 1.2 litres. At the second, and final, visit, the true mean in active treatment group is 1.5 litres, but that in the placebo arm remains 1.2 litres. However, suppose that 50 of the patients in the active arm withdrew, completely at random.

A complete case analysis is sensible. The mean for the active group, estimated from the 50 patients who complete, is around 1.5 litres; that in the placebo group around 1.2 litres. The estimated treatment effect is 0.3 litres.

Now consider the LOCF analysis. In the active group, we observe 50 patients, with a mean FEV_1 of around 1.5 litres. However, LOCF carries forward the first visit responses of the 50 who withdrew. These are around 1.2 litres. So the LOCF average response at the final visit is around $(50 \times 1.2 + 50 \times 1.5)/100 = 1.35$ litres. As no patients drop out of the placebo arm, the mean response at the final visit is the same under LOCF, around 1.2 litres. So the estimated treatment effect is 0.15 litres.

Thus LOCF is not sensible, even when data are MCAR. Further, it is hard to see how the LOCF analysis is a meaningful sensitivity analysis to the complete case analysis. □

[4]Such an assumption of exchangeability is rarely appropriate for longitudinal data, irrespective of any missing data issues.

In fact, as pointed out by Heyting *et al.* (1993), it is possible to see from the observed data whether LOCF is plausible. We first make the assumption that responses are not missing due to a MNAR missingness mechanism which is totally unrelated to prior responses. In other words we assume that the missingness mechanism is not too far from MAR. Now consider a group of patients with similar measurements. At each observation time, a proportion of them withdraw. As the missingness mechanism is approximately MAR, the unseen responses of these patients are distributed roughly according to the observed responses of patients in the group who have not yet withdrawn. Figure 2.2 illustrates this graphically. Only in the right panel, where *individual patients' responses* are virtually constant, is the LOCF assumption plausible. However, in *both* panels of Figure 2.2 a MAR analysis is sensible. Just because one may use LOCF as a "poor man's" MAR analysis in situations like the right panel is not sufficient to justify it — although it is the probable source of occasional anecdotes that LOCF tends to agree with MAR analysis. In this case, though it would be addressing the per-protocol not ITT hypothesis!

Note too that *it does not follow that if the mean profile is approximately constant, individual patient profiles are approximately constant*. In real life, approximately constant individual patient profiles are rarely seen.

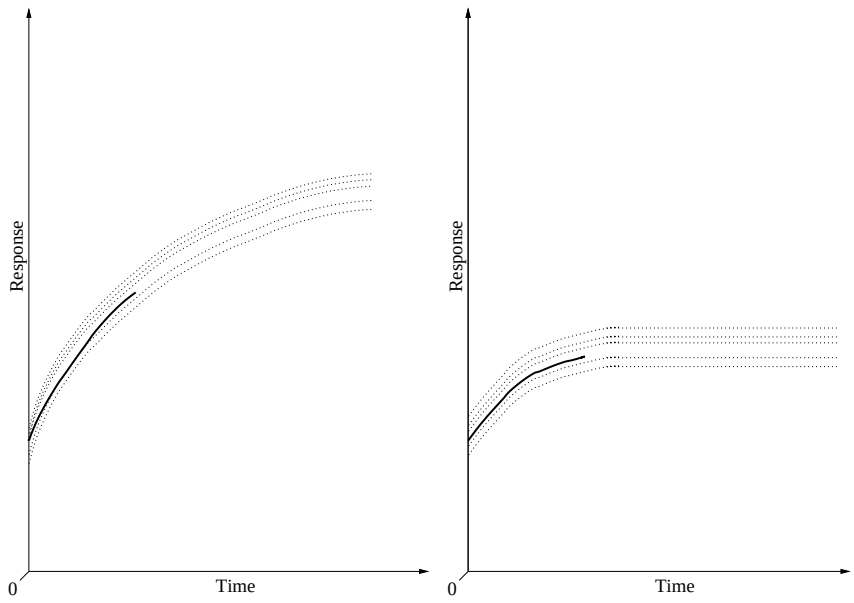

Figure 2.2: Panels show a group of patients with similar responses (dashed lines), one of whom (solid line) drops out. In the left panel, the group responses suggest the LOCF assumption is false. In the right panel, the group responses suggest it is less implausible

Another point sometimes made in favour of LOCF is that if there is no treatment effect it preserves the 'Type I error' (*i.e.* the chance of finding a statistically significant treatment effect when none in fact exists) at 5%. Although in a limited sense this is true (if both groups have identical distributions of response *and* withdrawal) the problem lies rather under the alternative hypothesis. There are many possible patterns of treatment effect and withdrawal for which the power of the LOCF test is the *same* as the test size. Further, merely maintaining the test size is not sufficient to justify a test procedure: if it were we could use the throw of a 20-sided die to calculate a test statistic with perfect nominal 5% type I error! Clearly we also need to consider the behaviour of the statistic under the range of alternative hypotheses. In this the LOCF test falls down badly, as it is unable to detect a wide range of actual treatment effects (Carpenter

et al., 2004). Further, in situations where the treatment effect *can* be detected by the LOCF procedure, the likelihood based analyses described Chapter 3 will usually be more powerful.

In summary, if we really wish to 'carry forward' information after withdrawal, then the appropriate distribution should be carried forward, not the observation. This is not difficult to do, and will give valid inference much more generally than carrying forward the last observation. While one could attempt to delineate specific circumstances where LOCF may perform reasonably, as we are writing generally (as this seems the best way of being relevant to most analyses) we do not attempt to do this.

As LOCF is neither valid under general assumptions nor based on statistical principles, it is not a sensible method, and should not be used. It is therefore unfortunate that Wood *et al.* (2004) found that LOCF is commonly used as a sensitivity analysis when the principal analysis is complete cases. In effect, LOCF is actually just an analysis of each patient's last observed value (so called Last Observation Analysis, LOA). If LOA is really of interest then by definition the last observed measurement needs to be analysed, but in this setting it is equally obvious that the time to this *event* must also be relevant, yet this is almost never considered in such analyses. When estimating treatment effects at the end of a trial, though, LOA is not useful, as it may well reflect misleading transient effects. Although there has been some confusion on this point (Shao and Zhong, 2003), seeing LOCF in this light helps expose its lack of credibility (Carpenter *et al.*, 2004). It is definitely not a sensitivity analysis in the sense described in Chapters 1 and 6. Lastly, Lavori (1992) comprehensively refutes LOCF in the context of psychiatry, and Pocock (1996) reinforces Heyting *et al.* (1992), noting 'it is doubtful whether this [LOCF] actually answers a scientifically relevant question'.

2.4 Missing indicator method

Sometimes, replacing the missing values with a value indicating 'missing data' is proposed. Once this is done the 'full data' analysis can be performed. This is known as the *missing indicator* method. This method can potentially be applied a range of settings. We consider pre-randomisation (i.e. baseline) variables first.

2.4.1 Missing indicator method with pre-randomisation variables

Suppose we wish to adjust an analysis for baseline response. This is often desirable as it gives a more precise estimate of the treatment effect. Now suppose some baseline responses are missing. Although this is unlikely in some settings, in others — for example psychiatric trials — it happens quite often.

First, consider possible reasons for missing baseline variables. By definition, these are measured before randomisation. Thus — assuming randomisation has been adequate — it is implausible that intervention allocation, or response to intervention, are causing missing baseline values.

Thus, in contrast to missing responses, it is often quite plausible that unseen baseline values are MCAR. In this case, analysis restricted to those with observed baseline will be unbiased, though as we shall see some efficiency may be gained by either (i) a full likelihood analysis (Chapter 3) or (ii) by using the missing indicator method and weighting — which we discuss below.

These methods both gain information from the correlation of baseline and post-randomisation variables in the model.

When unseen baseline values are MAR the argument above shows the MAR mechanism must depend on other baseline variables. However, because of randomisation, patients with any combination of missing and observed values must be equally likely in each intervention arm. Indeed, this holds true even if unseen baseline values are NMAR. Thus, as in the previous paragraph, an analysis restricted to those with observed baseline data will still be unbiased, though as before some some efficiency may be gained by a full likelihood analysis or by using the missing indicator method and weighting.

However, some more information can also be obtained by including in the analysis those fully observed baseline variables that are predictive of whether we observe the partially observed baseline variables. Again, this can be done by a direct likelihood analysis, or by using a conditional mean imputation with weighting (Section 2.5).

EXAMPLE 2.3 *Indicator for missing categorical baseline*

Consider hypothetical asthma data in which we wish to estimate the effect of a treatment (vs placebo) on 12 week post-randomisation FEV_1, adjusted for for baseline lung function (categorised as low or high). With no missing baseline, the model is

$$12 \text{ week } FEV_1 = \beta_0 + \beta_1 1[\text{active treatment}] + \beta_2 1[\text{high baseline lung function}], \qquad (2.1)$$

where

$$1[\text{active treatment}] = \begin{cases} 1 & \text{if the patient is on active treatment} \\ 0 & \text{if the patient is on placebo} \end{cases},$$

and so on.

	Variables	
Patient	Baseline lung function 0=low, 1=high	week 12 FEV_1(litres)
1	0	5.67
2	0	4.81
3	0	4.93
4	0	6.21
5	?	6.83
6	1	5.61
7	1	5.45
8	1	4.94
9	?	5.73
10	?	5.58
⋮	⋮	⋮

	Variables	
Patient	Baseline lung function 0=low, 1=high	week 12 FEV_1(litres)
1	0	5.67
2	0	4.81
3	0	4.93
4	0	6.21
5	2	6.83
6	1	5.61
7	1	5.45
8	1	4.94
9	2	5.73
10	2	5.58
⋮	⋮	⋮

Table 2.6: Replacing missing categorical baseline data with an additional category. Left, observed data; right, after replacing missing values with an additional category, '2'

However, suppose that we do not observe baseline lung function on all patients; rather the top of the data looks like the left column of Table 2.6. The missing indicator method replaces the missing baseline values by an additional category, '2', as in the right column, and then fits model (2.1), with the extra category:

$$\begin{aligned}\text{12 week lung function} = \beta_0 + \beta_1 1[\text{active treatment}] + \beta_2 1[\text{high baseline lung function}] \\ + \beta_3 1[\text{missing baseline lung function}].\end{aligned}$$

Assuming the randomisation worked, the reason for the missing baseline does not depend on post-randomisation measures. So the proportion of patients with a '2' should be balanced across the two groups. Thus it is not a confounder, in the epidemiological sense.

Therefore the estimated treatment effect is unbiased, whether baseline lung function is MCAR, MAR (given other fully observed baselines) or NMAR. Whichever mechanism is operating, some additional efficiency may be recovered by weighting, as described below. □

Suppose now both the partially observed baseline and the fully observed response are quantitative. Denote patient i's (baseline, response) data by (X_i, Y_i) and let $R_i = 1$ if X_i is observed, and 0 if X_i is missing. Further let $T_i = 1$ if patient i is on active treatment and $T_i = 0$ if they are on placebo. We wish to estimate the treatment effect adjusted for baseline. Among those with observed baseline, let $\mu_x = \mathrm{E}\,X$ and $\mu_y = \mathrm{E}\,[Y|T = 0]$ (i.e. the mean placebo response). Following White and Thompson (2005), the usual regression model for the observed data is

$$Y_i|X_i, R_i = 1, T_i \sim N[(\mu_y - \beta\,\mu_x) + \beta X_i + \gamma T_i, \sigma^2(1-\rho^2)], \tag{2.2}$$

where $\sigma^2 = \mathrm{Var}\,Y$, and ρ is the correlation of (X, Y). For those with missing baseline, we have

$$Y_i|X_i, R_i = 0, T_i \sim N[(\mu_y - \beta\,\mu_x + \delta) + \gamma T_i, \sigma^2]. \tag{2.3}$$

Here, δ represents the difference in the mean of Y between the groups with observed and missing baseline.

Notice that the treatment effect is the same in both (2.2) and (2.3), suggesting we could fit them together. This can be done if we define two new variables: $M_i = 1 - R_i$ and

$$Z_i = \begin{cases} X_i \text{ if } R_i = 1 \\ \bar{x} \text{ if } R_i = 0 \end{cases}.$$

These two covariates are defined for all patients, regardless of whether x is observed. So we simply regress Y on T, M and Z to obtain the estimated treatment effect.[5] However, as the variance is different in (2.2) and (2.3), the estimated treatment effect will be inefficient. This can be addressed by weighting the regression, with weights

$$w_i = \begin{cases} 1 \text{ if } R_i = 1 \\ (1-\rho^2) \text{ if } R_i = 0 \end{cases},$$

where the sample estimate of ρ from the observed (X_i, Y_i) pairs can be used to calculate the weights.

In fact, White and Thompson (2005) show that with 30% missing baselines and $\rho < 0.6$, the relative efficiency of an unweighted versus a weighted analysis is over 95%. Although ideally

[5] In fact, looking carefully at the model shows we can set Z_i to be any value if baseline is missing — provided it is the same value for each patient.

the uncertainty in estimating the weights should also be taken into account, their empirical study suggests this often makes little difference. If this is a concern, a likelihood analysis with corrected standard errors and degrees of freedom (Kenward and Roger, 1997) should be used, as described in Chapter 3.

	Variables	
Patient	FEV_1(litres)	
	baseline	6 months
1	0.98	1.30
2	?	0.98
3	0.84	0.82
4	?	1.42
5	?	1.31
6	0.65	?
7	?	2.04
8	2.40	2.42
9	1.47	?
10	?	1.63

	Variables		
Patient	Missing base-	FEV_1(litres)	
	line indicator	new baseline	6 months
1	0	0.98	1.30
2	1	999	0.98
3	0	0.84	0.82
4	1	999	1.42
5	1	999	1.31
6	0	0.65	?
7	1	999	2.04
8	0	2.40	2.42
9	0	1.47	?
10	1	999	1.63

Table 2.7: *Isolde* trial. Replacing missing quantitative baseline data with an additional category. Left, original data. Right, after creating a new indicator variable that is '1' if baseline FEV_1 is missing and '0' otherwise, and replacing missing baseline FEV_1 values by '999' (any value gives the same treatment estimate)

EXAMPLE 2.4 *Indicator for missing quantitative baseline*

For this example, we use baseline and 6 month follow-up data from the *Isolde* trial. Excluding one patient with missing baseline, we have 750 patients (376 active drug, 374 placebo) of whom 586 (298 active drug, 288 placebo) have FEV_1 observed at 6 months.

We first analyse the observed data, and then make some of the baseline FEV_1 values MAR given body mass index, and compare analyses of the remaining observations using (i) the remaining observed data, (ii) the missing indicator method and (iii) maximum likelihood.

In the observed data, we see that patients with higher body mass index (BMI) are more likely to be observed. This suggests the following model for observing baseline FEV_1 ($i \in 1,\ldots,750$):

$$\text{logit}\, p_i = \text{logit}\{\Pr(\text{Observe baseline FEV}_{1i})\} = -0.5 + 0.05 \times \text{body mass index}_i. \qquad (2.4)$$

The probability of observing baseline under (2.4) is shown in Figure 2.3.

For each patient we then draw a random number u_i from a $U[0,1]$ distribution and observe their baseline FEV_1 if $u_i \leq p_i$. Table 2.8 shows the data available for analysis after making some baseline FEV_1 missing in this way.

We now estimate the treatment effect at 6 months by (i) fitting an ANCOVA to the original data before any baseline values were made missing (586 patients); (ii) fitting an ANCOVA using data from the 400 patients with no missing data, (iii) using the missing indicator method (with and without weighting) on the $400 + 186 = 586$ patients with at most a missing baseline and

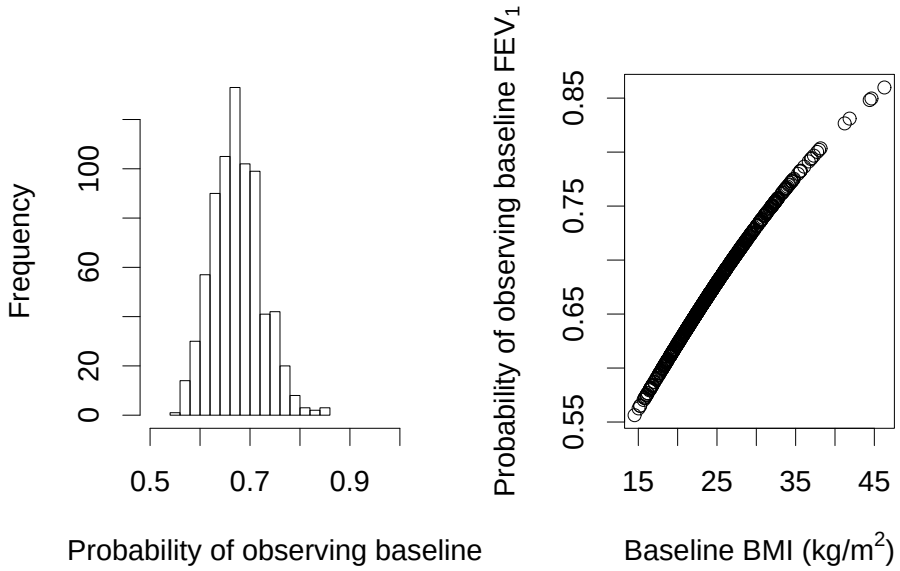

Figure 2.3: Left panel: histogram of probabilities generated by (2.4). Right panel: how these probabilities increase with baseline BMI

Treatment arm	Data available				Total
	Neither	Baseline only	6 month only	Both	
Active	31	47	96	202	376
Placebo	27	59	90	198	374
Total	58	106	186	400	750

Table 2.8: Number of patients with data available after making some baseline values missing using (2.4)

(iv) using direct maximum likelihood on the $586+106=692$ patients with some data. The data arrangement for the missing indicator method is shown in Table 2.7. The maximum likelihood analysis uses the same data arrangement as Table 3.10, and the code of Example 3.4.

Table 2.9 shows the results. Comparing analyses (i) and (ii) notice that, unlike when responses are MAR, as missing baselines are balanced across treatment arms by randomisation they cause remarkably little bias. In this example at least, they also result in little loss of information. Indeed, even the most efficient analysis (iv) barely gets more information here. Analysis (iiia) gives a slightly different point estimate here, but is strikingly inefficient. This is because there is a strong correlation between baseline and 6-month FEV_1 ($\hat{\rho}=0.93$ from the 400 patients with both observed). White and Thompson (2005) show we need to weight in this case; the weighted analysis (iiib) gives a point estimate close to (i) and is almost as efficient as (ii). Lastly, analysis (iv) allows us to include a little extra information from the 106 patients with baseline only. The result is a tiny gain in information — satisfyingly here we do better than the complete data analysis.

In summary, even though a non-trivial number of baseline values were made missing in this

Analysis	Treatment estimate	Standard error	d.f.	t-statistic	p-value
(i) Original data (n=586 observations)	0.0692	0.0135	583	5.12	4.2×10^{-7}
(ii) Missing some baseline values (n=400 observations)	0.0698	0.0164	397	4.26	2.6×10^{-5}
(iiia) Missing indicator analysis (n=586 observations)	0.0635	0.0264	582	2.40	0.017
(iiib) Weighted indicator analysis (n=586 observations)	0.0686	0.0167	582	4.10	4.7×10^{-5}
(iv) Maximum likelihood (n=186 6-month only + n=106 baseline only + n=400 with both)	0.0686	0.0160	433	4.28	2.3×10^{-5}

Table 2.9: Estimated 6 month treatment effect, adjusted for baseline. Row 1: all observed data; row 2: after making baselines missing according to (2.4); rows 3 & 4: missing indicator analysis, and row 5: maximum likelihood analysis using `SAS PROC MIXED` (code of Example 3.4)

analysis, and baseline is highly correlated with response, this causes negligible bias in the treatment estimate (unsurprisingly) but also negligible loss of power. This suggests that analysis (ii) is likely to be sufficient if only a small number of baseline values are missing. The missing indicator method is computationally simple, but needs weighting — at least in this example. In practice, given the cost of the trial relative to the difficulty of weighting, the weighted analysis is always preferable. Whether the maximum likelihood analysis is superior is debatable. If the assumption of bivariate normality is appropriate, we believe the more complex maximum likelihood analysis using the small sample correction to the standard error and degrees of freedom available in `SAS PROC MIXED` has a slight advantage. It also implicitly adjusts for the fact the weights in the weighted indicator analysis are estimated. □

2.4.2 *Other settings*

The missing indicator method can potentially be used for post-randomisation variables, or in non-randomised studies with missing covariates. Unfortunately, in these broader settings it is rarely sensible.

Consider Example 2.3 again. The ‘missing data’ category, ‘2’, does not represent a homogeneous group of patients. In general, a missing indicator category represents an unknown mix of the other categories. Without the protection of randomisation, we have no guarantee that missing values are balanced with respect to intervention/exposure. Therefore, if we wish to include the covariate to adjust for confounding, including the extra ‘missing data’ category can lead to severe bias in estimated intervention/exposure effects, and this bias can be in any direction. The exception is when the mechanism causing baseline to be missing is independent of outcome

given true (but possibly unobserved in some cases) baselines[6]. Thus in general the missing indicator method should be avoided (Greenland and Finkle, 1995).

2.4.3 Summary

When the response is quantitative, the missing indicator method is a convenient way to handle missing baseline values in clinical trials. Although weighting a missing indicator analysis will often give little gain, it is preferable given the relatively little extra effort involved. Provided randomisation has worked, this method can be used whether baseline is MCAR, MAR or MNAR.

Maximum likelihood methods can also be used in this case. If the missing baseline is discrete, then the missing indicator method has the edge, because available maximum likelihood methods usually rely on multivariate normality of the data, and thus model the categorical variable as continuous. However, if baseline is quantitative and we are prepared to assume multivariate normality, arguably maximum likelihood methods have a slight advantage (see Example 2.4). Maximum likelihood analysis also provides protection if randomisation is suspect (see Example 3.4).

When the response is discrete, as far as we are aware the missing indicator method has not been carefully investigated. It appears less attractive, not least because in logistic or Poisson regression introducing the missing indicator variable changes the interpretation of the estimated treatment effect. In practice, with a few missing baselines and a discrete response, omitting patients with missing baselines is probably satisfactory. Multiple imputation could be used, although this is not straightforward with non-monotone missing discrete data (see Chapter 5).

Away from the protection of randomisation (i.e. for response data in trials or in non-randomised studies) the missing indicator method should not be used.

2.5 Marginal and conditional mean imputation

These methods again replace each missing observation by a single value, leading to a 'completed' data set. The originally intended complete data analysis is then used. As above, we call these replaced values *imputed values*.

Marginal mean imputation, as its name suggests, ignores other variables. Missing values are imputed by the average of the observed values for that variable. It is also sometimes referred to as *simple mean* imputation or just *mean* imputation.

Clearly, marginal mean imputation is problematic for categorical variables, where the 'average category' has no meaning. However, the problems go far beyond this. As marginal mean imputation ignores all the other variables in the data set, using it reduces the associations in the data set. Also, imputing all the missing observations to the same value is clearly wrong, and will underestimate the variability in the unseen data. It further goes against the principles of Chapter 1, where we saw the best we could hope for was a good estimate of the *distribution* of the missing observations.

[6] In other words baseline may be MAR or NMAR, but in neither case, given baseline data, must the missingness mechanism additionally depend on outcome.

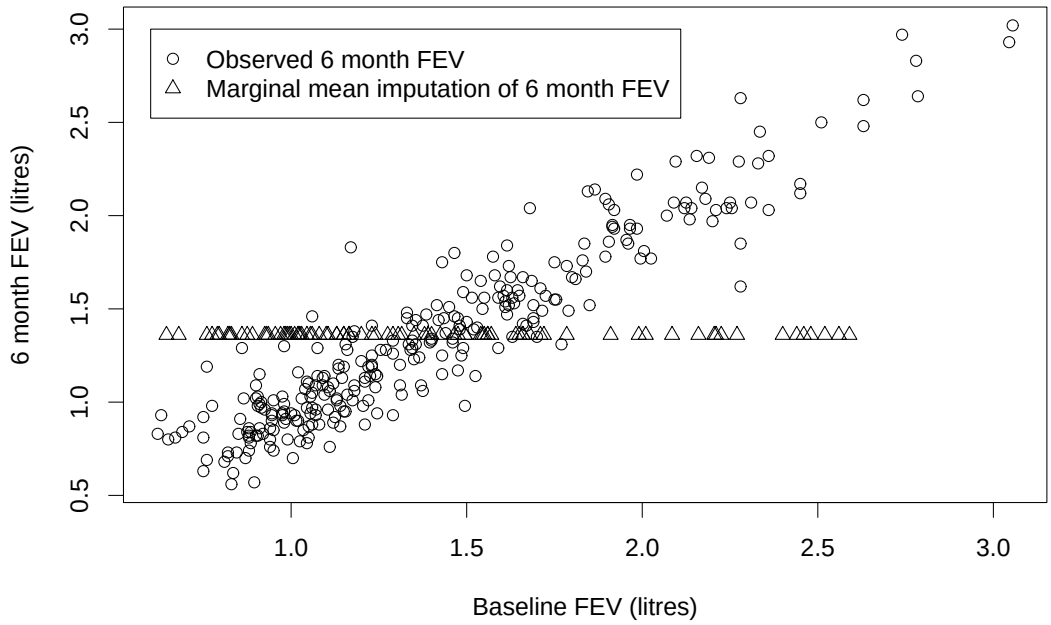

Figure 2.4: *Isolde* trial, placebo arm: plot of baseline FEV_1 against 6 month FEV_1 with missing 6 month FEV_1's imputed by the marginal mean

EXAMPLE 2.1 *Isolde trial (ctd)*

Consider the FEV_1 response 6 months after randomisation for the 375 patients in the placebo group. Eighty seven have a missing response. The mean FEV_1 of the remaining 288 is 1.36 litres. Marginal mean imputation sets each of the missing values equal to 1.36.

Figure 2.4 shows, for the 375 placebo patients, a plot of baseline FEV_1 against 6 month FEV_1. The 87 patients with marginal mean imputed values are shown with a '△'. The shortcomings of marginal mean imputation are immediately obvious. Unless a patient's baseline FEV_1 is close to the mean baseline FEV_1, the marginal mean is very unlikely to be close to the unobserved value. □

We now consider *conditional mean imputation*. In the simplest case, suppose we have one fully observed variable, x, linearly related to the variable with missing data, y. Using the observed pairs, (x_i, y_i), $i \in (1, \ldots, n_1)$, fit the regression of y on x:

$$\text{average value of } y_i = \alpha + \beta x_i, \tag{2.5}$$

obtaining estimates $(\hat{\alpha}, \hat{\beta})$ of (α, β). Then, for the missing y_i's, $i \in (n_1 + 1, \ldots, n)$, impute them as $y_i = \hat{\alpha} + \hat{\beta} x_i$.

EXAMPLE 2.1 *Isolde study (ctd)*

Consider again the baseline (denoted x) and 6 month (denoted y) FEV_1 measurements for the 375 placebo patients. Fitting (2.5) to the 288 patients with both values observed gives

$$\text{average value of } y_i = 0.024 + 0.947 \times x_i. \tag{2.6}$$

We can use this to calculate the mean imputation for each patient with missing 6 month FEV_1.

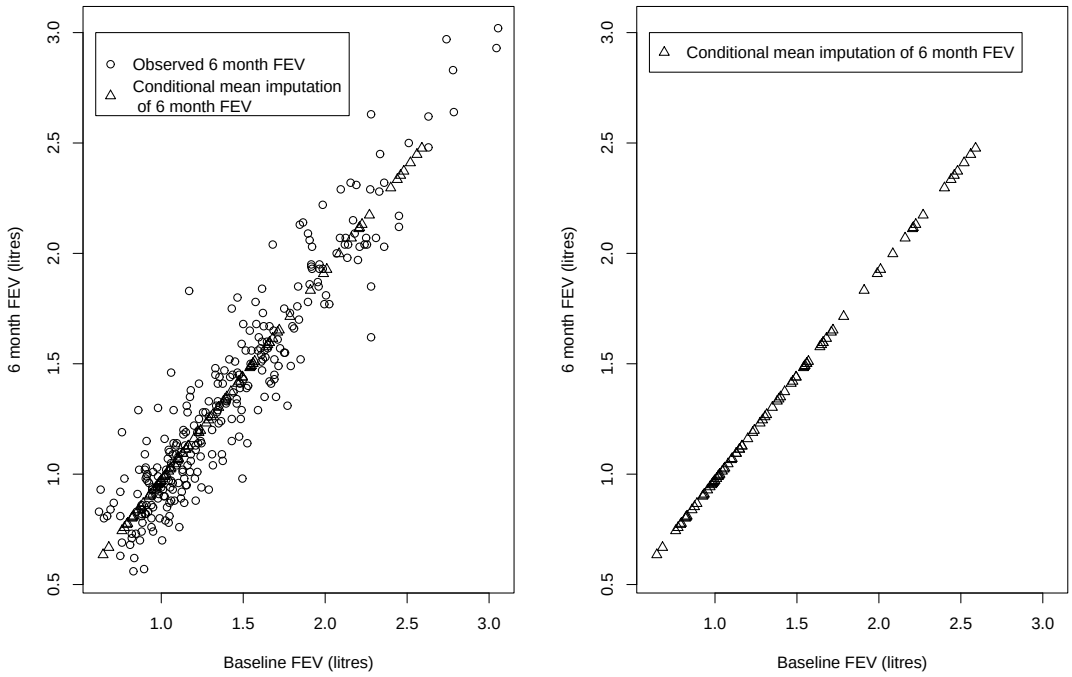

Figure 2.5: *Isolde* trial, placebo arm: plots of baseline FEV_1 against 6 month FEV_1 with missing 6 month FEV_1's imputed by the conditional mean (2.6). Left panel: Observed and imputed data; right panel: imputed data only

For example, a patient with baseline 0.645 litres is imputed a 6 month value of $0.024 + 0.947 \times 0.645 = 0.635$ litres.

Figure 2.5 shows the results of using (2.6) for the 88 placebo patients with missing 6 month FEV_1. Conditional mean imputed values are shown with a '△'. It's clear that the conditional imputations are much more plausible than the marginal imputations. However, as the right panel indicates, they are much less variable than the observed data. Thus, regarding the conditional mean imputations as 'observed data' and using them in an analysis will generally lead to underestimated standard errors, and p-values. □

One setting where the underestimation of the variance with conditional mean imputation may not be such a problem is when we have a quantitative response and missing baseline values. As with the missing indicator method, this is because randomisation ensures that baseline is not a confounder. As with the missing indicator method, we may need to weight as the variance of response given baseline will different in the group whose missing baselines have been replaced by the conditional mean imputations. To estimate the weights:

1. Using data from patients with both baseline and response observed, regress response on baseline and treatment. Note the residual standard error; call this $\hat{r}_b$.

2. Using data from patients with observed response but missing baseline, whose missing baselines are replaced by their conditional mean imputations, regress response on baseline and treatment. Note the residual standard error; call this $\hat{r}_m$.

3. Weights for patients with baseline and response observed are $\hat{r}_m^2$, and those for patients with missing baseline replaced by the conditional mean imputation are $\hat{r}_p^2$. Note $\hat{r}_m^2$ is used in the weights for those *with both baseline and response observed* and vice-versa.

As with the missing indicator method, weighting is probably advisable, if not always necessary.

EXAMPLE 2.4 *Missing baseline values (ctd)*

We revisit Example 2.4, where we artificially made some baseline FEV_1 missing. There we considered the missing indicator method. Now though, we use baseline BMI (which drives the missingness mechanism (2.4)) to conditionally impute missing baseline FEV_1.

The conditional mean imputation model for baseline FEV_1 is

$$\text{Expected baseline FEV}_{1i} = \alpha + \beta \times \text{baseline BMI}_i, \tag{2.7}$$

which we fit to the $i \in (1,\ldots,506)$ patients with baseline FEV_1 and BMI observed. (BMI is observed on all 750 patients). This gives estimates $(\hat{\alpha},\hat{\beta}) = (1.2268, 0.007542)$. For the 186 patients with only 6-month FEV_1 observed we then impute their baseline FEV_1 values as

$$1.2268 + 0.007542 \times \text{baseline BMI}_i.$$

For these data, $\hat{r}_b = 0.1638$ and $\hat{r}_m = 0.5007$, giving weights of 0.2507 for patients with both baseline and response observed and 0.02683 for those with missing baseline.

We then perform three analyses: (a) ANCOVA with conditional mean imputation for missing baseline values; (b) weighted ANCOVA with conditional mean imputation for missing baseline values, and (c) maximum likelihood analysis. Analysis (b) uses the weights calculated above. However, as the variance of the conditional mean imputations of baseline FEV_1's is very small compared to the variance of the observed baseline FEV_1's, normalised weights are virtually identical to those used in the weighted missing indicator method (analysis (iiib) in Table 2.9). Analysis (c) includes baseline BMI, but does not condition the treatment estimates on it. Effectively, it assumes that baseline and 6-month FEV_1 are MAR given fully observed BMI. Such maximum likelihood analyses are discussed in detail Chapter 3; this example uses the data arrangement in Table 3.12 and the code for Example 3.6.

Table 2.10 shows the results. The big differences are in the standard errors; because of the high correlation between baseline and 6-month FEV_1, weighting is essential. The weighted analysis (b) is very similar to the weighted analysis (iiib) in Table 2.9, but the point estimate is fractionally closer to the original data analysis (i) and the standard error is slightly smaller, possibly indicative of a little gain through conditional imputation with missing baselines. Analysis (c) has similar efficiency but a slightly different point estimate. This is probably because it makes the slightly different assumption that *both* 6-month and baseline FEV_1 are MAR given BMI.

The results suggest that if we wish to avoid a maximum likelihood analysis, the weighted missing indicator method — which gives estimates from a single model fit — is likely to be sufficient in practice. □

The conditional imputation above just used one variable. In general, we can use as many variables as we like, and form complicated, possibly non-linear imputation models. These can

Analysis	Treatment estimate	Standard error	d.f.	t	p-value
(a) Conditional imputation (n=586 observations)	0.0641	0.0261	583	2.46	0.0143
(b) Weighted conditional imputation (n=586 observations)	0.0689	0.0160	583	4.30	2.0×10^{-5}
(c) Maximum likelihood (n=186 6-month only + n=106 baseline only + n=400 with both)	0.0680	0.0161	434	4.24	2.7×10^{-5}

Table 2.10: Estimated 6 month treatment effect, adjusted for baseline. Row 1: missing baselines (made missing according to (2.4)) imputed using conditional imputation; row 2: weighted conditional imputation, and row 3: maximum likelihood analysis using `SAS PROC MIXED` (same code as Example 3.6). Note degrees of freedom for the maximum likelihood analysis are from option `ddfm=kr` in `SAS PROC MIXED`

improve the accuracy of the prediction. This is particularly so if the data is assumed MAR and we include in our imputation model all the variables, conditional on which the response is MCAR. In this case, the mean of the imputed data will be sensible. However, we are still imputing single values for the missing data, when as we have seen what we need to do is to estimate the distribution of the missing data.

We therefore need an additional step to correctly estimate the variability of quantities estimated from a 'completed' data set obtained using conditional mean imputation. It is possible, but often non-trivial, to do this on a case-by-case basis. Alternatively, the attraction of Multiple Imputation (MI) (Rubin, 1987) is that it provides a simple, yet both general and sufficient, approach for accounting for the variability of the estimated distribution of the missing data given the observed data.

To do this, MI does not treat any one set of imputations as the true 'unobserved' values of the missing data. Rather, taking into account the uncertainty in estimating both (i) the relationship between y and x variables (*i.e.* $\hat{\alpha},\hat{\beta}$ in (2.5)), and (ii) the residual variability, several 'complete' data sets are imputed. These then provide a convenient representation of the distribution of the missing data given the observed. Each is analysed using the method intended had there been no missing data. Then, in a key second stage, the results are combined in order to give sensible results, which are unbiased and have approximately the correct standard error. Rubin derived rules for doing this, and it is the generality and simplicity of these rules that has placed multiple imputation at the centre of methods for handling missing data.

EXAMPLE 2.1 *Isolde study (ctd)*

We now refine the conditional mean imputations above, to reflect (i) the variability in our estimates $(0.024, 0.947)$ of (α,β) and (ii) the variability of 6 month FEV_1 given baseline FEV_1.

Taking (i), statistical theory shows that (α,β) are normally distributed about $(0.024, 0.947)$,

and gives an estimate of their variance and covariance. We can then draw from this distribution. Statistical theory also shows that 6 month FEV_1 has a conditional normal distribution given baseline FEV_1, and estimates the variance of this distribution. This enables us to address (ii).

An algorithm that uses this information to generate the imputed data sets is described in Chapter 4. Here, we simply note that taking into account (i) and (ii) in generating a series of imputed data sets, gives a representation of the estimated distribution of the missing data given the observed data. Each of these imputed 'complete' data sets can be analysed using the method intended for the fully observed data. The results must then be combined, using Rubin's rules, to give sensible inferences. □

Notice that mean imputation, and its multiple imputation counterpart, can be used at any stage in the longitudinal trial follow-up. Further, the imputation model can include baseline, and other variables, which we do not wish to include in the eventual trial analysis.

2.6 Conclusions

This Chapter has reviewed a number of imputation methods for missing data. With the exception of complete case analysis, these methods impute a single value for the missing data. This is their common weakness, for this alone cannot provide an adequate estimate of the distribution of the missing data given the observed.

On top of this, we saw that LOCF makes a strong, and to us inappropriate, assumption about the expected behaviour of a patient post-withdrawal. Besides almost never being supported by the data, this does not correspond to any meaningful, well-defined, statistical model. Indeed, even under the assumption that data are MCAR, LOCF gives biased treatment estimates and the bias depends on the unknown treatment effect. It is therefore not sensible.

Likewise, the missing category method makes the strong assumption that the true category for all the missing observations is the same. Again, this is most unlikely. Thus, away from the special case of missing baselines in randomised clinical trials, this method generally leads to unpredictable biases in estimated intervention effects.

Apart from complete case analysis, the only method that uses the information in an assumed missingness mechanism is conditional imputation. Even there, we showed that, away from the special case of missing baselines in randomised clinical trials, a single conditional imputation is not sufficient. However, repeated conditional imputations, drawn to reflect the uncertainty in estimating the conditional imputation model and the residual variation, can lead to a sensible representation of the distribution of the missing data given the observed. This leads directly to multiple imputation.

The remainder of this book is concerned with likelihood and multiple imputation based methods for the analysis of partially observed data sets when (a) data are assumed to be MAR (Part II) and (b) data are assumed to be MNAR (Part III).

In Part II, Chapter 3 discusses model based methods for quantitative data, and Chapter 4 multiple imputation for quantitative data. The use of both approaches for discrete data is described in Chapter 5. Throughout, we aim to bring out the relationships between the methods. In particular we outline why, when model based approaches and MI are both applicable, they can be approximately equivalent.

Part II

Chapter 3

MAR Methods for Quantitative Data

3.1 Introduction

In Part I we reviewed the issues raised by missing data in clinical trials and outlined a systematic approach for appropriate analyses. We saw that when data are missing, additional assumptions are needed for an analysis to be sensible. These we summarise in the the missing data model. One way of understanding the implications of a missing data model for the intended full-data analysis is to consider the missingness mechanism it implies. As described in Chapter 1 these can be broadly classified as MCAR, MAR and MNAR. In Chapter 2 we reviewed a number of well known but ad-hoc methods for imputing data, and concluded that none of them generally lead to sensible analysis if data are MAR, although some (e.g. missing indicator) give valid inference for missing baseline data. However, some of them (such as LOCF) are not sensible even if the data are MCAR.

In Part II we focus on methods for sensible analyses when data are MAR. In this Chapter we describe a model based approach for quantitative data. It will become apparent that this approach is much more flexible than sometimes supposed. Chapter 4 describes multiple imputation, and outlines how the approaches described here relate to multiple imputation, particularly as implemented in `SAS`. Then, Chapter 5 describes the analogue of these approaches for discrete response data.

As discussed in Part I, before modelling the data it is important for the statisticians involved to take time (together with the investigators) to understand why observations might be missing, and to uncover any information in the data that helps explain this. As Chapter 1 shows, a sensible MAR analysis must condition or adjust for variables predictive of withdrawal. Some useful exploratory techniques are using t-tests or cross tabulations to investigate the association between baseline variables and withdrawal. It can also be useful to look, at each time point, whether there is a difference in response between patients who do, and do not, return for further visits. More formally, logistic regression and/or survival (withdrawal) analysis can be useful to establish key independent predictors of withdrawal. See, for example, Carpenter *et al.* (2002). In this Chapter, we assume that such exploratory data analysis has been done.

Table 3.1 provides an overview of various missing data scenarios and where they are discussed in this Chapter. For estimating treatment effects, we start with a simple situation with only one follow-up visit, and describe how to handle missing response, missing baseline, and additional variables predictive of withdrawal. We then show how these ideas extend naturally to handle the more usual setting of longitudinal follow-up. Readers unfamiliar with the ideas would benefit from following the development rather than jumping straight to the Section most relevant for their problem. A more detailed description of the ideas underpinning this Chapter, set in a trials context, is given in the Appendix A. This additional material may be useful to develop a deeper intuition.

