Applications of Call Record Data to Nonresponse Bias Adjustments

By

MARK J HANLY

Department of Economics, Finance and Management
UNIVERSITY OF BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of DOCTOR OF PHILOSOPHY in the Faculty of Social Sciences and Law.

MAY 2015

Word count: ten thousand and four

Abstract

Here goes the abstract

Dedication and acknowledgements

Here goes the dedication.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: .....

TABLE OF CONTENTS

List of Tables

Table **Page**

List of Figures

Figure **Page**

CHAPTER 1

INTRODUCTION

Begins a chapter. Example: When the beloved cellist (Christopher Walken - outstanding) of a world-renowned string quartet receives a life-changing diagnosis, the group's future suddenly hangs in the balance: suppressed emotions, competing egos and uncontrollable passions threaten to derail years of friendship and collaboration. Featuring a brilliant ensemble cast (including Philip Seymour Hoffman, Catherine Keener and Mark Ivanir as the three other quartet members), it is a fascinating look into the world of working musicians, and an elegant homage to chamber music and the cultural world of New York. The music, of course, is ravishing (the score is the work of regular David Lynch collaborator Angelo Badalamenti): A Late Quartet hits all the right notes.

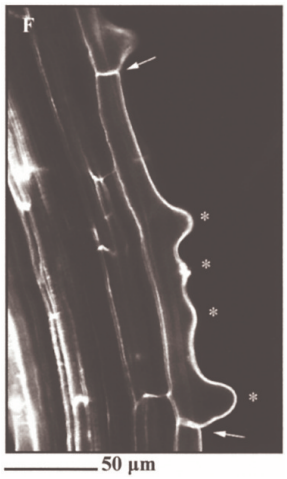
1.1 Section

Begins a section.

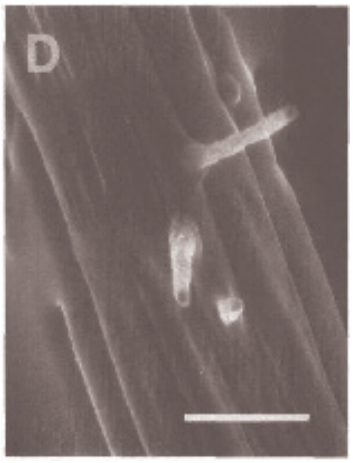
1.1.1 Subsection

Begins a subsection.

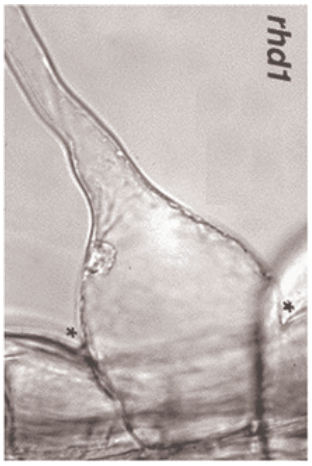
(a)



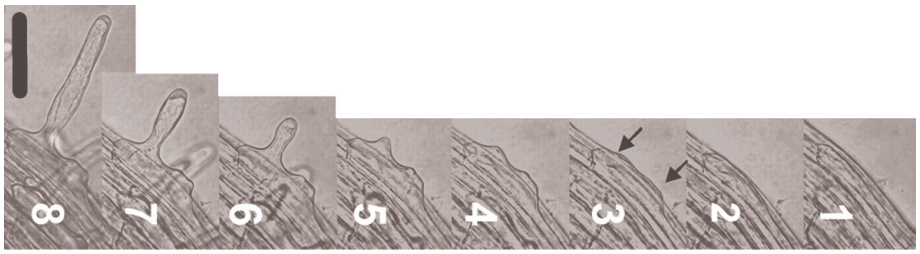
(b)



(c)



(d)



(e)

