Module 14: Missing Data
Concepts

Jonathan Bartlett & James Carpenter
London School of Hygiene & Tropical Medicine

Supported by ESRC grant RES 189-25-0103 and MRC grant G0900724

Pre-requisites

- Module 3
- For the section on multilevel data, Modules 4 and 5

Online resources:
www.missingdata.org.uk

Contents

Introduction **1**

Introduction to the Class Size Data **2**

C14.1 The Model of Interest **4**

C14.2 Investigating Missingness **5**

C14.2.1 Missing completely at random 5
C14.2.2 Missing at random 6
C14.2.3 Missing not at random 6
C14.2.4 Investigating missingness 7
C14.2.5 Missingness in multiple variables and monotone missingness 9

C14.3 Ad-hoc Methods **11**

C14.3.1 Missing category method 11
C14.3.2 Mean imputation 11
C14.3.3 Regression (conditional) mean imputation 12
C14.3.4 Random or stochastic regression imputation 13

C14.4 Complete Records Analysis **16**

C14.5 Multiple Imputation **18**

C14.5.1 Creating the imputations 18
C14.5.2 Choosing variables to include in the imputation model 20
C14.5.3 Proper imputation 21
C14.5.4 Using the imputed datasets 21
C14.5.5 How many imputations should we generate? 22
C14.5.6 Class size data 22
C14.5.7 Assumptions made by multiple imputation 23
C14.5.8 Software 23

C14.6 Inverse Probability Weighting **24**

C14.6.1 Implementation 24

C14.6.2 Assumptions and validity 24
C14.6.3 Class size data 25
C14.6.4 Inference after IPW 26
C14.6.5 IPW vs MI 26

C14.7 Multilevel and Longitudinal Studies 28

C14.7.1 Direct likelihood approaches....................28
C14.7.2 Multiple imputation 28
C14.7.3 Inverse probability weighting 30

C14.8 Summary and Conclusions 31

C14.8.1 What we have not covered....................32

References 34

Acknowledgements 36

Introduction

Missing data are ubiquitous in economic, social, and medical research. In this module we aim to introduce the issues raised by missing data and the central assumptions on which any analysis of partially observed data must rest. We then review commonly used ad-hoc approaches for handling missing data, before going on to introduce and contrast two increasingly widely used methods for the analysis of partially observed data: multiple imputation and inverse probability weighting.

The aims of this module, and the associated practicals, are to:

- Give you an understanding of the issues raised by missing data, and how to explore them in a particular data set
- Introduce the key concepts of Missing Completely At Random, Missing At Random and Missing Not At Random, which describe the assumptions on which analysis of partially observed data rests
- Critically review commonly used ad-hoc methods for 'handling' missing data
- Introduce, and give an intuition for, two increasingly widely used methods for the analysis of partially observed data
- Illustrate the use of these methods using Stata and MLwiN/REALCOM.

At the end of this module, the objective is that you will be able to frame the analysis of partially observed data in terms of the key concepts above, apply multiple imputation in relatively straightforward settings, and have the basic tools to critically review the analysis of partially observed data by others. For more information, please refer to our website, www.missingdata.org.uk, where among other things you will find information on forthcoming courses.

By missing data, we mean data that we intended to collect on units, which will often be individuals, but for some reason were unable to collect. We stress that the underlying, unobserved value exists; we just did not observe it. Thus we are not concerned here with counterfactual data, such as what an individual's educational attainment would have been if they had attended a different school.

Throughout this module, we assume that in the absence of any missing data, we have in mind the statistical model we would fit to address our research question, in other words the response variable and the key covariates whose association with the response we wish to explore. Although this will often involve fitting one model, given the response and set of covariates we term the most general model the *model of interest*. If we can obtain valid inference for the model of interest, we can simplify it if appropriate.

As with all statistical models, the model of interest contains one or more covariates, whose regression parameters we wish to draw inference for. These typically describe the mutually adjusted associations between explanatory variables and the response variable.

When some values that we intended to collect are missing, our model of interest does not generally change. However, the presence of missing values makes fitting the model of interest to the full data as originally intended impossible; reducing to the subset of complete records will often result in bias. It is also inefficient to discard valuable (and costly to collect) partial information on the units with one or more missing values.

The ubiquity of missing data means there has been a vast amount of research into its theoretical and practical implications over the last 35 years. Consistent with the aims above, we seek to summarise (some of) this and present it in an accessible way for the typical quantitative researcher, who has not had advanced statistical training.

The plan for the remainder of this module is therefore as follows. We begin by introducing a set of educational data which we use repeatedly to illustrate the discussion. Then, 0 outlines the centrality of assumptions, and gives an intuitive introduction to the key concept of missingness mechanisms, and their implications for the analysis. In the light of this, we critically review commonly used ad-hoc methods, and complete records analysis, in 0 and 0. We then consider two more theoretically well grounded methods for the analysis of partially observed data, namely multiple imputation (0) and inverse probability weighting (0). In 0 we outline extensions to multilevel structures, before summarising the module and presenting a strategy for the analysis of partially observed data in 0.

Introduction to the Class Size Data

To illustrate our discussion, we use data from the *class size study* (Blatchford *et al* 2002), kindly made available by Peter Blatchford. We use a subset of the data, so that our analyses are merely illustrative; substantive conclusions should not be drawn.

Table 14.1 describes the variables in the dataset, which consists of data from 4,873 children before and during their reception (i.e. first full time year of primary education) year. We have full data on each child’s numeracy test score before the reception year (**nmatpre**) and literacy test score at the end of the reception year (**nlitpost**). These originally quantitative scores have been standardised to have mean 0 and variance 1. We also have a binary indicator of special educational needs (**sen**). The dataset also contains **nmatpre_m**, which is the same as **nmatpre** except that some values have been made artificially missing.

Table 14.1. Variables in the class size data

Variable	Description
nmatpre	Pre-reception maths score
nmatpre_m	Pre-reception maths score, with 2,493 values missing
nlitpost	Post-reception literacy score
Sen	Special educational needs (1=yes, 0=no)

C14.1 The Model of Interest

We assume throughout this module that our aim is to fit some model of interest, which will enable us to answer our scientific or research question. To motivate our discussion, we focus on the linear regression model for **nlitpost** with **nmatpre** and **sen** as covariates, or explanatory variables. This is our model of interest. Table 14.2 shows the estimated regression coefficients based on the full data (n = 4873).

Table 14.2. Estimated regression coefficients and standard errors based on fit to full data

Explanatory variable	**Coefficient (SE)**
constant	0.045 (0.012)
nmatpre	0.585 (0.012)
sen	-0.432 (0.043)

Based on the full data, there is strong evidence (the coefficient divided by its standard error is much larger than 1.96) that pre-reception maths score and whether the child has special educational needs are independently predictive of post-reception literacy score. Given special educational needs, each standardised unit increase in pre-reception maths score increases post-reception literacy by 0.6 of a standard deviation on average. For a given pre-reception maths score, special educational needs reduces post-reception literacy score by just under half a standard deviation on average.

We will use these estimates, based on the full data, as our 'gold standard'. We will perform analyses with **nmatpre_m**, and using various methods for handling missingness in the variable, compare the estimates with those in Table 14.2.

While this example is much simpler than any substantive analysis, it nevertheless is sufficient to illustrate the key concepts and methods.