Data structure	What's missing?	Analysis aim	See
baseline and longitudinal follow-up	some baseline values, some follow-up responses	summarise data by mean, SD, at each follow-up time	§3.3
baseline and 1 follow-up	some follow-up data	estimate treatment effect	§3.4.1, §3.4.2
baseline and 1 follow-up	both baseline and follow-up data	estimate treatment effect	§3.4.3
baseline and 1 follow-up plus additional baseline variable predictive of withdrawal	both baseline and follow-up data	estimate treatment effect	§3.5.1
baseline and 1 follow-up plus post-randomisation variable predictive of withdrawal	baseline and follow-up data	estimate treatment effect	§3.5.2
baseline and longitudinal follow-up	some baseline values, some follow-up responses	estimate treatment effect at final follow-up visit	§3.6
baseline, longitudinal follow-up and additional follow-up data predictive of withdrawal	some baseline values, some follow-up responses	estimate treatment effect at final follow-up visit	§3.6

Table 3.1: Overview of Chapter 3. In each case, we discuss the estimation of treatment effects with and without baseline adjustment

3.2 Some modelling issues

Throughout this Chapter, we model the data with the multivariate normal distribution.[1] This presupposes the response data approximately follow this distribution, or can be transformed to do so. For longitudinal follow-up, it also raises the issue of the appropriate choice of covariance matrix. We advocate in the settings considered here the use of an *unstructured* covariance matrix. This provides a natural way to handle variables predictive of withdrawal on which we do *not* wish to condition (*i.e.* adjust) our treatment estimate. As will become clear by the end of the

[1] One advantage of multiple imputation is that individual variables can be transformed so the joint distribution is approximately multivariate normal, and imputation carried out assuming multivariate normality, before back transforming (see p. 91).

Chapter, a structured covariance matrix would be awkward in this context, as different structures would be required for different parts of the response. Using an unstructured covariance matrix also means analyses are almost equivalent to those using SAS `PROC MI`, as described in Chapter 4. One objection sometimes raised to the use of an unstructured matrix is potential inefficiency. We show in the next Section that the loss of power under an unstructured covariance matrix as compared with a more parsimonious choice is negligible for a final time point analysis with withdrawal.

All the analyses in this Chapter use SAS `PROC MIXED`. We always use the adjustment to the standard errors and degrees of freedom derived by Kenward and Roger (1997). This gives more accurate standard errors when the sample size is small, and corrects the default estimate of the degrees of freedom. One desirable consequence of this correction is, if no data are missing, the degrees of freedom for treatment estimates from SAS `PROC MIXED` will be identical to those from standard analyses, such as t-tests or ANCOVA.

3.2.1 Comparative power under different covariance structures

In advocating the use of the unstructured covariance matrix for analysing repeated measurements data we need to be sure that this is not an excessively inefficient procedure. Here we show that the loss of power for the final time point treatment comparison is negligible when moving from a structured to an unstructured covariance matrix. Such a comparison can only be made meaningfully for tests with the correct nominal size and to ensure this we need each test to be valid under a given *true* covariance matrix. This implies that all structures compared directly must be nested, with the true matrix the most parsimonious. We consider here a set of four such nested structures, with the number of parameters expressed in terms of the number of follow-up times (T):

AR(1)	first order autoregressive, 2 parameters;
ARH(1)	heterogeneous first order autoregressive, $T+1$ parameters;
AD(1)	first order ante-dependence structure, $2T-1$ parameters, and
UN	unstructured, $T(T+1)/2$ parameters.

It is assumed that the data actually follow an AR(1) structure with a correlation of 0.5. To compare the power of the final time point comparison under the different structures we consider the set of alternative hypotheses which generate a power of 80% under the AR(1) structure. The other structures must have less power than this (under the same alternatives): we are interested in the size of these differences. Two values of T are considered, 5 and 10; and four sample sizes (N): 12, 24, 48 and 96. We are assuming that 1/3 of subjects drop out in the following patterns (in sets of 12, '.' implies missing) with withdrawal distributed evenly between the two groups (Table 3.2).

Table 3.3 shows the results, as a percentage, under all combinations of structure, number of times, and number of subjects.

Apart from the very small sample size ($N = 12$) the loss of power is clearly negligible. It should also be noted that the assumption of a true AR(1) structure is unrealistically simple for most settings so the differences will effectively be even smaller in practice than those observed here.

$T = 5$:

Time					
1	2	3	4	5	replication
X	X	X	X	X	8
X	X	X	X	·	1
X	X	X	·	·	1
X	X	·	·	·	1
X	·	·	·	·	1

$T = 10$:

Time										
1	2	3	4	5	6	7	8	9	10	replication
X	X	X	X	X	X	X	X	X	X	8
X	X	X	X	X	X	X	X	X	·	1
X	X	X	X	X	X	X	·	·	·	1
X	X	X	X	X	·	·	·	·	·	1
X	X	X	·	·	·	·	·	·	·	1

Table 3.2: Power calculations: withdrawal pattern in the two treatment groups

		N			
T	Structure	12	24	48	96
5	AR(1)	80	80	80	80
	ARH(1)	69	76	78	79
	AD(1)	68	76	78	79
	UN	68	76	78	79
10	AR(1)	80	80	80	80
	ARH(1)	68	76	78	79
	AD(1)	67	75	78	79
	UN	67	75	78	79

Table 3.3: Estimated power, as a percentage, under all combinations of structure, number of times, and number of subjects

3.3 Summary statistics

Suppose we wish to summarise the quantitative response data in different treatment arms with means and standard deviations (SDs) at baseline and subsequent follow-up visits. If no data were missing, then for each treatment arm in turn, at baseline and subsequent follow-up times, we would calculate the sample mean and SD.

When some of the observations are missing, such marginal estimates are generally biased. We describe how to use a simple model to obtain unbiased estimates for each treatment arm in turn. This assumes missing data are MAR. That is to say, for each patient, we assume their missing data are MCAR conditional on their observed data.

Consider baseline and response data from a single treatment arm, and suppose we wish to estimate the mean and SD at each follow-up time.

3.3.1 Approach

We assume the data can be modelled by the multivariate normal distribution. For illustration, suppose there are baseline and two follow-up times, and write data from patient i, $i \in (1,\ldots,n)$ as (x_i, y_{i1}, y_{i2}). Using *all the observed data*, fit a multivariate normal distribution with an unstructured mean and covariance matrix. The estimated means, variances and covariances are all sensible under MAR.

Specifically, for the subset of patients with complete data, we fit the full distribution,

$$\begin{pmatrix} x_i \\ y_{i1} \\ y_{12} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & & \\ \sigma_{01} & \sigma_1^2 & \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{pmatrix} \right\}.$$

For patients with missing data, we fit the appropriate marginal distribution. For example, if a patient has only baseline observed, it is $x_i \sim N(\mu_0, \sigma_0^2)$. If a patient is missing the first follow-up observation, it is

$$\begin{pmatrix} x_i \\ y_{12} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu_0 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \\ \sigma_{02} & \sigma_2^2 \end{pmatrix} \right\}.$$

Other patterns of observed/missing data follow in the obvious way. For a more detailed argument, see Appendix A.

3.3.2 Further details and examples

In practice, using SAS, or other packages, we do not need to isolate and fit the different marginal models, for the different missing data patterns, separately. We simply put all the observed data into the analysis; the software handles the rest automatically.

As we regard each patient as independent of all the others, the key assumption is that, for each patient, their missing observations are MAR given their observed data. From a theoretical view, there is no need for the MAR mechanism to be the same for different patients, whether or not they have the same pattern of missing data. In summary, the analysis assumes nothing about the withdrawal mechanism for each patient *except it depends, at most, on their observed data*.

Specifically, this analysis is sensible with interim missing data assumed MCAR and after withdrawal, MAR. However, it is also valid if interim missing data are assumed MAR. This raises an interesting issue. For example, at first glance, it is difficult to justify mechanisms whereby baseline is MAR given subsequent responses, for we have the future affecting the past. However, what is actually being argued is that, for patients with these responses, the chance of observing his/her particular baseline was random. So it is not inherently illogical, although it may be practically unlikely. For this reason, MAR is often considered more plausible if missing data are mainly due to withdrawal, with a scattering of interim missing data.

A different issue arises if the data cannot be regarded as multivariate normal. In this case, we need to transform the data. The easiest approach (though it is not always possible) is to find a transformation f such that $f(x)$, $f(y_1)$, *etc.* are approximately normal. Working with the transformed data, we can then use the above approach to find the means and SDs of the transformed data, and back transform using the 'delta' method (see (A.16) for an example of this).

Patients observed at							N (%)
Baseline	0.5 years	1 year	1.5 years	2 years	2.5 years	3 years	
X	X	X	X	X	X	X	95 (25%)
X	X	X	X	X	X		39 (10%)
X	X	X	X	X			19 (5%)
X	X	X	X				24 (6%)
X	X	X					19 (5%)
X	X						42 (11%)
X							74 (20%)
Completers, but with sporadic interim missing data							39 (10%)
Withdraw before end, but with sporadic interim missing data							24 (6%)

Table 3.4: Pattern of missing data in placebo arm of *Isolde* trial. Observed data denoted by 'X'

EXAMPLE 3.1 *Isolde data, placebo arm*

We illustrate this approach with data from the placebo arm of the *Isolde* trial. The missing data pattern is summarised in Table 3.4. Withdrawal predominates, but there are a non-trivial number of interim missing values.

To estimate the means and SDs at each time point, assuming observations are MAR, we fit a 7-dimensional multivariate normal distribution with an unstructured covariance matrix, using `SAS PROC MIXED`. The resulting means and SDs are compared with those from the observed data in Table 3.5. In this example, both methods give similar estimates for the SDs. This is not always the case; sometimes under MAR the observed data underestimates the variance. Comparing the means, the methods give similar results at baseline (when there is virtually no missing data), but diverge later in the trial. The data are clearly not MCAR; further investigation shows patients with low FEV_1 are more likely to withdraw. Therefore sample estimates, which assume missing data are MCAR, are markedly biased. Assuming data are MAR, if patients continued taking placebo as per the protocol, we would expect a reduction in FEV_1 of 190 ml. □

	Mean (SD) of FEV_1 (litres)						
	Baseline	0.5 years	1 year	1.5 years	2 years	2.5 years	3 years
S	1.41 (0.49)	1.36 (0.49)	1.36 (0.50)	1.34 (0.50)	1.33 (0.50)	1.34 (0.54)	1.30 (0.49)
E	1.41 (0.49)	1.35 (0.49)	1.33 (0.49)	1.29 (0.49)	1.27 (0.48)	1.25 (0.52)	1.22 (0.49)

Table 3.5: Isolde trial, placebo arm: mean (SD) FEV_1 (litres), at baseline and follow-up visits. Top row: *S*ample values using all observed data (valid assuming MCAR); bottom row: *E*stimated using joint multivariate normal model (valid assuming MAR)

3.4 Estimating treatment effects when follow-up and/or baseline values are missing

We now consider the simplest realistic problem: two treatment groups, baseline and response measured once at a single follow-up visit. We suppose that all the baseline values are observed, but that some responses are missing. Initially we assume the responses are MCAR given treatment, then MCAR given both treatment and baseline. In each case we describe how to estimate the effect of treatment (i) unadjusted for baseline and (ii) adjusted for baseline.

3.4.1 Follow-up MCAR given treatment

As motivated in Chapter 1 regression analysis gives sensible answers if the response is MCAR given the covariates included. Since we always wish to include treatment, if the response is MCAR given treatment, simply omitting from the analysis all patients with missing responses gives sensible answers. If, further, we wish to adjust the treatment estimate for baseline, we just include this as another covariate.

EXAMPLE 3.2 *Isolde trial: baseline and 6 month response data*

Consider the baseline and 6 month response data, from both treatment arms. We have 750 baseline values, 269/376 active treatment group responses and 240/374 placebo group responses. A logistic regression confirms 1-year response depends on treatment, but not baseline FEV_1. We therefore analyse the data assuming it is MCAR given treatment.

To estimate the effect of treatment, we regress response on treatment for the 509 patients with both observed. The adjusted estimate is obtained from the same data, simply by including baseline as a covariate. Both estimates are shown in Table 3.6; as expected given the strong correlation between baseline and follow-up FEV_1, the conditional analysis has a much reduced standard error. □

Model	Estimated gain in FEV_1 (litres) due to treatment	std. error	p-value
Unadjusted	0.058	0.0429	0.175
Adjusting for baseline	0.074	0.0149	< 0.001

Table 3.6: Estimated effect of treatment, marginal and conditional on baseline, assuming 1-year response is MCAR given treatment

3.4.2 Follow-up MCAR given treatment and baseline

Still considering only missing responses, suppose that the missingness mechanism can depend on both treatment and baseline value. Following the argument in §3.4.1, regressing response on both baseline and treatment, using data from all patients with observed responses, gives a sensible estimate of the treatment effect adjusted for baseline.

Obtaining an estimate unadjusted for baseline in this setting is a little more tricky. As discussed in more detail in Appendix A, implicit in the approach for estimating marginal means and SDs (§3.3), where we used a multivariate response model, is that a patient's missing responses are MCAR given their observed responses. In other words, it is not necessary to have the variables that the missingness mechanism depends on as covariates; they can also be additional responses. Here, therefore, we need to include baseline as a second response, and treatment as a covariate.

Letting $T = 1, 0$ respectively denote the active and placebo treatments, y the response and x the baseline, the model is

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left\{ \begin{pmatrix} \alpha_0 + \alpha_1 T \\ \beta_0 + \beta_1 T \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \right\}, \tag{3.1}$$

where β_1 is the estimated treatment effect marginal to baseline. By analogy with §3.3.1, for patients whose response is missing, we fit (3.1) marginalised over y, *i.e.*

$$x \sim N(\alpha_0 + \alpha_1 T, \sigma_x^2). \tag{3.2}$$

If there were no missing data, and we fitted (3.1), the estimated treatment effect, and corresponding test, would be the same as a t-test using a pooled estimate of variance.

EXAMPLE 3.3 *Isolde data: baseline and 2.5 year follow up*

Two and a half years after randomisation, logistic regression indicates the probability of response depends on both treatment (p=0.002) and, to a lesser extent, baseline (p=0.05). In the placebo group, 173/374 patients remain, while in the active group 216/376 remain.

To estimate the treatment effect conditional on baseline, we use data from the (173+216)=389 patients with observed 2.5y response, adjusting for baseline and treatment. This can be done using standard least squares regression. The first row of Table 3.7 shows the results.

Model	Estimated treatment effect (l)	Standard error	p-value
Conditional on baseline (n=389 patients with complete data)	0.064	0.0214	0.003
Marginal to baseline (n=389 patients with complete data)	0.018	0.0514	0.725
Marginal to baseline (joint model, all observed data)	0.081	0.0391	0.038

Table 3.7: Isolde data: estimates of treatment effect 2.5 years after randomisation, assuming response is MCAR given baseline and treatment

We now use the approach above to estimate the effect of treatment *marginal* to baseline, when data are MCAR given baseline and treatment. Thus we fit model (3.1). To fit this model in `SAS PROC MIXED`, the data need to be arranged as shown in Table 3.8. Note the missing response. In order to fit four means, we need to define both treatment and time to be class variables (factors)

	Variable		
Patient identifier	Time (1=baseline, 2=2.5 years)	Treatment (1=active)	FEV_1 (litres)
1	1	0	0.980
1	2	0	Missing
2	1	1	0.890
2	2	1	0.910
3	1	0	1.770
3	2	0	1.150
⋮	⋮	⋮	⋮

Table 3.8: Isolde trial: arrangement of baseline and 2.5 year response data for estimating treatment effect marginal to baseline. The 2.5 year response data is assumed MCAR given baseline and treatment

with two levels, and fit the interaction between them. `SAS PROC MIXED` code for this model is given in Appendix C. The resulting treatment estimate is shown in the bottom row of Table 3.7. Finally, for comparison, Table 3.7 shows the estimated treatment effect from an OLS regression of response on treatment, omitting baseline. The latter uses data from the 389 patients with both observed. Were the data to be MCAR given treatment alone, this result should be similar to the estimate from the joint model. It is not, because, as we noted at the start of this example, response at 2.5 years is strongly predicted by baseline.

This example underlines the importance of adjusting (either by conditioning or having as an additional response) using all variables predictive of withdrawal. In practice, we usually wish to condition on baseline. However, frequently there will be other variables, predictive of withdrawal, which we do not wish to condition on. This example motivates a general approach to such settings. □

3.4.3 *Missing baseline and follow-up*

We continue to consider the simple setting of two treatment groups, with baseline and a single response. Here we suppose that some patients are missing baseline, while others are missing the response. As usual, we assume the data are MAR. Here, this means that for patients with missing response, this is MCAR given baseline and treatment, while for those with missing baseline, this is MCAR given response. Although theoretically this could also depend on treatment, this is implausible as it would imply a failure of randomisation.

First consider estimating the treatment effect unadjusted for baseline. Model (3.1) enables this to be done directly. Before, when we only considered missing responses, such patients

contributed information through the marginal distribution (3.2). However, the approach extends directly to patients with missing baseline. They contribute information through the marginal distribution $y \sim N(\beta_0 + \beta_1 T, \sigma_y^2)$. Thus, we can use exactly the same data arrangement as in Table 3.8, and exactly the same SAS code as before.

Second, consider the more usual situation where an estimate of treatment effect conditional on baseline is required. We need to fit a joint model to baseline and response, using all the observed data, and use this to estimate the conditional mean of interest. This can be done using the following model (Roger, 2005). The key is to fit the same mean to baseline, x, regardless of treatment group. This is equivalent to (3.1) with $\alpha_1 = 0$, that is modelling placebo patients as

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu_x \\ \mu_y^p \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \right\},$$

and patients receiving intervention as

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu_x \\ \mu_y^t \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \right\}. \tag{3.3}$$

Fitting this model to intervention and treatment group in SAS PROC MIXED gives estimates of all three μ parameters. In particular, as we show in Appendix A.4, estimates of μ_y^t and μ_y^p are conditional on baseline. Thus if we use the LSMEANS option to calculate the estimated intervention effect, $\mu_y^t - \mu_y^p$ and its standard error, this is equivalent to the baseline adjusted estimate when there are no missing data. With some data MAR, each patient contributes information through the appropriate marginal distribution for their observations.

EXAMPLE 3.4 *Isolde study: baseline and 6 month follow-up*

We illustrate this approach with the baseline and 6 month data from the *Isolde* study. Using a different mechanism in the placebo and active arms, we make some of the baseline values MCAR given 6 month response, and compare various analyses.

For the placebo patients with observed 6 month response, let the probability of observing baseline be:

$$\Pr(\text{observe baseline from patient } i) = p_i = \frac{1}{1 + e^{0.4 \times (6 \text{ month FEV}_1)}}. \tag{3.4}$$

The effect of this is shown in Figure 3.1. Then, for each placebo patient, generate a uniform variable on $[0,1]$ and set the baseline response to be missing if $u_i > p_i$. The resulting pattern of missing data is shown in Table 3.9 (right column).

In the active arm, we use the model

$$\Pr(\text{observe baseline from patient } i) = p_i = \frac{1}{1 + e^{-5 \times (6 \text{ month FEV}_1)}}, \tag{3.5}$$

which gives a high probability of baseline values being observed, and use the same method to set baseline observations missing.

Such a differential mechanism between the treatment arms is artificial. However, it serves to indicate that (i) when baseline is MAR, the modelling approach above gives sensible results, and (ii) the modelling approach is valid when different patients have different MAR mechanisms.

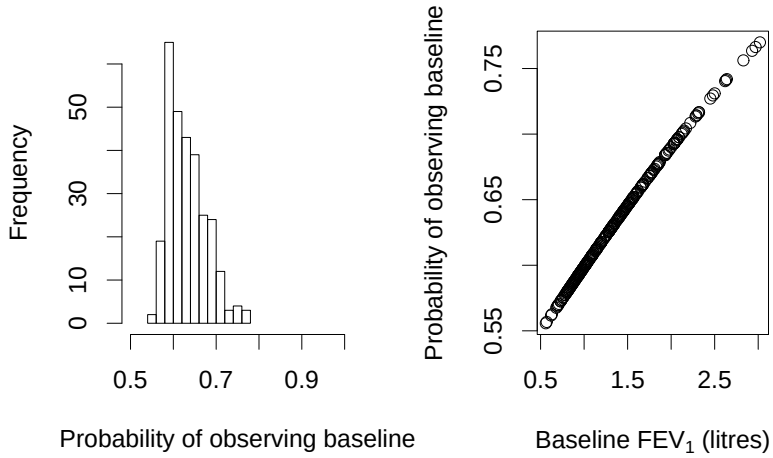

Figure 3.1: Left panel: histogram of probabilities generated by (3.4); right panel: how these probabilities increase with 6-month FEV_1

Note that in contrast to the missing baselines in Chapter 2, as the chance of seeing baseline depends on treatment, the missing indicator method (§2.4) and conditional imputation method (2.5) *will not* work here.

Using this mechanism, the resulting data available for analysis are summarised in Table 3.9. In order to fit this model, we arrange the data as shown in Table 3.10 and use the code shown in Appendix C.

Table 3.11 shows the results of estimating the treatment effect, conditional on baseline, using (i) the original data, before any of the baseline values were artificially set to be missing; (ii) using only patients with both baseline and response observed, and (iii) using all the observed data and model (3.3). Relative to the analysis of the original data, a complete case analysis is markedly biased, so even though the standard error increases, the t-statistic is greater than for the original data! Conversely, the treatment estimate from model is close to the full data estimate, and the increased standard error reflects the information lost by the missing baseline values. □

3.4.4 Summary

1. Variables with missing data need to be included as additional responses in the model alongside — or together with — the primary outcome.
2. If we wish to estimate a treatment effect conditional on such a response variable, that variable needs to have the same mean across treatment groups.
3. To estimate treatment effect marginal to such a response variable, it needs a different mean in each treatment group.

Data available	Treatment arm Active	Placebo
Baseline only	78	86
6 month only	1	187
Both	297	101
Total	376	374

Table 3.9: Number of patients with data available for fitting (3.3)

Variables			
`id`	`treat`	`time`	FEV_1
1	2	1	0.98
1	2	2	1.30
2	1	1	1.46
2	1	2	missing
3	2	1	missing
3	2	2	2.97
⋮	⋮	⋮	⋮

Table 3.10: Data arrangement for fitting model (3.3). Baseline is indicated by `time`=1; follow-up by `time`=2. Placebo patients are `treat`=2

3.5 Missing baseline/follow-up: handling additional covariates predictive of missing data

We now apply the above principles to two situations that frequently occur, namely:

1. The existence of a baseline variable, aside from baseline response and treatment, which is predictive of patient withdrawal. Under MAR we need to include such a variable in the analysis, but we may not wish to condition our estimate of treatment on it.

2. The existence of additional post-randomisation data, predictive of withdrawal. Again, under MAR such information needs to be included in the analysis, but we do not wish to condition our estimate of treatment on it.

EXAMPLE 3.5 *Isolde study: estimating treatment effect at last follow-up*

A natural analysis of the *Isolde* data is to use the 3 year data (final follow-up) to estimate the effect of treatment conditional on baseline (ANCOVA). However, a substantial number of patients withdrew (e.g. Table 3.4). We use logistic regression to relate the probability of a patient completing to baseline values. As might be expected, this shows that patients in the placebo group are much more likely to withdraw. However, in addition to treatment, baseline body mass index (BMI) and age are independently associated with withdrawal. Both these findings make clinical sense. Patients in this trial are chronically ill, and low BMI indicates a more serious phase of the disease, associated with an increased chance of withdrawal. Older patients are also more likely to withdraw.

Model/Data	Estimated treatment effect	Standard error	t-statistic
Original data	0.07	0.014	5.12
Complete data	0.13	0.019	6.79
Model (3.3), all observed data	0.08	0.018	4.46

Table 3.11: Results of various analyses of 6 month and baseline data when some baseline data are made missing

However, there are also some additional post-randomisation variables, one of which is the number of asthma exacerbations a patient has experienced since the last follow-up visit. This is also highly predictive of withdrawal, with a high rate of exacerbation associated with withdrawal.

Assuming data from patients who withdraw is MAR, we wish to obtain the original 'intended' estimate of the effect of treatment, adjusted only for baseline. However, to be valid, our model has to take into account the information about withdrawal in these additional variables. □

We extend our modelling approach above to answer this question. We do this in two stages. First, we show how to include information from baseline variables predictive of withdrawal. Then we consider how to use information from post-randomisation variables predictive of withdrawal, which turns out to be exactly the same problem.

3.5.1 Additional baseline variables predictive of withdrawal

Suppose initially we only have one additional variable to include. The extension to this case is straightforward, following the same pattern adopted when we moved from having baseline as a covariate to having baseline as a response. Recall from §3.4.3 that, when we did this, if we wanted an estimate of treatment conditional on baseline, then we fitted a common mean to baseline across treatment groups. However, when we wanted an estimate marginal to baseline, we fitted different means for each treatment group.

Therefore, here we bring in the additional baseline variable as a response. As we do not (usually) want to condition on it, we fit a separate mean to it, for each treatment group. The following example illustrates how this works. Additional baseline variables predictive of withdrawal are handled in exactly the same way.

EXAMPLE 3.6 *Isolde trial: Including BMI in MAR estimates of treatment*

We have seen that baseline BMI is predictive of patient withdrawal. We now show how to obtain an estimate of treatment conditional on baseline and marginal to BMI. As discussed above, we have to bring baseline BMI in as a response. Further, we have to fit a separate mean for the two treatment groups. Table 3.12 shows the arrangement of the data, and the new treatment variable `newtreat`. The 'response indicator' variable was previously the observation time. Now, values of 1, 2 and 3 indicate respectively baseline BMI, baseline FEV_1 and 3 year FEV_1. The variable `newtreat` has (i) separate values for BMI in each treatment group (1, 2) (ii) a shared value for

baseline FEV_1 (3) and (iii) separate values for post-randomisation FEV_1's (4, 5), depending on treatment arm.

Variables				
patient identifier	treatment group	response indicator	response	`newtreat`
1	2	1	22.3	1
1	2	2	0.98	3
1	2	3	1.30	4
2	1	1	25.3	2
2	1	2	1.46	3
2	1	3	missing	5
3	2	1	23.4	1
3	2	2	missing	3
3	2	3	2.97	4
⋮	⋮	⋮	⋮	⋮

Table 3.12: Data arrangement for estimating treatment effect, assuming missing data are MAR and allowing for the dependence of withdrawal on BMI. Treatment group 2 is the placebo group

We fit this model using `SAS PROC MIXED`, setting `newtreat` as a class (factor) variable. The difference between the estimated means for `newtreat=4` and `newtreat=5` is the estimated treatment effect, conditional on baseline, allowing for the fact that seeing the response can depend on BMI (as well as adjusting for baseline and treatment).

Table 3.13 shows the results of fitting three models. In the first row, we use data from only 308 patients with baseline and 3 year FEV_1 observed. The t-statistic for the baseline adjusted estimate of treatment therefore has 305 degrees of freedom. The second model is fitted in `SAS PROC MIXED`. Again, we use data from the 309 patients with FEV_1 observed at baseline and 3y. However, we also include baseline BMI, for these 309 patients, as a response too. As expected, the results are virtually identical to the first row. The final model is fitted *using exactly the same* `SAS PROC MIXED` *code*, merely including in the data baseline FEV_1 and BMI from the (751-309)=442 patients who withdrew before 3 years. In this example, the treatment estimate is only slightly changed (by about 20% of its standard error). We also have fractionally more information, as indicated by the slight decline in standard error, and the increase in degrees of freedom for the t-statistic. As discussed in §3.2, it is important to estimate these degrees of freedom using the method of Kenward and Roger (1997); the default method in `SAS PROC MIXED` gives quite different values.

Note that for this model `SAS` does not output the regression coefficient for baseline, as we are modelling baseline as a response. If the estimate of the regression coefficient for baseline is desired, this can readily be computed from the `SAS PROC MIXED` output. This is the ratio of the covariance of baseline and response divided by the variance of baseline. Using the data from

Model	Estimated treatment effect	Std. error	d.f.	t-statistic
OLS	0.089	0.0231	305	3.85
SAS 1	0.089	0.0231	306	3.85
SAS 2	0.085	0.0230	310	3.70

Table 3.13: Estimated 3 year treatment effect, adjusted for baseline. Row 1: all patients with observed baseline and 3 month treatment, estimated using OLS; row 2: estimates from SAS PROC MIXED, including BMI as a response, but using same data as row 1; row 3: using all observed data (*i.e.* additional BMI and baseline data for patients who withdraw)

model 2, from SAS PROC MIXED, this is estimated to be $0.2044/0.2264 = 0.9028$. The standard error is estimated as discussed in Appendix A, equation (A.17). Substituting the estimates of $\nu_1, \nu_{12}, \nu_2, \sigma_x^2, \sigma_{xy}^2$, from this analysis gives a variance of

$$\frac{1}{0.2264^2}(0.000302 - 2 \times 0.9028 \times 0.000301 + 0.000334 \times 0.9028^2) = 0.0245^2.$$

The corresponding coefficient and standard error from analysis 1 are 0.9028 and 0.0242^2, showing good agreement (as we have only taken 3 significant figures from the estimates of ν_1 *etc.*) □

Although the effect of adjusting for BMI does not alter the conclusions in this example, our approach shows how a baseline predictor of withdrawal may be included in the model, in such a way that (i) if there are no missing data, the estimated treatment effect is equivalent to what we would get from the intended 'full data' analysis (ii) if there are missing data, the estimate is valid under MAR. The approach is simple and straightforward computationally. It extends readily to include other covariates predictive of withdrawal.

3.5.2 Post-randomisation variables predictive of withdrawal

We could develop the above model to show the inclusion of more than one covariate predictive of withdrawal. Rather than do this, however, we show this is equivalent to extending the model to allow for a post-randomisation variable that is predictive of withdrawal.

The key observation is that our models are all built around assuming the joint distribution of all variables is multivariate normal. We parameterise this distribution in such a way that the effect of treatment adjusted for baseline is readily estimable. However, the multivariate normal distribution does not depend on time in any way. Therefore there is no difference between including baseline variables predictive of withdrawal and post-randomisation values predictive of withdrawal. We simply include them as another response in the model, estimating a separate mean for each treatment group. As in the example above, if the data are complete, our model gives the same estimated treatment effect as an ANCOVA fitted to the response and baseline data. If the data are not complete, our model gives a sensible estimate of the treatment effect under MAR.

Variable	Estimate	Std. Error	t-value	$\Pr(>\|t\|)$
Intercept	0.642	0.233	2.753	0.006
treatment: placebo	-0.047	0.039	-1.205	0.229
baseline FEV_1	-0.034	0.045	-0.750	0.453
BMI	0.014	0.004	3.138	1.78×10^{-3}
mean exacerbation rate	-0.069	0.015	-4.750	2.53×10^{-6}
sex: male	-0.072	0.047	-1.528	0.127
age	-0.004	0.003	-1.365	0.173

Table 3.14: Log odds ratios from a logistic regression of patient withdrawal (0=withdrawal) on baseline variables and exacerbation rate

EXAMPLE 3.7 *Isolde data: adjusting for post-randomisation exacerbation rate*

For each patient, we calculate their mean exacerbation rate as their total number of exacerbations before withdrawal divided by their time on trial. We then included this in a logistic regression of withdrawal before the end of the study, together with treatment, BMI, sex, age and baseline FEV_1. As Table 3.14 shows, patients with high exacerbation rates are much more likely to withdraw; this is therefore an important variable to adjust for in a MAR analysis. Figure 3.2 shows that exacerbation rate is quite non-normal. As our model is multivariate normal, and we are not concerned with interpreting changes in exacerbation rates directly, we use $\sqrt{\text{exacerbation rate}}$, which is more normally distributed (right panel, Figure 3.2).

To fit this model, we follow the previous arrangement of data (Table 3.12), giving Table 3.15. Note the extra response (4), for mean exacerbation rate, which has to have a different mean for the two treatment groups (`newtreat= 6, 7`). Fitting this model, using in turn mean exacerbations and their square root, gives the results in Table 3.16. Comparing the second row with Table 3.13, we see the estimated treatment effect is closer to that obtained before adjusting for BMI, but we now have fractionally more information. The difference between the results in Table 3.16 is very small; the assumption of normality for exacerbation rate does not appear important in this example. Nevertheless, it is preferable, where possible, to transform variables to approximate normality. Finally, the results underline the importance of adjusting for all the key predictors of withdrawal, including post-randomisation ones. □

So far we have obtained estimates of treatment conditional only on baseline. If we wish to condition on variables that we have included as responses, as previously observed, we simply fit a single mean for that variable across treatment groups, following the logic of (3.3). Of course, fully observed baseline values can be included as covariates in the model in the usual way. We will generally need an interaction of such a covariate with `newtreat`. The precise form this takes must be carefully chosen to ensure the desired conditioning on (adjustment for) each of the response variables.

The difference in handling baseline variables predictive of withdrawal, and post-randomisation variables such as exacerbations, is that with baseline variables we can either condition on (*i.e.*

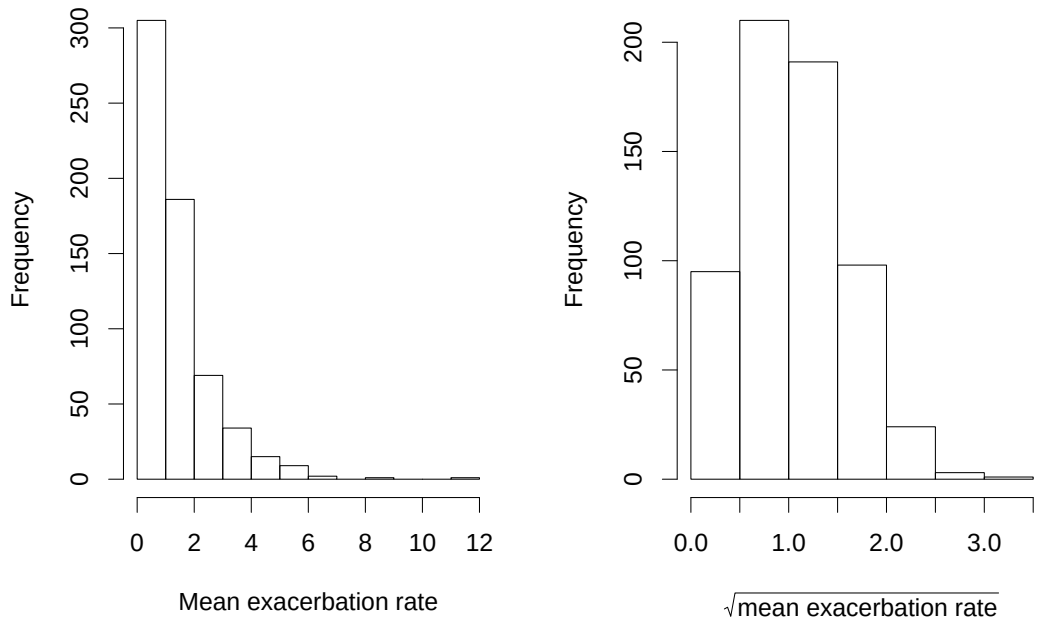

Figure 3.2: Histograms of mean exacerbation rate, and its square-root

adjust for) them, or marginalise over them. Conditioning on them may sometimes provide more precise estimates. Further, usually there are relatively few baseline values missing and the reason for them being missing does not depend on treatment. Thus conditioning on them by including them as a covariate in the analysis (and thus implicitly excluding from the analysis patients with them unobserved) is unlikely to bias the estimate of treatment effect. However, we do not usually wish to condition on post-randomisation responses. E.g. in *Isolde*, we do not want an estimate of the effect of treatment on FEV_1 conditional on exacerbations. Thus, we have to marginalise over post-randomisation responses, by including them in the model as shown above.

3.5.3 Summary

1. Baseline and post-randomisation variables predictive of withdrawal can be handled in the same way.
2. To obtain treatment estimates unadjusted for them, we include them as additional responses with separate means for each treatment group.
3. Usually, baseline variables predictive of withdrawal have relatively few missing values. Therefore, to adjust treatment estimates for them, it is best to include them as covariates.
4. If they have many missing values, include them as an additional response, but with the same mean across treatment arms.

Variables				
patient identifier	treatment group	response indicator	response	`newtreat`
1	2	1	22.3	1
1	2	2	0.98	3
1	2	3	1.30	4
1	2	4	1.3	6
2	1	1	25.3	2
2	1	2	1.46	3
2	1	3	missing	5
2	1	4	0.3	7
3	2	1	23.4	1
3	2	2	missing	3
3	2	3	2.97	4
3	2	4	0.5	6
⋮	⋮	⋮	⋮	⋮

Table 3.15: Data arrangement for estimating treatment effect, extending Table 3.12 to include mean exacerbation rate

3.6 Extension to longitudinal follow-up

We now illustrate how the framework we have set up extends naturally to include more follow-up visits. Although the primary analysis (were no data missing) is often ANCOVA at the last follow-up visit, when patients withdraw before the end of the study there is clearly valuable information in the previous observed values. Specifically, the previous observed values can provide valuable information about the reason for withdrawal, for example those who are performing poorly are more likely to withdrawal.

We can simply include the additional responses in the model, exactly the same way as mean exacerbation rate was included above, extending our `newtreat` variable accordingly. Assuming data are MAR, such an analysis is attractive because

1. if there were no missing observations, the estimated treatment effect, standard error, *etc.* would agree exactly with estimates obtained from an ANCOVA using only data from the final follow-up visit;

2. when end of follow-up data are missing, we make the best use of other measurements, under the missing at random assumption, and

Estimated treatment effect	Std. error	d.f.	t-statistic	p-value
0.08794	0.02306	310	3.81	0.0002
0.08856	0.02301	312	3.85	0.0001

Table 3.16: Estimated 3 year treatment effect, including mean exacerbation in the model. Row 1: results using exacerbation rate; row 2: results using square-root exacerbation rate

3. interim missing observations, and observations missing after patient withdrawal, are both handled without any additional work.

EXAMPLE 3.8 *Isolde study: analysis including all follow-up data*

We illustrate this approach with *Isolde*. Our estimate of interest is the treatment effect at 3 years, adjusted for baseline. We have already seen that withdrawal depends on baseline BMI and exacerbations. Even so, including 6 month FEV_1 in the model shown in Table 3.14 shows the odds of not withdrawing are 3% higher for each 100 ml increase in 6 month FEV_1 (p=0.03).

We therefore extend the analyses above to include data from all follow-up measurements, together with the number of exacerbations reported at each visit. For patients who withdrawal, this variable is sometimes reported after their last per-protocol FEV_1 measure. As a patient's withdrawal is often triggered by an exacerbation leading to a visit to the GP who advises withdrawal, this provides potentially valuable information in support of MAR withdrawal. Note, though, that exacerbations are not that close to normally distributed, even with a square-root transformation. In the absence of joint non-linear and linear modelling (which itself raises further questions about appropriate covariance structures), as before we use the square-root of exacerbation data.

As only one baseline FEV_1 value is, in fact, missing, we use baseline as a covariate in this analysis. This means we do not need a variable that estimates a joint mean for baseline and different means for different treatment groups for other variables. The data are arranged as shown in Table 3.17. Here, `newtreat` shows the order of the response variable: 1 = BMI; 2–7 = number of exacerbations since last follow-up, recorded half yearly from 6 months to 3 years, and 8–13 = FEV_1 recorded at half yearly follow-up visits from 6 months to 3 years. In the `SAS PROC MIXED` analysis, we include `newtreat` as a class variable, together with 'treatment'. We include a `newtreat` 'treatment' interaction to obtain the estimate of treatment at each follow-up visit. We further include a `newtreat` 'baseline' interaction to get a different adjustment for baseline at each time point. Likewise, to adjust for on age and sex, again with a different adjustment at each time point, we fit a `newtreat` 'sex' and `newtreat` 'age' interaction.

Table 3.18 shows the results of fitting three models. Each has all the post-randomisation FEV_1 measures as responses (6 scheduled for each patient). In addition, as BMI is predictive of withdrawal, but we do not wish to condition on it, all the models have BMI as an additional response. Models 1 and 2 are simplified. They replace the 6 follow-up exacerbation readings with a single

			Variables			
patient identifier	sex (1=male)	age (years)	baseline FEV_1 (litres)	treatment (2=placebo)	`newtreat`	response
1	1	63.98	0.98	2	1	22.3
1	1	63.98	0.98	2	2	0
1	1	63.98	0.98	2	3	1
1	1	63.98	0.98	2	4	1
1	1	63.98	0.98	2	5	2
1	1	63.98	0.98	2	6	0
1	1	63.98	0.98	2	7	0
1	1	63.98	0.98	2	8	1.30
1	1	63.98	0.98	2	9	1.15
1	1	63.98	0.98	2	10	1.03
1	1	63.98	0.98	2	11	0.98
1	1	63.98	0.98	2	12	0.96
1	1	63.98	0.98	2	13	1.10
2	2	64.33	0.89	1	1	25.3
2	2	64.33	0.89	1	2	0
2	2	64.33	0.89	1	3	0
2	2	64.33	0.89	1	4	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 3.17: Data arrangement for estimating treatment effect, including longitudinal follow-up data on exacerbations and FEV_1

value, $\sqrt{\text{mean exacerbation rate}}$, derived as described in Example 3.7. Thus models 1 and 2 fit an 8-dimensional normal distribution. Model 1 has the same covariance matrix for both treatment groups (36 parameters); but, as there is some evidence of more variability in the placebo group, model 2 fits a separate covariance matrix for each treatment group (72 parameters).

Model 3 attempts to use all possible information about withdrawal in the exacerbation rate data. Instead of using mean exacerbation rate, we have as 6 additional responses the exacerbation rates observed at each of the clinic visits. Combined with the 6 FEV_1 measures and BMI this makes 13 responses per patient. Again, we fit an unstructured covariance matrix, this time with 91 parameters.

Reassuringly, all three models give similar results, suggesting that, after including mean exacerbation rate in the model, there is little gained by including the number of exacerbations at each visit. As $\sqrt{\text{mean exacerbation rate}}$ is roughly normal, but the number of exacerbations is very far from normal, models 1 and 2 are slightly preferable.