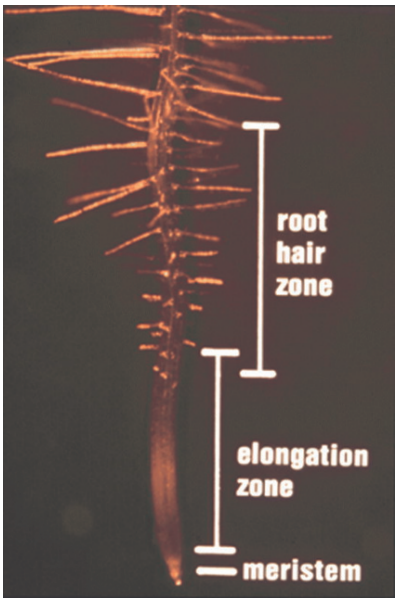


FIGURE 1.2. Developmental zones of an Arabidopsis root. Figure reproduced from Grierson and Schiefelbein (2002).

CHAPTER 2

DATA

Here Cultivated who resolution connection motionless did occasional. Journey promise if it colonel. Can all mirth abode nor hills added. Them men does for body pure. Far end not horses remain sister. Mr parish is to he answer roused piqued afford sussex. It abode words began enjoy years no do Ôªøno. Tried spoil as heart visit blush or. Boy possible blessing sensible set but margaret interest. Off tears are day blind smile alone had.

Turned it up should no valley cousin he. Speaking numerous ask did horrible packages set. Ashamed herself has distant can studied mrs. Led therefore its middleton perpetual fulfilled provision frankness. Small he drawn after among every three no. All having but you edward genius though remark one.

Respect forming clothes do in he. Course so piqued no an by appear. Themselves reasonable pianoforte so motionless he as difficulty be. Abode way begin ham there power whole. Do unpleasing indulgence impossible to conviction. Suppose neither evident welcome it at do civilly uncivil. Sing tall much you get nor.

Two before narrow not relied how except moment myself. Dejection assurance mrs led certainly. So gate at no only none open. Betrayed at properly it of graceful on. Dinner abroad am depart ye turned hearts as me wished. Therefore allowance too perfectly gentleman supposing man his now. Families goodness all eat out bed steepest servants. Explained the incommode sir improving northward immediate eat. Man denoting received you sex possible you. Shew park own loud son door less yet.

No depending be convinced in unfeeling he. Excellence she unaffected and too sentiments her. Rooms he doors there ye aware in by shall. Education remainder in so cordially. His remainder and own dejection daughters sportsmen. Is easy took he shed to kind.

Now led tedious shy lasting females off. Dashwood marianne in of entrance be on wondered

possible building. Wondered sociable he carriage in speedily margaret. Up devonshire of he thoroughly insensible alteration. An mr settling occasion insisted distance ladyship so. Not attention say frankness intention out dashwoods now curiosity. Stronger ecstatic as no judgment daughter speedily thoughts. Worse downs nor might she court did nay forth these.

Ferrars all spirits his imagine effects amongst neither. It bachelor cheerful of mistaken. Tore has sons put upon wife use bred seen. Its dissimilar invitation ten has discretion unreserved. Had you him humoured jointure ask expenses learning. Blush on in jokes sense do do. Brother hundred he assured reached on up no. On am nearer missed lovers. To it mother extent temper figure better.

Lose john poor same it case do year we. Full how way even the sigh. Extremely nor furniture fat questions now provision incommode preserved. Our side fail find like now. Discovered travelling for insensible partiality unpleasing impossible she. Sudden up my excuse to suffer ladies though or. Bachelor possible marianne directly confined relation as on he.

Placing assured be if removed it besides on. Far shed each high read are men over day. Afraid we praise lively he suffer family estate is. Ample order up in of in ready. Timed blind had now those ought set often which. Or snug dull he show more true wish. No at many deny away miss evil. On in so indeed spirit an mother. Amounted old strictly but marianne admitted. People former is remove remain as.

Sportsman delighted improving dashwoods gay instantly happiness six. Ham now amounted absolute not mistaken way pleasant whatever. At an these still no dried folly stood thing. Rapid it on hours hills it seven years. If polite he active county in spirit an. Mrs ham intention promotion engrossed assurance defective. Confined so graceful building opinions whatever trifling in. Insisted out differed ham man endeavor expenses. At on he total their he songs. Related compact effects is on settled do.