A further question of interest to the investigators was whether the treatment effect could be summarised by a straight line, for each treatment group, and whether the slope of these lines was different. All these patients have declining FEV_1, but a slower rate of decline in the active

Model	Estimated treatment effect (litres increase in FEV_1)	Std. error	df	t-value	p-value
1	0.090	0.02040	436	4.44	$< .0001$
2	0.091	0.02032	432	4.48	$< .0001$
3	0.088	0.02049	435	4.29	$< .0001$

Table 3.18: *Isolde* data: Estimated treatment effect 3 years after randomisation, obtained using the full longitudinal follow-up, adjusting for baseline, sex, age and including BMI and exacerbations as additional responses to obtain valid estimates assuming MAR

treatment would be interesting to know about. Appendix C gives code for fitting this model. The treatment effect can be summarised by a straight line for each group but there is no evidence of a different slope between the groups. The estimated difference in rate of decline is 4 ml per year (se 0.35), giving $t = 1.14$ (435 df) and p=0.25. □

3.7 Inverse probability weighting methods

The methods described in this Chapter are likelihood based. An alternative approach uses inverse probability weighting. For example, suppose that we are interested in the treatment effect at the final time point. Suppose too that given baseline variables (e.g. treatment and baseline response) the probability of withdrawal does not depend on post-randomisation responses.

Then we can proceed as follows:

1. Fit a logistic regression of a binary indicator for the final response being observed (1 if it is observed, 0 if not) on baseline variables. Obtain the predicted probabilities for each patient, p_i, $i \in (1,\ldots,n)$.
2. Using data, and fitted probabilities p_i, from those patients whose final response is observed, fit a weighted regression of response on treatment (adjusting for baseline if desired) weighting by $1/p_i$.

Assuming our weighting model is correct, this will give a consistent estimate of the treatment effect. However, to estimate the standard errors we need to take into account the fact that we have estimated the weights. Such standard errors are not produced automatically by most regression software, which assumes the weights are known without error. One option is to use the bootstrap. The other is to use a sandwich estimator of variance, which incorporates uncertainty due to the weights.

Unfortunately, such inverse probability weighting methods are usually quite inefficient relative to likelihood methods. This is because — unlike in the likelihood methods described above — no use is made of information from patients who have some post-randomisation responses but not the final post-randomisation response. Further, it sometimes happens that some individuals

can get relatively high weights. The conclusions can then be sensitive to these weights, which is a concern if we are not confident they are estimated accurately.

Inverse probability weighting methods can be extended to allow for post-randomisation variables, provided we have no interim missing data (so that patients are observed till they withdraw). Suppose the response is measured at three time points after randomisation, and let R_{i1}, R_{i2}, R_{i3} be indicators for seeing the first, second and third response from patient i.

Then the weight required for an estimate of the treatment effect at the final time point is the inverse of $\Pr(R_{i1} = 1 \text{ and } R_{i2} = 1 \text{ and } R_{i3} = 1 \mid \text{baseline data})$, since all patients with data in the final regression are observed at all three time points. This can be estimated by noting that

$$\begin{aligned}
&\Pr(R_{i1} = 1 \text{ and } R_{i2} = 1 \text{ and } R_{i3} = 1 | \text{baseline data}) \\
&= \Pr(R_{i3} = 1 | R_{i2} = 1 \text{ and } R_{i1} = 1 \text{ and baseline data}) && (3.6) \\
&\quad \times \Pr(R_{i2} = 1 | R_{i3} = 1 \text{ and baseline data}) && (3.7) \\
&\quad \times \Pr(R_{i1} = 1 | \text{baseline data}). && (3.8)
\end{aligned}$$

The first term of the product on the right hand side (*i.e.* adjacent to (3.6)) can be estimated from a logistic regression using all patients whose responses are observed at times 1 and 2. The second term of the product (*i.e.* adjacent to (3.7)) can be estimated from a logistic regression using all patients observed at time 1, and the third term of the product (*i.e.* adjacent to (3.7)) can be estimated from a logistic regression using all the patients with baseline data.

Hence the weights for inverse probability weighting the regression can be obtained. However, all the issues discussed four paragraphs above remain — indeed they are of greater concern here. In practice, the lack of efficiency of IPW methods relative to likelihood methods is the greatest concern. Although methods to improve efficiency and robustness have been proposed, they are not yet sufficiently developed to cope with more than a few special situations (Carpenter *et al.*, 2006). Therefore, we do not pursue inverse probability weighting methods further.

3.8 Summary

1. Under MAR, it is desirable to include all the longitudinal follow-up data on the primary response of interest, up to the end point of the analysis.
2. Likewise, it is desirable to include any post-randomisation data predictive of this response being missing.
3. This can be done by direct extension of the approach in §3.5.1, §3.5.2.

3.9 Conclusions

In this Chapter, we have considered quantitative outcomes. Using the multivariate normal model and assuming unseen data are MAR, we have shown how sensible estimates of treatment effects, adjusted for other variables if desired, in most of the commonly occurring settings (Table 3.1). We have shown how the direct modelling approach using the multivariate normal distribution is far more flexible than it may first appear.

One possible drawback of this approach is that mixed models sometimes encounter convergence difficulties. Such difficulties are more likely to occur as the number of parameters in the mixed model increases, especially if this is combined with non-trivial missing information on some of these parameters. However, we did not encounter any convergence problems with any of the models in this Chapter (we used SAS 9). Note that missing baselines, if handled as discussed in this Chapter, do not cause any particular computational problems as they are effectively included in the model as another outcome. Of course, analyses involving multiple follow-up times are much more processor and memory intensive. Thus we needed to increase the default memory allocation for Example 3.8, but the analysis was still completed in much less time (the most complex model fitted in 10 minutes) than that required for the large MI models described in the next Chapter.

In the next Chapter we will give an intuitive review of multiple imputation (MI). As available in SAS, and most other software, this assumes a multivariate normal distribution for the data. We show how MI can be used in a trials setting, and re-analyse some of the examples in this Chapter.

However, it is worth noting that the imputation model is multivariate normal, as are the models we fit in this Chapter. Thus the treatment effects can always, in principle, be estimated directly through modelling. The advantages of modelling are that it is quicker, involves fewer judgements (such as whether the MCMC sampler has converged) and yields a unique maximum likelihood estimate. By contrast, inferences from MI are slightly different each time. Where the precise answer is critical for decision making, a substantial number of imputations may be necessary to get the Monte-Carlo variability acceptably low.

We therefore advocate direct modelling, if possible. Our hope is that the principles, examples and SAS code in this Chapter will enable readers to set up appropriate models for their data, assuming missing observations are MAR. Of course, there are other issues, such as model checking, we have not discussed here. In particular, it may be that some variables predictive of missingness are not at all normal, like exacerbations in the *Isolde* data. In day-to-day practice, the best one can do is to transform them to approximate normality. The examples considered here suggest this is an acceptable approach.

Chapter 4

Multiple imputation for quantitative data

4.1 Introduction

In Chapter 3 we assumed data were MAR and discussed how, using a direct modelling approach with the multivariate normal distribution, sensible estimates of treatment effects could be obtained when baseline and/or responses were missing. We further showed how the models could be extended to include variables predictive of withdrawal, but for which we do not wish to adjust our estimated treatment effects.

A feature of all the models was that variables with missing values were treated as responses, and the model was parameterised to enable the treatment effect to be estimated. The underlying justification for this approach followed from (A.9) which showed that joint modelling gave sensible estimates provided all the variables predictive of withdrawal were included either as responses (if partially observed) or as either responses or covariates.

Although this approach is flexible, and computationally relatively straightforward, one disadvantage is that the underlying simplicity of the model of interest is often obscured. For example, in the *Isolde* trial, to estimate the effect of treatment at 3 years, we ended up including in the model all responses from randomisation to 3 years, baseline BMI and the post-randomisation exacerbation rate. Our approach also relied critically on the flexibility of the multivariate normal distribution, which has no analogue for discrete data.

At the end of Chapter 2 we discussed conditional mean imputation, and gave an intuitive motivation for multiple imputation. We now develop this further. The following example shows how multiple imputation relates to the modelling approach of the previous Chapter.

EXAMPLE 4.1 *Multivariate normal modelling and multiple imputation*

Consider again the setup in §3.3. There we had data from 375 placebo patients. On 288 of these, baseline (x) and response (y) were observed, while on the remaining $(375-288)=87$ only baseline was observed. Suppose that y is MCAR given x, and we wish to estimate the marginal (*i.e.* unadjusted) mean of y.

Multiple imputation divides this problem into two parts. Broadly, speaking, the first part

1. uses 288 patients with response and baseline observed to obtain a sensible estimate of $[y|x]$, and then

2. for each patient with missing response draws (or imputes) K values from this estimated distribution. Here we take $K=5$. These are put together with the observed data as shown in Table 4.1 in order to make K imputed or 'completed' data sets.

Observed Data	Imputed data set				
	1	2	3	4	5
(0.980, 1.300)	(0.980, 1.300)	(0.980, 1.300)	(0.980, 1.300)	(0.980, 1.300)	(0.980, 1.300)
(1.770, 1.310)	(1.770, 1.310)	(1.770, 1.310)	(1.770, 1.310)	(1.770, 1.310)	(1.770, 1.310)
⋮	⋮	⋮	⋮	⋮	⋮
(0.630, 0.930)	(0.630, 0.930)	(0.630, 0.930)	(0.630, 0.930)	(0.630, 0.930)	(0.630, 0.930)
(0.645, ?)	(0.645, *0.769*)	(0.645, *0.877*)	(0.645, *1.720*)	(0.645, *0.789*)	(0.645, *0.399*)
(0.980, ?)	(0.980, *0.776*)	(0.980, *1.060*)	(0.980, *1.126*)	(0.980, *0.661*)	(0.980, *1.118*)
⋮	⋮	⋮	⋮	⋮	⋮
(1.660, ?)	(1.660, *1.560*)	(1.660, *1.566*)	(1.660, *1.624*)	(1.660, *1.641*)	(1.660, *1.487*)

Table 4.1: *Isolde* data: imputation of 6 month FEV_1 (l) (imputed observations in italics)

In the second step, we

1. perform the analysis we would have carried out, were the data complete, on each of the K imputed data sets. This gives K estimates and their corresponding standard errors.
2. Combine these estimates into an overall estimated effect and standard error, using certain rules.

Conversely, the approach of Chapter 3 seeks to parameterise the model for the missing and observed data in such a way that we obtain the desired estimate of the marginal mean of y as a by-product of estimating $[y|x]$ in the initial step above. Proceeding through the rest of the multiple imputation steps is then unnecessary; we merely arrive at a less precise version of the same estimate. □

Of course, in many cases the approach of Chapter 3 is not possible. This may be because the model of interest and the imputation model assume different distributions (one may be logistic, the other multivariate normal) or it may be because the model of interest includes non-linear adjustments for the covariates. Such situations arise frequently in the survey setting for which MI was originally developed, but less often in modelling quantitative outcomes in clinical trials. Nevertheless, MI is a very useful tool for analysis of clinical trials, especially so for sensitivity analysis.

In the remainder of this Chapter, we give some more details of MI, and then re-visit some of the examples from Chapter 3. Besides illustrating MI, our aim is to describe

1. when MI gives the same answer as a direct modelling approach, and hence is unnecessary, and
2. when the additional flexibility of MI makes it preferable.

In this regard, the first point to note is that MI as implemented in SAS, is based around the multivariate normal model. Thus, using MI in SAS does not avoid assuming variables (such as mean exacerbation rate in Example 3.7) are normally distributed. Of course the user can turn to other implementations of multiple imputation where other distributions are incorporated or program the necessary steps him/herself.[1] The advantage of multiple imputation is that it can, given some approximation, be comparatively straightforward to program. A fairly recent review of implementations of MI is given by Horton and Lipsitz (2001).

4.2 Brief outline of multiple imputation

4.2.1 The MI procedure

We first describe the MI procedure in a very simple setting, that of missing baseline measurements with a completely observed post-randomisation response. We then provide a brief, intuitive, justification for MI. The details on which this is based were developed by Rubin and are brought together in Rubin (1987).

As above, let x_i denote the baseline and y_i response. Suppose x_i is MCAR given response, y_i, so that we can ignore the model for the missingness mechanism. Our model of interest, which we would fit directly if no data were missing, is

$$y_i = \alpha + \beta x_i + \theta t_i + e_i, \quad e_i \overset{iid}{\sim} N(0, \sigma^2_{y|x}), \tag{4.1}$$

where $t_i = 1$ for patients receiving the active treatment and 0 for those on placebo and interest focuses on θ, the baseline adjusted treatment effect. For MI we also need an *imputation* model, which describes the conditional distribution of the potentially missing observations (here x_i) in terms of other variables in the data model (here y_i and t_i) and possibly other variables predictive of missingness, whether pre- or post-randomisation. Here the imputation model is another simple regression model:

$$x_i = \delta + \eta y_i + \xi t_i + f_i, \quad f_i \overset{iid}{\sim} N(0, \sigma^2_{x|y}). \tag{4.2}$$

For the MI procedure we (i) draw, 'appropriately' the missing x_i from (4.2) to make a 'completed' data set and then (ii) repeat this process K times giving K 'completed' data sets. We define 'appropriate' more exactly below, but here we just note two important points.

1. We need estimates of the parameters in the imputation model (4.2). As we saw in Chapter 3, under MAR we can obtain consistent estimators of these using patients with complete data only. Assuming MAR, this holds in all imputation models, however complex, which makes MI very attractive in practice.

2. Proper multiple imputation requires that the imputed data are drawn from the Bayesian posterior distribution, with likelihood defined by (4.2) and uninformative priors for the parameters. This means in practice that each set of imputed data will be based on a different set of model parameters, themselves drawn for each imputation from the Bayesian posterior. We return to this in more detail on p. 80.

[1] This is likely to be necessary for less routine settings, such as zero inflated data, where the imputation model is a mixture model.

A key assumption of MI is that if there were no missing data, the sampling distribution of $\hat{\theta}$ is normal. Thus, if the model of interest is a generalised linear model, we need to work with the parameters in the linear predictor, not their transforms. So for logistic regression we would apply the MI procedure to log-odds ratios, not odds ratios. Here, given sufficient patients, if there were no missing observations, the assumptions of model (4.1) give $\hat{\theta}$ a normal distribution. We can therefore proceed with MI.

Suppose we have appropriately constructed our K 'completed' sets of data. We fit the model of interest, here (4.1), to each in turn, and denote by $\tilde{\theta}_k$ the estimate of θ from the kth completed set, and by V_k the corresponding conventional estimate of the variance of $\tilde{\theta}_k$, calculated as though the 'completed' dataset was actually observed. Thus, in practice, V_k is the usual estimate of the variance of $\tilde{\theta}_k$ produced by the software.

The MI estimator of θ is the average of the individual estimators

$$\tilde{\theta}_{MI} = \frac{1}{K}\sum_{k=1}^{K}\tilde{\theta}_k. \tag{4.3}$$

The estimated variance of this combines *between-* and *within-imputation* variability as follows

$$V_{MI} = \frac{1}{K}\sum_{k=1}^{K}V_k + \left(1+\frac{1}{K}\right)\left(\frac{1}{K-1}\right)\sum_{k=1}^{K}(\tilde{\theta}_k - \tilde{\theta}_{MI})^2. \tag{4.4}$$

This is a a very intuitive expression. On the right hand side, the left hand term,

$$\frac{1}{K}\sum_{k=1}^{K}V_k = W, \text{ say,} \tag{4.5}$$

is an estimate of the variability that would be obtained from a single complete sample. To this is added a term which represents the variability across the imputations, the between-imputation variance:

$$\left(\frac{1}{K-1}\right)\sum_{k=1}^{K}(\tilde{\theta}_k - \tilde{\theta}_{MI})^2 = B, \text{ say.} \tag{4.6}$$

This introduces the increase in variability due to the incompleteness of the data. The multiplier $(1+1/K)$ arises because inference actually conditions on the finite number, K, of imputed data sets used (Rubin, 1987, p. 88 and ff). What is remarkable about (4.3)–(4.6) is that they only involve what are termed 'complete data quantities', that is, statistics which are calculated after fitting the model of interest to each of the 'completed' data sets in turn. Together we refer to (4.3) and (4.4) as called 'Rubin's combination rules' or simply 'Rubin's rules' for MI.

It turns out (Rubin, 1987) that

$$\frac{\tilde{\theta}_{MI} - \theta}{\sqrt{V_{MI}}}$$

has an approximate t_v distribution where

$$v = (K-1)\left(1+\frac{W}{B}\right)^2. \tag{4.7}$$

Under this distribution, hypothesis tests, and confidence intervals, can then be carried out in the usual way.

4.2.2 Quantification of the information lost with missing data

If there were no missing data, and we 'used' multiple imputation, all the 'imputed' data sets would be the same, so there would be no between-imputation variance, *i.e.* B would be equal to zero. In this sense we can say that the percentage increase in variance due to the missing data is

$$\left(\frac{W+B}{W}\right)100\% = \left(1+\frac{B}{W}\right)100\%. \tag{4.8}$$

Alternatively, recalling that 'information' is 1/variance, the percentage of missing information is

$$\left(\frac{W+B}{W}\right)^{-1}100\% = \frac{W}{W+B}100\%. \tag{4.9}$$

In fact, it turns out (Rubin, 1987) that a better estimate of this quantity is

$$\frac{B/W+2/(v+3)}{1+B/W},$$

but we find (4.9) more intuitive and sufficient for practical work.

Given that the unseen data are MAR, this therefore quantifies the lost information. Notice that this quantity can only be calculated after the model of interest has been fitted to each imputed data set. In other words, it depends on what we are estimating. Thus, as noted in Chapter 1, away from a particular data set, quantity of interest and assumed missingness mechanism, the question 'how many observations should be missing before we have to worry' cannot be meaningfully answered.

4.2.3 Justifying the MI procedure

Although the MI procedure is very intuitively appealing and comparatively simple, a rigorous justification of the sampling distribution of the MI estimator from the frequentist viewpoint is surprisingly subtle. For details see Rubin (1987); Wang and Robins (1998); Robins and Wang (2000).

Arguably, MI is at heart a Bayesian procedure; it is certainly easiest to understand from this perspective. The following justification assumes data are MAR. Suppose we have two parameters γ_1, γ_2 about which we wish to draw inferences using the Bayesian paradigm. Given a joint prior distribution for these, the observed data, y, and the data model, we then have a posterior distribution for γ_1 and γ_2 which we write $[\gamma_1, \gamma_2 \mid y]$. Now suppose that our focus is on γ_2, with γ_1 being regarded as a nuisance. The posterior can be partitioned as

$$[\gamma_1, \gamma_2 \mid y] = [\gamma_1 \mid y][\gamma_2 \mid \gamma_1, y,]$$

so that the marginal posterior for γ_2 can be written

$$[\gamma_2 \mid y] = \mathrm{E}_{\gamma_1}([\gamma_2 \mid \gamma_1, y]).$$

In particular, using the standard formulae for conditional expectations and variances, the posterior mean and variance for γ_2 can be expressed

$$\mathrm{E}(\gamma_2 \mid y) = \mathrm{E}_{\gamma_1}\{\mathrm{E}_{\gamma_2}(\gamma_2 \mid \gamma_1, y)\}$$

and

$$\mathrm{V}(\gamma_2 \mid y) = \mathrm{E}_{\gamma_1}\{\mathrm{V}_{\gamma_2}(\gamma_2 \mid \gamma_1, y)\} + \mathrm{V}_{\gamma_1}\{\mathrm{E}_{\gamma_2}(\gamma_2 \mid \gamma_1, y)\}.$$

These can be approximated using empirical moments. Let γ_1^k, $k = 1, \ldots, K$, be draws from the marginal posterior distribution of γ_1, then approximately:

$$\mathrm{E}(\gamma_2 \mid y) \simeq \frac{1}{K}\sum_{k=1}^{K}\{\mathrm{E}_{\gamma_2}(\gamma_2 \mid \gamma_1^k, y)\} = \tilde{\gamma}_2 \text{ say}, \tag{4.10}$$

and

$$\mathrm{V}(\gamma_2 \mid y) = \frac{1}{K}\sum_{k=1}^{K}\{\mathrm{V}_{\gamma_2}(\gamma_2 \mid \gamma_1^k, y)\} + \frac{1}{K-1}\sum_{k=1}^{K}\{\mathrm{E}_{\gamma_2}(\gamma_2 \mid \gamma_1^k, y) - \tilde{\gamma}_2\}^2. \tag{4.11}$$

The parallel between (4.10)–(4.11) and (4.3)–(4.4) is clear. The final link between these expressions and the MI procedure is then to use γ_2 to represent the parameters of the original model of interest and γ_1 to represent the unobserved measurements. Thus, in the example on p. 77, we need to identify γ_2 with the parameters of model (4.1), y with the observed data and γ_1 with the missing baseline data.

We have already noted that MI assumes the full data distribution of the parameter of interest is normal. This is more than sufficient to ensure that the conditional posterior moments for γ_2 can be approximated by maximum likelihood, or equivalent efficient, estimators from the completed data sets.

Notice that the MAR assumption means that we can impute the missing data directly from the marginal posterior of the imputation model. This above argument also indicates why, for Rubin's formulae, (4.3)–(4.4), to hold, we need to use imputation draws from a proper Bayesian posterior. It follows that (4.3) and (4.4) approximate the mean and variance of the posterior distribution in a fully Bayesian analysis. Assuming this is normal, the mean and variance define the distribution uniquely.

However, as we commented above, justifying the *frequentist* tests and confidence intervals rigorously is non-trivial. Wang and Robins (1998); Robins and Wang (2000) give details and in the process compare the implications of proper and improper methods of imputation, providing a very careful analysis of the properties of Rubin's variance expression. This level of detail is probably not necessary for routine practical application of MI however.

4.2.4 Proper imputation

We have seen that for MI to work successfully we need to make *proper* imputations. In particular these need to be (or approximate) draws from a Bayesian posterior distribution in which uncertainty about the parameters in the imputation model is properly represented. This is rarely established rigorously in applications, but typically the following broad guidelines, extracted from Rubin (1987), p. 126–127, are followed:

1. 'Draw imputations following the Bayesian paradigm as repetitions from a Bayesian posterior distribution of the missing values under the chosen models for nonresponse and data, or an approximation to this posterior distribution that incorporates appropriate between-imputation variability'

2. 'Choose models for the data that are appropriate for the complete-data statistics likely to be used — if the model for the data is correct, then the model is appropriate for all complete-data statistics'

In other words point (1) tells us that *both the missing data and parameters of the imputation model* have distributions, and we must not condition on any particular value of the imputation model when drawing our imputations. If we impute using procedures or software for fitting Bayesian models, then it is hard to get this wrong. However, as the example below shows, if we approximate the Bayesian approach, we can go wrong if we are not careful.

EXAMPLE 4.1 *Estimating the mean and variance of a partially observed response (ctd)*

Suppose we have fully observed baseline, x, and partially observed response, y. We wish to estimate μ_y, and its standard error. We have already seen how this can be done by direct modelling. Here we use MI and compare the results. We again use data from the 374 patients in the placebo arm of the *Isolde* trial, of whom 288 have 6 month FEV_1.

First, we need a suitable imputation model *i.e.* in this case for $[y|x]$. We use the linear regression

$$y_i = \alpha + \beta x_i + \varepsilon_i, \;\; \varepsilon_i \sim N(0, \sigma^2_{y|x}).$$

Fitting this model to the 288 patients with both 6 month and baseline FEV_1 gives

$$y_i = 0.024 + 0.947 x_i + \varepsilon_i, \;\; \varepsilon_i \sim N(0, 0.0281). \tag{4.12}$$

Recall that under MAR this is a consistent estimate of the imputation model.

We first describe an 'improper' imputation algorithm, then an approximation to the proper imputation algorithm and finally proper Bayesian imputation. In each case we create $K = 5$ imputations.

Improper imputation algorithm

To impute a 'complete' data set, for each patient, $i = 289, \ldots, 374$, with missing 6 month follow up,

1. draw $\varepsilon_i^\star$ from $N(0, 0.0281)$;
2. impute the 6 month response as $0.024 + 0.947 x_i + e_i^\star$.

We repeat this process 5 times to create our 5 imputed data sets.

Approximate proper imputation algorithm

To impute a 'complete' data set,

1. Draw $\sigma^{2\star}$ from its estimated sampling distribution, $0.0281 \times 286/\chi^2_{286}$
2. Define X as the 288×2 matrix with first column all 1 and second column $x_1, \ldots, x_{288}$. Draw $(\alpha^\star, \beta^\star)$ from their estimated sampling distribution

$$N\left(\begin{pmatrix} 0.024 \\ 0.947 \end{pmatrix}, \sigma^{2\star}(X^T X)^{-1}\right)$$

3. draw $\varepsilon_i^\star$ from $N(0,\sigma^{2\star})$;
4. for each patient, $i = 289,\ldots,374$, with missing 6 month follow up impute the 6 month response as
$$\alpha^\star + 0.947\beta^\star x_i + \varepsilon_i^\star.$$

We repeat the whole process to obtain imputations $2,\ldots,5$.

Bayesian (proper) imputation

1. Choose an uninformative prior distribution for the mean and variance of the bivariate normal distribution of (baseline, 6 month) FEV_1, denoted $[\mu,\Sigma]$ (this implicitly includes the parameters $\alpha,\beta,\sigma^2_{y|x}$).
2. Obtain the posterior distribution,
$$[\mu,\Sigma,\{y_i\}_{i=289}^{374}|\{y_i\}_{i=1}^{288},\{x_i\}_{i=1}^{374}]$$
3. Sample $\mu,\Sigma,\{y_i\}_{i=289}^{374}$ from this distribution; discard μ,Σ and combine $\{y_i\}_{i=289}^{374}$ with the observed data $\{y_i\}_{i=1}^{288},\{x_i\}_{i=1}^{374}$ to impute the 'complete' data set.
4. Repeat the previous step to create imputed data sets 2–5.

Usually, we have to use Markov Chain Monte Carlo (MCMC) methods to obtain and sample from the posterior distribution. For bivariate normal models, this is implemented in `SAS PROC MI`. By default this uses an uninformative prior for μ,Σ, known as the Jeffreys prior (Schafer, 1997, p. 154). The missing values in Table 4.1 have been filled in using `SAS PROC MI`.

Notice that neither the full Bayesian model, nor the maximum likelihood approximation, condition on μ,Σ,σ^2. Rather, new values are drawn for each imputation. This is one key element in making them proper.

Suppose, as in Chapter 3 we are interested in the marginal mean 6 month FEV_1. In this case, the multiple imputation rules are applied as follows. For each completed data set we have
$$\tilde{\theta}_k = \frac{1}{374}\sum_{i=1}^{374} y_i, \quad \text{and} \quad V_k = \frac{1}{373}\sum_{i=1}^{374}(y_i-\tilde{\theta}_k)^2.$$
giving the results in Table 4.2. From (4.3),
$$\tilde{\theta}_{MI} = \frac{1}{5}(1.354+1.362+1.357+1.361+1.354) = 1.358.$$
From (4.5), the within imputation variance is
$$W = \frac{1}{5}(0.02553^2+0.02526^2+0.02543^2+0.02614^2+0.02582^2) = 0.02564^2$$
and from (4.6), the between imputation variance is
$$B = \frac{1}{4}(\{1.354-1.358\}^2+\{1.362-1.358\}^2+\{1.357-1.358\}^2+\{1.361-1.358\}^2 + \{1.354-1.358\}^2) = 0.0037815^2.$$

Imputation, k	$\tilde{\theta}_k$	V_k
1	1.354	0.02553^2
2	1.362	0.02526^2
3	1.357	0.02543^2
4	1.361	0.02614^2
5	1.354	0.02582^2

Table 4.2: Estimates of mean and its variance from each of the 5 imputed data sets

Thus, from (4.4) the estimated variance is

$$0.02564^2 + (1 + 1/4) \times 0.0037815^2 = 0.0260^2.$$

From (4.7), the degrees of freedom for the t-distribution for inference are

$$4 \times \left[1 + \frac{0.02564^2}{1.2 \times 0.0037815^2}\right]^2 = 6181.5.$$

From (4.8), in the increase in variability due to the missing data is estimated as

$$\frac{0.02564^2 + 1.2 \times 0.0037815^2}{0.02564^2} = 1.026,$$

so that the fraction of information missing is $(1 - 1/1.026) \times 100 = 2.5\%$.

To interpret this, suppose, for some constant, A, the variance of $\tilde{\theta}_{MI}$ is approximately A/n. If there are no missing data, $n = 374$. However, with missing data assumed MAR, the relative increase in variance is $1.026 = 374/\tilde{n}$, where $\tilde{n}$ is the new effective sample size. Therefore $\tilde{n} = 365$, so, using multiple imputation, the loss of precision in estimating 6 month FEV_1 is equivalent to roughly $374 - 365 = 9$ patients. Compare the complete case analysis (which needs the stronger MCAR to be sensible), where the loss of information is $374 - 288 = 86$ patients, and the value of using multiple imputation (or equivalently maximum likelihood, as in Chapter 3) becomes clear. Note that the same comparison can be derived from the maximum likelihood (ML) approach described in Chapter 3 by comparing the (Kenward-Roger) estimate of degrees of freedom associated with the two estimates. Fitting the model to observed baseline and 6 month responses by ML gives 368 degrees of freedom for the 6 month mean FEV_1 suggesting the loss of information is equivalent to $374 - 368 = 6$ patients, slightly less than by multiple imputation, as expected.

Table 4.3 compares estimates of the (marginal) mean 6 month response from maximum likelihood and the three imputation methods. They all agree very well here. `SAS PROC MI` agrees well with maximum likelihood, because the between imputation variance is small relative to the within, so a small number of imputations is sufficient. Improper imputation is similar here too, for the same reason.

In practice, we rarely need to combine imputed estimates and variances manually; most software packages for MI perform this automatically. Finally note that the distribution of the multiple

imputation estimator is different from that of the maximum likelihood estimator. Hence the degrees of freedom from (4.7) tend to increase with the number of imputations. Thus they bear no direct relation to the degrees of freedom for an effect in `SAS PROC MIXED`, which, as discussed two paragraphs above, relate to the information available to estimate a parameter.[2] □

Method	Estimate	standard error	95% confidence interval
Maximum likelihood	1.356	0.026	(1.305, 1.407)
`SAS PROC MI`	1.358	0.026	(1.307, 1.409)
Improper MI	1.358	0.026	(1.307, 1.410)
Approximate proper MI	1.354	0.026	(1.304, 1.404)

Table 4.3: Estimates of 6 month marginal mean FEV_1, from various methods. Maximum likelihood (ML) uses the Kenward-Roger estimate of the degrees of freedom. Note ML estimates are slightly different from Table 3.5 because only baseline and 6 month observations are used here

The second aspect of proper imputations concerns the choice of imputation model. Broadly speaking, this must be at least as complex as the substantive model (from which the complete data statistic is calculated). For example, if we are interested in the effect of baseline on response, then baseline must be included in our imputation model. Likewise, if we are interested in a treatment-by-time interaction, then the treatment-by-time interaction must be included in our imputation model. In particular, a treatment effect must always be included in the imputation model, and when imputing a covariate, the response variable must always be included. This makes sense; if we ignore a relationship between the variables when creating imputed data, we cannot hope to accurately estimate that relationship from the imputed data.

4.2.5 Multi-dimensional estimators

So far we have only considered using MI to estimate a single parameter and its variance. However, Rubin's rules generalise in the expected way to parameter vectors. Thus, suppose now that $\tilde{\theta}_k$ is a $p \times 1$ vector of parameter estimates calculated from imputation k, with estimated $p \times p$ variance covariance matrix V_k. Then

$$\begin{aligned} \tilde{\theta}_{MI} &= \frac{1}{K}\sum_{i=1}^{K}\tilde{\theta}_k, \\ W &= \frac{1}{K}\sum_{i=1}^{K}V_k, \\ B &= \frac{1}{K-1}\sum_{i=1}^{K}(\tilde{\theta}_k-\tilde{\theta}_{MI})(\tilde{\theta}_k-\tilde{\theta}_{MI})^T, \end{aligned} \tag{4.13}$$

[2] For a correction to the MI degrees of freedom when the sample size (of the data we intended to collect) is small — the MI degrees of freedom would otherwise tend to increase indefinitely with the number of imputations — see Barnard and Rubin (1999).

where both W and B are now $p \times p$ matrices. By analogy with (4.4),

$$\mathrm{V}(\tilde{\theta}_{MI}) = V_{MI} = W + \left(1 + \frac{1}{K}\right) B.$$

Inferences are now based on the F rather than t distribution through

$$(\tilde{\theta}_{MI} - \theta) V_{MI}^{-1} (\tilde{\theta}_{MI} - \theta)^T \sim F_{p,v},$$

where, as above, p is the dimension of θ, and the residual degrees of freedom v are calculated using formulae given by, for example, Schafer (1997), p. 115. Usually when we are interested in multidimensional problems we will be looking at a *subset* of the parameters in θ, say of dimension $q < p$, and then a similar expression to that above can be constructed in an obvious way.

In practice, handling a multidimensional θ is slightly more difficult than the above analogy would suggest. This is because accurately estimating (4.13) is problematic with small K. Indeed, our estimate will be singular (so not invertible) if $q \geq K$. Clearly considerably larger values of K are required in such settings. Thus, when a multidimensional problem is of interest, one of the attractions of MI — that relatively few imputations are needed — is compromised to some extent.

Various approaches for improving the estimate of Σ_{MI} are described by Schafer (1997), p. 115 onwards. Our advice is, if possible, to avoid these and instead choose K to be about $10 \times q(q+1)/2$. Schafer (1997), Chapter 4, also gives rules for combining p-values and likelihood ratio statistics across imputations. However, none of these have become established like the approach described here (which is based on Wald test statistics).

4.2.6 *Non-parametric multiple imputation*

When there is a substantial quantity of data available, one can consider a non-parametric approximation to the imputation distribution. In a trials context, the idea is as follows. Suppose, as usual, we observe baseline and response on patients $1, \ldots n_1$, but only baseline on patients $n_1 + 1, \ldots n$. For $i = n_1 + 1, \ldots n$,

1. find a group of patients with (i) observed response and (ii) the same, or similar, baseline values to patient i.

2. Draw a response at random from this group, and allocate this to patient i.

In this way we form our first imputed data set; we repeat this process to form subsequent imputed data sets.

This is a version of 'hot-deck' imputation (Reilly, 1993). The attraction is that, compared with a parametric model, fewer constraints are placed on the distribution of the missing data; these now depend on the particular definition of 'similar' used. Some additional resampling is required to provide the necessary Bayesian formulation, and in this context the approach is known as the 'approximate Bayesian bootstrap' (Rubin and Schenker, 1987). The method can be viewed as using a smoothed non-parametric estimate of the imputation distribution. Another approach,

discussed further on p. 115, is the 'propensity score' method for multiple imputation (Little and Rubin, 2002).

These approaches can work well in large survey problems, for which they were developed, especially if the observed values are discrete (e.g. sex, social class) for then the matching step, (1) above, is likely to produce a reasonable size group to draw the missing data from. However, most clinical trials are an order of magnitude smaller than this, which makes non-parametric methods much more variable than their parametric counterparts. We do not consider them further here, or recommend them for routine use in the analysis of trials.

4.2.7 Some further issues

With multiple imputation, a fair amount of discussion focuses around the question of how many imputations are required. Consider the case of a single parameter. First note that, provided we use two or more imputations, our estimates, confidence intervals and inferences will be sensible. This is because, as we noted on p. 78, Rubin's variance formulae *condition* on the number of imputations (Rubin, 1987, p. 88 and ff), so are valid (provided the number of imputations is greater than the number of variance parameters to be estimated). Nevertheless they will be imprecise. Rubin (1987), p. 114, shows that the relative variance of using only K imputations instead of an infinite number is approximately

$$(1+\lambda/K),$$

where λ is the rate of missing information (4.9), expressed as a fraction, not a percentage. So, with 20% missing information, using 2 imputations this gives a standard deviation that is about $\sqrt{1+0.2/2} \approx 1.05$ that obtained with an infinite number of imputations. Hence, as commented by Schafer (1999), 10 imputations is likely to be enough for practical purposes. Nevertheless, as the example below shows, for close agreement with results from mixed modelling rather more may be needed.

Recently, there has been considerable interest in the 'chained equations' method for imputing missing data. Implementations of this vary in certain details, but the following example illustrates the central idea. Suppose we have two variables, x, y, both with some missing values. We proceed as follows:

1. preparatory step: fill in the missing values of x with randomly chosen observed values from x;
2. regress *observed* y on 'filled-in' x, then fill in the missing y values using regression imputation (*i.e.* the 'approximate proper imputation' algorithm on p. 81);
3. regress *observed* x on 'filled in' y, then replace the previously imputed missing x values using regression imputation;
4. regress *observed* y on 'filled in' x, then replace the previously imputed missing y values using regression imputation, and
5. repeat steps 3-4 typically 10 times (to obtain convergence) then a further K times to obtain K imputed data sets.

This approach is described by Raghunathan *et al.* (2001), for data which are approximately *monotone missing*. In clinical trials, a monotone missing pattern corresponds to patients being observed until they withdraw, then no longer (baseline and/or interim missing values are not allowed). An epidemiological example is given by Taylor *et al.* (2002); this approach is due originally to van Buuren *et al.* (1999). These articles show how the approach extends to discrete and other forms of data, basically by replacing the linear regression steps with a more appropriate response model. A version of this approach can be validly used with monotone missing data. When data is non-monotone missing, however, although a computationally attractive approach, it lacks a well established theoretical basis, so that even those who propose it suggest it is used cautiously. Thus, van Buuren *et al.* (2006) write in their abstract

> The theoretical weakness of this approach is that the specified conditional densities can be incompatible, and therefore the stationary distribution...may not exist...[Nevertheless it] appears that, despite the theoretical weaknesses, the actual performance of conditional model specification for multivariate imputation can be quite good, and therefore deserves further study.

As the above quote suggests, the key issue is that the sequence of regression imputations defines many conditional distributions, and these do not guarantee the existence of a unique joint distribution (Gelman and Raghunathan, 2001). As the theoretical basis of this approach is not well understood, and theoretically validated MCMC methods (such as programmed in `SAS PROC MI`) are suitable for most trials settings, we do not think this approach is ready for definitive (e.g. regulatory) analyses at this time. Of course, if we have non-monotone missing discrete data, then the multivariate normal approximation of `SAS PROC MI` may be inappropriate. Some suggestions for this setting are discussed in Chapter 5; results from these methods could usefully be cross-checked with chained equations multiple imputation.

4.3 Application to examples in Chapter 3

In this section we re-analyse some of the examples obtained in Chapter 3 using multiple imputation, and compare the results with those obtained earlier.

EXAMPLE 4.2 *Isolde analyses revisited*

First, we consider again the example on p. 63, where we sought to estimate the 3 year treatment effect conditional on baseline. Recall we found that BMI was predictive of withdrawal, and included it in our analysis as a marginal variable. To make our multiple imputation comparable, we therefore include baseline and follow-up data, together with BMI. `SAS PROC MI` fits a 3-dimensional multivariate normal model to these data using MCMC, and then samples from the posterior of this to impute the missing 3 year responses. It uses the EM algorithm to obtain starting values for the MCMC process. `SAS PROC MI` does not allow any structure in the multivariate normal model. So we cannot have treatment as a covariate. Rather, we must carry out the imputation separately in the two treatment groups. This obviously fits a separate covariance matrix to each treatment group. The effect of this is to allow a different adjustment for BMI in each treatment group.[3]

[3] Arguably this is not ideal, but in most examples it will probably not make much difference.

Method	Estimated treatment effect (l)	Std. error	Model based df	MI df, ν	95% CI	CI length
MI, 5 imputations burn=100; between=200	0.079	0.0199	N/A	22	(0.038, 0.120)	0.082
MI, 5 imputations burn=1000; between=1000	0.087	0.0200	N/A	18	(0.045, 0.130)	0.085
MI, 10 imputations burn=1000; between=1000	0.088	0.0295	N/A	16	(0.026, 0.151)	0.125
MI, 10 imputations burn=5000; between=5000	0.088	0.0204	N/A	39	(0.047, 0.129)	0.082
MI, 50 imputations burn=5000; between=5000	0.088	0.0233	N/A	132	(0.042, 0.134)	0.092
Maximum likelihood, baseline as response	0.085	0.0230	310	N/A	(0.040, 0.130)	0.090
Maximum likelihood, baseline as covariate	0.086	0.0229	305	N/A	(0.041, 0.131)	0.090
MI, 200 imputations burn=5000; between=5000	0.087	0.0229	N/A	579	(0.042, 0.132)	0.090

Table 4.4: *Isolde* data: estimates of effect of treatment at 3 years, adjusting for baseline and marginal to BMI

The results of MI are shown in Table 4.4, together with likelihood analyses using multivariate linear models, which we discuss further below. For a sharper comparison, the seed was the same for each run of `SAS PROC MI`. In each row, we show the values of 'burn' and 'between'. These are, respectively, the number of MCMC updates allowed for the sampler to converge to the posterior, and the number of updates between each saved draw from the posterior, *i.e.* each imputation. The greater the value of 'burn', the more likely the MCMC sampler has converged to the true posterior distribution before the first imputation is drawn. The number of updates between each imputation needs to be sufficient for the imputations to be statistically independent. Again, the greater the value of 'between', the more likely the imputations are independent.

Comparing the estimated treatment effect in the first row, obtained using the default burn in and between imputation values, with the others, we see the MCMC sampler has not converged. The results are not reliable. Note, though, that no warning to this effect is issued by `SAS`!

Increasing the burn in and between imputation values stabilises the estimated treatment effect at 0.088 (rows 2–5). However, with 5 or 10 imputations, the estimated standard error is quite

variable (rows 2–4). Only when the number of imputations is increased to 50 does the standard error appear to settle down at a value slightly above that for the mixed model, which is what we would expect.