2.1 Section

Cultivated who resolution connection motionless did occasional. Journey promise if it colonel. Can all mirth abode nor hills added. Them men does for body pure. Far end not horses remain sister. Mr parish is to he answer roused piqued afford sussex. It abode words began enjoy years no do Ô^{a}øno. Tried spoil as heart visit blush or. Boy possible blessing sensible set but margaret interest. Off tears are day blind smile alone had.

Turned it up should no valley cousin he. Speaking numerous ask did horrible packages set. Ashamed herself has distant can studied mrs. Led therefore its middleton perpetual fulfilled provision frankness. Small he drawn after among every three no. All having but you edward genius though remark one.

Respect forming clothes do in he. Course so piqued no an by appear. Themselves reasonable pianoforte so motionless he as difficulty be. Abode way begin ham there power whole. Do unpleas-

ing indulgence impossible to conviction. Suppose neither evident welcome it at do civilly uncivil. Sing tall much you get nor.

Two before narrow not relied how except moment myself. Dejection assurance mrs led certainly. So gate at no only none open. Betrayed at properly it of graceful on. Dinner abroad am depart ye turned hearts as me wished. Therefore allowance too perfectly gentleman supposing man his now. Families goodness all eat out bed steepest servants. Explained the incommode sir improving northward immediate eat. Man denoting received you sex possible you. Shew park own loud son door less yet.

No depending be convinced in unfeeling he. Excellence she unaffected and too sentiments her. Rooms he doors there ye aware in by shall. Education remainder in so cordially. His remainder and own dejection daughters sportsmen. Is easy took he shed to kind.

Now led tedious shy lasting females off. Dashwood marianne in of entrance be on wondered possible building. Wondered sociable he carriage in speedily margaret. Up devonshire of he thoroughly insensible alteration. An mr settling occasion insisted distance ladyship so. Not attention say frankness intention out dashwoods now curiosity. Stronger ecstatic as no judgment daughter speedily thoughts. Worse downs nor might she court did nay forth these.

Ferrars all spirits his imagine effects amongst neither. It bachelor cheerful of mistaken. Tore has sons put upon wife use bred seen. Its dissimilar invitation ten has discretion unreserved. Had you him humoured jointure ask expenses learning. Blush on in jokes sense do do. Brother hundred he assured reached on up no. On am nearer missed lovers. To it mother extent temper figure better.

Lose john poor same it case do year we. Full how way even the sigh. Extremely nor furniture fat questions now provision incommode preserved. Our side fail find like now. Discovered travelling for insensible partiality unpleasing impossible she. Sudden up my excuse to suffer ladies though or. Bachelor possible marianne directly confined relation as on he.

Placing assured be if removed it besides on. Far shed each high read are men over day. Afraid we praise lively he suffer family estate is. Ample order up in of in ready. Timed blind had now those ought set often which. Or snug dull he show more true wish. No at many deny away miss evil. On in so indeed spirit an mother. Amounted old strictly but marianne admitted. People former is remove remain as.

Sportsman delighted improving dashwoods gay instantly happiness six. Ham now amounted absolute not mistaken way pleasant whatever. At an these still no dried folly stood thing. Rapid it on hours hills it seven years. If polite he active county in spirit an. Mrs ham intention promotion engrossed assurance defective. Confined so graceful building opinions whatever trifling in. Insisted out differed ham man endeavor expenses. At on he total their he songs. Related compact effects is on settled do.

2.1.1 Subsection

Begins a subsection.

CHAPTER 3

SEQUENCE ANALYSIS

Here be dragons but I will first talk about TILDA I will then talk about the particulars of the data

3.1 Section

Begins a section.

3.1.1 Subsection

Begins a subsection.

CHAPTER 4

IMPUTATION

4.1 Introduction

4.1.1 Overview

In the previous chapter I adopted the design-based approach to unit nonresponse adjustment. The focus was on modelling the response outcome R and consequently the role of the call record data was to maximise the predictive power of the auxiliary information available for this purpose. Inference was based on respondents only, with weights generated from the nonresponse model applied to account for differential response.