Rows 6–7 give the results for two multivariate linear models. The first is that described on p. 63. There we had baseline as a response, as well as BMI. This model allows a different (marginal) adjustment for BMI in each treatment group, but otherwise constrains the variance to be the same in both treatment groups. The second model takes baseline as a covariate, and fits a separate covariance matrix for each treatment group. Thus this model is exactly the same as the imputation model.

We see that both models give similar results, though the second gives a treatment estimate slightly closer to that obtained with MI. Both have standard errors below that for MI with 50 imputations. Finally, the last row of the table gives the results from 200 imputations. In line with theory, these are virtually identical to those from the mixed model above.

We conclude the following. First, for practical applications, the default burn in and the number of between imputation draws should be substantially increased from the `SAS` defaults. The extra computational time this takes is negligible. Secondly, in line with our discussion above, 5–10 imputations is sufficient to get a reasonably accurate answer. However, different runs may still vary noticeably. While this does not invalidate the MI inference, which takes account of this extra uncertainty, it may be considered undesirable, especially if any resulting decisions are finely balanced. Lastly, to be sure of getting results which are virtually equal to those from a likelihood analysis with just a single MI analysis, substantially more imputations are needed. This number would have to be increased by a further order of magnitude if we were looking a joint test rather than a single parameter. □

EXAMPLE 4.3 *Longitudinal Isolde analysis revisited using multiple imputation*

We now revisit the full analysis of the *Isolde* study, as described in Example 3.8, using multiple imputation. Again, baseline FEV_1 is used as a covariate[4]. As before, we include body mass index (BMI) as a marginal predictor of missing FEV_1 (but we do not adjust our treatment estimate for it). Unlike in Example 3.8, in this analysis, we do not adjust the treatment estimate for sex and age.

As in the above example, `SAS PROC MI` is used to create the multiple imputations, implying a joint multivariate normal distribution for the variables, which are baseline FEV_1, the six follow up FEV_1 measurements, baseline BMI and the mean exacerbation rate (see p. 66 for details). As before, the square root transform is used for the mean exacerbation rate.

Table 4.5 shows baseline-adjusted treatment estimates, from various MI runs and maximum likelihood, together with their measures of precision and degrees of freedom. For MI, we ran analyses with 5, 10, 50 and 200 imputations respectively. Each analysis burned in the sampler for 5000 updates and updates a further 5000 times between drawing each imputation. The corresponding maximum likelihood analysis is shown in the final row (with a separate covariance matrix for each treatment arm).

As we are using a joint multivariate normal distribution for the incomplete variables we should expect similar results from the two approaches, with a slight decrease in precision for the multiple imputation based analysis. This is indeed what we see. There is also a suggestion that there

[4]The single patient with a missing baseline FEV_1 is omitted from all analyses.

is some change from 10 to 50 imputations, suggesting that 10 may be too few. The differences in the treatment estimate itself between the multiple imputation and likelihood vanish as the number of imputations increases, in line with theory. The conclusions are identical: adjusting for baseline FEV_1, patients on the active treatment have a FEV_1 that is 86 ml higher after 3 years. □

Method	Estimated treatment effect (l)	Std. error	Model based df	MI df, ν	95% CI	CI length
MI, 5 imputations burn=5000; between=5000	0.093	0.021	N/A	18	(0.049, 0.137)	0.088
MI, 10 imputations burn=5000; between=5000	0.095	0.020	N/A	56	(0.055, 0.134)	0.079
MI, 50 imputations burn=5000; between=5000	0.089	0.022	N/A	190	(0.046, 0.132)	0.086
MI, 200 imputations burn=5000; between=5000	0.090	0.021	N/A	924	(0.049, 0.131)	0.082
Maximum likelihood	0.090	0.020	433	N/A	(0.050, 0.129)	0.079

Table 4.5: *Isolde* data: various estimates of effect of treatment at 3 years, each using all longitudinal follow-up, conditional on baseline and marginal to BMI and mean exacerbation rate

4.4 Conclusions

From the analyses presented above we see little difference between the likelihood and multiple imputation approaches. This similarity is anticipated. As long as we confine ourselves to analyses — by either maximum likelihood or multiple imputation — which are based on the multivariate normal distribution, we know from the underlying theory that the two approaches are closely related, indeed there is a sense is which the multiple imputation is providing an approximation to the likelihood analysis. In these circumstances there is little point in using multiple imputation and we would recommend that the likelihood approach be the first choice.

This does not mean that multiple imputation has no additional role to play. We see four settings where its use might be considered, although the latter three raise further questions that do not, as yet, have definitive answers. The first is when we wish to take advantage of the relative ease of using more complex models with multiple imputation. This is useful when we wish to increase the complexity of the imputation model to make the MAR assumption more plausible. Model criticism is also easier with the imputed data. The second is non-parametric imputation, discussed in §4.2.6. Here questions revolve around the definition of 'similar', the sample

size necessary for the method to work well, and when it is likely to substantially improve the accuracy of inferences.

The third setting is when, because of one or more variables, the joint distribution of the data is not multivariate normal. For example, we have seen with the measure of exacerbation in the *Isolde* example above that the use of a normal distribution is not necessarily convincing for all variables, even after transformation. It has been argued that such approximations are less of an issue when used within multiple imputation rather than likelihood, because the approximation is applied only to the distribution of the imputed values, rather than all the values, observed or not. This appears plausible when we use a form of regression imputation, so fully observed variables are treated as fixed (see also the discussion of the chained equation method in §4.2.7). However, it remains to be demonstrated theoretically. The argument does not apply, though, when using, for example `SAS PROC MI` with the MCMC option (as we have in this Chapter), for then all the variables, fully observed or not, are treated as responses in a multivariate normal distribution. To an extent such concerns are academic; our own experience is that both approaches tend to give very similar results.

The fourth setting is sensitivity analysis. One important strength of multiple imputation is that the imputation model need not be formally consistent with the substantive model. That is, it can contain structure or variables not in the substantive model. In this way it can be used to do analyses that cannot be readily accommodated within a formal likelihood or Bayesian framework. Under the assumption of MAR there is little advantage to be gained in this, provided we can manage to fit an appropriate joint model for the relevant variables. When we move to exploring sensitivity to nonrandom withdrawal however, multiple imputation does provide a simple and attractive route for this. For example, it allows us to formulate appropriate ITT analyses when withdrawal is associated with treatment termination. This is an example of MNAR missing data in the sense that the model for the data for a subject who continues in a trial is not the same as one who drops out. A range of problems like this can be handled by fitting a pattern-mixture model to the observed data, and imputing from models derived from this. We outline such approaches when we look at sensitivity analysis in Chapter 6.

Chapter 5

Discrete data

5.1 Introduction

In this Chapter we consider the analysis of discrete data, assuming missing data are missing at random. We use mixed models and multiple imputation. Once again we have in mind that the primary analysis, if no data were missing, would use data from one of the follow-up visits (usually the last), and estimate the intervention effect adjusted for baseline. However, when data are missing at random, we seek to use additional follow-up data and baseline to obtain sensible estimates of the intervention effect. We show how to do this in a series of steps, which mirror those in Chapter 3 but are more involved because of the additional issues that arise with discrete data.

We develop and illustrate our approach using the logistic model for binary and binomial data. This is a very common example of discrete data and the ideas carry over directly to common models for ordinal data such as the proportional odds, and for nominal data such as the Poisson log-linear.

Unfortunately, the analysis of discrete data is complicated by the fact that there is no natural, analytically tractable, analogue of the multivariate normal distribution. Thus some of the properties of the multivariate normal distribution used in Chapter 3 — for example to estimate an intervention effect conditional on baseline when some baseline values are missing — cannot be used here. There are instead several alternative ways of modelling dependent non-normal data and these lead, in general, to parameter estimates with quite different interpretations. An important manifestation of this is the distinction between so-called *subject-specific* (henceforth SS)[1] and *population-averaged* (henceforth PA) models. Apart from certain special cases, the estimated treatment effects that result from these two modelling approaches are *not* equivalent. This distinction, which does not arise with linear models (*i.e.* in Chapters 3, 4) has important implications for the handling of missing values, so we begin with this.

5.2 Subject-specific versus population averaged models

As the name suggests, a subject-specific (SS) treatment effect represents the effect of treatment on a particular subject's outcome. For example, if the outcome is binary (pain/no pain), the odds of no pain in the placebo group might be 1. Now if a subject in the placebo group has the treatment, their odds of no pain might be 2, giving an odds-ratio of 2. However, there is another way to define an odds-ratio for the effect of treatment. For this we take the proportion of subjects with no pain under placebo, and the associated odds of this, and the corresponding

[1] In our context subject-specific means *patient-specific*, but we refer to the former for consistency with the literature.

proportion and associated odds under treatment. The ratio of these two odds, the *population averaged* (PA) odds-ratio is different from the SS odds-ratio, indeed, we know that it will be closer to one (Zeger and Liang, 1986). Further, because it involves averaging over a specific population, if we move from one population to another, the PA odds-ratio may well change, even when the SS odds-ratio does not.

In other words, a SS model estimates the effect *on a subject* of changing his or her treatment, while the PA model estimates the effect *averaged over a population* of changing the treatment. Under a multivariate normal linear model for quantitative data, the two things are the same. For discrete data they are not.

To illustrate this difference, suppose y_{ij} is the response from subject $i \in (1,\ldots,I)$, at time $j \in (1,\ldots,J)$. Let $T_i = 1$ if subject i is treated, and 0 otherwise. Table 5.1 contrasts a simple model for continuous y_{ij} with a logistic model for binary y_{ij}. For the normal model (left column), the expected SS treatment effect is equal to the PA treatment effect. This is because the response, y_{ij}, is not a function of $\alpha + \beta T_i + u_i + e_{ij}$, but is exactly equal to it. This means that SS and PA estimates are always the same, which is why we did not distinguish between them in Chapter 3.

By contrast, for the logistic model (right column) this is not the case, because the response is now a non-linear function of $\alpha + \beta T_i + u_i$. Thus the SS and PA estimates of treatment only agree if all $u_i = 0$, *i.e.* $\sigma_u^2 = 0$. As this is a consequence of the non-linear relationship between the covariates and expected value of the response, this result applies to almost all models for discrete data using non-linear relationships between response and covariates (e.g. logistic, complementary log-log, probit). An exception is models for count data using a log link (Zeger and Liang, 1986; Zeger *et al.*, 1988).

A further complication for analysing discrete data is that different statistical procedures are commonly used for the two types of model. SS models are typically expressed in the form of Generalised Linear Mixed Models (GLMMs) a natural extension of the class of likelihood-based random effects models used in Chapter 3. There we saw that likelihood methods are relatively convenient to use and, provided the appropriate variables are included in the model, they give sensible parameter estimates when data are MAR. However, for PA models, likelihood methods are rather inconvenient and usually non-likelihood methods, called Generalised Estimating Equations (GEEs) are used (Liang and Zeger, 1986). Roughly speaking, GEEs estimate the parameter values to match the mean and variance of the observed data. To do this, a working assumption is usually made about the variance/covariance structure. For example, we may assume independence or a auto-regressive order 1 correlation structure. It turns out that even if this working assumption is wrong, the estimates will still be consistent (although they may be somewhat inefficient).

The standard errors of these estimates are then subsequently calculated in a way that adjusts for any incorrect working assumptions about the variance/covariance structure. These so-called 'robust' or 'empirical' standard errors are obtained from an empirical estimate, known as a *sandwich estimator*, of the covariance matrix. For a comprehensive discussion of the details of these types of models see, e.g. Diggle *et al.* (2002). Unfortunately, because of their non-likelihood nature, such estimating equation methods raise additional issues when data are missing. We return to this in §5.2.3.

Normal response	Binary response, logistic model
Model: $y_{ij} = \alpha + \beta T_i + u_i + e_{ij}$, $u_i \sim N(0, \sigma_u^2)$, $e_{ij} \sim N(0, \sigma_e^2)$, u_i independent of e_{ij}	*Model:* $\text{logit}\{\Pr(y_{ij} = 1)\} = \alpha + \beta T_i + u_i$, $u_i \sim N(0, \sigma_u^2)$
Expected value of response given subject-specific u_i: $\mathrm{E}[y_{ij} \mid T_i = 1, u_i] = \alpha + \beta + u_i$	*Expected value of response given subject-specific u_i:* $\mathrm{E}[\Pr(y_{ij} = 1) \mid T_i = 1, u_i]$ $= 1/\{1 + \exp[-(\alpha + \beta + u_i)]\}$ $= \text{expit}(\alpha + \beta + u_i)$, say†
$\mathrm{E}[y_{ij} \mid T_i = 0, u_i] = \alpha + u_i$	$\mathrm{E}[\Pr(y_{ij} = 1) \mid T_i = 0, u_i]$ $= 1/\{1 + \exp[-(\alpha + u_i)]\}$ $= \text{expit}(\alpha + u_i)$, say.
Expected subject-specific treatment effect $\mathrm{E}[y_{ij} \mid T_i = 1, u_i] - \mathrm{E}[y_{ij} \mid T_i = 0, u_i]$ $= \beta$	*Expected subject-specific risk-difference due to treatment* $\mathrm{E}[\Pr(y_{ij} = 1) \mid T_i = 1, u_i] - \mathrm{E}[\Pr(y_{ij} = 1) \mid T_i = 0, u_i]$ $= \text{expit}(\alpha + \beta + u_i) - \text{expit}(\alpha + u_i)$
Expected value of response averaged over subject-specific u_i (population averaged): $\mathrm{E}_u[\mathrm{E}[y_{ij} \mid T_i = 1, u_i]]$ $= \mathrm{E}_u[\alpha + \beta + u_i]$ $= \alpha + \beta$; $\mathrm{E}_u[\mathrm{E}[y_{ij} \mid T_i = 0, u_i]]$ $= \mathrm{E}_u[\alpha + u_i]$ $= \alpha$	*Expected value of response averaged over subject-specific u_i (population averaged):* $\mathrm{E}_u[\mathrm{E}[\Pr(y_{ij} = 1) \mid T_i = 1, u_i]]$ $= \mathrm{E}_u[\text{expit}(\alpha + \beta + u_i)]$; $\mathrm{E}_u[\mathrm{E}[\Pr(y_{ij} = 1) \mid T_i = 0, u_i]$ $= \mathrm{E}_u[\text{expit}(\alpha + u_i)]$
Population averaged treatment effect $\mathrm{E}_u[\mathrm{E}[y_{ij} \mid T_i = 1, u_i]] - \mathrm{E}_u[\mathrm{E}[y_{ij} \mid T_i = 0, u_i]]$ $= \beta$	*Population averaged risk-difference due to treatment* $\mathrm{E}_u[\text{expit}(\alpha + \beta + u_i)] - \mathrm{E}_u[\text{expit}(\alpha + u_i)]$ $\neq$ subject-specific risk-difference due to treatment

Table 5.1: Comparison of model for quantitative and binary response, y_{ij}, illustrating the implications for SS and PA estimates of treatment effect. For details, see §5.2. (†) – *expit* is the inverse of the *logit* function

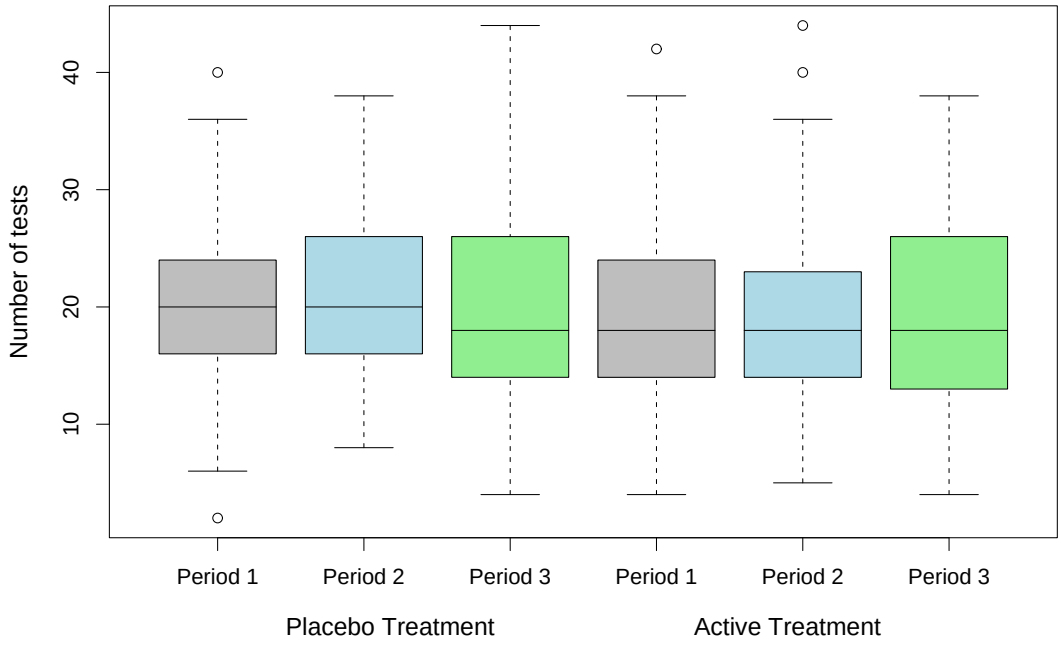

Figure 5.1: Longitudinal binary data: distribution of number of tests undertaken by each subject in each period

EXAMPLE 5.1 *Longitudinal binary data*

To illustrate these differences in practice we use longitudinal binary data, which we simulated from a real clinical trial. Two hundred and forty one subjects were randomised to receive either a placebo or active treatment. There are three treatment periods, each several weeks in length. In each of these periods, each subject undergoes a series of tests (the number varying from subject to subject) with each test having a binary outcome (1=success). Figure 5.1 shows the distribution of the number of tests undertaken by each subject in each period. As Table 5.2 indicates, a number of patients withdrew before the end of the study.

We now compare three estimates of the treatment effect using data from period 3 alone. The first uses logistic regression, assuming that the results of the repeated tests undertaken by an individual are independent. Standard errors are obtained from the information matrix in the usual way. Let y_{ij} be the response from patient i (189 at period 3) to test j (ranging from 4 to 44). The model, adjusting for baseline (base_i) and treatment ($\text{treat}_i = 1$ for active, 0 for placebo), is

$$\begin{aligned} \mathrm{E}\, y_{ij} &= \mu_{ij}, \\ \mathrm{Var}\, y_{ij} &= \mu_{ij}(1-\mu_{ij}), \\ \mathrm{logit}(\mu_{ij}) &= \alpha_0 + \alpha_1 \text{base}_i + \alpha_2 \text{treat}_i, \\ \mathrm{Cor}\,[(y_{ij}, y_{ij'})] &= 0. \end{aligned} \tag{5.1}$$

The second model is a PA model, allowing a fixed correlation between the different test results from the same subject (*i.e.* an exchangeable correlation structure). This model uses the empiri-

	Treatment	
	Placebo	Active
Withdrew before period 1	0	0
Withdrew before period 2	10	5
Withdrew before period 3	25	12
Completers	82	107

Table 5.2: Longitudinal binary data: patient withdrawal by treatment arm

Model	Estimates (std. errors) of		
	β_0	β_1	β_2
(5.1)	1.54 (0.141)	0.02 (0.003)	0.64 (0.145)
(5.2)	1.66 (0.399)	0.02 (0.009)	0.57 (0.372)
(5.3)	3.16 (0.667)	0.03 (0.011)	1.10 (0.574)

Table 5.3: Parameter estimates from fitting (5.1), (5.2) and (5.3) to the data from period 3

cal sandwich estimate of variance discussed above. The superscript p differentiates coefficients in this model from those in the subject specific model below:

$$\begin{aligned} \mathrm{E}\, y_{ij} &= \mu_{ij}, \\ \mathrm{Var}\, y_{ij} &= \mu_{ij}(1-\mu_{ij}), \\ \mathrm{logit}(\mu_{ij}) &= \beta_0^p + \beta_1^p \mathrm{base}_i + \beta_2^p \mathrm{treat}_i \\ &= \eta^p, \text{ the population averaged linear predictor} \\ \mathrm{Cor}\,[(y_{ij}, y_{ij'})] &= \rho\,. \end{aligned} \tag{5.2}$$

The third model is a SS model, with linear predictors from the same subject sharing a common SS random effect:

$$\begin{aligned} \mathrm{E}\,[y_{ij}|u_{0j}] &= \mu_{ij}, \\ \mathrm{Var}\,[y_{ij}|u_{0j}] &= \mu_{ij}(1-\mu_{ij}), \\ \mathrm{logit}(\mu_{ij}) &= \beta_0 + u_i + \beta_1 \mathrm{base}_i + \beta_2 \mathrm{treat}_i \\ &= \eta, \text{ the subject specific linear predictor} \\ u_i &\sim N(0, \sigma_u^2), \\ \mathrm{Var}\,\mathrm{logit}\mu_{ij} &= \sigma_u^2. \end{aligned} \tag{5.3}$$

We call the inverse of the logistic function *expit*. Notice that in model (5.2)

$$\mathrm{E}\,[y_{ij}] = \mathrm{expit}(\beta_0^p + \beta_1^p \mathrm{base}_i + \beta_2^p \mathrm{treat}_i),$$

whereas in model (5.3)

$$\mathrm{E}\,[y_{ij}] = \mathrm{E}_u[\mathrm{E}\,[y_{ij}|u]] = \mathrm{E}_u[\mathrm{expit}(\beta_0 + u_i + \beta_1 \mathrm{base}_i + \beta_2 \mathrm{treat}_i)].$$

For the fitted means, μ_{ij} from (5.2) and (5.3) to be equal, the parameter estimates will be different in the two models (unless $\sigma_u^2 = 0$). The differences will increase with σ_u^2. Also, while the $\mathrm{Cor}\,(y_{ij}, y_{ij'})$ is a parameter in (5.2), to estimate this from (5.3) involves taking expectations over u. Further, in (5.3) the correlation will vary with the covariates, e.g. baseline.

Table 5.3 gives the results of fitting these three models to the data from period 3 alone. The parameter estimates from (5.1) are similar to those from (5.2). However, the standard errors are much larger for (5.2), as it allows for the correlation between test results from the same subject, estimated to be 0.30. By contrast, the parameter estimates from (5.3) are considerably larger, but notice as well how the standard errors have also increased although not quite as much. Consequently the treatment effect becomes significant at the 5% level. In part, this reflects the fact that maximum likelihood estimates extract more information from the data than GEE estimates.

This large increase in absolute size of the parameters estimates and the accompanying increase in standard errors reflects the substantial between subject variability ($\sigma_u^2 = 7.60$). The larger this variance, the larger the difference in the size of the corresponding parameters from the two approaches. □

5.2.1 *Obtaining population-averaged coefficients from subject-specific coefficients*

We have seen already, and will see again below, that there are advantages from a missing value perspective in using a likelihood analysis. In the current setting this implies that there are advantages in using the SS model. However, it may well be in some settings that PA effects are of more clinical interest. On some occasions therefore we may wish to derive the appropriate PA estimates from the corresponding SS estimates. Strictly (from a mathematical viewpoint) the PA and SS models cannot both be correct (*i.e.* logistic), except in some very special cases. This creates difficulties in moving from SS to PA odds-ratios in some settings. Fortunately, for the type of generalised linear mixed model we consider here, logistic models *can* be applied approximately for both types of model and we show now how we can use this to move from SS to PA estimates.

For a simple random subject effect model, that is, one of the form of (5.3) where there is a single common subject effect, or intercept, (u_i) with variance σ_u^2, the k^{th} PA parameter, β_k^p, and the corresponding SS parameter, β_k, (typically log odds-ratios in the logistic setting) are approximately related as follows (Zeger *et al.*, 1988):

$$\beta_k^p \approx \frac{\beta_k}{\sqrt{1 + 0.34584 \times \sigma_u^2}}. \tag{5.4}$$

More generally, if we have a set of q random effects $u_i = \{u_{i1}, \ldots, u_{iq}\}^T \sim \mathrm{N}(0, \Sigma_u)$ with coefficients $z = \{z_1, \ldots z_q\}^T$ in the linear predictor, *i.e.*:

$$z^T u = \sum_{k=1}^{q} z_k u_k$$

the same expression can be used except that the variance σ_u^2 is replaced by the variance of the ensemble, $\mathrm{V}(z^T u) = z^T \Sigma_u z$. An impression of the accuracy of this approximation can be gained from Figure 5.2. This shows the relationship between the PA linear predictor, η^p, and the transformed SS linear predictor, under a random intercepts model with a range of variances, σ_u^2. We see that the approximation is excellent for small variances and worsens as the variance increases.

These approximations depend on the unknown (co)variance parameter(s), Σ_u, which must be estimated in practice. This introduces a further degree of approximation. Standard errors are

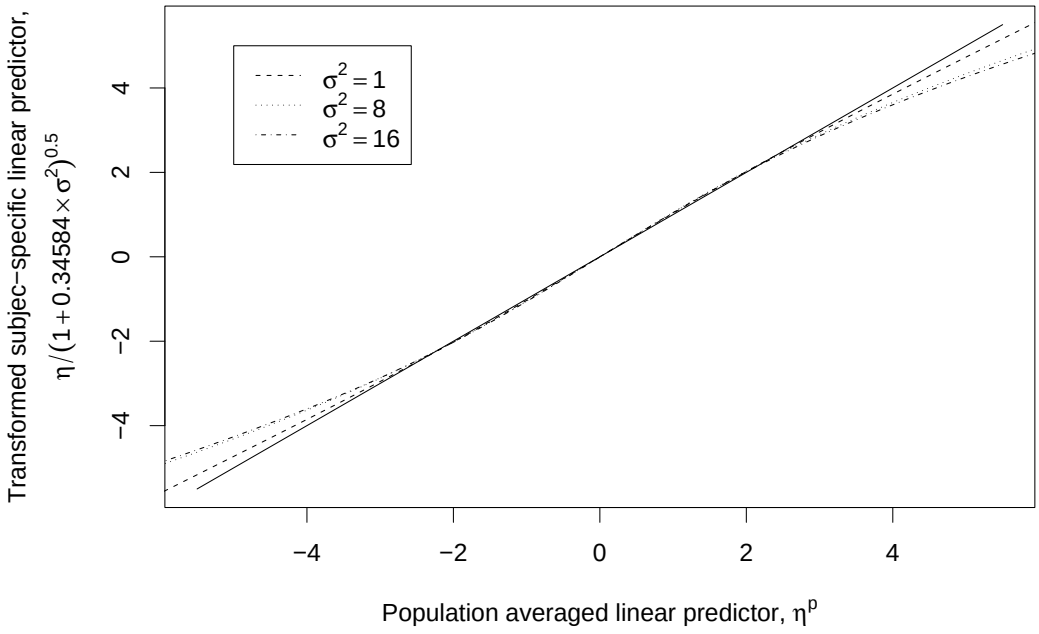

Figure 5.2: Accuracy of approximation of population averaged linear predictor, η_p, by transformation of subject-specific linear predictor, η, using (5.4). Solid line is equality

also needed for the estimates, $\hat{\beta}_k^p$. To a first order of approximation these can be obtained by scaling standard errors from the SS model in the same way we scaled the coefficient estimates. In other words we divide the SS standard error by the denominator of (5.4). This has the advantage of preserving test statistics and relative length of confidence intervals, but ignores the uncertainty in the estimated (co)variance parameter(s).

EXAMPLE 5.2 *Moving from SS to PA coefficients*

Using the data from the previous example, we fit model (5.2), and compare the resulting PA coefficients with those obtained by fitting model (5.3) and then using the approximation (5.4).

In this case — recalling from the previous example that σ_u^2 was estimated as 7.60 — using (5.4) we get

$$1.10/\sqrt{1+0.34584\times 7.60} = 0.577.$$

This is compared with the other estimates in Table 5.4[2]. The agreement is not perfect, but may nevertheless be adequate in many settings. The Table also compares the robust standard errors from the PA model (5.2) with the transformed SS standard errors. The effect of treatment is not significant at the 5% level under the PA model, but it is under the SS model. This may partly reflect the fact that likelihood models extract more information from the data, but note too that rescaling the SS standard errors by dividing by the denominator of (5.4) is only first order approximate, as we do not take into account the variability in estimating σ_u^2. □

[2]The difference in the SS to PA conversion is due to retaining higher precision in the calculations.

Method	PA, model (5.2)		PA (Transformed SS)		SS, model (5.3)	
	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)
	0.57	(0.372)	0.57	(0.301)	1.10	(0.574)

Table 5.4: Comparison of PA treatment effect estimates from a GEE with those obtained by transforming SS estimates, using (5.4)

5.2.2 *Population Averaged or Subject-Specific models*

In the absence of missing data, should a PA or SS model be specified as the primary analysis? The debate between advocates of PA and SS models is ongoing, and not always informative. A useful discussion is given by Diggle *et al.* (2002), p. 131–140. We note the following. Suppose the intended full data analysis estimates a baseline adjusted treatment effect using data from the end of the study, by fitting a GLM to data from the final follow-up. This will give the same estimated treatment effect as fitting a GEE with an independence correlation structure to the longitudinal follow-up data, including the full treatment-time and baseline-time interactions. The robust standard errors from the GEE should also be similar, but will not agree precisely.

SS models allow the variance structure to be modelled, rather than just regarded as a nuisance parameter. If the variance is well modelled, then SS models give more precise estimates of (SS) treatment effects, because they are likelihood based. However, the down-side is that the estimated treatment effects may be sensitive to the chosen variance model. Unfortunately, for many trials with binary outcome there is insufficient information for a detailed assessment of the variance structure.

If the focus is on the response of a subject to treatment, then SS models are appropriate. PA models are arguably more appropriate in population based studies in epidemiology. Of course, PA treatment effects depend on the characteristics of the population over which averaging is done. As we remarked above, the *same* SS treatment effect (same underlying physiological effect of treatment) will give different PA estimates of treatment effect over populations with different heterogeneity (Zeger *et al.*, 1988).

In summary, the appropriate analysis depends on (i) the precise question and (ii) to what extent the subjects in the trial are 'representative' of a population. Below, we therefore consider both PA and SS analyses when data are missing.

5.2.3 *Implications for missing data*

As SS models are typically estimated by maximum likelihood, for the reasons given in Chapters 3 and 4 inferences are still valid if some responses are MAR. However, the GEE based estimating procedures that are normally used for PA models are moment based estimators in which the estimates of the β^p's are chosen to minimise the difference between the marginal mean of the data and that of the model, weighted by the variance. However, we have already seen that marginal means (and variances) are not sensible estimators if data are MAR. Under MAR, like-

lihood estimation implicitly uses the dependence structure to correct for the bias caused by the missing data. With GEEs the dependence structure is only a working approximation (*i.e.* known to be mis-specified), so this correction is not available. Although corrections can be made using suitable weighting, such methods, in their simple form at least, are very inefficient (Carpenter *et al.*, 2006). We do not pursue such approaches here.

It follows in our context that if data are MAR, it is not sensible to use GEEs directly for parameter estimation. Depending on the analysis we would have performed were no data missing, there are several options. These are summarised in Figure 5.3 which also outlines the structure of the rest of this Chapter.

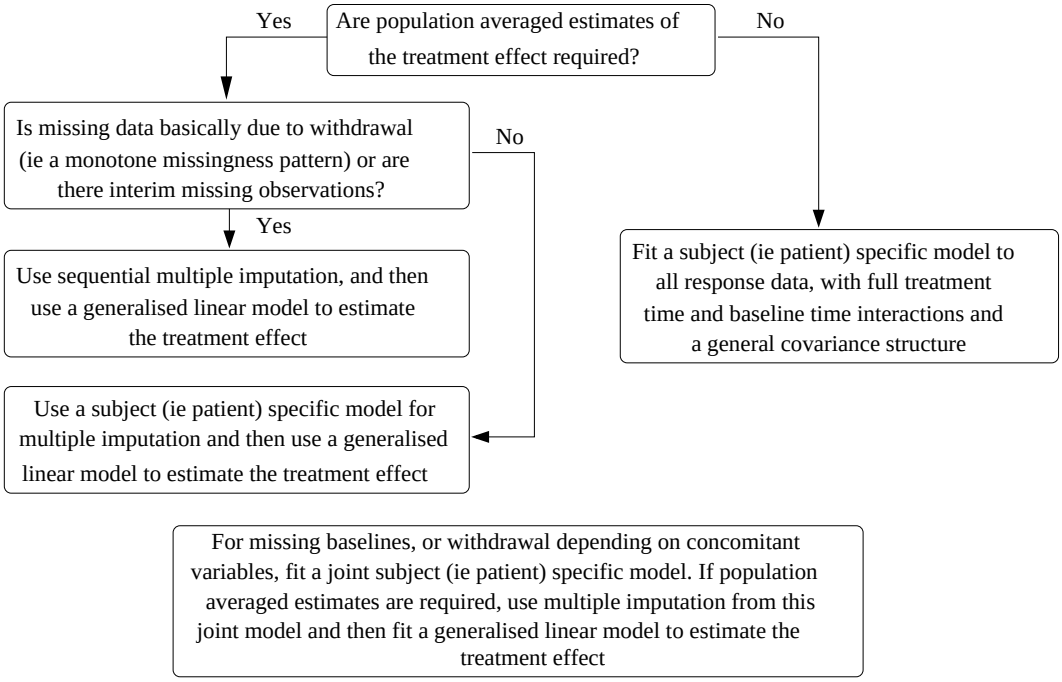

Figure 5.3: Overview of MAR methods for discrete data

First, if the intended analysis with no missing data was SS, then we can proceed in a similar way to that outlined in Chapter 3. In other words, we begin by looking for covariates that are predictive of withdrawal, and then either condition on these (by fitting them as covariates) or, if such covariates are post-randomisation so we cannot condition on them, they can be jointly modelled with the outcome variable or incorporated in a multiple imputation procedure. Assuming data are MAR, using this likelihood approach will give sensible parameter estimates and inferences.

Secondly, if the intended analysis with no missing data was either to (i) use the observations at the end of the trial and fit the discrete-data equivalent of an ANCOVA, or (ii) fit a GEE to estimate longitudinal PA treatment effects, then we have two options:

(i) obtain estimates from a SS model, as in the previous paragraph, and convert them to PA estimates as described in §5.2.1, or

(ii) fit a SS model and then use multiple imputation. That is, use the fitted SS model to obtain (approximate) proper imputations for the missing data and then fit the original PA model of interest to each imputed data set, combining the estimates using the MI rules.

If data are monotone missing (so that each subject is observed till he or she withdraws) we need not fit a repeated measures SS model to all the data and then impute from this. Rather, we can fit a sequence of GLM's, logistic regressions in the binary/binomial setting, and impute from these.

In the remainder of this Chapter, we discuss and illustrate all these approaches in more detail.

5.3 Subject-specific analyses with missing data

When considering quantitative data in Chapter 3, we used the multivariate normal distribution with an unstructured covariance matrix. There is no direct analogue of this for discrete data. Our approach is to use normally distributed SS random effects, chosen with the aim of minimising the variance structure imposed on the data. In other words, as in Chapter 3, we wish to let the data 'speak for itself' on the variance structure. As with quantitative data (§3.2.1), we anticipate minimal loss of power for moderate sample sizes.

However, this is less straightforward in the current setting than with the multivariate normal linear model. First, unless we are assuming that the data are over-dispersed, the variance of each observation is determined by its mean. For example, a binary observation with success probability π has variance $\pi(1-\pi)$ irrespective of other aspects of the dependence structure. Secondly, the correlation between two observations also depends on their means. So the concept of a particular correlation or covariance structure, independent of the mean structure, does not apply here. Using the structure of a generalised linear mixed model we instead focus on the covariance structure of the *random effects*. We can then think of these as underlying latent variables that are thresholded to give the binary responses. Although features of the dependence structure of the random effects carry through, in a qualitative sense, to that of the binary/binomial observations, there is no simple equivalence.

The covariance structure of the random effects can be treated *in principle* like that in the linear model setting. In practice however, with binary data, we have much less information on this structure and usually only very specific and parsimonious structures can be successfully used. Two are frequently considered, (i) random intercepts (simple subject effects) and (ii) random intercepts and slopes.

As in §5.2, let i index patients and j index observation times. Let $\delta_i = 1$ if patient i is randomised to active treatment, and 0 otherwise. Let base_i be patient i's baseline response. A logistic model for binary data, in which the mean is equal to the success probability, and which has a different treatment effect and baseline adjustment at each time j, is given by:

$$\begin{aligned} \mathrm{E}(y_{ij}) &= \mu_{ij}, \\ \text{logit}(\mu_{ij}) &= \alpha_j + \delta_i\beta_j + \text{base}_i\gamma_j + u_i, \\ u_i &\sim N(0,\Sigma_u). \end{aligned} \tag{5.5}$$

The corresponding random intercepts and slopes model is:

$$\begin{aligned} \mathrm{E}(y_{ij}) &= \mu_{ij}, \\ \text{logit}(\mu_{ij}) &= \alpha_j + \delta_i\beta_j + \text{base}_i\gamma_j + u_{0i} + ju_{1i}, \\ \begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} &\sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u_0} & \\ \sigma_{u_0u_1} & \sigma^2_{u_1} \end{pmatrix} \right\}. \end{aligned} \tag{5.6}$$

The simple random intercepts model (5.5) will often suffice when the response is fairy stable over the duration of the trial. The introduction of the random slopes allows some more flexibility in the dependence structure, in particular correlations will tend to be higher for pairs of measurements that are closer together. In practice, the fit of models (5.5) and (5.6) can be compared using the change in twice the log-likelihood ratio, which follows a χ^2_2 distribution approximately[3].

EXAMPLE 5.3 *Longitudinal binary data: fitting random intercepts and random intercepts and slopes models*

We now compare the results of fitting (5.5) and (5.6) to the data discussed in Example 5.2. Recall there are 241 subjects, two treatment groups and three periods. Before we restricted ourselves to period 3 data only, but now we use data from all 3 periods. We omit the baseline period interaction, as it is not present in these data (which we simulated — see Example 5.1).

	Model			
	Random intercepts		Random int./sl.	
Parameter	Estimates	(SE)	Estimate	(SE)
intercept	2.65	(0.43)	3.28	(0.53)
baseline	0.03	(0.001)	0.03	(0.01)
treatment	1.15	(0.41)	1.25	(0.53)
period 1	−0.97	(0.14)	−1.12	(0.49)
treatment × period 1	1.04	(0.22)	1.56	(0.65)
period 2	−0.39	(0.14)	−0.75	(0.30)
treatment × period 2	0.55	(0.23)	0.39	(0.41)
$\mathrm{Var}\, u_0$	5.90	(0.94)	18.25	(3.76)
$\mathrm{Var}\, u_1$	—		2.46	(0.49)
$\mathrm{Cov}\,(u_0, u_1)$	—		−0.80	(0.053)
$-2\times$ log-likelihood	4941.6		4671.2	

Table 5.5: Longitudinal binary data: results of fitting random intercepts model (5.5) and random intercepts and slopes model (5.6)

Table 5.5 shows the results. It is clear from the log-likelihoods that the random intercepts and slopes model fits substantially better. The variance/correlation matrix of the linear predictor implied by the random components of the model is

$$\begin{pmatrix} 9.99 & & \\ 0.86 & 6.76 & \\ 0.46 & 0.83 & 8.29 \end{pmatrix}. \tag{5.7}$$

[3] Approximately, because under the null hypothesis $\sigma^2_{u_1} = 0$, on the boundary of values for a variance.

Notice how the correlation declines as the time between observations increases; with random intercepts alone it is constant. Further notice that the random effects variance under this model is greater at each time point than under the random intercepts model, (5.5). This is interwoven with the fact that, relative to the random intercepts model, the parameter estimates are larger under this model, implying the fitted probabilities are more extreme, so that the component of variance due to the binomial distribution, $\mu_{ij}(1-\mu_{ij})$, is smaller. However, although the estimated treatment effects are larger under the random intercepts and slopes model, their significance is reduced in all cases.

Now compare the estimated treatment effect (std. error) for period 3 from the random intercepts and slopes model, 1.25 (0.532), with that from fitting the random intercepts model to data from the last period alone, 1.10 (0.574) (last row of Table 5.3). If missing patient responses are MAR, then the estimate in Table 5.3, based on only the observed data at the last time period, will be biased and the standard error wrong. However, as described in Chapter 3, a model for all the data, allowing for the correlation between observations on the same person, will remove this bias, and recover some of the lost information. This is exactly what we see here; there is a slight increase in the treatment effect, and a slight decrease in the standard error and the z-statistic changes from 1.92 to 2.34. □

The model (5.6) does not require repeated observations on each subject at each time. Rather, it can be fitted with a single observation from each subject at each time. For SS analyses with missing data, and one observation on each subject at each time, we therefore recommend fitting (5.5) and extending it to (5.6) if possible.

Sometimes, however, we have sufficient repeated observations, $k = 1, \ldots K_i$ on each subject at each time to allow us to fit a more general covariance structure than random intercepts and slopes. This is shown in (5.8).

$$
\begin{aligned}
\mathrm{E}\left(y_{ijk}\right) &= \mu_{ij}, \\
\mathrm{logit}(\mu_{ij}) &= \alpha_j + \delta_i\beta_j + \mathrm{base}_i\gamma_j + u_{ij}, \\
\begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iJ} \end{pmatrix} &\sim N\left\{ \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u_1} & & & \\ \sigma_{u_0u_2} & \sigma^2_{u_2} & & \\ \vdots & \vdots & \ddots & \\ \sigma_{u_1u_J} & \cdots & \sigma_{u_{J-1}u_j} & \sigma^2_{u_J} \end{pmatrix} \right\}.
\end{aligned}
\tag{5.8}
$$

To have sufficient information to fit this model, we need a sufficient number of observations on each subject at each time; thus we have explicitly included the subscript k in the response, although the mean is the same for each k. Alternatively, suppose we only have one observation on each subject at each time, but that the times are reasonably close together, so that the responses can be considered to have the same mean. Then, if we group observations close together in time into the same 'periods', so we can estimate a variance term for each period, we can again fit (5.8).

In practice, a simpler form of (5.8), where we assume $\sigma^2_{u1} = \sigma^2_{u2} = \cdots = \sigma^2_{uJ}$, is often much easier to fit. Unless the variance is changing markedly with time, which is unusual in most clinical trials, such a simplified model is likely to prove adequate. Likewise, whereas in Chapter 3, we advocated fitting a separate correlation matrix to each treatment arm in general, while this remains preferable, it will usually be impractical in the current setting because of lack of information, which often causes computational difficulties.

EXAMPLE 5.4 *Longitudinal binary data*

We fit two versions of (5.8) to these data. The second constrains $\sigma^2_{u1} = \sigma^2_{u2} = \sigma^2_{u3}$, which makes fitting considerably faster. The estimated variances from random intercepts and slopes (5.7) suggest this is plausible. Table 5.6 shows the results. The difference in $-2\times$log-likelihood of 1.3 supports this.

Parameter	Model: Unstructured variance	Model: Restrict $\sigma^2_{u1} = \sigma^2_{u2} = \sigma^2_{u3}$
intercept	3.06 (0.575)	3.21 (0.550)
baseline	0.03 (0.008)	0.03 (0.008)
treatment	1.24 (0.543)	1.38 (0.586)
period 1	−0.91 (0.53)	−1.25 (0.444)
treatment × period 1	1.42 (0.675)	1.18 (0.659)
period 2	−0.38 (0.489)	−0.56 (0.391)
treatment × period 2	0.74 (0.582)	0.64 (0.586)
Var u_1	10.0 (1.95)	8.87 (1.27)
Var u_2	8.50 (1.85)	8.87 (1.27)
Var u_3	7.26 (1.83)	8.87 (1.27)
Cor (u_1, u_2)	0.72 (0.066)	0.72 (0.066)
Cor (u_1, u_3)	0.45 (0.107)	0.46 (0.105)
Cor (u_2, u_3)	0.66 (0.091)	0.68 (0.089)
−2 log likelihood	4580.7	4582.0

Table 5.6: Results of fitting (5.8) to the longitudinal binary data

Note the sensitivity of the estimated treatment effects to the estimated variance/covariances. As the variance increases, the treatment effects tend to increase. Thus, when modelling data of this kind, it is important to spend some time considering an appropriate random effects structure, and it is important this structure does not impose inappropriate constraints. □

5.3.1 *Concomitant variables predictive of withdrawal*

We now consider how to extend this subject-specific model to include a concomitant variable (and by direct extension several such variables) that is predictive of withdrawal, but which we do not want to adjust (condition) our treatment estimate on (if we did, then it would simply be included in the linear predictor in the usual way). Our approach is to jointly fit (i) a SS model for the concomitant variable, and (ii) a SS model for the response, and allow them to be correlated. Then, if the response is MAR given the concomitant variable, the resulting estimated treatment effects will be valid, provided we include in the model subjects who only have the concomitant variable observed.

Suppose the concomitant variable from subject i is denoted v_i. Possible examples might be subject age, disease history, concomitant diseases. We suppose v_i can be modelled as normal, or transformed so it is approximately normal. This makes computation easier, and does not affect the interpretation of the treatment estimates for the response. To keep the algebra simple suppose we only have binary responses, y_{ij}, from one period, with (common) binary indicator of treatment effect, treat_i, as before. The model is

$$\begin{aligned} \text{logit}\{\Pr(y_{ij}=1)\} &= \beta_0 + u_i + \beta_1 \text{treat}_i \\ v_i &= \alpha_0 + w_i + \alpha_1 \text{treat}_i \\ \begin{pmatrix} u_i \\ w_i \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \\ \sigma_{uw} & \sigma_w^2 \end{pmatrix}\right] \end{aligned} \tag{5.9}$$

Note that, for some subjects with missing responses, we will only have v_i. These data must be included in the analysis for the treatment effects to be valid under MAR, though.

EXAMPLE 5.5 *Longitudinal binary data: analysis of period 1 responses with baseline as a concomitant variable*

We illustrate this approach by fitting (5.9) using data from period 1 of the longitudinal binary data set, with baseline as the concomitant variable, v. As there is no possible treatment effect at baseline, we set $\alpha_1 = 0$.

Table 5.7 shows the results, with the covariance matrix for (u, v) parameterised in terms of the correlation. For comparison, column 2 shows the results of fitting a random intercepts model to the binary data alone. As there are no missing data at period 1, we expect the parameter estimates to be very similar, and indeed they are. □

Parameters	Estimates (std. errors)	
	logistic model only	joint model
β_0	3.77 (0.422)	3.73 (0.418)
β_1	2.89 (0.626)	2.80 (0.587)
α_0	—	51.1 (1.77)
σ_u^2	12.1 (2.36)	11.7 (2.43)
σ_v^2	—	758 (69.0)
ρ	—	0.27 (0.078)

Table 5.7: Results of fitting random intercepts model only (column 2) and joint model (5.9) (column 3) to baseline and period 1 responses from the longitudinal binary data

5.4 Population-averaged analyses with missing data

We now assume that, were the data complete, the desired analysis would be a logistic regression using data from the final time point, in which the effect of treatment is adjusted for baseline. As

we have already discussed, in contrast to the analysis of continuous data, the results of such an analysis cannot be taken directly from an appropriate SS model, even if the data are complete.

Here we discus two alternatives, both of which involve multiple imputation (MI). In each case an imputation model will be set up to create appropriate sets of imputations at the final time point, and the completed datasets are then analysed using the original intended method (*i.e.* logistic regression), and the results combined in the conventional MI manner.

The first option applies when we have no interim missing data; that is to say subjects withdrawal and we have no subsequent data on them. The second approach is valid when we have interim missing data as well, and extends to handling missing baseline values.

5.4.1 No interim missing data

If, after subjects withdraw, there are no further data on them — that is to say there are no interim missing data — then (p. 87) we say we have a *monotone* withdrawal pattern. In this case we do not need to specify a joint model for the imputed data. Rather we can specify a series of conditional models and impute from these.

Specifically, suppose we intend to observe 50 subjects in the placebo arm at times $j = 1, 2, 3$ but that we have monotone withdrawal as in Table 5.8. Before describing the imputation strategy,

Subject identifiers	Last observation at time	Observation time 1	2	3	No. of subjects
1–35	3	X	X	X	35
36–45	2	X	X	.	10
46–50	1	X	.	.	5

Table 5.8: Monotone missing data due to subject withdrawal. An 'X' denotes the observation is seen, and a '.' that it is missing

we note that all our imputations are assumed to be proper, or asymptotically proper, in the sense described in Chapter 4, p. 80. What we call *monotone regression imputation* proceeds as follows, with K sets of imputations.

1. Use data from the 45 subjects observed at times 1 and 2 to estimate the logistic regression

$$\text{logit}\Pr(y_{i2} = 1) = \alpha_0 + \alpha_1 y_{i1}. \tag{5.10}$$

   Then, for each of the 5 subjects missing at time 2, (approximately) properly impute K missing values. Denote these by y_{i2}^k, $k = 1, \ldots, K$, and $i = 46, \ldots 50$.

2. Use data from the 35 subjects with no missing data to estimate the logistic regression

$$\text{logit}\Pr(y_{i3} = 1) = \beta_0 + \beta_1 y_{i1} + \beta_2 y_{i2}. \tag{5.11}$$

3. For each k,

   Using data from the 10 subjects observed only at time 1 and 2, together with the k^{th} set of imputations, $y^k_{46,2},\ldots y^k_{50,2}$, the estimated parameters from (5.11) and their variance/covariance matrix to give a single (approximately) proper imputation of the missing data at time 3, denoted $y^k_{36,3},\ldots,y^k_{50,3}$.

4. Putting these imputations together gives K imputed data sets.

If we have a monotone withdrawal pattern over more time points, this algorithm extends in the obvious way.

Note that we should adjust for any baseline covariates predictive of withdrawal in our imputation models (5.10) and (5.11), in order to ensure the imputations are valid under MAR. Further, as treatment allocation is usually predictive of withdrawal, we should have different imputation models for each treatment arm. In practice, if there are sufficient data, we would therefore recommend fitting a full interaction with treatment, baseline and each of the regressors in models (5.10) and (5.11).

Thus, if there were two treatment groups, 0 and 1, (5.10) becomes

$$\text{logit}\Pr(y_{i2}=1)=\alpha_0+\alpha_1 y_{i1}+\alpha_2\times 1[\text{treat}_i=1]+\alpha_3\times 1[\text{treat}_i=1]y_{i1}, \tag{5.12}$$

where $1[.]$ is an indicator for the event in brackets. Likewise (5.11) becomes

$$\begin{aligned}\text{logit}\Pr(y_{i3}=1)=\beta_0+\beta_1 y_{i1}+\beta_2 y_{i2}+\beta_3\times 1[\text{treat}_i=1]+\beta_4\times 1[\text{treat}_i=1]y_{i1}\\ +\beta_5\times 1[\text{treat}_i=1]y_{i2}.\end{aligned} \tag{5.13}$$

EXAMPLE 5.6 *Dental pain data*

Three hundred and sixty six subjects who had moderate or severe post-surgical pain following extraction of their third molar were randomised to receive a single dose of one of five increasing doses of a test drug, or an active control, or a placebo. The response was degree of pain relief, measured on an ordinal scale from 0 (none) to 4 (complete). This was measured before the extraction, and 18 times in the 24 hours following extraction. In the latter part of the trial, many subjects withdrew, particularly in the low dose and placebo arms.

Here, we focus on pain relief 6 hours after randomisation. To illustrate the use of the `monotone logistic` option in `SAS PROC MI` (v. 9.1), we dichotomise the pain relief score to 0 if the original scale was 0, 1, 2 and 1 if the original scale was 3, 4. Thus a response of 1 means some, or complete, pain relief. We now impute the missing values using the algorithm above.

Subjects reported their degree of pain 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4, 5 and 6 hours after randomisation. For these analyses we ignore all subsequent measurements (which were increasingly missing). Five out of 366 subjects had interim missing values and are excluded from this analysis. All subjects were observed up to 1.5 hours. Subsequently, Table 5.9 shows the drop out pattern.

At 6 hours, 179 out of 212 subjects remaining had a response of 1. Further, subjects often keep the same response for several visits. Indeed, 162 have the same response until withdrawal

No. of subjects	Hours after tooth extraction 2	3	4	5	6
212	X	X	X	X	X
11	X	X	X	X	.
8	X	X	X	.	.
10	X	X	.	.	.
33	X	.	.	.	.
87	.	.	.	.	.
361 patients in total					

Table 5.9: Withdrawal pattern for dental data, for observations up to 6 hours after extraction. Unseen observations are denoted '·'. Five patients with interim missing data are excluded

(152 always 0), and of the remainder 80% make only 1 transition. Thus there is not enough information in the data to fit the appropriate extensions of (5.12) and (5.13) to several time points and treatments. Rather, we fit a simpler model. Let $\delta_{ik} = 1$ if subject i has treatment k. At each time $j = 2, \ldots, 6$, hours, the model is

$$\text{logit}\Pr(y_{ij} = 1) = \alpha_{jk}\delta_{ik} + \beta_{jk}\delta_{ik}y_{i(j-1)}, \tag{5.14}$$

in other words, a different linear dependence on the previous observation, on the logistic scale, for each treatment group.

After imputing the missing data 6 hours after tooth extraction, we fit a logistic regression to estimate the treatment effects. Table 5.10 compares the results of a complete data analysis, and multiple imputation using `SAS PROC MI` with 5, 50 and 500 imputations. Looking first at the complete case analysis, there is little to choose between the treatments; the only borderline significant comparison is between drug A at 1800mcg and drug A at 450mcg. The degree of pain relief increases with each increase in dose of A with the exception of the highest dose when there is a suggestion it falls back.

Turning to the `SAS PROC MI` results in the 3rd column, it is clear that using 5 imputations is nowhere near enough for these data. The standard errors are very large (and the degrees of freedom for the reference t-distribution are small) and the parameter estimates are erratic. The latter two columns show the results of using 50 and 500 imputations respectively. Compared with the complete case analysis, the probability of pain relief now increases steadily with dose, as expected. However, the standard errors are all about twice those for the complete case analysis. The number of imputations needed for the results to settle down and the increase in the standard errors are both surprising.

Further analysis by Daniel (2007) showed that problems occur because, at certain observation times in certain treatment groups, when we stratify by previous observation to fit (5.14) the

Parameter	Observed data (n=212)		Multiple imputation, with K imputations $K=5$		$K=50$		$K=500$	
	Estimate	(se)	Estimate	(se)	Estimate	(se)	Estimate	(se)
A, 450mcg	−0.43	(0.92)	0.54	(2.14)	−1.03	(1.72)	−0.90	(1.83)
A, 900mcg	0.43	(0.94)	1.09	(2.96)	−0.41	(2.05)	−0.07	(2.03)
A, 1350mcg	0.54	(0.94)	0.62	(2.90)	−0.27	(1.93)	−0.01	(1.97)
A, 1800mcg	1.46	(1.08)	1.79	(2.70)	0.46	(2.12)	0.69	(2.07)
A, 2250mcg	1.23	(1.00)	2.07	(2.95)	0.80	(2.19)	1.01	(2.10)
C, 400mcg	0.05	(0.89)	0.82	(3.38)	−0.53	(2.20)	−0.33	(2.10)
Placebo, log(odds)	1.25		0.27		1.53		1.29	

Table 5.10: Results of multiple imputation (using SAS) for estimation of the treatment effects 6 hours after tooth extraction. All parameter estimates are log-odds ratios vs the placebo (i.e. not adjusted for baseline)

maximum likelihood estimate of the probability of relief at the current observation time is 0 or 1. In other words for some k and j one or both of α_{jk}, β_{jk} should be estimated as $\pm\infty$. PROC LOGISTIC's attempt to do this triggers a warning message, but unfortunately the default option in PROC MI is to hide this from the user, and continue.

Besides giving a large (or small) parameter estimate, when this occurs the corresponding standard error is very large. The effect of this is that the resulting imputed data can sometimes be very wrong. Looking at Table 5.10, this is the cause of both the variation in the results as the number of imputations increases, and also the increase in the standard error.

To address this, we did the following. For each time, j, before fitting (5.14) we checked if any of the estimates α_{jk}, β_{jk} were $\pm\infty$. If they were, for the group defined by that time, j, treatment group, k, and previous observation $y_{i(j-1)}$, we added a one-off observation with a 0 or 1 outcome as appropriate. This is sufficient to prevent the problems described in the previous paragraph occurring. Having thus avoided the numerical problems in fitting (5.14), multiple imputation proceeds in the usual way.

The results of this analysis are shown in Table 5.11. Again, they show we need far more than 5 imputations. As before, after MI for treatment A the effect estimates increase with dose — in line with our intuition. Now, though, the MI and observed standard errors are similar, in line with what we might expect.

A natural next step would be to use monotone regression with an ordinal model, such as the proportional odds. This can be readily done with the monotone option of SAS PROC MI, which detects if the response has more than two categories and substitutes the proportional odds model for the logistic automatically. Additional options permit a more general multinomial model to be fitted. □

Parameter	Observed data (n=212)		Multiple imputation, with K imputations $K=5$		$K=50$		$K=500$	
	Estimate	(se)	Estimate	(se)	Estimate	(se)	Estimate	(se)
A, 450mcg	−0.43	(0.92)	−0.33	(0.90)	−0.29	(0.90)	−0.41	(0.95)
A, 900mcg	0.43	(0.94)	0.57	(0.99)	0.53	(0.95)	0.37	(0.96)
A, 1350mcg	0.54	(0.94)	0.65	(1.03)	0.65	(0.96)	0.48	(0.97)
A, 1800mcg	1.46	(1.08)	1.21	(1.30)	1.04	(1.04)	0.92	(1.02)
A, 2250mcg	1.23	(1.00)	1.19	(1.03)	1.13	(0.96)	1.00	(0.98)
C, 400mcg	0.05	(0.89)	−0.04	(1.12)	−0.01	(0.88)	−0.13	(0.91)
Placebo, log(odds)	1.25		0.93		0.92		1.06	

Table 5.11: Results of multiple imputation for estimation of the treatment effects 6 hours after tooth extraction when we add observations to avoid $(\hat{\alpha}_{jk}, \hat{\beta}_{jk})$ being $\pm\infty$. Details in the text. All parameter estimates are log-odds ratios vs the placebo. The same starting random number seed was used for $K = 5, 50, 500$

5.5 Interim missing data

The problem caused by interim missing data is more difficult to handle in the discrete case. Consequently if — as in Example 5.6 above — the withdrawal pattern is predominantly monotone, so that omitting the small number of subjects with interim missing data is unlikely to be misleading, we would do this.

However, if a substantial number of subjects have interim missing data, this approach is unsatisfactory. A practical alternative is as follows. Suppose, as above, the analysis of interest is the estimated treatment effect at the end of the trial, adjusted for baseline. We fit an appropriate SS model to the data, as discussed above; this gives sensible parameter estimates under the MAR assumption. Then we can use MI, imputing the missing data at the end of the trial from the mixed model, to obtain the marginal estimate of treatment at the end of the trial.

In effect, we are using the SS model to draw appropriate imputations for the missing data. We need the SS model because the data are discrete, so imputations from software that assumes multivariate normality are likely to be inappropriate, particularly if the data are binary and the fitted probabilities are close to 0 or 1.

The difficulty with this approach is that, unlike in the continuous case, it is no longer entirely clear how to draw approximate proper imputations for the missing observations, aside from fitting a Bayesian model and sampling from the appropriate posterior. This is because the joint distribution for a subject's responses implied by the SS model is no longer analytically tractable, so unlike with the multivariate normal distribution for continuous data, the appropriate conditional distributions cannot be readily derived. Instead, an approximate approach for binary data is as follows.

1. Fit the mixed model to the data, obtaining estimates of the fixed parameters, $\hat{\beta}$, their covariance matrix, $\hat{\Sigma}_\beta$, the subject-specific random effects, $\hat{u}_i$, and their covariance matrix, $\hat{\Sigma}_{u_i}$, $i \in 1,\ldots I$.

2. For each of the K imputations:

   (a) Draw $\beta^\star$ from $N(\hat{\beta}, \hat{\Sigma}_\beta)$.

   (b) For each subject i with missing data at the end of the study, draw $u_i^\star$ from $N(\hat{u}_i, \hat{\Sigma}_{u_i})$. Denote the i^{th} patient's column-vectors of covariates for the fixed and random effects at the end of the study (time J) by x_{iJ} and z_{iJ} respectively. Then calculate the predicted probability of a positive response at the end of the study as

   $$\pi_{iJ}^\star = \text{expit}(x_{iJ}'\beta^\star + z_{iJ}'u_i^\star).$$

   Finally draw the imputed y_{iJ} from the binomial distribution

   $$y_{iJ} \sim \text{Bin}(\pi_{iJ}^\star, 1) \tag{5.15}$$

   (c) Put the imputed and observed data together to obtain the k^{th} imputed data set.

3. Analyse each of the K imputed data sets, and combine the results using the usual rules for multiple imputation.

This procedure is only approximate, for a number of reasons. The principal one is that $\hat{u}_i$ and $\hat{\Sigma}_{u_i}$ are a function of the data and $\hat{\beta}$. Thus, when we draw $\beta^\star$, we should really update the estimate of $\hat{u}_i$ and $\hat{\Sigma}_{u_i}$ before drawing the $u_i^\star$. In practice, doing this entails considerable additional programming. Our intuition is that the benefit from this extra work may often be small — especially if (as will often be the case) the variance of $\hat{\beta}$ is small relative to that of $\hat{u}_i$. However, this remains to be formally investigated.

If we are imputing missing data from another discrete distribution, this approach can still be used. Clearly the underlying SS model needs to be for data from that distribution. In addition, we replace the binomial distribution in (5.15) with the appropriate discrete distribution.

Before illustrating this approach on the longitudinal binary data, consider how this algorithm works if we fit an unstructured covariance matrix for the random effects over time (5.8). Suppose subject i is observed at time 1 and 2 but not at time 3. Then we will not have $\hat{u}_{3i}$ and $\hat{\Sigma}_{u_{3i}}$. By contrast, if we fitted a model without a separate random effect at each period, such as random intercepts and slopes, we would have an estimate of the random effect for each subject, regardless of whether they were observed on all occasions.

Fortunately, `SAS PROC NLMIXED` will predict $\hat{u}_{3i}$, regardless of whether the subject was observed at period i. Otherwise, we can proceed as follows. As we have fitted the unstructured covariance matrix for the random effects over time (5.8), we can draw $u_{3i}^\star$ from the conditional normal distribution of u_{3i} given $\hat{u}_{2i}, \hat{u}_{1i}$, knowing that the joint distribution is

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \sim N(0, \hat{\Omega}_u),$$

where $\hat{\Omega}_u$ is the estimated variance covariance matrix of the random effects.

EXAMPLE 5.7 *Longitudinal binary data*

Recall the longitudinal binary data has a monotone withdrawal pattern at the subject level, but that each subject has a different number of tests in each period. This means that the missing data is not monotone. For this example, we assume our intended analysis, were the data complete, would be to use only the data from period 3, and to fit a marginal, or population-averaged, estimate of the treatment effect adjusted for baseline, using generalised estimating equations with an exchangeable correlation structure.

As the number of tests varies in each period for each subject, when subjects withdraw, we do not know how many tests they have missed. To cope with this we need to insert an extra step into the algorithm above, as shown below.

To use the approach described above, we first fit model (5.8) to the data, giving the parameter estimates shown in Table 5.6, right column. In addition we request the variance/covariance matrix of these parameter estimates. We also ask `SAS PROC NLMIXED` to predict the u_{i3} and its standard error, which it will do for every subject regardless of whether they were observed at period 3 or not.

Then, as described above, for the 52 subjects missing data at period 3 we proceed as follows. For each of the K imputations, we use the above algorithm to impute the probability of a positive response, $\pi^{\star}_{i3}$. Before imputing a subject's missing response, we first impute the number of tests they underwent.

We impute the number of tests within treatment arm; if desired this imputation model could be refined to take account of possible dependence on previous responses. In the active treatment arm, the number of tests is approximately $N(19, 7.75^2)$; in the placebo arm it is approximately $N(20, 7.07^2)$, irrespective of period. To impute the number of visits, we draw from the appropriate normal distribution and round to the nearest integer. Values less than 2 are replaced by 2.

Suppose in imputation k we impute n_{ik} tests for subject i in period 3. Implementing the algorithm above, We then

1. draw β's
2. draw u's
3. calculate linear predictor
4. impute the n_{ik} missing y's

We use this algorithm to impute completed data sets at time 3, then analyse each one by fitting a GEE, to obtain an estimate of treatment adjusted for baseline. These estimates are then combined in the usual way. Note the GEE program gave robust standard errors, and these were used to calculate the 'within' component of variance. Table 5.12 shows the results.

As the number of tests a subject undergoes varies between imputations, the results can vary considerably. Thus, 5 imputations is not nearly sufficient. In this case, 5 similar data sets give an over-estimate of the treatment effect and an underestimate of its variance. With 50 imputations, the results have settled down. Compared with the observed data, the treatment estimate is increased, as is its statistical significance. With 100 imputations, the treatment effect is borderline significant at the 5% level.

Model/method	Estimated treatment effect	Std. Err.	D.F.	Z	p-value
Observed data	0.57	0.334	N/App	1.71	0.09
5 imputations	0.95	0.345	959	2.76	0.006
10 imputations	0.87	0.389	150	2.23	0.03
50 imputations	0.73	0.417	361	1.76	0.08
100 imputations	0.78	0.401	966	1.94	0.052
Rescaled SS (details in text)	0.69	0.293	N/App	2.35	0.02

Table 5.12: 'Population averaged' estimates of treatment effect (log odds ratio) at period 3, obtained using multiple imputation from the SS model. 'N/App': not applicable for the model/method

We compare this with the re-scaled SS treatment effect, derived from the estimated treatment coefficient in column 3, row 3 of Table 5.6. Here, the scale parameter is $\sqrt{1+0.3458\times 8.87}\approx 2$, so the point estimate is similar. The standard error is a little smaller. This is probably due to a combination of likelihood methods being more efficient and not correcting for the variance of the estimate of σ_u^2. □

5.5.1 Extension to missing baseline

In principle the methods described above can be readily extended to handle missing baselines, although in practice fitting the models may not be easy. If a baseline is missing, it can be included in a mixed model as a response, and then the mixed model used for imputation, along the lines described in the previous Section. Alternatively, if baseline is continuous, the model described in §5.3.1 can be used, and missing baselines imputed.

5.6 Additional issues

In this Chapter we have considered methods for binary and binomial data, which are the most common form of discrete data that arise. In principle, the methods extend directly to other discrete data, by simply changing the response distribution and link function. Such SS models can be fitted in `SAS PROC NLMIXED`. Further, if data are monotone missing, `SAS PROC MI` will automatically detect if the response has more than two levels and replace the logistic regression model with the proportional odds model (McCullagh, 1980). Additional options give the general multinomial model.

Currently, the only option to impute in the non-monotone setting available in `SAS PROC MI` uses the multivariate normal distribution. After imputation, this procedure will also round imputations to the nearest integer for discrete data. While this makes the imputations more palatable for

some 'consumers', and may well be an essential pre-requisite to fitting the substantive model, it does not make the underlying assumption of normality more plausible. In particular, problems are likely to arise where the link function is not approximately linear. For logistic models, therefore, bias may be induced when many of the fitted probabilities are < 0.1 or > 0.9. Using `SAS PROC MI` with a multivariate normal imputation model for binary data is the same as using it for continuous data, so we have not discussed it here.

`SAS PROC MI` also offers a discriminant imputation method and a propensity score imputation method for binary data with a monotone missing pattern. Discriminant imputation can be shown to be a close approximation to logistic regression (Carpenter, 1983), so we prefer the latter. Regarding propensity score imputation, our experience chimes with the comment in the `SAS` v9 manual,

> The propensity score method does not use correlations among variables and is not appropriate for analyses involving relationship among variables, such as a regression analysis (Schafer (1999), p. 11). It can also produce badly biased estimates of regression coefficients when data on predictor variables are missing (Allison, 2000).

As discussed in more detail in Chapter 4, there is a routine in `stata`, called `ice`, which implements the method for monotone missing data when the missing pattern is not monotone. A similar macro is also available in `SAS`. We reiterate our comments in Chapter 4, that this method is still in its infancy and lacks a theoretical basis. Until it is more fully understood, and as there exist theoretically sound alternatives, it cannot be recommended for regulatory analysis.

5.7 Conclusions

In this Chapter we have developed methods for handling missing continuous data from Chapters 3 and 4 for discrete data, illustrating the various methods with longitudinal binary data. All the methods are valid under the assumption the data are MAR; only one of the methods strictly requires a monotone pattern of missingness.

Figure 5.3 gives an overview of our conclusions. Broadly speaking, if SS estimates are required, then fitting the appropriate SS model will give sensible answers even if subjects have non-monotone missing observations. However, care needs to be taken over the random effects component of the model, as this can affect estimated treatment effects quite substantially. As for continuous data, we favour as little structure as possible. Based on results for continuous data, §3.2.1, with more than 50 subjects in each group any loss of power is likely to be negligible.

For PA estimates, if the missing data follow a monotone pattern, we recommend monotone sequential imputation. If this is not desirable, for example because excluding subjects with non-monotone withdrawal potentially causes serious bias, multiple imputation with a SS imputation model is a natural alternative.

Finally, the examples in this Chapter suggest that re-scaling SS point-estimates to obtain approximate PA ones can work quite well, but does ignore uncertainty in the *estimated* scale factor. It does however have the advantage of preserving the inferences obtained from the likelihood based subject-specific analysis.

Part III

Chapter 6

Sensitivity analysis

In part II we discussed the analysis of data under the 'Missing At Random' (MAR) assumption. Here we look at some methods for investigating the sensitivity of inference to this assumption. This usually means introducing into the problem, by one route or another, departures from this assumption and this in turn means that we need to consider MNAR models. There are three main routes for this, and we begin by outlining the generic forms of these.

Let Y_O denote the observed data, Y_M data we intended to collect but could not, and R the indicator of whether data were collected or not. Suppose we are interested in some parametric quantity $\theta(Y_O, Y_M)$ which, ideally (were all the data observed), would be estimated from the observed and missing data. This might for example be the principal treatment comparison. To estimate θ in practice using maximum likelihood applied to the *observed* data we need to integrate (*i.e.* average) out the missing observations from the joint likelihood of the observed and missing data:

$$\hat{\theta} = \int \theta(Y_O, Y_M)[Y_M, Y_O, R]\, dY_M.$$

As we are no longer assuming MAR, we cannot eliminate the withdrawal mechanism from this. We can approach this problem by factoring $[Y_M, Y_O, R]$ in one of two ways:

$$\begin{aligned}
\hat{\theta} &= \int \theta(Y_O, Y_M)[Y_M, Y_O, R]\, dY_M. \\
&= \int \theta(Y_O, Y_M)[Y_M, Y_O | R][R]\, dY_M \qquad (6.1) \\
&= \int \theta(Y_O, Y_M)[R | Y_M, Y_O][Y_M, Y_O]\, dY_M \qquad (6.2)
\end{aligned}$$

Option (6.1) is called a 'pattern mixture' model. In contrast to the MAR setting, the distribution (or pattern) of the data is different for the observed and unobserved portions. In the simplest case, when R is a scalar, this approach models the two portions (or patterns) separately, and then calculates a weighted average (or mixture), with weights $\Pr(R = 1)$ and $\Pr(R = 0)$.

Option (6.2) is known as a selection model. Suppose we draw a candidate Y_M from the distribution of $Y_M | Y_O$. As data are MNAR, this draw is weighted by its probability of being observed (or *selected* for observation).

A third approach, which we discuss a little further in §6.2.4, introduces latent variables upon which both the measurement and withdrawal process depend. Integration is then required over this latent structure. For a discussion of these three views of the modelling problem, and for a description of a framework that incorporates them all, see Diggle (1998).

Looking at (6.1) and (6.2), it is clear that there is enormous scope for different kinds of pattern mixture and selection models. For example, a natural starting point for the pattern in the unobserved data is the pattern in the observed data, but there are many possible modifications to go from one to the other. Figure 6.1 illustrates this. We suppose that we have a trial with longitudinal follow-up, and if a patient is observed at each follow-up visit we will fit a straight

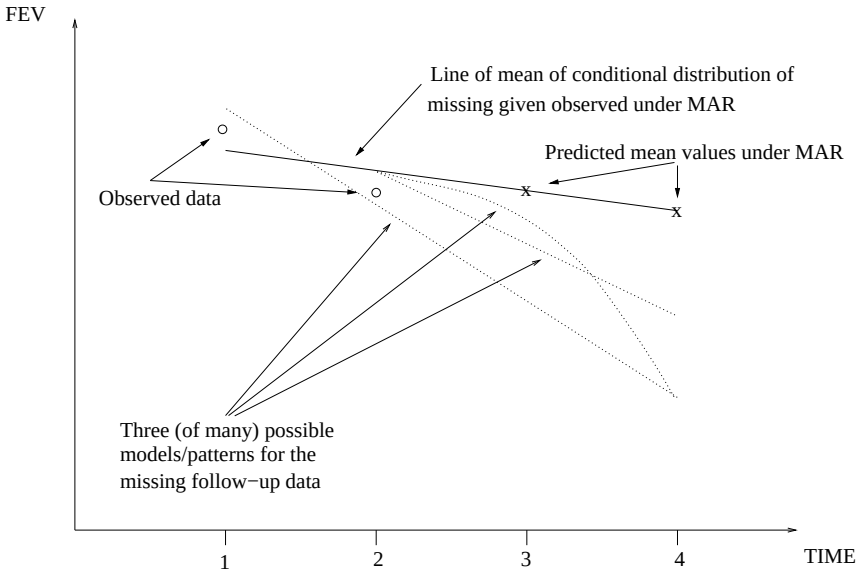

Figure 6.1: Hypothetical asthma trial: illustration of three of the many models/patterns possible for the missing data, when a patient withdraws after the second follow-up

line to their data. The figure shows a patient who withdrew after their first two visits, together with four options for the 'pattern' of their response after withdrawal: the MAR option (same pattern as for those who do not withdraw) and three others.

Figure 6.1 makes explicit the key issue — with no data, we cannot say which pattern is more likely. The problem is not alleviated if we adopt a 'selection model' instead of a 'pattern mixture' approach. This is because the relationship between response and the unseen data can only be estimated subject to uncheckable modelling and distributional assumptions (Kenward, 1998). Again, once we start thinking about models for the withdrawal process, there are many possibilities. Any suggestion that the selection model is a less arbitrary framework for sensitivity analysis is therefore illusory. In the light of this, it is no surprise that the missing data literature contains hundreds of proposals for sensitivity analysis.

In order for sensitivity analyses not to be completely arbitrary, some guiding principles are needed. The following may be useful:

1. sensitivity analyses should be pre-defined, addressing the impact of clinically plausible departures from MAR;

2. sensitivity analyses should be as transparent as possible to clinical investigators and regulators, and

3. the statistical methods should be applicable to a wide range of settings.

To address the first point, possible sensitivity analyses should be discussed with investigators and regulators as part of the trial planning, and described in the protocol where possible. This helps avoid the evidence provided by a trial being devalued by post-hoc, arbitrary sensitivity analyses. As part of this process, it may be desirable to formally collect opinion on the differences between responders and non-responders. We describe an approach for this below.

The second point is very important; the greater the understanding of the methods amongst investigators and regulators, the more likely (i) they will be able to inform/direct the analysis to address their concerns and (ii) they will accept the results.

The third issue is related to the second, and concerns the necessity of becoming familiar with how an approach performs in a variety of settings if one is to be aware of problems that may arise and how they can be addressed. Such awareness is essential if one is to have confidence in a method.

The methods we describe in this Chapter are ones we think can be used in the principled way described above. We note that what we term the 'transparency' of the method should not be equated with technical simplicity. The selection models we describe below are an example of this: while we believe they are fairly transparent, quite complex statistical machinery is needed to fit them.

For other approaches that have been suggested for sensitivity analysis in a clinical trial setting, related in varying degrees to those suggested here, see Molenberghs and Kenward (2007) part V, Molenberghs *et al.* (2006), Verbeke and Molenberghs (2000), Ch. 19–20 and Scharfstein *et al.* (1999).

The plan for the remainder of this Chapter is as follows. First, we briefly discuss the CPMP guideline on sensitivity analysis. Then we discuss sensitivity analysis via (i) selection models and (ii) pattern mixture models. In each case we give examples and code to illustrate our approach.

6.1 A note on the CPMP guideline

The CPMP points to consider on missing data (Committee for Proprietary Medicinal Products (CPMP), 2001) stress the importance of sensitivity analysis and agreeing its nature and scope in advance. However, their focus is very much on comparing the results of certain 'methods', rather than comparing the sensitivity of the conclusions to varying the assumptions about the missingness mechanism.

This distinction is potentially important, as it is possible to implement a range of methods, all of which make very similar assumptions about the missingness mechanism. This both misses the point of sensitivity analysis, and can lead to misleading conclusions.

Several of the methods discussed in the guidelines are 'degenerate'. In other words, they explicitly or implicitly replace missing values by a single value, *i.e.* a degenerate statistical distribution. LOCF and best/worst case imputation are examples of these. All the statistical methodology surrounding trials has been developed to allow correct inference in the presence of uncertainty. It is therefore strange to adopt the 'certainty' of a best or worst value when it comes to sensitivity analysis. Once we view such methods as examples of degenerate imputation, it follows that the imputations are extremely implausible, and therefore the conclusions are likely to be misleading. This applies both to resulting estimates of treatment effects and standard errors.

Further, when data are continuous, allocation of a 'best' or 'worst' value in inherently difficult. Should it be the worst possible physiological value, or the worst one observed in that particular

intervention arm, or the worst observed in the trial? Fruitless debate over such artificial questions is avoided when we impute distributions, when discussion moves to the relative likelihood of such values (amongst others). As we show below, such information can be obtained from investigators, synthesised, and combined with the trial information, all in a scientifically valid way.

We therefore do not consider LOCF, best or worst case imputation further in this Chapter.

6.2 Selection models

If we used a mixed model for the MAR analysis of the responses, as in Chapter 3, this is a very natural approach for sensitivity analysis. This is because we can continue with the same model for the responses. The difference is now we have to add a model for the reason for missing data. Then the two models need to be fitted together, which is most readily done using MCMC methods in `winBUGS` (Spiegelhalter *et al.*, 1999).

Depending on the context, different forms may be more appropriate for the model for withdrawal or non-response. We term these *selection models* to differentiate them from the model of interest for Y_{ij}. We now discuss some selection models. Suppose we have a trial which intends to collect data at follow-up visits $j = 1,\ldots,J$ on each of $i = 1,\ldots,I$ patients. Let $R_{ij} = 1$ if patient i attends follow-up visit j.

6.2.1 Model I: no withdrawal — observing a patient is always possible

Our first model assumes that patients do not 'withdraw', rather there is always the chance they may be observed at the next follow-up visit. In this case at each time point, we can use a logistic model for response. A simple model is:

$$\text{logit}\Pr(R_{ij} = 1) = \alpha_j + \delta Y_{ij}, \quad i = 1,\ldots I, j = 1,\ldots J. \tag{6.3}$$

In words, the log odds of observing patient i at visit j depends on the visit, (α_j), but also on the response, Y_{ij}. Thus a positive value of δ implies the log-odds of observing the response increases with the value of the response.

Clearly, model (6.3) cannot be fitted alone, for $R_{ij} = 1$ if Y_{ij} is observed. However, using numerical integration (over the unseen Y_{ij}'s), it can be fitted in conjunction with a mixed model for the response that we used for the MAR analysis. The numerical integration can either be done in a frequentist framework (Diggle and Kenward, 1994) or using MCMC methods, such as in `winBUGS` (Carpenter *et al.*, 2002; Spiegelhalter *et al.*, 1999). Further although δ can be estimated, the estimated value and its standard error depend critically on the distributional assumption made for the missing data. Unfortunately, we cannot assess the plausibility of this assumption, as we have not seen these data. Thus, we conclude (cf Kenward (1998); Carpenter *et al.* (2002)) it is better not to estimate δ. Rather, from the observed data and in discussion with investigators and regulators, we identify a set of possible values for the (log) odds of response per unit change in Y_{ij}. We then fit the model with each of these values for δ in turn, and explore how sensitive our estimated treatment varies.

Follow-up visit	Response seen from patient i?	Value of R_{ij} under model 1	Value of R_{ij} under model II	Response seen from patient i?	Value of R_{ij} under model 1	Value of R_{ij} under model II
1	Yes	1	1	Yes	1	1
2	Yes	1	1	No	0	1
3	Yes	1	1	Yes	1	1
4	No	0	0	Yes	1	1
5	No	0	–	No	0	0
6	No	0	–	No	0	–

Table 6.1: Values of R_{ij} under models I and II when (left) a patient is observed at each follow-up visit until they withdraw and (right) a patient has an interim missing value, is subsequently seen and then withdraws

Note that if we constrain $\delta = 0$, then model (6.3) can be fitted alone; this corresponds to a MAR assumption, that the probability of response does not depend on the unseen observation. This shows that the model is too simple, for we have seen that MAR means that given the observed data, there is no further dependence of withdrawal on the unseen data. However, the only 'observed' data in (6.3) is current response, Y_{ij}. A more plausible model includes treatment and previous visit:

$$\begin{aligned}\operatorname{logit}\Pr(R_{ij}=1) = {} & \alpha_j + \beta_k 1[\text{Patient } i \text{ has treatment } k] \\ & + \gamma Y_{i,j-1} + \delta(Y_{ij} - Y_{i,j-1}), \quad i = 1,\ldots I, j = 1,\ldots J. \end{aligned} \tag{6.4}$$

In this parameterisation, conditional on visit, treatment and response at last visit, δ is the additional change in the log odds of a patient's response per 1 unit increase/decrease in Y_i from the last visit. While, from the the point of view of the resulting treatment estimates, having $\{\delta(Y_{ij} - Y_{ij-1}) + \gamma Y_{ij-1}\}$ in (6.4) is equivalent to having $\{\delta Y_{ij} + \gamma' Y_{ij-1}\}$, we find (6.4) easier to explain to investigators and regulators.

In many applications we may want to extend (6.4) to include other variables too. Theoretically, any variables that are included (conditionally or as responses) in the mixed model to maximise the plausibility of MAR should be included in the response model. This is likely to make treatment estimates less sensitive to MNAR. To see this, note that δ is the change in the log-odds of response per 1 unit change in Y_{ij} *conditional* on all the other variables in the model. Thus, if (aside from Y_{ij}), the model gives good predictions for the probability of withdrawal, then it is in turn plausible that the residual dependence of response Y_{ij} on withdrawal is likely to be small. In other words plausible values of δ are likely to be close to 0. Thus the treatment estimates are less likely to be very sensitive to MNAR.

6.2.2 *Model II: no data available after patient withdrawal*

Model I always has a non-zero probability that a patient might be seen at a visit. This is obviously wrong if the patient has withdrawn from follow-up altogether. Adopting the following

model addresses this:

$$\begin{aligned}\text{logit}\Pr(R_{ij}=1|R_{i(j-1)}=1) &= \alpha_j+\beta_k 1[\text{Patient } i \text{ has treatment } k]+\gamma Y_{i,j-1}+\delta(Y_{ij}-Y_{i,j-1}),\\ \Pr(R_{ij}=1|R_{i(j-1)}=0) &= 0, \quad i=1,\ldots,I; j=2,\ldots J. \end{aligned} \tag{6.5}$$

Note that, for simplicity, we start from follow up visit 2, $j=2$. Otherwise, we need to include baseline, or drop the dependence on the previous observation at $j=1$. Computationally, as the code for the example below shows, this model is no more difficult than model I. However, for many trials, where interim withdrawal is a secondary issue to patient withdrawal, it is more satisfactory.