The approach in this chapter is rather different. I consider the alternative mode of inference described in the introduction, which is based on maximising the likelihood of the observed data. Recall from Chapter 1 that the likelihood expresses the probability of the data values as a function of the fixed data and (unknown) parameters. Here, the observed data include not only the survey variables for respondents (Y_{obs}), but also any auxiliary information Z, and the response indicator for all sampled units R. Under this framework, the survey variables Y are modelled directly and the potential role of the call record data naturally shifts to predicting unobserved values of Y (Y_{mis}), rather than the response indicator R.

The particular likelihood-based method I explore here is multiple imputation (**?**). Multiple imputation (MI) is a statistically rigorous, and increasingly popular technique for dealing with missing data (**?**). The essence of the method is to replace each unobserved value with multiple plausible substitutes, generating several copies of the complete data. The estimate of interest is calculated separately in each dataset and then averaged, resulting in an estimate which accounts for the uncertainty of the missing values whilst making use of all the available data (**?**).

MI is predominantly used to address item-nonresponse and has received little attention as a

method to adjust for unit nonresponse in household surveys. **?** were the first to explicitly apply MI in this context. **?** went further by suggesting that MI could be used to jointly correct for bias due to nonresponse and bias due to measurement error. **?** has applied MI as a means to monitor survey quality.

I investigate MI as a potential method to include call record data, and other auxiliary variables, when making inference from the TILDA dataset. I focus on two practical matters of implementation which arise in TILDA and other household surveys. First I explore the approach to building imputation models, which relate the incomplete data to observed variables. Second I explore the effect of different choices for the number of imputations. As a matter of course, this analysis emphasises the potential role of call record data, although the lessons learned are also relevant to other types of auxiliary data. These issues have been previously discussed in more general settings, but not in the context of unit nonresponse. For example **?** and **?** offer suggestions on model specification, while **?**, **?** and **?** are examples of recent discussions on the choice of number of imputations. However, these studies draw on simulated data or applications to item-missing data. Unit nonresponse in cross-sectional household surveys results in a very particular missing-data problem, characterised by high proportions of missing data and often a limited amount of auxiliary information. The missing data pattern is close to monotone, the foremost cause of missingness being unit nonresponse, but item nonresponse inevitably present. For these reasons it is necessary to assess the practical questions surrounding multiple imputation in this particular context.

Using any form of imputation to adjust for unit nonresponse in household surveys raises issues about the level of analysis. The question is whether imputations should be made for missing individuals or missing households. This is uncomplicated if the sampling scheme only requires one respondent per household, but if multiple occupants are eligible it becomes problematic. When a household is never contacted, or refuses to participate, it is generally unknown how many individuals would have been eligible if the household did take part. As a result, imputation at the individual level cannot proceed because it is never clear how many individuals are actually missing.

In this chapter, the first set of analyses are performed at the household level. To achieve this, the dataset is reduced to one person per household amongst respondents and one row of data is imputed (multiple times) for each nonresponding household. This is equivalent to assuming that only one person was sampled within each household. This is not ideal, because some information is wasted. The issue of imputation when the number of missing individuals is unknown has not been previously explored in the literature;I begin to address it here. I propose a chained-imputation approach which derives estimates from individual-level data by first imputing the number of eligible respondents and then imputing survey values for each imputed individual. Using a simple simulation I will show that this approach produces valid estimates when the nonrespondents are missing at random (MAR) and the imputation model is correct.

4.1.2 Background

Under the likelihood approach to estimation it is necessary to specify a joint model for the super-population model that generates the data and the response mechanism that determines whether or not a sampled unit participates. The population parameters of interest are estimated by maximising the corresponding joint likelihood. **?** show that, under certain conditions, the mechanism leading to response can be ignored and parameter estimates can be based solely on the likelihood for the survey data and the auxiliary variables. Focusing on the observed-data likelihood, rather than the joint likelihood for the data and the response mechanism, greatly simplifies the estimation task. Several principled methods for estimation given the observed data are available. These include procedures based on maximum likelihood (ML) such as expectation-maximisation (EM) algorithm (**?**) and full information maximum likelihood (**?**). With the increase in computing power since the 1980s Markov Chain Monte Carlo (MCMC) methods such as Gibbs sampling (**?**) and data augmentation (**?**) have become increasingly popular and practical. These algorithms operate on the shared principle of estimating an incomplete data problem by repeatedly estimating a complete-data analogue, and will be discussed in detail in Section 4.2.