If we change the logit link to the 'complementary log-log' link, $\log(-\log\Pr(R_{ij}=1))$, then the model turns out to be a discrete time proportional hazards model (Aitkin *et al.*, 1989), and the coefficients can be interpreted as log-hazard ratios. In practice, they are likely to be similar as the logistic and complementary log-log functions are similar unless probabilities are close to 0 or 1.

Although this is a model for withdrawal, it is possible to include patients with interim missing values. A simple approach is to assume that these interim missing values are MAR, while withdrawal is MNAR. Depending on the context, this may or may not be plausible. To include such patients, we just have to (re-)define R_{ij} to be 1 until the visit when the patient withdraws: see Table 6.1. Note that, under model II, for visits following withdrawal there is no point in including R_{ij} in the model as the probability of seeing the patient is 0.

6.2.3 Comparison of models I and II

Suppose that a trial has no interim missing data, and that the higher a response, the less likely the patient is to withdraw. Then model II, with $\delta > 0$, is appropriate. If we instead fit model I, then we will tend to impute lower values for the missing observations. This follows because at each visit after withdrawal model I has a non-zero probability of observing the patient. If they are not observed, their imputed response will be lower. This effect will increase over time as the model tries to fit the true probability of observing the response, 0, for that patient.

Thus, if treatment improves response, and withdrawal is predominantly in the treatment arm, model I might yield a conservative estimate of treatment effect. If, as is more likely, withdrawal is predominantly in the placebo arm, the model will tend to over estimate the treatment effect. Neither is desirable in practice.

6.2.4 Some other models

We have encountered trials where the response is highly variable and withdrawal depends on the trend over follow-up rather than the last couple of observations. In this case, on option is to model the covariance of Y_{ij} with random intercepts and slopes, and then replace Y_{ij} in (6.5) with the random slope for patient i, or preferably the random slope multiplied by time since randomisation (since otherwise the effect of the slope on the probability of withdrawal is the same at all follow-up times). While these models can be slightly easier computationally, they

are usually more difficult to explain to investigators and regulators, who are often unfamiliar with the implications of random intercepts and slopes in continuous and discrete models. They are special cases of the latent variable models mentioned earlier; for an example see Verzilli and Carpenter (2002), and a more general discussion see Diggle (1998).

6.2.5 *Software*

All these models require integration over the missing data. This can be done using conventional numerical integration. Molenberghs *et al.* (2006) provide SAS code for fitting selection models using the EM algorithm with numerical integration as part of the E step. Some simpler models can also be fitted using `SAS PROC NLMIXED` and some of the random coefficient models can also be fitted in `MLwiN` (Rasbash *et al.*, 2004). In practice, we have found the most convenient route is to adopt a Bayesian model, with vague priors, so that posterior means will approximate maximum likelihood estimates. Markov Chain Monte Carlo (MCMC) methods (Gilks *et al.*, 1996), as implemented in `WinBUGS`, provide a convenient tool for fitting such models, especially if one can work from code for a similar model in a different setting. We therefore give the code for the examples below in Appendix C.

MCMC methods work by setting up a simulation process which converges over time to the true Bayesian posterior distribution. The speed of convergence can be greatly increased by mean-centreing quantitative covariates in the model. This can be important, as realistic models for even moderate trials can take a considerable time to fit.

6.2.6 *Bells and whistles*

There are many possible refinements of the selection model that could be included. While each makes the model more plausible, each also makes the model more complex, so less interpretable by non-statisticians. As this devalues the point of sensitivity analysis, in practice such additional refinements should be avoided, or kept to a minimum.

For example, we have already noted that variables included in the response model to maximise the plausibility of missing at random should ideally also be included in the selection model. In practice, including every past observation as a predictor for response at visit j is likely to lead to large selection models which are complicated to fit and interpret. Thus we recommend concentrating on visit, treatment, previous and current response, plus perhaps one or two other key variables.

Likewise the effect of variables on the probability of response could vary as the trial progresses. However, unless such interactions are identified as highly likely before the data are collected, it is best not to try and estimate them as part of the sensitivity analysis.

Again, one can imagine that, within the same trial, some patients' response pattern might be best described by Model I above, but others by Model II. Although possible, fitting both models simultaneously as part of a sensitivity analysis is more tricky computationally. We therefore recommend focusing on one of the models, which can be chosen either because it best describes the behaviour of the majority of patients or because it will give more conservative estimates of treatment efficacy.

EXAMPLE 6.1 *Sensitivity analysis with the Isolde data*

We illustrate the selection model approach on the Isolde data, comparing Models I and II. Recall this dataset has two treatment groups, and 6 follow-up visits. As in Chapter 3, we fit a normal model to FEV_1, fitting time as a class (factor) variable with a full treatment×time and baseline×time interaction. As before we have an unstructured covariance matrix. Thus the model is:

$$\begin{aligned} \text{E}\, Y_{ij} &= \phi_j + \eta_j 1[\text{patient } i \text{ has active treatment}] + \nu_j(\text{baseline FEV}_1)_i; \quad j = 1,\ldots,6 \\ \text{Var} \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{i6} \end{pmatrix} &= \Sigma, \text{a } 6 \times 6 \text{ matrix}, \end{aligned}$$

where we denote coefficients by ϕ, η, ν to differentiate them from those in (6.5). Thus this model has 18 parameters for the mean of Y and 21 in the covariance. We will focus on the baseline adjusted estimate of treatment at the final visit, η_6.

In this data set, around 45% of patients have interim missing values; however in terms of the total number of missing values, those due to withdrawal predominate. It is therefore of interest to compare the results of fitting model (6.4) with model (6.5) for the selection process. For the latter model, we keep patients with interim missing observations within the model, assuming these values are MAR and defining R is as shown in the right half of Table 6.1.

We fit the model in `winBUGS` (Appendix C). To improve convergence we mean centred baseline FEV_1 and response. For all the models we had a 'burn in' of 3000 samples. Then to get a reasonably uncorrelated sample from the posterior, we then sampled a further 50,000 but only included every 10th sample in the estimates of the posterior means, SDs and 95% credible intervals in Table 6.2.

As a check on the `winBUGS` code we first fitted the model with $\delta = 0$, and compared the results with `SAS PROC MIXED`. They were very similar. In this model, the coefficient estimating the log-odds of response on FEV_1 at the previous visit was estimated as 0.31 (std. error 0.12). Recalling that FEV_1 is measured in litres, this means that, after adjusting for treatment group and visit, the odds of responding at the next visit are 3% higher for every 100 ml increase in FEV_1. For the sensitivity analysis, for both models, we therefore set $\delta = 0.1$. This gives an additional 1% increase in the odds of response per 100ml increase in FEV_1 between the last and current visit.

Table 6.2 gives the results. As noted in the previous paragraph, the MAR estimates from `SAS PROC MIXED` and `winBUGS` agree closely, as do the other MAR parameter estimates (not shown). At visit 6, adjusting for baseline, patients with active treatment have an FEV_1 90 ml higher on average, with $p < 0.001$. Under MNAR both models show an increased effect of treatment. This is because more patients withdraw, and they withdraw earlier, in the placebo arm. Their imputed values are lower under MNAR than MAR, and hence the final treatment effect is greater. As discussed above, the fact that the estimated treatment effect is fractionally greater with model I (6.4) is expected; we anticipate the difference between the models to increase with greater values of δ. Note the standard error is little changed here. If however, we sampled δ from a distribution (possibly obtained from experts), with known mean and variance (akin to the approach in §6.5) we would expect the standard error to increase.

Coefficient	MAR estimates ($\delta = 0$)		MNAR estimates ($\delta = 0.1$)	
	SAS PROC MIXED	winBUGS	winBUGS Model (6.5)	winBUGS Model (6.4)
Baseline	0.888 (0.021)	0.888 (0.021)	0.887 (0.021)	0.888 (0.022)
Treatment	0.092 (0.020)	0.092 (0.021)	0.093 (0.021)	0.094 (0.020)

Table 6.2: Isolde data: estimated coefficients (standard errors) for baseline and treatment at visit 6, from MAR and MNAR models

6.3 Extension to discrete data

The models fitted in winBUGS above can be extended for longitudinal discrete data, by changing the mixed model for the response to a non-linear mixed model. Binomial, Poisson and ordinal responses can all be handled this way. One caution is that experience shows that this approach works best when the MAR model can be fitted fairly simply by maximum likelihood (using SAS PROC NLMIXED) or penalised quasi-likelihood methods (for a comparison see Ng *et al.* (2006)). If the MAR model is hard to fit, then obtaining convergence of the MCMC sampler for the MNAR model will be more tricky.

6.4 Pattern mixture models

We now discuss an alternative approach to sensitivity analysis, via the pattern mixture approach. We shall see below that this approach can be readily combined with multiple imputation. It can also be readily used with other approaches.

A pattern mixture model conditions the joint density, $[Y_M, Y_O, R]$ as

$$[Y_M, Y_O, R] = [Y_M, Y_O|R][R].$$

Under MAR, $[Y_M, Y_O|R = 1] = [Y_M, Y_O|R = 0]$, that is to say the joint density of patient's responses is the same for patients those with full and partially observed data. Under MNAR, the two are different.

A natural way to perform sensitivity analysis is to model the observed data, then assume the model for the missing data is a slight modification of this. By varying the modification, we can often rapidly perform a range of sensitivity analyses. We describe a method for doing this below. Further, we show how expert opinion, from investigators or regulators, on differences between responders and non-responders can be incorporated into the analysis.

Our approach is based on that described by White *et al.* (2007), who give the technical details. Here, we describe the central ideas, and illustrate the method using data from a trial of interventions to improve the quality of peer review. The application to clinical data is direct.

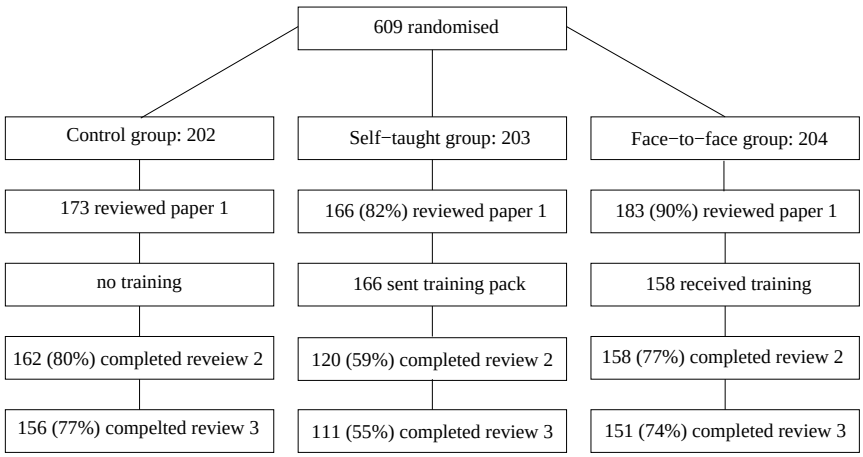

Figure 6.2: Progress of patients through the peer review trial

EXAMPLE 6.2 *Peer review trial*

While there are many studies illustrating the inadequacies of peer review (Rennie, 1999), there have been few evaluations of interventions that try to improve peer review and no randomised controlled trials of the effects of training Callaham *et al.* (2002, 1998). This motivated Schroter *et al.* (2004) to carry out a single blind randomised controlled trial among reviewers for a general medical journal, comparing two different types of training (face-to-face training or a self-taught package) with a control group.

Each participating reviewer was sent a baseline article to review (paper 1). If this was returned, one intervention group received a full day's face-to-face training, and the other group was mailed a self-taught training package. Two to three months later participants who had completed their first review were sent a further article to review (paper 2); if this was returned a third paper was sent three months later (paper 3).

Reviewers were sent manuscripts in a similar style to the standard BMJ request for a review, but were told these articles were part of the study and were not paid. The three articles were based on three previously published papers, with original author names, titles and location changed. In addition 9 major and 5 minor errors were introduced.

The principal outcome considered here is the mean Review Quality Index (RQI) (van Rooyen *et al.*, 1999a). This is a 7-item validated instrument, with each item ranging between 1 and 5 (van Rooyen *et al.*, 1998, 1999b; Walsh *et al.*, 2000). Each review was rated independently by two editors and then the mean score of the 7 items, averaged over the two editors, was used. Full details are given in Schroter *et al.* (2004).

The progress of participants through the trial is summarised in Figure 6.2. Note that a substantially higher proportion of reviewers dropped out in the self-taught arm. Here, we focus on the results for paper 2. This is because analysis of the complete data for paper 2 suggested the postal intervention was beneficial (first row Table 6.3) whereas the review of paper 3 (six months after intervention) did not show any significant differences between the arms in terms of RQI or number of major errors detected.

Table 6.4 shows the mean RQI at paper 1 for those who did, and did not, review the second paper. It suggests that the missing observations may not be missing completely at random;

		Postal vs. control				Face-to-face vs. control			
		Posterior		95% credible		Posterior		95% credible	
		mean	sd	interval		mean	sd	interval	
Complete cases		0.291	0.077	0.140	0.442	0.160	0.071	0.021	0.299
c=0	Approximation	0.246	0.153	-0.053	0.545	0.144	0.100	-0.052	0.341
	MCMC	0.246	0.151	-0.042	0.564	0.144	0.100	-0.050	0.344
c=0.5	Approximation	0.246	0.140	-0.028	0.520	0.144	0.091	-0.033	0.322
c=1	Approximation	0.246	0.126	-0.001	0.493	0.144	0.080	-0.013	0.301
	MCMC	0.246	0.126	0.004	0.505	0.145	0.080	-0.014	0.302

Table 6.3: Peer review trial: posterior mean intervention effect, standard deviation and 95% credible intervals. Results are unadjusted for covariates. Uncertainty in posterior means due to Monte Carlo estimation is less than 0.0006

furthermore, in contrast to the control group, in the intervention groups reviewers who dropped out tended to have worse RQI scores for paper 1. This is particularly important in the postal arm, because the number of reviewers dropping out in this arm is substantially greater.

When the investigators saw these results, they decided to approach BMJ editorial staff to try and elicit prior information on the difference between non-responders and responders in this study, with a view to performing a Bayesian sensitivity analysis to assess the impact of non-response. Further details about how this was done are described below. □

6.4.1 Pattern mixture model for MNAR

For simplicity, we first consider the case of a randomised trial with two arms (intervention and control), a single, normally distributed outcome and no covariates. The extension to more arms involves no new concepts. We outline how to include covariates below.

Suppose n_C subjects are randomised to the control arm. Let Y_{iC} be the response for the i^{th} subject in the control arm, $i \in (1,\ldots,n_C)$, and let $R_{iC} = 1$ if this subject completes the trial. Conversely, let $R_{iC} = 0$ if this subject drops out, so that Y_{iC} is unseen. Denote by π_C the probability of withdrawal in the control arm, so $\pi_C = \Pr(R_{iC} = 0)$. For the intervention arm, define n_I, Y_{iI}, R_{iI} and π_I analogously.

For the control arm, our model is that those responses that are observed come from a distribution with mean μ_C, and variance σ^2, while those that are unobserved come from a shifted distribution, with mean $(\mu_C + \delta_C)$ and variance σ^2_M. We write this as

$$\begin{aligned} Y_{iC}|R_{iC} = 1 &\sim (\mu_C, \sigma^2) \\ Y_{iC}|R_{iC} = 0 &\sim (\mu_C + \delta_C, \sigma^2_M). \end{aligned} \tag{6.6}$$

For the intervention arm, we define μ_I, δ_I analogously, and assume the variances are equal to those in the control arm (although this can readily be relaxed if desired).

Under our model, the average response in the control arm is $(1-\pi_C)\mu_C + \pi_C(\mu_C + \delta_C)$. Likewise, the average response in the intervention arm is $(1-\pi_I)\mu_I + \pi_I(\mu_I + \delta_I)$. We see that δ_C

		Control group	Postal group	Face-to-face group
Returned review of paper 2	n	162	120	158
	mean	2.65	2.80	2.75
	SD	0.81	0.62	0.70
Did not return review of paper 2	n	11	46	25
	mean	3.02	2.55	2.51
	SD	0.50	0.75	0.73

Table 6.4: Review Quality Index of paper 1 by whether or not paper 2 was reviewed

and δ_I govern the degree of informative missingness. In this case, with no covariates, the MAR assumption corresponds to the case $\delta_C = \delta_I = 0$. Of course, δ_I and δ_C will generally differ. For example, missingness may well be more informative among individuals who have been encouraged to change their behaviour than among controls.

The average effect of the intervention is then

$$\begin{aligned} \Delta &= (1-\pi_I)\mu_I + \pi_I(\mu_I+\delta_I) - \{(1-\pi_C)\mu_C + \pi_C(\mu_C+\delta_C)\} \\ &= (\mu_I - \mu_C) + (\delta_I\pi_I - \delta_C\pi_C). \end{aligned} \tag{6.7}$$

Note that $(\mu_I - \mu_C)$, the average treatment effect amongst completers, can be estimated using the usual complete case analysis. If we denote this by Δ^{CC}, then we simply have

$$\Delta = \Delta^{CC} + (\delta_I\pi_I - \delta_C\pi_C), \tag{6.8}$$

i.e. treatment effect = treatment effect in completers +bias due to informative withdrawal.

We take a Bayesian approach, which allows us to assume a prior distribution for the informative missingness parameters. For the control arm, we assume $\delta_C \sim N(m_C, s_C^2)$. Likewise, for the intervention arm we assume $\delta_I \sim N(m_I, s_I^2)$. Usually, δ_C and δ_I will be correlated, so let

$$c = \mathrm{Cor}\,(\delta_C, \delta_I). \tag{6.9}$$

6.4.2 Analysis

To obtain the posterior distribution we assume non-informative priors, independent of δ, for $(\mu_I, \mu_C, \pi_I, \pi_C, \sigma^2)$. Note it turns out that σ_M does not appear in expressions for the posterior mean and variance of the overall treatment effect, and therefore requires no prior.

By replacing the observed prior with a normal approximation, White *et al.* (2007) derive the following formulae for the posterior mean and variance, enabling the calculations to be performed as a simple modification of an analysis of complete cases. They also present an exact Bayesian analysis (with code) using the `winBUGS` program, which uses the discrete prior distribution implicit in the questionnaire. Both analyses assume normality for the trial response data.

The formulae below give the posterior mean and posterior standard deviation (standard error) in intervention arm compared to the control arm. These are modified in the obvious way if comparisons between other interventions are required.

First, estimate the probabilities of non-response, π_I, π_C, by the observed fractions of non-responders in the intervention and control arms. Denote these fractions p_I, p_C. Recalling the treatment effect (6.8), and the definitions of the mean, variance and correlation of δ_I, δ_C, given above equation (6.9), we can estimate the posterior mean and variance of the treatment effect by *correcting* the usual complete-cases estimates to take account of informative withdrawal:

$$\text{Posterior mean} = \hat{\Delta}^{CC} + C \tag{6.10}$$

$$\text{Posterior variance} = se(\hat{\Delta}^{CC})^2 + V_1 + V_2, \tag{6.11}$$

where $se(\hat{\Delta}^{CC})$ is the estimated standard error of the 'complete cases' treatment estimate. From (6.8), the correction term for the point estimate in (6.10) is

$$C = m_I p_I - m_C p_C \tag{6.12}$$

which uses the experts' best estimates of the mean of δ_C and δ_I, together with the observed withdrawal proportions. The first variance correction term in (6.11) is

$$V_1 = p_I^2 s_I^2 - 2cs_C s_I p_C p_I + p_C^2 s_C^2, \tag{6.13}$$

which allows for uncertainty about δ_C and δ_I using the prior variances s_C, s_I and their correlation c. The second variance correction term in (6.11) is

$$V_2 = (m_I^2 + s_I^2)\frac{p_I(1-p_I)}{n_I} + (m_C^2 + s_C^2)\frac{p_C(1-p_C)}{n_C}, \tag{6.14}$$

which allows for uncertainty about p_C and p_I. This term decreases with sample size and will often be negligible compared with V_1. Approximate posterior credible intervals can be calculated as 'posterior mean' $\pm$ 1.96 $\times \sqrt{\text{'posterior variance'}}$.

In the special case where the missing data are known to be equally informative in each trial arm ($\delta_C = \delta_I = \delta$), the mean correction C becomes $m(p_I - p_C)$ and the main variance correction V_1 becomes $s^2(p_C - p_I)^2$, so the adjustments depend only on imbalance in missingness probabilities in the two arms. If δ_C and δ_I are allowed to be unequal, but assumed to have the same prior variance s^2, then the second term of the posterior variance is $s^2[(p_I - p_C)^2 + 2(1-c)p_C p_I]$. This is typically very sensitive to the correlation c, and increases as the correlation decreases.

6.4.3 *Eliciting priors*

Several writers have argued that priors for (δ_I, δ_C) should be elicited from experts in the field (Kadane and Wolfson, 1998; O'Hagan, 1998). Ideally, prior beliefs would express the viewpoint of an uncommitted observer or consumer of the research, suggesting the choice of experts unconnected with the trial. In practice, it is the trial investigators who are best informed about the circumstances of the trial and most likely to be prepared to have their opinions elicited, and they are therefore the most realistic source of expert opinion.

We consider the elicitation of the prior in two stages. We begin by considering how to elicit prior beliefs in a single arm of a trial, for example for the parameter δ_C in (6.6).

Recall from (6.6) that δ_C is the difference in the control arm between the true mean of the unobserved data and the true mean of the observed data. As there is uncertainty about this parameter, we need to elicit prior information from experts on the distribution of plausible values. There are two principal sources of confusion here. First, we need opinions on the difference in true means, not the difference in the sample means. Secondly, we require information on the distribution of possible values of the parameter, not the distribution of the unobserved data. These points must be stressed to the experts.

We used the questionnaire shown in Appendix B to elicit this information. We asked experts to assign a total weight of 100 over 9 categories in which the difference in the average outcome (review quality index) between responders and non-responders varied from ≤ 1 to ≥ 1 in steps of 0.25.

Most of the questionnaires were completed at a BMJ research seminar, following a short description of the trial methods and a brief presentation explaining the reason for eliciting prior information and introducing the questionnaire.

The questionnaire in Appendix B does not attempt to collect information on the correlation, c, between δ_C and δ_I. Our pilot questionnaire attempted to do this with a further table for investigators to complete. This sought information on the average difference between responders and non-responders in the control arm for each average difference between responders and non-responders in an intervention arm. Unfortunately, it quickly became apparent that this was too unfamiliar for investigators to readily tackle. We therefore had to make do with the same prior for δ in each arm. While this is a limitation, we can still examine the sensitivity of results to a range of values of c.

An alternative, simpler, approach for obtaining an estimate of the correlation that subsequently occurred to us is the following. We ask the expert first to suppose that δ_C (the value in the control arm) takes a specific small value — say -0.5 — and then ask for their best guess of the mean of δ_I. Secondly, we ask the expert to suppose that δ_C takes a specific large value — say $+0.5$ — and then ask for their best guess of the mean of δ_I.

If we denote the proposed values by L_C and H_C, and the expert's conditional mean values by L_I and H_I, then an estimate of $c = \mathrm{Cor}\,(\delta_C, \delta_I)$ is

$$\frac{H_I - L_I}{H_C - L_C} \times \frac{s_C}{s_I},$$

where s_C, s_I are defined above equation (6.9).

6.4.4 *Some additional issues*

White *et al.* (2007) describe how the analysis above should be modified to adjust for covariates, both in the prediction of withdrawal and in the estimation of the treatment effect. They also discuss the subtle, and in practice usually minor, implications this has for obtaining and interpreting the prior.

EXAMPLE 6.2 *Application to peer review trial (ctd)*

Twenty two questionnaires were completed, 2 by investigators who had seen the data, 12 by BMJ editors and editorial staff who had not seen the data and 8 by various other BMJ staff, who

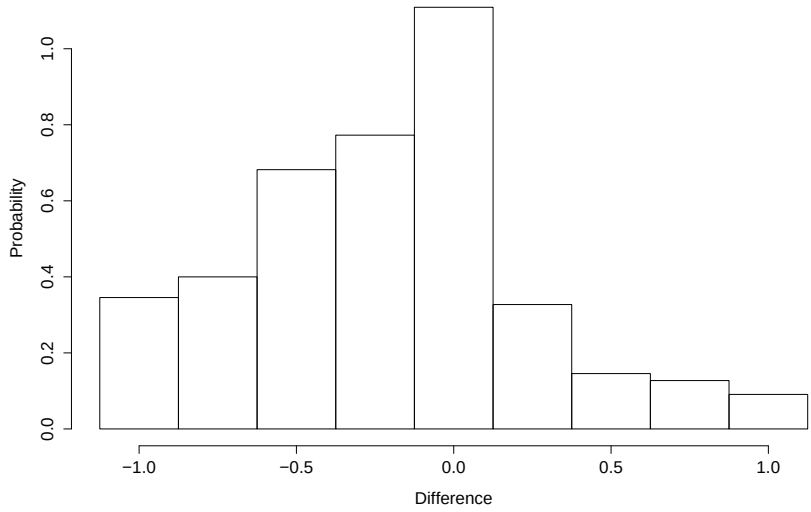

Figure 6.3: Editors' prior distribution for δ, the difference between mean RQI of non-responders and responders

had likewise not seen the data. The two investigators who had seen the data before completing the questionnaire gave priors broadly in line with the others, so they were retained in the sample. 95% of priors had a mean ≤ 0; of these 50% were between -.3 and -0.05. It was therefore reasonable to pool the experts' opinions. The resulting pooled prior distribution is shown in Figure 6.3: it has mean -0.21 and variance 0.46^2. The average within questionnaire variance is 0.4^2, so an approximate intra-class correlation is 0.76, suggesting reasonable consensus among the experts.

As we were unable to elicit information about the correlation between δ's (see §6.4.3), for the approximate analysis we explored the sensitivity of the results to correlations of 0, 0.5 and 1.

Figure 6.4 shows the estimated effect of the postal intervention as c ranges from 0 to 1. Numerical results for $c = 0, 0.5$ and 1, are shown in Table 6.3, where for comparison we include the results of the full Bayesian analysis. This only used correlations of 0 or 1, since fractional values are difficult to implement. `winBUGS` code and further details (including baseline adjusted estimates) are given in White *et al.* (2007).

The unadjusted complete case analyses shows a significant difference at the 5% level in favour of both interventions compared with control. Adjusting for MNAR, with a correlation, $c = 1$, the lower end of the 95% credible interval touches 0. As c declines to zero, the interval width increases further, as expected.

Including prior information on differences between responders and non-responders has two main effects. Firstly, it reduces all the estimated effects, because of the greater degree of missingness in the intervention groups (Table 6.4) together with the prior belief that missing outcomes were on average worse than observed outcomes (Figure 6.3). Secondly, it increases the standard deviation, reflecting the uncertainty about non-response bias. The results are relatively insensitive to the assumed value of c because most of the missing data are in the intervention arm.

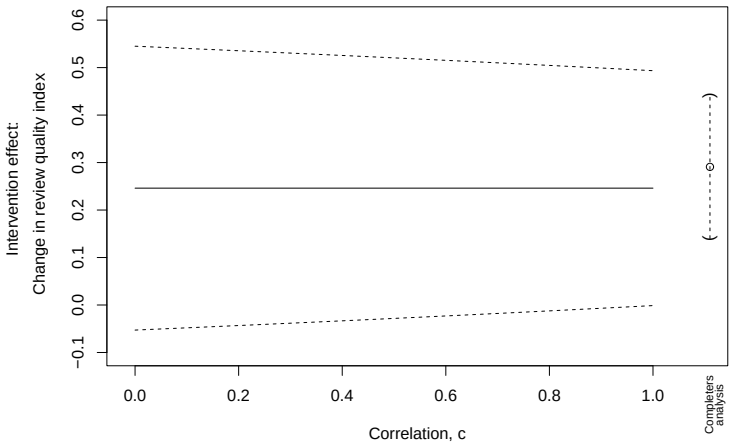

Figure 6.4: Estimated effect of postal intervention compared with control: complete cases analysis (right hand end of figure) and adjusted for informative missingness, showing effect of varying the correlation c between informative missingness in control and postal arm. This analysis uses the approximate method and is unadjusted for covariates

Reassuringly, the results for the approximate method, which summarises Figure 6.3 by a normal distribution, are very similar to the MCMC results, which treat the prior as a discrete distribution. We would expect differences to occur mainly in the 95% credible intervals, and the results in Table 6.3 support this view. In practice, it would appear that the analytic 'exact method' is unlikely to be misleading unless the prior is exceptionally skew.

We conclude that the experts we consulted would interpret this trial as failing to reject the null hypothesis of no effect of intervention. □

6.4.5 Pros and cons of prior elicitation

This method allows the sensitivity of conclusions to MNAR to be investigated relatively transparently. Investigators and regulators usually have a clear idea of the differences between non responders and responders, so this approach is accessible to them. Further, in our experience the full implementation of this model in `winBUGS` is not necessary; the approximate method works well. We have programmed an EXCEL spreadsheet to implement it. If no prior information is available, we can adopt a working value of c, usually 0 as this gives the widest confidence intervals, and find δ which causes the treatment effect to be non-significant. The plausibility of this value of δ can then be assessed.

Note that, relative to an MAR analysis, this method always widens credible intervals, and in this respect provides a more stringent test of a treatment effect than the common analysis which assumes MAR.

Here we have discussed quantitative outcomes, but the method could also be applied to binary outcomes. Here, Magder (2003) has measured informative missingness via the response probability ratio, while Higgins *et al.* (2006) have proposed a similar Bayesian analysis based on an 'Informative Missing Odds Ratio'[1].

[1] *i.e.* the ratio of the odds of response in patients whose data are observed to the odds of response in patients

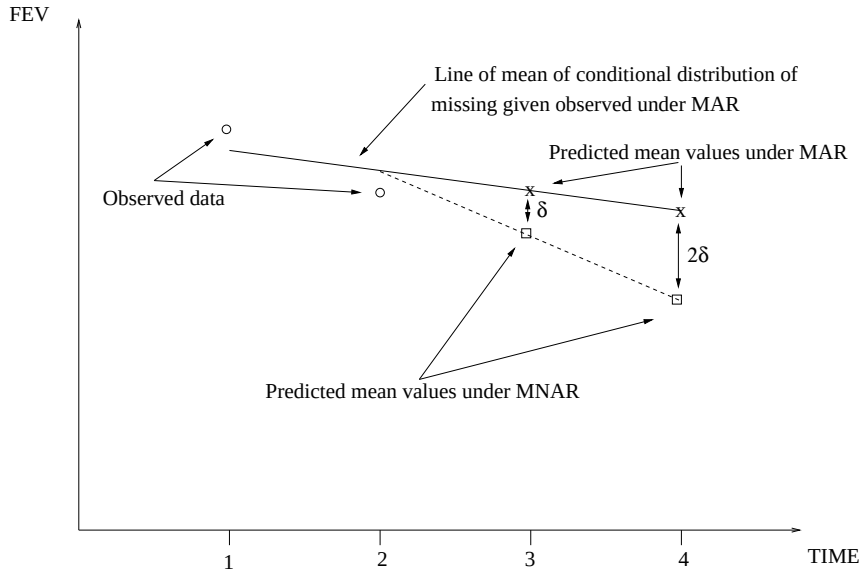

Figure 6.5: Schematic illustration of increasing the rate of decline by δ after withdrawal

6.5 Pattern mixture approach with longitudinal data via MI

The above approach has the potential to be extended to longitudinal data in several ways. Perhaps the easiest is to start with the conditional distribution of the missing given the observed data under MAR, and then modify this after a patient has dropped out.

A natural approach is to suppose that patients who withdrawal have a different, usually poorer response than predicted by MAR. For example in an asthma trial, we might suppose that FEV_1 improves more slowly (or declines more quickly) after withdrawal. If the change in rate of decline is denoted δ, then the conditional mean for the first response after withdrawal is reduced by δ, the second by 2δ and so on. This is schematically illustrated in Figure 6.5.

As discussed above, if possible we can elicit from experts the mean and variance of δ_l in the treatment group l, which is assumed to be normally distributed, and $\text{Cor}(\delta_l, \delta_{l'})$, for all treatment groups l and l'. In practice, the theory above shows the widest confidence intervals occur when the correlation is zero (assuming it is not negative), and useful information can be obtained by assuming the distribution of δ is the same in both arms.

We can use this approach with multiple imputation as follows. First, we create the K imputations under MAR. Suppose we have two treatment groups. Then, for each imputation, k, we sample

$$\begin{pmatrix} d_{1k} \\ d_{2k} \end{pmatrix} \sim N\left(\begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right).$$

For each patient, in each treatment arm, $l = 1, 2$, for each imputation, we then decrease/increase the first MAR imputed observation by d_{lk}, the second by $2d_{lk}$ and so on. We then analyse the resulting datasets and combine the estimates using Rubin's rules. If the time between observations is not constant, we may want to change the multipliers of d from $1, 2, 3, \ldots$, to maintain a linear change. We can handle interim missing observations by decreasing them by d_l, or simply leaving them with their MAR imputed values. The latter is consistent with a different mechanism driving interim missing data and patient withdrawal.

whose data are missing. Under MAR, this would be 1.

Coefficient	Parameter estimates when $\delta =$				
	0 (MAR)	-0.01	-0.02	-0.03	-0.04
Treatment	-0.086 (0.021)	-0.088 (0.026)	-0.098 (0.026)	-0.095 (0.028)	-0.101 (0.029)
Baseline	0.89 (0.021)	0.89 (0.021)	0.89 (0.021)	0.89 (0.021)	0.89 (0.022)
Intercept	0.14 (0.042)	0.12 (0.047)	0.11 (0.048)	0.09 (0.051)	0.08 (0.051)

Table 6.5: Estimates of treatment effect at the final time point when patients who withdraw have each subsequent MAR imputed FEV_1 value reduced by $\delta, 2\delta$ and so on. The reference group is those on active treatment. All values are in litres

As the theory above suggests, the confidence intervals are going to be narrower the greater the correlation between the d's for different treatment arms. In practice, information on this correlation is unlikely to be available, so as the correlation is unlikely to be negative, a conservative approach is to set it to zero, *i.e.* set $\sigma_{12} = 0$. Likewise often, in the absence of prior information, we can set $\delta_1 = \delta_2$.

EXAMPLE 6.3 *Isolde data: sensitivity analysis through MI*

We illustrate the above approach which is implemented in a SAS macro by Roger (2006) (see Appendix C). Recall there is an active and placebo arm in this trial. We will use the above approach, sampling for imputation $k = 1, \ldots, 50$, (d_{1k}, d_{2k}) from

$$\begin{pmatrix} d_{1k} \\ d_{2k} \end{pmatrix} \sim N\left(\begin{pmatrix} \delta \\ \delta \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right).$$

We compare the results of multiple imputation under MAR with values $-10, -20, -30, -40$ ml for δ and $\sigma^2 = 0.005^2$. Note that the *Isolde* trial has a large number of patients (just under half) with some interim missing data. Imputed interim missing data is left unchanged by the macro (in other words it is assumed the MAR model is correct). Only when a patient has withdrawn do we start to change the MAR imputations.

Table 6.5 shows the results, for the treatment estimate at the final time point. We used $K = 50$ imputations. We see that the treatment estimate increases with δ, but the standard error also increases, so that the significance of treatment slightly declines ($z = -4.1$ for $\delta = 0$ and -3.5 for $\delta = 0.04$). The intercept also decreases as δ increases. This makes sense as the MNAR model reduces the FEV_1, but more patients withdraw, and withdraw earlier, in the placebo arm. Thus the effect of increasing the rate of decline when patients withdraw results in a greater reduction in final FEV_1 values in the placebo arm than the active arm, and hence a larger estimated treatment effect. The increase in standard error, resulting from the increased variability of the final FEV_1's, warns that this sensitivity analysis cannot be interpreted to say that the MAR analysis underestimates the significance of the treatment effect. □

6.5.1 Further points

This approach can be readily generalised to discrete data. In that case, though, we have to work with the mean of the underlying distribution (Binomial, Poisson *etc.*) rather than working

directly with the imputed data (both have the same effect for the normal distribution).

An alternative way to obtain prior information on the post-withdrawal distribution would be to present experts with a graph of the MAR predictions for a 'typical' patient in each treatment arm and ask them to sketch how they expect those who do not withdraw to differ. The information could then be summarised, either to estimate the parameters of (6.5) or non-parametrically. The latter option would require a modification to the `SAS` macro, or implementation in `winBUGS`, along the lines described by White *et al.* (2007).

6.6 Pattern-mixture models and intention to treat analyses

The above shows the pattern mixture approach to modelling in the presence of missing data is an intuitive framework when we wish to consider associations between withdrawal and subjects' post-withdrawal behaviour. This arises naturally when we consider the implication for analysis of a subject both withdrawing *and* discontinuing, or otherwise changing in an unplanned way, his or her treatment regime. It is implicit when using an MAR based analysis, as we have done in Chapters 3–5, that the same model for future behaviour — given the past — applies to subjects' outcomes whether they withdraw or not. This assumption is probably behind much of the unease that has been expressed towards such MAR modelling in the past. If withdrawal is likely to be associated with a change in treatment regime then the MAR analysis does not provide an estimate of the desired treatment comparison. As argued in §1.8.2, the MAR assumption is compatible with a *per protocol* analysis.

Further, treatment change at withdrawal, in particular discontinuation, is an MNAR process: a different model is needed for those who withdraw and those who do not. Thus, to provide an intention to treat (ITT) analysis it is necessary to specify how we assign treatment to withdrawals and how we believe that future behaviour under the new regime should be modelled. In any particular setting we may or may not actually have information on treatment adherence following withdrawal. In this way an ITT analysis can be viewed in the missing data setting as a particular MNAR analysis constructed under specific assumptions and, further, can therefore be considered as part of a sensitivity analysis.

As Little and Yau (1996) show, under moderately simple assumptions about the behaviour of withdrawals, it is straightforward to construct such ITT analyses using multiple imputation. Two key features are exploited.

1. For multiple imputation the imputation model need not be completely compatible with the model used for the analysis (the substantive model). The word "uncongenial" is used in this setting to describe such a difference. Here the substantive model uses the ITT rule that allocates subjects to their randomised groups irrespective of subsequent compliance, but the imputation model incorporates compliance (known or postulated).

2. If we make simple but reasonable assumptions about the consequence of treatment change on future behaviour, then the MNAR pattern-mixture imputation model can be constructed from components that can be estimated from an MAR model, so greatly simplifying the modelling process.

We illustrate these points with a simple example. Suppose we wish to estimate the ITT treatment effect at the final time point in a two-armed trial of an active drug versus placebo, under the

assumption that withdrawals discontinue treatment and the future behaviour of their outcomes (conditional on the past) is the same as those in the placebo group (with the same past). The appropriate pattern-mixture model for this setup can be constructed in an obvious way, by using the estimated model from the placebo group to represent future behaviour of withdrawals from the active group. This model is identical to the MAR model *for all observed data* and so can be consistently estimated from the observed data. The models differ only with respect to their implications for the unobserved behaviour of the withdrawals from the active group. Hence imputations for the *future* outcomes differ between the two models, and this in turn affects the estimated final treatment comparison. Using the same basic idea a variety of alternative pattern-mixture models can be constructed to examine in a very simple way the impact of different behaviours of the withdrawals on the results from ITT analyses. For further details see Little and Yau (1996); Kenward *et al.* (2003).

6.7 Conclusions

In this Chapter we have outlined methods for sensitivity analysis via (i) selection models and (ii) pattern mixture models. To an extent, the attraction of each approach depends on the problem at hand. However, as we have shown, the pattern-mixture approach lends itself to approximate inference either analytically (for a single post-randomisation measurement) or using MI (for longitudinal follow-up). As usual with multiple imputation, we can fit any model of interest to the imputed data — e.g. a mixed model to estimate a trend over time or a model that combines aspects from different treatment groups for an ITT analysis.

In conclusion, we would stress the advantages of pre-planning the sensitivity analysis. Besides allowing the time for appropriate code to be developed, pre-defining the scope of the sensitivity analysis makes it much easier to interpret the results. The aim should always be to assess the sensitivity of the conclusions to *plausible* departures from MAR, rather than to assert that a particular MNAR analysis is correct.

Appendix A

Justification for the approach in Chapter 3

Here, we give a slightly more detailed justification of the modelling approach in Chapter 3. To avoid unnecessary algebra, we do this initially in a very simple situation, namely data from one arm of a trial. We first consider the case of missing responses (§A.1), and then missing baseline and responses (§A.3). Lastly, in §A.4, we show in generality why the approach in §3.4.3 can be used to obtain estimates conditional on baseline when some baseline values are missing. This is sufficient to justify all the analyses in Chapter 3.

A.1 Key ideas: data from a single trial arm, missing responses

Consider data from a single trial arm, and let (x_i, y_i) be the baseline and response from patient i, $(i = 1, \ldots, n)$. For now we suppose that baseline, x_i, is always observed and that the response, y_i, is only observed on n_1 out of the n patients. When the response is missing we refer to it as Y_i. We introduce a further variable, r_i, which is 1 if y_i is observed, and 0 otherwise. We think of r_i as indicating whether y_i is missing, and refer to it as the *missingness indicator*. The model for $\Pr(r_i = 1)$ is the algebraic version of the missingness mechanism introduced in Chapter 1. Finally, we suppose the data have been ordered, so that the n_1 patients whose response is observed are grouped together at the top.