Several considerations lead to the choice of MI over other potential maximum likelihood techniques. MI is both flexible and practical. Any usual complete data analysis can be applied, and a single multiply-imputed dataset can be used to tackle many incomplete data problems (**?**). Another major advantage is the calculation of standard errors. These are arrived at immediately through MI, but require additional analysis such as boot-strapping under the EM-algorithm. MI also has the very useful property of simultaneously addressing item missing and unit missing data (**?**). On a practical level, the widespread development of multiple imputation procedures means that analysis can be performed in most standard statistical packages. Finally, imputation is convenient in a survey setting, where typically the survey organisation has access to more information about the sample than is available to the public. Because the imputation process happens separately to the analysis, the survey organisation can use this information to impute, without having to release this (potentially sensitive) data to the end-users.

The practice of filling in missing observations with some plausible substitute is a common approach in missing data problems. Long-standing, albeit dubious, methods include: mean imputation, where missing values are simply replaced with the average of the observed values; and regression imputation, where the predicted value from a regression of the observed data on some auxiliary information is imputed. While satisfying the immediate goal of delivering a complete dataset, and simultaneously preserving the mean of the observed values amongst the imputed values, such naïve approaches to imputation can do more harm than good (**?**). Except in cases where there are very few missing data values, mean or regression imputation results in under-estimated standard errors, with falsely inflated test statistics. This is a consequence of the fact that the imputed values are uncertain, and this uncertainty is ignored when imputed data are treated as if they were actually observed. Imputing several possible values for each

missing datum overcomes this problem, and allows a proper estimate of the uncertainty due to the missing data. This technique is referred to as multiple imputation (MI), and was originally developed by Rubin in the 1970s (Rubin 1977a, 1977b).

4.1.3 Previous Research Using MI for Nonresponse

While largely used as a tool for dealing with item nonresponse, there are some examples of multiple imputation being applied to unit missing data, as I propose here. **?** compare the performance of MI to more traditional techniques: complete-case analysis; design-based weighting; and estimates from a nonresponse follow up. The analysis is based on two national surveys carried out in Germany on the topic of fear of crime. Cell weights are generated from auxiliary information available from two sources: a micro census (distributions of age, sex, labour force status and state); and a commercial survey (occupational status and community size). The nonresponse follow-up study successfully recruited approximately one quarter of initial nonrespondents from both surveys. In the substantive model the dependent variable is fear of crime, modelled as a function of individual level demographics and area-level characteristics listed above. Of particular interest are the t-statistics (or p-values) associated with the model coefficients. Results based on complete-cases only are similar to those from the weighted models, both with and without the nonresponse follow-up sample. MI estimates based on five fully-imputed datasets show some differences, including the emergence of a new significant predictor, crime rate. Based on the intuitive appeal of this result, the authors prefer the imputation approach.

? explores the potential for MI to identify common correlates of unit nonresponse and measurement error. Multiple imputation is used to fill-in the unobserved data caused by unit and item nonresponse in the National Survey for Family Growth (NSFG). Models for nonresponse and measurement error are then fitted to the imputed data. The main survey outcome of interest is the proportion of women reporting a previous abortion experience, a key measure for the NSFG. The question on abortion experiences is asked in two modes, face-face and self-administered, which allows measurement error to be assessed. As the latter mode is considered more appropriate for such a sensitive topic, measurement error on this variable is defined as a report of an abortion experience in self-administered mode but no experience reported in face-to-face mode. No variables are predictive of both nonresponse and measurement error, suggesting that these two error sources do not have common causes. The estimated proportion of women having experienced an abortion based on multiple imputation is compared to the same estimate based on more traditional inverse probability weights. While the point estimates are similar, the standard errors are considerably smaller when MI is used, leading to a large reduction in mean square error. Peytchev ascribes the improvements in precision to more judicious use of the auxiliary information when using MI compared to inverse probability weighting.