EXAMPLE A.1 *Isolde trial*

Table A.1 illustrates the notation with $n = 4$ patients from the placebo arm of the Isolde trial (described at the beginning of Chapter 2). □

Patient identifier	baseline FEV_1, x_i, (litres)	6 month FEV_1, y_i, (litres)	Missingness indicator, r_i
1	0.98	1.30	1
3	1.77	1.31	1
2	1.66	missing	0
4	2.11	missing	0

Table A.1: Isolde trial: data from 4 placebo patients. Here $n = 4$, $n_1 = 2$ and note we have rearranged the order of patients so the n_1 with fully observed data appear first

We have already seen that the choice of sensible analyses when data are missing depends on assumptions about the missingness mechanism, $\Pr(r_i = 1)$. In this Subsection, we have baseline and a single response only. Thus, we have one of three possibilities:

1. The response is MCAR. So, the probability of observing the response, $\Pr(r_i = 1)$, is number between 0 and 1 and does not depend on y_i or x_i.
2. The response is MAR. So $\Pr(R_i = 1)$ depends on x_i.
3. The response is MNAR. Even after allowing for baseline, x_i, $\Pr(R_i = 1)$ depends on y_i too.

Since the model for $\Pr(R_i = 1)$ has such a key role, it follows that we should no longer think of a patient's data as just (x_i, y_i), $(i \in 1, \ldots, n)$, but include the missingness indicator too. The statistical model for the data is then a model for their joint distribution. Leaving out for a moment the index i, we write this as $[x, y, r]$, or $[x, Y, r]$ if the response is missing. In order to consider the implications for the analysis of the missingness mechanisms above, it is useful to write

$$[x, y, r] = [r|x, y][x, y], \tag{A.1}$$

in other words, the joint distribution of baseline, response and missingness indicator is the conditional distribution of missingness indicator given baseline and response, multiplied by the joint distribution of baseline and response.

We focus on estimating (i) the (marginal) distribution of the response, $[y]$, particularly its mean and variance, and (ii) the conditional distribution of response given baseline, $[y|x]$. We do this when some responses are MCAR, MAR and MNAR.

Response MCAR

For a patient with observed response, following (A.1) their density is $[r_i = 1|y_i, x_i][y_i, x_i]$. Under MCAR, $[r_i = 1|y_i, x_i]$, the probability of observing the response, is constant, say η. So their density is $\eta\,[y_i, x_i]$.

If response is missing, we have to integrate (or sum) (A.1) over the range of unseen response values:

$$\int [r_i = 0|Y_i, x_i][Y_i, x_i]\, dY_i.$$

Now, $[r_i = 0|Y_i, x_i]$ is the probability of not observing the response, which under MCAR is constant, and $(1 - \eta)$. So the density of the data is

$$\int (1 - \eta)[Y_i, x_i]\, dY_i = (1 - \eta)[x_i].$$

The log-likelihood for the data is simply the sum of the log-distributions for each patient (viewed as a function of the model parameters):

$$\ell = \sum_{i=1}^{n_1} \log(\eta\,[y_i, x_i]) + \sum_{i=n_1+1}^{n} \log((1 - \eta)[x_i]).$$

However notice that the joint distribution $[y_i, x_i]$ can be written as $[y_i|x_i][x_i]$. This gives

$$\ell = \left\{ \sum_{i=1}^{n_1} \log[y_i|x_i] \right\} + \left\{ \sum_{i=1}^{n} \log[x_i] \right\} + \{n_1 \log \eta + (n - n_1) \log(1 - \eta)\}. \tag{A.2}$$

The third term in curly brackets is the log-likelihood for the missingness model. None of the parameters of the distribution of baseline and response appear here, so when the log-likelihood

is differentiated to find maximum likelihood estimates of these parameters, this term vanishes. For inferences about response and baseline, we can therefore ignore it.

Now consider the first two terms of (A.2). The first tells us that the information about the distribution of response given baseline resides in data from the n_1 patients from whom both were observed. The second term tells us that if we are interested in the response distribution, marginal to baseline, we should average our estimated parameters of $[y|x]$ over $[x]$ with the latter estimated using baseline data *from all n patients*.

EXAMPLE A.2 *Isolde trial*

Consider baseline and 6 month data from the placebo arm of the Isolde study, available from 374 and 288 patients respectively[1]. Although the data have a slight negative skewness, we use a bivariate normal model. Assume the missingness mechanism is MCAR.

Following the above, we can use the data from the 288 patients to estimate the effect on 6 month FEV_1 of having a higher baseline FEV_1. Fitting the linear regression gives:

$$\begin{aligned} \text{6 month FEV}_1 \text{ given baseline FEV}_1 &= [Y|x] \\ &= 0.024 + 0.947 \times \text{baseline FEV}_1 + \varepsilon, \quad \varepsilon \sim N(0, 0.028). \end{aligned} \tag{A.3}$$

To obtain the average 6 month FEV_1 marginal to baseline FEV_1, we average (A.3) over all 374 baseline values. Letting $\bar{x} = (x_1 + x + 2 + \cdots + x_n)/n = 1.407$ we have

$$\text{average 6 month FEV}_1 = \text{E}_X \text{E}[Y|x] = 0.024 + 0.947\bar{x} = 0.024 + 0.947 \times 1.407 = 1.356. \tag{A.4}$$

Further, the marginal variance of Y is (using the conditional variance formula)

$$\begin{aligned} \text{Var}[Y] &= \text{E}_X \text{ Var}[Y|x] + \text{Var}_X \text{E}[Y|x] \\ &= \text{E}_X \, \sigma^2_{y|x} + \text{Var}_X [\alpha + \beta x] \qquad (\sigma^2_{y|x} \text{ is the residual variance of} \\ &= \sigma^2_{y|x} + \beta^2 \sigma^2_x. \qquad \text{the regression of } Y \text{ on } x) \end{aligned} \tag{A.5}$$

Now, β and $\sigma^2_{y|x}$ are estimated from the 288 patients with baseline and 6 month response. From (A.3) they are 0.947, 0.028 repectively. However, σ^2_x is estimated from 374 baseline values as 0.243. Putting in estimates from the data we see

$$\text{Var}[Y] = 0.028 + 0.947^2 \times 0.243 = 0.246.$$

The standard error of the mean of Y is thus $\sqrt{0.246/377} = 0.026$.

In practice, performing the conditional calculations above would be cumbersome in a more realistic model. Fortunately, it is not necessary. Any program for fitting repeated measures data will maximise (A.2). We ask for an estimate of the mean and variance at each time point, using all available data. Using `SAS PROC MIXED` gives the following estimates (standard errors of means in parentheses):

$$\begin{pmatrix} \text{baseline FEV}_1 \\ \text{6 month FEV}_1 \end{pmatrix} \sim N \left\{ \begin{pmatrix} 1.407(0.025) \\ 1.356(0.026) \end{pmatrix}, \begin{pmatrix} 0.243 & 0.230 \\ 0.230 & 0.246 \end{pmatrix} \right\}$$

[1] 1 patient is missing both baseline and 6 month FEV_1, though they later return to the trial.

Note the agreement with the results above. The off diagonal term in the matrix is the covariance of baseline and response, σ_{xy}. □

Does the above argument depend $P(R_i = 1)$ being the same for each patient, or can it vary? The answer is, *provided that it does not depend on* (x_i, y_i) it can be different for each patient. To see this, look at (A.2) and note that if η becomes η_i, these values still do not affect the likelihood for the parameters of the distribution of (x_i, y_i). In other words, one patient may be MCAR with probability 100%, another with probability 0%, another with probability 37%, and so on. Provided they are all MCAR, the above analysis is sensible.

How does the above relate to the 'obvious' MCAR analysis, where, if we want to estimate the marginal mean of Y we only use the n_1 observed values? Then, the log-likelihood (A.2) becomes

$$\ell = \left\{ \sum_{i=1}^{n_1} \log[y_i|x_i] \right\} + \left\{ \sum_{i=1}^{n_1} \log[x_i] \right\} + \{n_1 \log \eta\} \tag{A.6}$$

$$= \left\{ \sum_{i=1}^{n_1} \log[y_i] \right\} + \left\{ \sum_{i=1}^{n_1} \log[x_i|y_i] \right\} + \{n_1 \log \eta\}. \tag{A.7}$$

Note how, as there are only n_1 observations here, we can re-write (A.6) to obtain (A.7). It follows that to estimate the marginal distribution of Y we only need the n_1 observed values of y, and no longer the baseline observations x. To consider which is preferable, we re-visit the example.

EXAMPLE 2.1 *Isolde trial (ctd)*

If we just use the 288 complete cases, we can calculate the marginal mean and variance of 6 month FEV_1 directly from the observed y_i. However, in order to see how this compares with the previous analysis, we work through this again, but now with $n = n_1 = 288$. The mean baseline for the 288 patients whose 6 month reading is also seen is $\bar{x}_{288} = 1.411$. So (A.4) becomes

$$\text{average 6 month FEV}_1 = \text{E}_X\text{E}[Y|x] = 0.024 + 0.947\bar{x}_{288} = 0.024 + 0.947 \times 1.411 = 1.360. \tag{A.8}$$

As the regression of y on x estimated from the 288 patinets on whom both are observed passes through $(\bar{x}_{288}, \bar{y}_{288})$, $1.360 = (y_1 + y_2 + \cdots + y_{288})/288$ exactly. Under MCAR, any difference between $\bar{x}$ and $\bar{x}_{288}$ is down to chance, so the difference between (A.4) and (A.8) is chance.

Now consider the variance of Y. Before, (A.5), this was

$$\sigma^2_{y|x} + \beta^2 \sigma^2_x,$$

with $\sigma^2_{y|x}, \beta$ estimated from the 288 patients with (y, x) observed and σ^2_x estimated from the 374 patients with baseline. Now, though, we can only use data from the 288 patients to estimate σ^2_x. Doing this gives

$$\text{Var}[Y] = 0.028 + 0.947^2 \times 0.237 = 0.241,$$

which is exactly the usual sampling standard error,

$$\left\{ (y_1 - 1.360)^2 + (y_2 - 1.360)^2 + \cdots + (y_{288} - 1.360)^2 \right\} / 286.$$

Which is better? In this case the standard error is smaller using only the 288 patients with complete data. However, the estimate of σ^2_x from the complete data will likely be closer to the

true value, as it uses 86 more observations. This is reflected in the degrees of freedom of the t-distribution for estimating confidence intervals: 287 (observed data only) versus 373 (include information from 88 patients with only basline values). The resulting confidence intervals are:

$$\begin{array}{ll} \text{use only 288 observed values of } y: & (1.304, 1.417), \\ \text{in addition use 374 baseline values:} & (1.305, 1.407). \end{array}$$

Using the extra information gives a narrower confidence interval. We conclude that, even if we really believe data are MCAR, it is more sensible to use all observed data in an analysis than data from patients who have no missing values (complete cases). □

Response MAR

Here, the density $[r_i|x_i, y_i] = [r_i|x_i]$, that is to say the probability of observing the 6 month response depends on baseline. So, if $r_i = 1$, the distribution of the patient's data is $[r_i = 1|x_i][y_i, x_i]$. If $r_i = 0$, the distribution of a patient's data is

$$\int [r_i = 0|x_i][Y_i, x_i]\, dY_i = [r_i = 0][x_i].$$

Following (A.2), the log likelihood of the data is

$$\ell = \left\{ \sum_{i=1}^{n_1} \log[y_i|x_i] \right\} + \left\{ \sum_{i=1}^{n} \log[x_i] \right\} + \left\{ \sum_{i=1}^{n_1} \log[r_i = 1|x_i] + \sum_{i=n_1+1}^{n} \log[r_i = 0|x_i] \right\}. \quad \text{(A.9)}$$

Comparing with (A.2) we see the only difference is that the likelihood for the probability of missing data is no longer a constant (e.g. η), but now depends on x_i. However, provided the parameters of the model for $[y_i, x_i]$ are not shared by the model for the missingness mechanism, the terms involving r_i do not affect the maximum likelihood estimates of parameters of $[y_i, x_i]$.

EXAMPLE 2.1 *Isolde data: MAR model (ctd)*

As before, we take a bivariate normal model for $[y_i, x_i]$, with parameters $(\mu_y, \sigma_y^2, \sigma_{yx}, \sigma_x^2, \mu_x)$. Suppose the model for observing y_i is logistic:

$$\text{logit}\Pr(r_i = 1) = \gamma_0 + \gamma_1 x_i.$$

The parameters of this model, (γ_0, γ_1), are clearly different from the parameters of the model for $[y_i, x_i]$. So we can ignore the missingness model when estimating $(\mu_y, \sigma_y^2, \sigma_{yx}, \sigma_x^2, \mu_x)$.

However, suppose for some uncommon reason the model for observing y_i was

$$\text{logit}\Pr(r_i = 1) = \mu_y + \gamma_1 x_i. \quad \text{(A.10)}$$

Now there is a shared parameter, μ_y, between the missingness model and the model for (y_i, x_i). This means we *cannot* ignore the missingness model when estimating $(\mu_y, \sigma_y^2, \sigma_{yx}, \sigma_x^2, \mu_x)$. Fortunately, shared parameter models like (A.10) are not usually sensible for trials, let alone necessary. We do not consider them further. □

The above shows that the likelihood of the data with the missingness mechanism is MAR is effectively the same as when it is MCAR. Thus, when estimating the distribution of $[y_i|x_i]$, we again use data from the n_1 patients with both observed. When estimating the marginal distribution $[y_i]$ we use all the data, as shown in the example on p. 141.

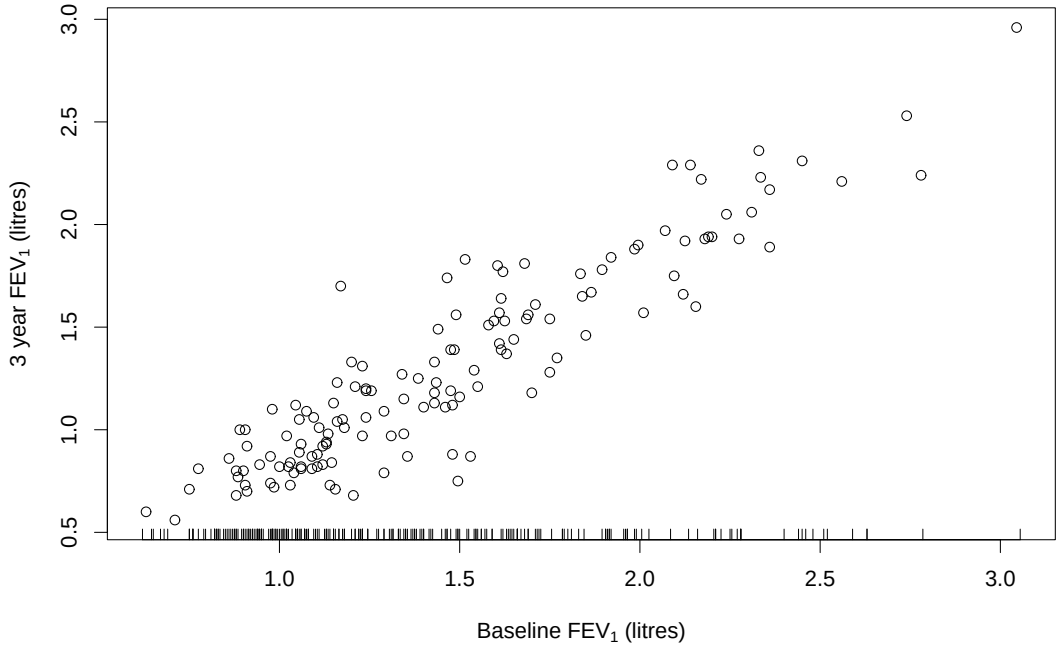

Figure A.1: Isolde trial, placebo arm: plot of 3 year FEV_1 against baseline FEV_1. 234 patients with missing 3 year FEV_1 have their baseline value shown by a '|'

Notice that, unlike the situation when data is MCAR, it is no longer sensible just to use data from n_1 patients with both baseline and response to estimate the distribution of $[y]$. Whereas with MCAR data, doing this just lost some information, with MAR data it introduces bias.

As with MCAR, it is not necessary that the missingness mechanism should be the same for all patients. For example, for different patients the chance of dropout can depend in different ways, and in different degrees, on baseline. All that is required is that, once we have used the baseline information, there is no further information about the missingness mechanism in the unseen response. Thus, some patients may be MCAR, and some MAR, but using the likelihood of all the observed data is still sensible.

EXAMPLE A.2 *Isolde trial (ctd)*

Figure A.1 shows 3 year FEV_1 against baseline FEV_1. Only 141 out of 374 patients have both measurements. Further, the 'rug' of the 234 baseline values for which there are no 3 year values suggest that patients with lower baseline FEV_1 are much more likely to withdraw. We therefore assume data are MAR, dependent on baseline.

The estimate of the mean and variance of 3 year FEV_1, obtained from the $n_1 = 141$ patients whose values are observed, are $(1.30, 0.239)$ respectively. Regressing the 141 3-year values on baseline values gives

$$\begin{aligned} \text{3 year FEV}_1 \text{ given baseline FEV}_1 &= [Y|x] \\ &= -0.06 + 0.927 \times \text{baseline FEV}_1 + \varepsilon, \quad \varepsilon \sim N(0, 0.037). \end{aligned} \tag{A.11}$$

As we saw before, taking expectations over (A.11) using only data from the $n_1 = 141$ gives the marginal mean and variance for y obtained from the 141 observed 3-year responses. However, whereas before, when MCAR was plausible, $\bar{x}_{n_1} \approx \bar{x}$ and $\hat{\sigma}^2_{x,n_1} \sim \hat{\sigma}^2$, this is no longer the case under MAR (Table A.2).

Parameter	Estimate obtained using data from	
	141 patients with y_i observed	all 374 placebo patients
μ_x	1.47	1.41
σ^2_x	0.235	0.243

Table A.2: Estimates of mean and variance of baseline obtained with and without including patients with 3-year response

Such results are to be expected when data are MAR, because the patients from whom y_i is observed are selected non-randomly. However, likelihood (A.9) shows that, as under MCAR, using a conditional expectation and variance approach on (A.11) to estimate the marginal mean and variance of y gives the same answer as fitting a a bivariate normal model. SAS proc MIXED gives:

$$\begin{pmatrix} \text{baseline FEV}_1 \\ \text{3 year FEV}_1 \end{pmatrix} \sim N\left\{ \begin{pmatrix} 1.41(0.025) \\ 1.24(0.029) \end{pmatrix}, \begin{pmatrix} 0.243 & 0.225 \\ 0.225 & 0.247 \end{pmatrix} \right\}$$

For estimating marginal means and variances (*i.e.* mean and variance of the response) we have seen that fitting the bivariate normal model is preferable if data are MCAR, and essential if data are MAR. We have also noted it has the further advantage that different patients can have different missingness mechanisms, so long as none of them are MNAR. We conclude the bivariate normal model is the best way to estimate marginal means and variances. □

Before moving on, we consider the effect of missing data on the difference between using the observed and expected information to estimate the variance of parameter estimates. If we denote all the parameters of the model for $[y,x]$ by θ, the observed information can be written

$$-\frac{\partial^2}{\partial \theta^2}\ell(\theta),$$

whereas the expected information is

$$-\mathrm{E}_{X,Y,R}\frac{\partial^2}{\partial \theta^2}\ell(\theta).$$

As the full data are $[y,x,r]$, the expectation must be taken over this distribution. Write this as

$$\mathrm{E}_{X,Y,R}\frac{\partial^2}{\partial \theta^2}\ell(\theta) = \left\{\mathrm{E}_{X,Y|R=1}\frac{\partial^2}{\partial \theta^2}\ell(\theta)\right\}\Pr(R=1) + \left\{\mathrm{E}_{X,Y|R=0}\frac{\partial^2}{\partial \theta^2}\ell(\theta)\right\}\Pr(R=0),$$

and note that, under MAR, the distributions $[Y,X|R=1]$ and $[Y,X|r=0]$ are different. It follows that, to calculate the expected information, we must *explicitly* specify a model for $[R]$. This

becomes yet more difficult if we wish to maintain the option of allowing different patients to have different missingness mechanisms.

Thus, with missing data, the observed information should *always* be used. It is worth checking that your statistical package uses this, to avoid the possibility of misleading variance estimates. This point was noted by Jennrich and Schluchter (1986) and developed by Kenward and Molenberghs (1998).

Response MNAR

We very briefly discuss the MNAR case, which we focus on in Chapter 6. If y_i is MNAR, then we cannot simplify $[r_i|y_i,x_i]$ further. So, if $r_i = 1$, the distribution of the patient's data is $[r_i = 1|y_i,x_i][y_i,x_i]$. If $r_i = 0$, the distribution of a patient's data is

$$\int [r_i = 0|Y_i,x_i][Y_i,x_i]\, dY_i \neq [r_i = 0|x_i] \text{ as before.} \tag{A.12}$$

The missingness mechanism is now inseperable from $[y,x]$, so we cannot express the likelihood as (A.9). Thus, in order to obtain maximum likelihood estimates of the parameters of $[y,x]$ we must also specify a model for $[r|y,x]$, and, for those with $r = 0$, obtain their contribution to the likelihood by calculating the integral on the left of (A.12).

However, even if one suspects that some patients are MNAR, the degree of remaining dependence of $r|x$ on y may be small, so that an MAR analysis will not be too misleading. Thus, an MAR analysis is usually a good starting point.

A.2 Summary of findings

We have considered a single follow-up visit, which may be missing, and baseline. We found the following analyses were most sensible, if some responses are MCAR and some MAR:

1. To estimate the marginal mean and variance of the response, use a joint model, using all available data. The marginal means of the partially observed response are 'corrected' through their correlation with the fully observed baseline, and
2. To estimate the distribution of the response conditional on fully observed baseline, use only the data from patients with both observed.

Of course, the parameters of the conditional distribution in (2) can be estimated from the parameter estimates in (1), but this gives the same answer as fitting (2), which is usually simpler.

A.3 Missing baselines and responses

Here we suppose that some baseline observations are missing, and some 6 month responses are missing. We suppose that, where it is missing, baseline is MCAR given response. Likewise, where it is missing, response is MCAR given baseline. Then, as above, the likelihood for the missingness mechanism does not affect the likelihood of $[y,x]$. Suppose we re-order the data

so that patients $i = 1,\ldots,n_0$ have only $[y_i]$, patients $i = (n_0+1,\ldots n_1)$ have both and patients $i = (n_1+1,\ldots,n)$ have only $[x_i]$. Omitting the likelihood of the missingness mechanism, we can adapt (A.2) to see the log likelihood is

$$\ell = \left\{\sum_{i=1}^{n_0}\log[y_i]\right\} + \left\{\sum_{i=n_0+1}^{n_1}\log[y_i|x_i]\right\} + \left\{\sum_{i=n_0}^{n}\log[x_i]\right\} + \left\{\sum_{i=1}^{n_1}\log[r_i=1|x_i] + \sum_{i=n_1+1}^{n}\log[r_i=0|x_i]\right\}. \tag{A.13}$$

Here, though, we cannot use only data from the (n_1-n_0) patients with both observations to estimate the distribution of $[y|x]$. This is because the parameters of $[y]$ are shared with those of $[y|x]$. Assuming the y's are *not* missing randomly, ignoring this extra information leads to bias in the estimates of parameters of $[y|x]$. Rather, in general, we must proceed by maximising (A.13). We may need to use the parameterisation dictated by the software, and then calculate the parameters of interest from this. Often, though, we shall see that this can be avoided.

EXAMPLE A.2 *Isolde study: missing baseline and 6 month data (ctd)*

Recall that of the 374 patients in the placebo arm, 86 have missing 6 month data. We now make some of the baseline values missing for the remaining 288, and then illustrate the use of the methods above to estimate the regression of y on x.

First, for these 288 patients we say the probability of observing baseline is

$$\Pr(\text{observe baseline from patient } i) = p_i = \frac{1}{1+e^{0.4\times(6\text{ month FEV}_1)}}. \tag{A.14}$$

The effect of this is shown in Figure A.2. Then, for each patient, we generate a uniform variable on $[0,1]$ and set the baseline response to be missing if $u_i > p_i$. We already had 86 patients with baseline only, of the 288 with both we now have 194 with both baseline and response, and 94 with response only.

Maximising the log-likelihood (A.13) using `SAS PROC MIXED` estimates the parameters as:

$$\begin{pmatrix}\text{baseline FEV}_1\\ \text{3 year FEV}_1\end{pmatrix} \sim N\left\{\begin{pmatrix}1.41\\1.35\end{pmatrix}, \begin{pmatrix}0.243 & 0.230\\0.230 & 0.246\end{pmatrix}\right\}. \tag{A.15}$$

From the properties of the conditional normal distribution it follows that all the parameters of the regression model

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0,\sigma^2_{y|x}),$$

can be estimated from (A.15). For instance, using the notation of previous examples,

$$\hat{\beta} = \hat{\sigma}_{xy}/\hat{\sigma}^2_{xx} = 0.230/0.243 = 0.947.$$

To obtain the standard error, we use a 1-term Taylor expansion (the 'δ-method'):

$$\operatorname{Var}\frac{\hat{\sigma}_{xy}}{\hat{\sigma}^2_x} = \operatorname{Var} f(\hat{\sigma}_{xy},\hat{\sigma}^2_x) \approx \left(\tfrac{\partial}{\partial\sigma_{xy}}f, \quad \tfrac{\partial}{\partial\sigma^2_x}f\right) \operatorname{Var}(\hat{\sigma}_{xy},\hat{\sigma}^2_x) \begin{pmatrix}\frac{\partial}{\partial\sigma_{xy}}f\\ \frac{\partial}{\partial\sigma^2_x}f\end{pmatrix}. \tag{A.16}$$

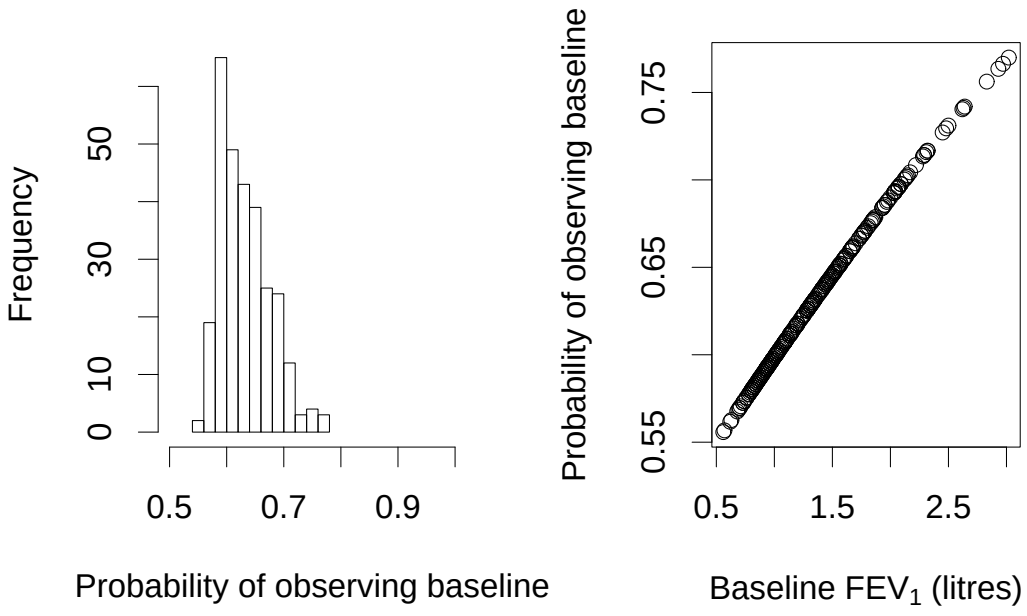

Figure A.2: Left panel: histogram of probabilities generated by (A.14). The right panel how these probabilities increase with 6-month FEV_1

Let $\text{Var}\,\hat{\sigma}_{xy} = \nu_1$, $\text{Var}\,\hat{\sigma}_x^2 = \nu_2$ and $\text{Cov}\,(\hat{\sigma}_x^2, \hat{\sigma}_{xy}) = \nu_{12}$. From `SAS PROC MIXED`, $\hat{\nu}_1 = 0.000311$, $\hat{\nu}_2 = 0.000335$ and $\hat{\nu}_{12} = 0.000306$. Evaluating the derivatives and multiplying up, (A.16) gives

$$\begin{aligned} \text{Var}\,\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} &= \frac{1}{(\hat{\sigma}_x^2)^2}\left(\hat{\nu}_1 - 2\hat{\beta}\hat{\nu}_{12} - \hat{\beta}^2\hat{\nu}_2\right) \\ &= \frac{1}{0.243^2}\left(0.000311 - 2\times 0.947\times 0.000306 + 0.947^2\times 0.000335\right) \\ &= 0.023^2. \end{aligned} \tag{A.17}$$

Note that, as the distribution of $\hat{\beta}$ is nearly normal, the likelihood is close to quadratic. Thus the δ-method approximation is accurate.

Table A.3 compares estimates of $\hat{\beta}$ (i) before the extra 94 baseline values were made missing, (ii) after they have been made missing, but just using data from the 194 patients with both baseline and response, and (iii) after they have been made missing, deriving $\hat{\beta}$ from the model for the joint data as above. Relative to (i) we see that (ii) is biased, as expected. Analysis (iii) removes this bias. The standard error is much wider for analysis (ii) reflecting the lost information. It is fractionally less (beyond shown precision) with analysis (iii), reflecting the extra observations used in this analysis.

We conclude patients with missing baseline but observed follow-up measurement should be included in the analysis, for estimating both $[y]$ and $[y|x]$. □

Data	$\hat{\beta}$ (std. error)
Original data, before extra 94 baselines made missing	0.947 (0.020)
Only using 194 with both baseline and 6-month observed	0.958 (0.024)
Using all observations, as described in the text	0.947 (0.024)

Table A.3: Estimates of β using various subsets of data

A.4 Justification of using model in 3.4.3 to obtain conditional treatment estimates

We now develop the argument of the previous Section more formally to show the equivalance of REML (GLS) and analysis of covariance in a repeated measurements setting with baseline observation. Note first that it is sufficent to show the equivalence in the case of no missing data. With missing data assumed MAR, each patient simply contributes through the marginal distribution of their observations.

First a relationship between generalized least squares and covariate adjustment is shown in a generic multivariate normal setting. It is then shown how this applies to the repeated measurement setting with a baseline observation.

I. Generic result

Suppose that we have one observation $\mathbf{Y}$ from a T dimensional multivariate normal distribution with covariance matrix $\mathbf{\Sigma}$. Further, suppose that we can partition $\mathbf{Y}$ such that

$$\mathrm{E}\{\mathbf{Y}\} = \mathrm{E}\left\{\begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{pmatrix}\right\} = \begin{pmatrix}\mathbf{0}\\ \boldsymbol{\mu}\end{pmatrix}$$

for $\mathbf{Y}_1$ and $\mathbf{Y}_2$ with dimensions p and q respectively, $p+q=T$, and $\boldsymbol{\mu}$ unconstrained. Then the generalized least squares estimator of $\boldsymbol{\mu}$ for $\mathbf{\Sigma}$ *known* is

$$\tilde{\boldsymbol{\mu}} = \mathbf{Y}_2 - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{Y}_1 \tag{A.18}$$

for

$$\mathbf{\Sigma} = \begin{pmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22}\end{pmatrix}.$$

That is, $\tilde{\boldsymbol{\mu}}$ is the covariate adjusted estimator of $\boldsymbol{\mu}$ with $\mathbf{Y}_1$ as covariate.

Now suppose that $\mathbf{\Sigma}$ must be estimated. Assume that there are an additional m independent observations, whose mean can be assumed 0 without loss of generality,

$$\mathbf{Z}_i = \begin{pmatrix}\mathbf{Z}_{1i}\\ \mathbf{Z}_{2i}\end{pmatrix} \sim \mathrm{N}(\mathbf{0},\mathbf{\Sigma}),\ \ i = 1,\ldots m,$$

where the partition conforms to that for $\mathbf{Y}$. Define

$$\mathbf{A} = \begin{pmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{pmatrix} = \sum_{i=1}^{m}\mathbf{Z}_i\mathbf{Z}_i^T.$$

Applying standard REML theory to the complete dataset, $\mathbf{Y}$ and $\{\mathbf{Z}_i\}$, $i = 1, \ldots m$, we find that the REML estimator of $\boldsymbol{\Sigma}_{11}$ is

$$\hat{\boldsymbol{\Sigma}}_{11} = \frac{1}{m+1}\left(\mathbf{A}_{11} + \mathbf{Y}_1\mathbf{Y}_1^T\right).$$

Similary, the REML estimator of $\boldsymbol{\Sigma}_{21}$ is

$$\hat{\boldsymbol{\Sigma}}_{21} = \mathbf{A}_{21}\hat{\boldsymbol{\Sigma}}_{11}\mathbf{A}_{11}^{-1},$$

which implies that the REML estimator of the regression coefficient in (A.18) can be written

$$\hat{\boldsymbol{\Sigma}}_{21}\hat{\boldsymbol{\Sigma}}_{11}^{-1} = \mathbf{A}_{21}\hat{\boldsymbol{\Sigma}}_{11}\mathbf{A}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{11}^{-1}.$$

In the special case that $\mathbf{Y}_1$ has dimension $p = 1$ this reduces to

$$\hat{\boldsymbol{\Sigma}}_{21}\hat{\boldsymbol{\Sigma}}_{11}^{-1} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}$$

which is equivalent to the estimator that we would obtain from an analysis of covariance of $\mathbf{Y}_2$ and $\{\mathbf{Z}_{2i}\}$ on $\mathbf{Y}_1$ and $\{\mathbf{Z}_{1i}\}$, $i = 1, \ldots, m$. In conclusion, in the generic setting given here, the REML (GLS) estimator of $\boldsymbol{\mu}$ is identical to the covariate adjusted estimator, provided that $\mathbf{Y}_1$ has dimension 1.

II. Application to the repeated measurements setting

Suppose that the ith subject in a two group trial supplies a baseline measurement Y_{i0} and T repeated measurements $Y_{1i}, \ldots, Y_{1T}$. There are n_1 and n_2 subjects respectively in the two arms. Assume that the subjects are ordered according to treatment group, so that we can define for $t = 0, 1, \ldots, T$, for treatment group 1:

$$S_{1t} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{it}$$

and for treatment group 2:

$$S_{2t} = \frac{1}{n_2}\sum_{i=1+n_1}^{n} Y_{it}.$$

From these we define

$$D_t = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2}(S_{1t} - S_{2t}), \quad t = 0, 1, \ldots, T,$$

the scaled mean treatment differences, for which

$$\mathrm{E}[D_0, D_1, \ldots, D_T] = (0, \delta_1, \ldots, \delta_T),$$

where the δ_t are assumed to be unconstrained. Similarly we define

$$M_t = \left(\frac{1}{n_1 + n_2}\right)^{-1/2}(n_1 S_1 + n_2 S_2), \quad t = 0, \ldots, T$$

for which

$$\mathrm{E}[M_0, M_1, \ldots, M_T] = (\psi_0, \psi_1, \ldots, \psi_T),$$

with the ψ_t assumed to be unconstrained.

Let $\mathbf{X}$ be the $n \times 2$ design matrix for (D_t, M_t) that applies to any time point for $t > 0$ and choose any $n \times (n-2)$ matrix $\mathbf{H}$ that satisfies

$$\mathbf{H}^T\mathbf{X} = \mathbf{0} \text{ and } \mathbf{H}^T\mathbf{H} = \mathbf{I}.$$

Define

$$\mathbf{V}_t = \mathbf{H}^T \begin{pmatrix} Y_{1t} \\ Y_{2t} \\ \vdots \\ Y_{nt} \end{pmatrix}, \quad t = 0, 1, \ldots, T.$$

The sets of derived variables $\{D_t, S_t, \mathbf{V}_t\}$ are mutually independent within time points and have a common (unstructured) covariance matrix across times. The transformation from the original data to these is of full rank, and the treatment effects of interest are proportional to the δ_t's.

As a final step we just need to map these quantities onto those in the part I. First, because the means of the M_t are saturated, and because of the mutual independence, these contribute nothing to the estimation of the other parameters and so can be ignored. Second, we identify the D_t's above with the Y_t's in part I. Third, the ith element of $\mathbf{V}_t$ ($i \in 1, \ldots, n-2$) is identified with the tth element of $\mathbf{Z}_i$ in I (where $m = n-2$). Fourth, and finally, the scaled treatment effects δ_t are equated with the elements of $\boldsymbol{\mu}$ in part I. We can now apply the result from part I to show that the baseline adjusted treatment difference for any time $t > 0$ will be identical to that obtained from a REML analysis applied to the whole data set treating all the observations including the baseline as responses, provided we ensure the same mean (across treatment groups) for baseline.

A.5 Summary

1. When baseline and responses are missing, and assumed MAR, we need to fit a bivariate normal response model.
2. Patients from whom we only observe the response contribute the marginal normal likelihood for the responses.
3. Patients from whom we only observe baseline contribute the marginal normal likelihood for the baselines.
4. Means and variances are estimated by default; other parameters can be estimated from these (see the example on p. 147) and standard errors estimated using the delta method (e.g. (A.16). Alternatively, as described in Chapter 3, the model can often be parameterised so the estimates of interest are obtained directly.

Appendix B

Prior eliciting questionnaire (Subsection 6.4.3)

Questionnaire on differences between responders
and non-responders in reviewer trial
James Carpenter & Stephen Evans, 14th Jan 2003

The trial

Compares the effect of reviewer training by face-to-face workshop or postal tuition pack with control (i.e. no intervention). Participating reviewers were sent three papers (one baseline and two post randomisation) each of which contained a number of errors. Principal outcomes are quality or review (a score summed over 7 items) and number of major errors identified. Here we focus on the quality of review.

Missing outcome data

In any trial, missing outcome data are likely. If non-responders differ systematically from responders then we may get biased estimates of the benefit of the intervention.

Many statisticians now advocate performing sensitivity analysis to allow for systematic differences between responders and non-responders.

We wish to use a Bayesian method that requires researchers to guess the magnitude of this systematic difference, even though it cannot be estimated from data. We will use this information both in the main analysis of the trial and in methodological work for a more statistical audience.

What we want to do

We wish to ask all the investigators to think about plausible values of the mean difference between responders and non-responders. We would like to stress that we are asking about the average differences that you would expect to observe over many thousands of responders and non-responders, not the random differences that you might expect to observe between individual responses. Naturally you are not sure what the differences between responders and non-responders are (if any). Nonetheless you may feel that some differences are more plausible than others. We therefore ask you to enter your weight of belief for each of the possible differences shown in the table overleaf. On a scale of 0 (impossible) to 100 (certainty), the more strongly you believe the plausibility of a particular difference, the greater should be your weight for that difference. Your weights should sum to 100.

If anything is not clear, or you have any comments then please contact James on (+44) (0)20 7927 2033 or by email on james.carpenter@lshtm.ac.uk

Please turn over

Questionnaire, page 1

Job title:.....

Question 1: Have you seen the data? Circle your answer: YES / NO

Question 2: Suppose the mean review quality for reviewers who respond to the second and third (i.e. final) paper is 3, with standard deviation 0.5, so that about 95% of these responders have values between 2 and 4. [These numbers are completely invented.]

What is your expectation for the mean review quality for those who do not respond to the either the second or the third paper?

To help you, the hypothetical example shows a rather idiosyncratic statistician who is convinced that non-responders will differ on average from responders by 1/2 or 3/4 points, but is not sure whether they will be better or worse than responders.

Mean review quality of reviewers who do not respond to either the 2nd or the 3rd paper (minimum 0, maximum 5)

	Non-responders								
	worse than responders by			same as responders	better than responders by				TOTAL
	1 or more	0.75	0.5	0.25		0.25	0.5	0.75	1 or more	
Hypothetical example	0	25	25	0	0	0	25	25	0	100
Your answers										

PLEASE CHECK YOUR WEIGHTS SUM TO 100

THANK YOU FOR YOUR HELP

Appendix C

Code for examples

The code for each of the examples in Chapters 3–6 is given below. With the exception of two examples, all the code is SAS.

C.1 Code for Chapter 3

Example 3.1

```
* Input placebo arm data only
data first;
   infile "...";
   input id time fev;
run;

proc mixed data=first;
    class id time;
    model fev = time/ ddfm=kr  htype=2 noint;
    repeated time/subject=id type=un ;
    lsmeans time;
run;
```

Example 3.2

```
* Adjusted for baseline
data second;
    infile "...";
    input fev fevbl treat;
run;

proc reg data=second;
    model fev=treat;
run;

proc reg data=second;
    model fev=treat fevbl;
run;
```

Example 3.3

```
* Adjusted for (=conditional on) baseline

data thirda;
    infile "...";
    input fev fevbl treat;
run;

prog reg data=thirda;
    model fev=fevbl treat;
run;

* Data MCAR given baseline and treatment.
* Note time indexes baseline and response
data third;
    infile "...";
    input id time fev treat ;
run;

*Treatment estimate marginal to baseline
proc mixed data=third;
    class time treat subject;
    model fev = time*treat/s ddfm=kr htype=2;
    repeated time/subject=id type=un;
run;
```

Example 3.4

```
data four;
    infile "...";
    input id treat time fev ;
    if time=1 then mytreat=0;
    else mytreat=treat;
run;

proc mixed data=four;
    class id time mytreat;
    model fev = mytreat/ s htype=2 noint ddfm=kr;
    repeated time/subject=id type=un ;
    lsmeans mytreat / diff;
run;
```

Example 3.6

```
data five;
    infile "...";
    input id treat rinidc fev newtreat;
run;

proc mixed data=five;
    class id time newtreat;
    model fev = newtreat/ s htype=2 noint ddfm=kr;
    repeated rinidc/subject=id type=un ;
    lsmeans newtreat / diff;
run;
```

Example 3.7

This uses the same code as Example 3.6, but with a different data arrangement, shown in Table 3.12.

Example 3.8

```
data six;
    infile "...";
    input id sex age baseline treat newtreat resp ;
run;

*Model with all exacerbations as separate responses
proc mixed data=six maxit=50 ;
    class id sex treat newtreat;
    model resp = newtreat treat age sex baseline
                 newtreat*treat newtreat*age
                 newtreat*sex newtreat*baseline   /s  Htype=2  Ddfm=Kr;
    repeated newtreat/Subject=Id Type=Un ;
run;
```

C.2 Code for Chapter 4

Example 4.1

```
*i) Maximum likelihood
* time indexes baseline and 6 month fev
data first;
    infile "...";
    input id time fev;
run;

proc mixed data=first;
    class id time;
    model fev = time/ ddfm=kr  htype=2 noint;
    repeated time/subject=id type=un ;
    lsmeans time;
run;

*ii) SAS proc MI
data second;
    infile "...";
    input id base fev;
run;

proc mi data=second seed=1 out=full;
    var pfev fev;
    mcmc impute=full;
run;

proc print data=full;
run;

proc means data=full;
    output out=full2 mean(fev)=m1 stderr(fev)=s1;
    by _Imputation_;
run;

proc mianalyze data=full2;
    stderr s1;
    modeleffect m1;
run;
```

Example 4.4

```
data third;
    infile "..";
    input id treat bmi base fev;
*   Make treatment 1/0 for simplicity
    treat=treat-1;
run;

proc sort data=third out=thirdsort;
    by treat;
run;

proc reg data=thirdsort;
    model fev = base treat;
run;

proc mi data=thirdsort seed=1 out=full nimpute=200;
    by treat;
    MCMC nbiter=5000 niter=5000;
    var bmi base fev;
    mcmc impute=full;
run;

proc sort data=full;
    by _imputation_ id;
run;

proc reg data=full outest=outreg covout noprint ;
     model fev= base treat;
     by _Imputation_;
run;

proc mianalyze data=outreg;
    modeleffects Intercept base treat;
run;

*Proc mixed alternative:
data fourth;
    infile "...";
*   order indexes response: BMI and 3 year FEV
    input id treat baseline order resp ;
run;

proc mixed data=fourth asycov;
    class id treat order;
    model resp = order treat baseline baseline*order
                 treat*order/ s htype=2  ddfm=kr;
    repeated order/subject=id type=un group=treat;
run;
```

Example 4.5

```
data fifth;
    infile "...";
    input id treat bmi base mexas fev1-fev6;
*   smexas is the square root of the mean exacerbation rate
    smexas=sqrt(mexas);
run;

proc sort data=fifth out=fifthsort;
    by treat;
run;

*   Model using observe data only
proc reg data=fifthsort;
    model fev6 = base treat;
run;

proc mi data =fifthsort seed=1 out=full nimpute=50;
    by treat;
    MCMC nbiter=5000 niter=5000;
    var bmi smexas base fev1-fev6;
    mcmc impute=full;
run;

proc sort data=full;
    by _imputation_ id;
run;

proc reg data=full outest=outreg covout noprint ;
    model fev6= base treat;
    by _Imputation_;
run;

proc mianalyze data=outreg;
    modeleffects Intercept base treat;
run;

* Maximum likelihood analysis
data fifthml;
    infile "...";
*   Note we need a different arrangement of the
*   data from multiple imputation
    input id treat base order resp ;
run;
```

```
proc mixed data=fifthml maxit=50 ;
    class id treat order;
    model resp = order treat base order*treat order*base
                 /s  Htype=2  Ddfm=Kr;
    repeated order/subject=id type=un group=treat;
run;
```

C.3 Code for Chapter 5

Example 5.1

```
data first;
    infile '..';
*   The data set only contains observations from period three
    input sub base period treat resp;
run;

*   Fit generalised linear model
proc genmod data=first descending;
    class sub;
    model resp = base treat  / d=bin;
run;

*   Fit population averaged model by GEE, with exchangeable correlation
proc genmod data=first descending;
    class sub;
    model resp = base treat  / d=bin;
    repeated subject=sub / type=exch corrw;
run;

*   Fit random intercepts model by maximum likelihood
*   Note model was fitted three times, each time starting from
*   the estimates from the previous fit. At the third time,
*   estimation left the parameters little changed.

proc nlmixed data = example2;
    parms interc=1.7 b=0.016 t=0.5443 lnsig=0.6535;
*   linear predictor:
    eta = interc + t*treat + b*base + u;
*   model:
    pi = probnorm(eta);
    model  resp ~ binomial(1,pi);
*   distribution of random effects
    s = exp(lnsig);
    random u ~ normal([0],[s]) subject=sub;
run;
```

Example 5.3

```
data second;
    infile '...';
    input sub base period treat resp;
*   Create dummy variables for the three periods
    if period=1 then pr1 = 1; else pr1=0;
    if period=2 then pr2 = 1; else pr2=0;
    if period=3 then pr3 = 1; else pr3=0;
    output;
run;

* Random intercepts model
* note log parameterisation of variance
proc nlmixed data = second;
*   parameter starting values; b is baseline, t is treatment and
*   pt1 pt2 are period treatment interactions
    parms interc=2.7 b=0.04 p1=-1 p2=-0.4 t=1.15 pt1=1.0 pt2=0.5
    lnsig=1.8;
*   linear predictor
    eta = interc +  t*treat + b*base + p1*pr1 + p2*pr2+ pt1*pr1*treat
          + pt2*pr2*treat + u;
*   model
    expeta = exp(eta); pi = expeta/(1+expeta);
    model  resp ~ binomial(1,pi);
*   random intercepts
    random u ~ normal(0,exp(lnsig)) subject=sub;
run;

*   Random intercepts and slopes model
*   Note parameterisation of correlation
proc nlmixed data = second;
*   starting values
    parms interc=3.3 b=0.03 p1=-1.12  p2=-0.75 t=1.25 pt1=1.56 pt2=0.39
    lnsig0=2.90 lnsig1=0.90 lncorr=-2.21;
*   linear predictor
    eta = interc +  t*treat + b*base + p1*pr1 + p2*pr2+ pt1*pr1*treat
          + pt2*pr2*treat + u0 + u1*period;
   expeta = exp(eta); pi = expeta/(1+expeta);
   model  resp ~ binomial(1,pi);
*  random intercepts and slopes
   s0=exp(lnsig0);
   s1=exp(lnsig1);
   r01=( -1 + exp(lncorr)  ) /( 1 + exp(lncorr) );
   random u0 u1~ normal([0,0],[s0,sqrt(s0*s1)*r01,s1])  subject=sub;
run;
```

Example 5.4

```
data third;
    infile '...';
    input sub bascov per trt resp;
*   generate dummy variables for the periods
    if per=1 then pr1 = 1; else pr1=0;
    if per=2 then pr2 = 1; else pr2=0;
    if per=3 then pr3 = 1; else pr3=0;
run;

*   Multivariate subject effect

proc nlmixed data = third;
    parms interc=2.65 b=0.026 p1=-0.972  p2=-0.39 t=1.15
          pt1=1.035 pt2=0.546 lnsig0=2 lgcr12=1
          lgcr13=1 lgcr23=1 lnsig1=2 lnsig2=2 ;
*   linear predictor
    eta = interc +  t*trt + b*bascov + p1*pr1 + p2*pr2+ pt1*pr1*trt
          + pt2*pr2*trt + u1*pr1 + u2*pr2 + u3*pr3;
    expeta = exp(eta); pi = expeta/(1+expeta);
    model  resp ~ binomial(1,pi);
*   random structure
    s0 = exp(lnsig0);
    s1 = exp(lnsig1);
    s2 = exp(lnsig2);
    r12 = ( exp(lgcr12) )/(1+exp(lgcr12));
    r13 = ( exp(lgcr13) )/(1+exp(lgcr13));
    r23 = ( exp(lgcr23) )/(1+exp(lgcr23));

     random u1 u2 u3 ~ normal([0,0,0],[s0,sqrt(s0*s1)*r12,s1,
                                       sqrt(s0*s2)*r13,
                                       sqrt(s1*s2)*r23,s2]) subject=sub;
run;

*    For the second analysis, we replace the random statement by
*    s0 = exp(lnsig0);
*    r12 = ( exp(lgcr12) )/(1+exp(lgcr12));
*    r13 = ( exp(lgcr13) )/(1+exp(lgcr13));
*    r23 = ( exp(lgcr23) )/(1+exp(lgcr23));
*
*    random u1 u2 u3 ~ normal([0,0,0],[s0,s0*r12,s0,s0*r13,
*                                      s0*r23,s0]) subject=sub;
```

Example 5.5

```
data fourth;
    infile '...';
    input id treat respind resp;
*   Note response indicator shows whether response is
*   baseline or test results for period 1.
run;

proc nlmixed data = fourth;
*   starting values
    parms intnorm=51.12 intbin=3.73 t=2.80 lgvbin=2.46
          trcorr=0.56 lgvnorm=6.63;

*   covariance parameters of the bivariate latent structure
    vbin = exp(lgvbin);
    vnorm = exp(lgvnorm);
    cov = sqrt(vnorm)*sqrt(vbin)*(exp(trcorr)-1)/(exp(trcorr)+1) ;

*   linear predictor for the normal observation, conditional on ubin
*   note third term on right: conditional mean of normal given ubin
    etanorm  = intnorm + ubin*cov/vbin;

*   linear predictor for the binary observations, conditional on ubin
    etabin  = intbin + t*treat + ubin;

*   User defined log-likelihood function
*   respind=0 indicates baseline
    if (respind=0) then do;
*   conditional variance of baseline
    v = vnorm-cov*cov/vbin;
      logl = -0.5*(log(v)+(resp-etanorm)*(resp-etanorm)/v);
    end;
    else if (respind=1) then do;
*   contribution for binary data
    pi = exp(etabin)/(1+exp(etabin));
    logl = resp*log(pi)+(1-resp)*log(1-pi);
    end;

    model resp ~ general(logl);

    random ubin ~ normal(0,vbin) subject=id;

*   For ease of interpretation, request back-transformed
*   estimates of variances and correlations

    estimate 'correlation' (exp(trcorr)-1)/(exp(trcorr)+1);
    estimate 'normal variance' exp(lgvnorm);
    estimate 'binary variance' exp(lgvbin);
run;
```

Example 5.6

```
data fifth;
    infile '...';
*   variables: wt is weight, pr1-pr10 is pain relief at observation
*   times 1-10
    input  id age wt treat base pr1-pr10;

*   make dummy variables for the treatment groups;
*   this seems to be necessary for PROC MIANALYZE to work properly
    if treat=1 then t1=1;
    else t1=0;
    if treat=2 then t2=1;
    else t2=0;
    if treat=3 then t3=1;
    else t3=0;
    if treat=4 then t4=1;
    else t4=0;
    if treat=5 then t5=1;
    else t5=0;
    if treat=6 then t6=1;
    else t6=0;

*   make data set monotone missing
    if (pr4=. and pr5 ne .) then delete;
    if (pr5=. and pr6 ne .) then delete;
    if (pr6=. and pr7 ne .) then delete;
    if (pr7=. and pr8 ne .) then delete;
    if (pr8=. and pr9 ne .) then delete;
    if (pr9=. and pr10 ne .) then delete;

    output;
run;

proc mi data=fifth seed=1 out=imputed nimpute=50;
    class treat pr1 pr2 pr3 pr4 pr5 pr6 pr7 pr8 pr9 pr10;

*   specification of imputation models
    monotone logistic( pr6 = pr5 treat pr5*treat  / details);
    monotone logistic( pr7 = pr6 treat pr6*treat / details);
    monotone logistic( pr8 = pr7 treat pr7*treat / details);
    monotone logistic( pr9 = pr8 treat pr8*treat / details);
    monotone logistic( pr10= pr9 treat pr9*treat  / details);
    var treat pr1 pr2 pr3 pr4 pr5 pr6 pr7 pr8 pr9 pr10;
run;
```

```
*   Model to estimate treatment effect at visit 10 from
*   each imputation
proc genmod data=imputed descending ;
    class id treat;
    model pr10 = t1 t2 t3 t4 t5 t6 / covb d=bin;
    by _Imputation_;
    ods output ParameterEstimates=gmparms
    ParmInfo=gmpinfo
    CovB=gmcovb;
run;

proc mianalyze parms=gmparms covb=gmcovb parminfo=gmpinfo;
    modeleffects Intercept t1 t2 t3 t4 t5 t6 ;
run;

proc genmod data=dental descending;
    class id treat;
    model pr10 = t1 t2 t3 t4 t5 t6 / covb d=bin;
    ods output ParameterEstimates=gmparms
    ParmInfo=gmpinfo
    CovB=gmcovb;
run;
```

Example 5.7

Fit model with common variance (cf code for example 5.4 above) and predict the u_3's and their standard errors.

```
data sixth;
    infile '...';

*   NB this version of the data includes the missing
*   responses, denoted with a .
    input sub base period treat resp;

*   Create dummy variables for different periods
    if period=1 then pr1 = 1; else pr1=0;
    if period=2 then pr2 = 1; else pr2=0;
    if period=3 then pr3 = 1; else pr3=0;
    output;
run;
```

```
proc nlmixed data = sixth cov cor;

*   starting values
    parms interc=2.65 b=0.026 p1=-0.972  p2=-0.39 t=1.15
    pt1=1.035 pt2=0.546
    lnsig0=2 lgcr12=1 lgcr13=1 lgcr23=1  ;

*   linear predictor
    eta = interc +  t*treat + b*base + p1*pr1 + p2*pr2+ pt1*pr1*treat
          + pt2*pr2*treat + u1*pr1 + u2*pr2 + u3*pr3;
    expeta = exp(eta); pi = expeta/(1+expeta);

    model  resp ~ binomial(1,pi);
    s0 = exp(lnsig0);
    r12 = ( exp(lgcr12) )/(1+exp(lgcr12));
    r13 = ( exp(lgcr13) )/(1+exp(lgcr13));
    r23 = ( exp(lgcr23) )/(1+exp(lgcr23));

    random u1 u2 u3 ~ normal([0,0,0],[s0,s0*r12,s0,
                                      s0*r13,s0*r23,s0]) subject=sub;

*   Predict the u3's and their standard errors for imputation
    predict u3 out=missthree;
run;

proc print data=missthree;
run;

*   The output is then saved and the remaining
*   calculations carried out in R.
*   Thus the following code is R code

#   Read in missthree
#   Variables of interest are pred and se
u.three<-data.frame(scan("missthree.dat",list(obs=0,sub=0,base=0,
                         period=0,treat=0,resp=0,pr1=0,pr2=0,
                         pr3=0,pred=0,se=0,df=0,
                         tval=0,probt=0,alpha=0,low=0,up=0)))

#   Only keep columns of interest
u.three<-u.three[,c(1,2,10,11)]
```

```
#   Rows are repeats of the same data (for each observation)
#   Only keep first of each set of 3
u.three.id<-unique(u.three$sub)
for (i in seq(along=u.three.id)) {
if (i==1) { u3.per3sim<-u.three[u.three$sub==u.three.id[i],][1,]
} else {
u3.per3sim<-rbind(u3.per3sim,u.three[u.three$sub==u.three.id[i],][1,])
} }
rm(u.three)
u3.per3sim<-data.frame(u3.per3sim)
# Now have the u3's and their SEs for each subject.

# For treatment three, the linar predictor is
intercept + baseline + treat

# Parameter estimates from NLMIXED run above
mean.pars<-c(3.2118,0.03040,1.3778)
vmat<-matrix(c( 0.3021,  -0.00264,  -0.1637,
                -0.00264,  0.000058, 0.000215,
                -0.1637 , 0.000215 ,0.3434 ),ncol=3 )

# Imputation function preliminaries:

# Note column names of data per3sim
#  "id"     "base"   "period" "treat"  "resp"

# vector miss.per3 has ids of the 52 subjects
# with missing data at period 3
# first extract their treatment group and baseline

treat.missper3<-rep(52,0)
treat.missbase<-rep(52,0)
for (i in seq(along=miss.per3)) {
treat.missper3[i]<-per3sim[per3sim$id==miss.per3[i],4][1]
treat.missbase[i]<-per3sim[per3sim$id==miss.per3[i],2][1]
}


# Imputation function

impute.per3sim<-function() {

# simulate the number of tests
no.tests<- round(rnorm(52,mean=19+treat.missper3,
                           sd=sqrt(60-10*treat.missper3)))
no.tests[no.tests<=2]<-2 # agrees with original data minimum
```

```
# Draw sample from multivariate normal distribution
# Create imputed linear predictor (length 52
# as 52 patients have missing data)
imp.pars<-mvrnorm(1,mu=mean.pars,Sigma=vmat)
imp.lp<-rep(0,52)
for (i in 1:52){
imp.lp[i]<-sum( imp.pars*c(1,treat.missbase[i],treat.missper3[i]) ) +
           rnorm(1, mean=u3.per3sim[u3.per3sim$sub==miss.per3[i],3],
                 sd=u3.per3sim[u3.per3sim$sub==miss.per3[i],4] )
}

# Expand it by the number of tests
expand.lp<-rep(imp.lp,no.tests)
s
# Sample from binomial distribution (n=1) for the responses
# to these tests
imp.response<-rbinom(n=sum(no.tests),
                     size=rep(1,sum(no.tests)),
                     prob=expit(expand.lp))

#
#make data frame for results
temp<-data.frame(cbind( rep(miss.per3,no.tests),
                        rep(treat.missbase,no.tests),
                        rep(rep(3,52),no.tests),
                        rep(treat.missper3,no.tests),
                        imp.response  ))

names(temp)<-names(per3sim)

# output the data frame at the end of the function
temp
}


# Create and analyse each imputed data
# set in turn using this function

set.seed(1)
no.imputations<-100
imp.treat<-rep(0,no.imputations)
imp.se<-rep(0,no.imputations)

for (j in 1:no.imputations) {

   temp.analysis<-data.frame(rbind(impute.per3sim(),
                             per3sim[per3sim$period==3,]) )
```

```
   gee.mod<- gee(resp~treat+base,data=temp.analysis,id=id,
                 family=binomial(logit),corstr="independence")

   imp.treat[j]<-gee.mod$coefficients[2]

   imp.se[j]<-sqrt(diag(gee.mod$robust.variance))[2]

}

# Combine the results using Rubin's rules
mi.est<-mean(imp.treat)
with.var<- mean(imp.se^2)
bet.var<-var(imp.treat)
mi.se<-sqrt( with.var + (1 + 1/no.imputations)*bet.var)
mi.df<- (no.imputations -1) * (
         1 + with.var/ ( (1+ 1/no.imputations)*bet.var) )^2
```

C.4 Code for Chapter 6

Example 6.1

```
winBUGS code for selction modelling.

First, model for the reponse alone:

model{  # model is enclosed in curly brackets

  for (i in 1: npatients) {  # 751 patients which we index by i

#    Model of for the 6 post randomisation FEV measurments is
#    multivariate normal

     fev[i, 1:6] ~ dmnorm(mu[i, ], R[ , ] )

#    Parameterise mean with full treatment time and
#    baseline time interaction

     for (j in 1:6) {
         mu[i,j]<- betacon.time[j] + betacon.treat[j]*treat[i]
                   + betacon.base[j]*base[i]
     }

  }

# Wishart prior for precision matrix R

  R[1:6 , 1:6] ~ dwish(lambda[1:6 , 1:6  ], 6)
```

```
# create an estimate of variance/covariance matrix:

  vcov.mat[1:6 ,1:6 ] <-inverse(R[ 1:6, 1:6 ])

# priors for coefficients: vague normal:

  for (j in 1:6) {
     betacon.time[j]~dnorm(0.0,1.0E-6)
     betacon.treat[j]~dnorm(0.0,1.0E-6)
     betacon.base[j]~dnorm(0.0,1.0E-6)
  }

}

Second, joint model for response and withdrawal, under
Model 1 - all misssing data treated as interim missing.

model {
  for (i in 1: npatients) {  # 751 patients which we index by i

#    Model of for the 6 post randomisation FEV measurments is
#    multivariate normal

     fev[i, 1:6] ~ dmnorm(mu[i, ], R[ , ] )

#    Parameterise mean with full treatment time and
#    baseline time interaction

     for (j in 1:6) {
         mu[i,j]<- betacon.time[j] + betacon.treat[j]*treat[i]
                   + betacon.base[j]*base[i]
     }
  }

# Model for observing the response

  for(j in 2:6) {
     resp[i,j] ~ dbin(p[i,j],1) # Response is 1 (yes) or
                                #0 (no) on each occasion

#    linear predictor depends on visit, treatment, previous reading
#    and the difference between the previous and current reading
     logit(p[i,j]) <- alpha[j] + betadrop*treat[i] + gamma*fev[i,j-1] +
                      0.1*(fev[i,j]-fev[i,j-1])
                      # last term is log odds ratio of difference
                      # which is fixed at 0.1
  }
```

```
# Wishart prior for precision matrix R

  R[1:6 , 1:6] ~ dwish(lambda[1:6 , 1:6  ], 6)

# create an estimate of variance/covariance matrix:

  vcov.mat[1:6 ,1:6 ] <-inverse(R[ 1:6, 1:6 ])

# priors for coefficients: vague normal:

  for (j in 1:6) {
     betacon.time[j]~dnorm(0.0,1.0E-6)
     betacon.treat[j]~dnorm(0.0,1.0E-6)
     betacon.base[j]~dnorm(0.0,1.0E-6)
     alpha[j]~dnorm(0.0,1.0E-6)
  }

  betadrop~dnorm(0.0,1.0E-6)
  gamma~dnorm(0.0,1.0E-6)


}

Third, model 2 for non-response: after a patient withdraws,
they do not return.

model {
  for (i in 1: npatients) {  # 751 patients which we index by i

#    Model of for the 6 post randomisation FEV measurments is
#    multivariate normal

     fev[i, 1:6] ~ dmnorm(mu[i, ], R[ , ] )

#    Parameterise mean with full treatment time and
#    baseline time interaction

     for (j in 1:6) {
         mu[i,j]<- betacon.time[j] + betacon.treat[j]*treat[i]
                   + betacon.base[j]*base[i]
     }

  }
```

```
# Model for observing the response

  for (j in 2:lvisit[i]) { # variable lvisit is time of
                           # list visit (2...6)

     resp[i,j] ~ dbin(p[i,j],1) # Response is 1 (yes) till
                                # withdrawal, then 0

#    linear predictor depends on visit, treatment,
#    previous reading and the difference between the
#    previous and current reading

     logit(p[i,j]) <- alpha[j] + betadrop*treat[i] + gamma*fev[i,j-1] +
                      0.1*(fev[i,j]-fev[i,j-1])
                      # last term is log odds ratio of difference
                      # which is fixed at 0.1
  }

# Wishart prior for precision matrix R

  R[1:6 , 1:6] ~ dwish(lambda[1:6 , 1:6  ], 6)

# create an estimate of variance/covariance matrix:

  vcov.mat[1:6 ,1:6 ] <-inverse(R[ 1:6, 1:6 ])

# priors for coefficients: vague normal:

  for (j in 1:6) {
     betacon.time[j]~dnorm(0.0,1.0E-6)
     betacon.treat[j]~dnorm(0.0,1.0E-6)
     betacon.base[j]~dnorm(0.0,1.0E-6)
     alpha[j]~dnorm(0.0,1.0E-6)
  }

  betadrop~dnorm(0.0,1.0E-6)
  gamma~dnorm(0.0,1.0E-6)


}
```

Example 6.1

Code given in White *et al.* (2007).

Example 6.3

```
*   First the SAS macro for adapting the imputations:
*   Author: Prof James H. Roger, james.h.roger@gsk.com
/*
Macro name: Modify
Parameters;
        Data=    Name of original data set
        Imp=     Name of imputed data set
        Out=  Name of Output data set
        Var=     List of variables as used in Var statement for MI
        Delta=   The amount to decrease imputed value by
        S=       SD for Normal distribution from which Change is
                 sampled with mean Delta for each imputation
        Trx=     Name of variable holding treatment classification
*/

%macro modify(data= ,imp= ,out= ,var= ,delta= ,s= ,trx= );
         %local i n;

         %* Get number of elements in the Var List as macro variable n;
         %let i=1;
         %let txt=%scan(&var, &i, %str( ));
         %do %while(%length(&txt)) ;
                  %let i=%eval(&i +1);
                  %let txt=%scan(&var, &i, %str( ));
         %end;
         %let n=%eval(&i -1);

         * Delete data sets before we use them;
         Proc datasets library=work;
         delete Temp1 Temp2 Temp3;
         quit;

         * Set up data set with indicator of whether data is missing;
         * This gets around issue of variables having same names in
         * original and imputed data sets;
         * Add Row number so we can merge this with every imputed
         * data set;
         data Temp1;
                  set &data;
                  array My_Var[1:&n] &Var;
                  array My_Ind[1:%eval(&n+1)] My_Ind1-My_Ind&n No;
                  keep My_row My_ind1-My_Ind&n;
                  My_Row=_N_;
                  * Note that No is My_Ind[&n+1];
                  No=0;
```

```
do i= &n to 1 by -1;
My_ind[i]= ( ((My_Var[i] > .z) + (My_ind[i+1])) > 0 );
 end;
 run;

* Add Row number of each record in the imputed data set
* within the Imputation number;
 data Temp2;
          set &imp;
          by _imputation_;
          retain My_Row 0;
          if first._imputation_ then My_Row=0;
          My_Row=My_Row+1;
 run;

 * Merge the two data sets based on this Row number;
 proc sql;
          create table Temp3 as
          select A.*, B.*
          from Temp2 A left join Temp1 B
          on A.My_Row = B.My_Row
          order by _imputation_, &Trx;
 quit;

 * Now calculate the required changes in imputed values;
 * My_Ind=1 if data is Real;
 * My_Ind=0 if data is Imputed;
 data &out;
          set  Temp3;
          by _imputation_ &trx;
          array My_Var[1:&n] &Var;
          array My_Ind[1:&n] My_Ind1-My_Ind&n;
          drop My_Row Change i My_ind1-My_Ind&n;
          retain delta;
          *** Here is where the modfication is done ***;
          * Change allows us to build up delta
          * within the subject;
          * Reset Delta for each Imputation * Treatment level;
          if first.&trx then do;
                   Delta=&delta+&s*rannor(0);
          end;
          Change=0;
          do i=1 to &n;
                   * If it is imputed then increase
                   * Change by delta;
                   if My_ind[i]=0 then do;
                            Change=Change + delta;
                   end;
```

```
                        * Do the change;
                        My_Var[i]=My_Var[i]-Change;
                end;
        run;

%mend modify;

*   Now we apply it to the example:

data two;
    infile "...";
    input id treat base fev1-fev6;
run;

proc sort data=two out=twosort;
     by treat;
run;

*   Multiple imputation
proc mi data =twosort seed=1 out=full nimpute=50;
    by treat;
    MCMC nbiter=5000 niter=5000;
    var fev1-fev6;
    mcmc impute=full;
run;


proc sort data=full;
    by _imputation_ id;
run;

*   Estimate treatment effect at final time point using each
*   imputed data set
proc reg data=full outest=outreg covout noprint ;
    model fev6= base treat;
    by _Imputation_;
run;

proc mianalyze data=outreg;
    modeleffects Intercept base treat;
run;

*   Now use modify macro (above) and redo analysis
*   delta is the mean of the difference in slope after withdrawal, with
*   standard error s (ctd...)
```

```
*   For example, to obtain same results as MAR, set delta=0
    and s to be tiny:

%modify(data=tensort, imp=full,
    out=james, var= fev1 fev2 fev3 fev4 fev5 fev6,
               delta=0.0, trx=treat, s=0.0000005);

proc reg data=james outest=outreg covout noprint ;
     model fev6= base treat;
     by _Imputation_;
run;

proc mianalyze data=outreg;
    modeleffects Intercept base treat;
run;
```

References

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989) *Statistical modelling in GLIM*. Oxford: Oxford University Press.

Allison, P. D. (2000) Multiple imputation for missing data: a cautionary tale. *Sociological methods and Research*, **28**, 301–309.

Barnard, J. and Rubin, D. (1999) Small-sample degrees of freedom with multiple imputation. *Biometrika*, pp. 948–955.

Brown, D. J. (2003) ICH E9 guideline 'Statistical principles for clinical trials': a case study. Response to A. Philips and V Haudiquet. *Statistics in Medicine*, **22**, 13–17.

Burge, P. S., Calverley, P. M. A., Jones, P. W., Spencer, S., Anderson, J. A. and Maslen, T. K. (2000) Randomised, double blind, placebo controlled study of fluticasone pripionate in patients with moderate to severe chronic obstructive pulmonary disease: the isolde trial. *British Medical Journal*, **320**, 1297–1303.

Busse, W. W., Chervinsky, P., Condemi, J., Lumry, W. R., Petty, T. L., Rennard, S. and Townley, R. G. (1998) Budesonide delivered by Turbuhaler is effective in a dose-dependent fashion when used in the treatment of adult patients with chronic asthma. *J Allergy Clin Immunol*, **101**, 457–463.

Callaham, M. L., Wears, R. L. and Waeckerle, J. F. (1998) Effect of attendance at a training session on peer reviewer quality and performance. *Annals for Emerging Medicine*, **32**, 318–322.

Callaham, M. L., Knopp, R. K. and Gallagher, E. J. (2002) Effect of written feedback by editors on quality of reviews. *Journal of the American Medical Association*, **287(21)**, 2781–2783.

Carpenter, J., Pocock, S. and Lamm, C. J. (2002) Coping with missing data in clinical trials: a model based approach applied to asthma trials. *Statistics in Medicine*, **21**, 1043–1066.

Carpenter, J., Kenward, M., Evans, S. and White, I. (2004) Letter to the editor: Last observation carry forward and last observation analysis by J. Shao and B. Zhong, Statistics in Medicine, 2003, **22**, 2429–2441. *Statistics in Medicine*, **23**, 3241–3244.

Carpenter, J. R., Kenward, M. G. and Vansteelandt, S. (2006) A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **169**, 571–584.

Carpenter, R. G. (1983) Scoring to provide risk-related primary health care: Evaluation and up-dating during use. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **146**, 1–32.

Chan, A. and Altman, D. G. (2005) Epidemiology and reporting of randomised trials published in PubMed journals. *The Lancet*, **365**, 1159–1162.

Committee for Proprietary Medicinal Products (CPMP) (2001) *Points to consider on missing data*. London: European Agency for the Evaluation of Medicinal Products; download from www.emea.eu.int (accessed January 10th 2006).

Daniel, R. (2007) Personal communication.

Diggle, P. J. (1989) Testing for random dropouts in repeated measurement data. *Biometrics*, **45**, 1255–1258.

Diggle, P. J. (1998) Dealing with missing values in longitudinal studies. In *Advances in the Statistical Analysis of Medical data* (Eds B. S. Everitt and G. Dunn), pp. 203–228. London: Arnold.

Diggle, P. J. and Kenward, M. G. (1994) Informative dropout in longitudinal data analysis (with discussion). *Journal of the Royal Statistical Society Series C (applied statistics)*, **43**, 49–94.

Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002) *Analysis of longitudinal data (second edition)*. Oxford: Oxford University Press.

Gelman, A. and Raghunathan, T. E. (2001) Using conditional distributions for missing-data imputation, in discussion of 'using conditional distributions for missing-data imputation' by Arnold *et al*. *Statistical Science*, **3**, 268–269.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov chain Monte-Carlo in practice*. London: Chapman and Hall.

Greenland, S. and Finkle, W. D. (1995) A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, **142**, 1255–64.

Heyting, A., Tolboom, J. T. B. M. and Essers, J. G. A. (1992) Statistical Handling of Drop-Outs in Longitudinal Clinical Trials. *Statistics in Medicine*, **11**, 2043–2061.

Heyting, A., Tolboom, J. T. B. M. and Essers, J. G. A. (1993) Response to Letter to the Editor from R. I. Harris. *Statistics in Medicine*, **12**, 2248–2250.

Higgins, J. P. T., White, I. R. and Wood, A. (2006) Missing outcome data in meta-analysis of clinical trials: development and comparison of methods, with recommendations for practice *Technical report, MRC Biostatistcs Unit, Cambridge UK* .

Hollis, S. and Campbell, F. (1999) What is meant by intention to treat analysis? Survey of published randomised controlled trials. *British Medical Journal*, **319**, 670–674.

Horton, N. J. and Lipsitz, S. R. (2001) Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, pp. 244–254.

ICH E9 Expert Working Group (1999) Statistical Principles for Clinical Trials: ICH Harmonised Tripartite Guideline. *Statistics in Medicine*, **18**, 1905–1942.

Jennrich, R. I. and Schluchter, M. D. (1986) Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, **42**.

Kadane, J. B. and Wolfson, L. J. (1998) Experiences in elicitation. *The Statistician*, pp. 3–19.

Kenward, M. G. (1998) Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in Medicine*, **17**, 2723–2732.

Kenward, M. G. and Molenberghs, G. (1998) Likelihood based frequentist inference when data are missing at random. *Statistical Science*, pp. 236–247.

Kenward, M. G. and Roger, J. H. (1997) Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**, 983–997.

Kenward, M. G., Molenberghs, G. and Thijs, H. (2003) Pattern-mixture models with proper time dependence. *Biometrika*, **90**, 53–71.

Lavori, P. W. (1992) Clinical trials in psychiatry: should protocol deviation censor patient data. *Neurosychompharmacology*, **6**, 39–48.

Lewis, J. A. (1999) Statistical Principles for Clinical Trials (ICH E9) An introductory note on an international guideline. *Statistics in Medicine*, **18**, 1903–1904.

Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical analysis with missing data (second edition)*. Chichester: Wiley.

Little, R. J. A. and Yau, L. (1996) Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, **52**, 471–483.

Magder, L. S. (2003) Simple approaches to assess the possible impact of missing outcome information on estimates of risk ratios, odds ratios, and risk differences. *Controlled Clinical Trials*, **24**.

McCullagh, P. (1980) Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B (statistical methodology)*, **41**, 109–142.

Moher, D., Schulz, K. F., Altman, D. G. and for the CONSORT group, L. L. (2001) The consort statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *The Lancet*, **357**, 1191–94.

Molenberghs, G. and Kenward, M. G. (2007) *Missing data in clinical studies*. Chichester: Wiley.

Molenberghs, G., Thijs, H., Kenward, M. G. and Verbeke, G. (2003) Sensitivity analysis of continuous incomplete longitudinal outcomes. *Statistica Nederlandica*, **57**, 112–135.

Molenberghs, G., Thijs, H., Jansen, I., Beunkens, C., Kenward, M. G., Mallinkrodt, C. and Carroll, R. J. (2004) Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, **5**, 445–464.

Molenberghs, G., Beunkens, C., Jansen, I., Thijs, H., van Steen, K., Verbeke, G. and Kenward, M. G. (2006) Analysis of incomplete data. In *Pharmaceutical Statistics with SAS* (Eds A. Dmitrienko, C. Chuang-Stein and R. D'Agostino). Cary, NC: SAS publishing.

Murray, G. and Findlay, J. G. (1988) Correcting for the Bias caused by Drop-outs in Hypertension Trials. *Statistics in Medicine*, **7**, 941–946.

Ng, E. S. W., Carpenter, J. R., Goldstein, H. and Rasbash, J. (2006) Estimation in generalised linear mexed models with binary outcomes by simulated maximum likelihood. *Statistical Modellling*, **6**, 23–42.

O'Hagan, A. (1998) Eliciting prior beliefs in substantial practical applications. *The Statistician*, pp. 21–25.

Peduzzi, P., Wittes, J. and Detre, K. (1993) Analysis as-randomized and the problem of non-adherence: an example from the veterans affairs randomized trial of coronary artery bypass surgery. *Statistics in Medicine*, **12**, 1185–1195.

Pocock, S. J. (1996) Clinical trials: A statistician's perspective. In *Advances in Biometry* (Eds P. Armitage and H. A. David), pp. 405–421. London: Wiley.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. and Solenberger, P. (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, **27**, 85–95.

Rasbash, J., Steele, F., Browne, W. and Prosser, B. (2004) *A user's guide to* `MLwiN` *(version 2.0)*. London: Institute of Education, 20 Bedford Way.

Reilly, M. (1993) Data analysis using hot deck multiple imputation. *The Statistician*, **42**, 307–313.

Rennie, D. (1999) Editorial peer review: its development and rationale. In *Peer review in health sciences* (Eds F. Godlee and T. Jefferson). London: BMJ books.

Robins, J. M. and Wang, N. (2000) Inference for imputation estimators. *Biometrika*, **85**, 113–124.

Roger, J. H. (2005) Personal communication.

Roger, J. H. (2006) SAS macro for sensitivity analysis via pattern mixture modelling. Email: james.h.roger@gsk.com.

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

Rubin, D. B. (1987) *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. B. and Schenker, N. (1987) Interval estimation from multiply-imputed data: a case study using agriculture industry codes. *Journal of the American Statistical Association*, **81**, 366–374.

Rubin, D. B., Stern, H. and Verhovor, V. (1995) Handling 'don't know' survey responses: the case of the slovenian plebiscite. *Journal of the American Statistical Association*, **90**, 822–828.

Savage, M. P., Douglas J. S. Jr, Fischman, D. L., Pepine, C. J., King S. B. 3rd, Werner, J. A., Bailey, S. R., Overlie, P. A., Fenton, S. H., Brinker, J. A., Leon, M. B. and Goldberg, S. (1997) Stent placement compared with balloon angioplasty for obstructed coronary bypass grafts. Saphenous Vein De Novo Trial Investigators. *New England Journal of Medicine*, **337**, 740–747.

Schafer, J. L. (1997) *Analysis of incomplete multivariate data*. London: Chapman and Hall.

Schafer, J. L. (1999) Multiple imputation: a primer. *Statistical Methods in Medical Research*, **8**, 3–15.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999) Adjusting for nonignorable drop-out using semi-parametric nonresponse models (with comments). *Journal of the American Statistical Association*, **94**, 1096–1146.

Schroter, S., Black, N., Evans, S., Carpenter, J., Godlee, F. and Smith, R. (2004) Effects of training on quality of peer review: randomised controlled trial. *British Medical Journal*, **328**, 673–675.

Shao, J. and Zhong, B. (2003) Last observation carry-forward and last observation analysis. *Statistics in Medicine*, pp. 3241–3244.

Shih, W. J. and Quan, H. (1997) Testing for treatment differences with dropouts present in clinical trials - a composite approach. *Statistics in Medicine*, **16**, 1225–1239.

Spiegelhalter, D. J., Thomas, A. and Best, N. G. (1999) *WinBUGS version 1.2 user manual*. Cambridge: MRC Biostatistics Unit.

Spiegelhalter, D. J., Abrams, K. R. and Myles, J. P. (2003) *Bayesian Approaches to Clinical Trials and Health Care Evaluation*. New York: Wiley.

Taylor, J. M. G., Cooper, K. L., Wei, J. T., Sarma, R. V., Raghunathan, T. E. and Heeringa, S. G. (2002) Use of multiple imputation to correct for non-response bias in a survey of urologic symptoms among African-American men. *American Journal of Epidemiology*, **156**, 774–782.

van Buuren, S., Boshuizen, H. C. and Knook, D. L. (1999) Multiple imputation of missing blood presure covariates in survival analysis. *Statistics in Medicine*, pp. 681–694.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M. and Rubin, D. B. (2006) Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, **76**, 1049–1064.

van Rooyen, S., Godlee, F., Smith, R., Evans, S. and Black, N. (1998) The effect of blinding and unmasking on the quality of peer review: a randomised trial. *Journal of the American Medical Association*, **280**, 234–237.

van Rooyen, S., Black, N. and Godlee, F. (1999a) Development of the review quality instrument (rqi) for assessing peer reviews of manuscripts. *Journal of Clinical Epidemiology*, **52**, 625–629.

van Rooyen, S., Godlee, F., Evans, S., Black, N. and Smith, R. (1999b) Effect of open peer review on quality of review and on reviewers' recommendations: a randomised trial. *British Medical Journal*, **318**, 23–27.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. New York: Springer Verlag.

Verzilli, C. and Carpenter, J. R. (2002) A Monte Carlo EM algorithm for random-coefficient-based dropout models. *Journal of Applied Statistics*, **29**, 1011–1021.

Walsh, E., Rooney, M., Appleby, L. and Wilkinson, G. (2000) Open peer review: a randomised controlled trial. *British Journal of Psychiatry*, **176**, 47–51.

Wang, N. and Robins, J. M. (1998) Large-sample theory for parametric multiple imputation procedures. *Biometrika*, **85**, 935–948.

White, I., Carpenter, J., Evans, S. and Schroter, S. (2007) Eliciting and using expert opinions about non-response bias in randomised controlled trials. *Clinical Trials*, **4**, 125–139.

White, I. R. and Thompson, S. G. (2005) Adjusting for partially missing baseline measurements in randomized trials. *Statistics in Medicine*, **24**, 993–1007.

White, I. R., Babiker, A. G., Walker, S. and Darbyshire, J. H. (1999) Randomisation-based methods for correcting for treatment changes: examples from the concorde trial. *Statistics in Medicine*, **18**, 2617–2634.

White, I. R., Carpenter, J. R., Pocock, S. J. and Henderson, R. (2003) Adjusting treatment comparisons to account for non-randomised interventions: an example from an angina trial. *Statistics in Medicine*, **22**, 781–793.

Wood, A. M., White, I. R. and Thompson, S. G. (2004) Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals. *Clinical Trials*, **1**, 368–376.

Zeger, S. L. and Liang, K.-Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.

Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.

Index