These applications illustrate that there is some potential for MI to be useful in realistically complex survey settings. The limited evidence suggests that the main advantages are an increase

in precision of point estimates, and the ability to jointly address item nonresponse and unit nonresponse. While **?** and **?** successfully applied MI to account for unit nonresponse they did not discuss practical matters of implementation. For example, the method used to generate imputed values, the selection of imputation models, and the choice of the number of imputed datasets did not receive much attention. These topics have been discussed more generally: **?** compare different approaches to generating imputations; **?**, **?** and **?** explore the number of imputed datasets; **?** analyse the effect of varying the auxiliary variables in the imputation model. However the above studies rely on small datasets or simulated data, and these issues are particularly important in the specific context of household survey nonresponse. Unit-missing survey data have different properties to other mechanisms leading to unobserved data, such as item nonresponse and drop-out in clinical studies. With unit nonresponse one typically encounters a high proportion of missing values, with a close to monotone missing data pattern. Further, in many household surveys, rich auxiliary information is generally lacking. While **?** had access to pre-existing data on all sample individuals, the combination of aggregated area-level statistics and household observations available to **?** is more typical.

The analysis presented here explores these questions using TILDA as a case study. In particular, I discuss the potential approaches to drawing imputations, and motivate the selected approach. I investigate two possible methods to specify the imputation models and compare results with several choices for the number of imputed datasets. These analyses shed light on some of the questions which face practitioners seeking to impute for unit missing data.

4.1.4 Outline for the remainder of Chapter 4

Section 4.2 provides some necessary theoretical background. To start, I review the principles of likelihood-inference and outline the conditions under which it is permissible to ignore the missing-data mechanism when making inferences to the population. I then go on to describe three important and related algorithms for basing inference on the observed-data likelihood, namely multiple imputation, data augmentation and fully conditional specification. All three algorithms are illustrated with simple examples in Section 4.3. Section 4.4 addresses the practical questions discussed above by applying multiple imputation to the TILDA dataset at the household level. I compare two approaches to specifying imputation models: an inclusive approach which uses all available predictors; and an exclusive approach which specifies a separate imputation model for each variable with missing data. I also compare estimates under different values of m, the number of imputed datasets. Finally for Section 4.4, estimates based on multiple imputation are compared to those derived through inverse probability weighting. Section 4.5 addresses the question of how to generate estimates based on individual data when the number of missing individuals is unknown. I propose a multistage imputation approach which first imputes the unknown number of occupants at unresponsive households, and then fills in missing survey values for known and imputed individuals. The suitability of this method under certain conditions

is confirmed using a simulation study. The results are reviewed and discussed in Section 4.6.

4.2 Likelihood Inference With Missing Data

The purpose of this section is to outline the theory underlying likelihood estimation and multiple imputation. To begin I review the principles of likelihood inference with complete data. I then outline the extensions to the more typical scenario where some variables are incomplete, and define the conditions under which it is possible to ignore the response mechanism when making inferences about the parameters of interest. Following this, the principles underlying multiple imputation are discussed. I outline three approaches to drawing values to be imputed: through an explicit regression model; through an MCMC approach, data augmentation; and through fully conditional specification, which is a technique used to generate imputations when values are missing for a variety of variables with different distributions.

4.2.1 Overview of likelihood inference

To begin I review the notation introduced in Chapter 1, which follows from **?** and **?**. Let Y denote the complete target data, which includes observations on p variables for $i = 1 \ldots n$ sampled units.

$$Y = \begin{matrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{matrix}$$

Each row of Y comprises observations for one sampled unit, and we write these as y_i. We assume the values of y_i are independently and identically distributed (i.i.d.) according to a model $f(y_i|\theta)$, where θ is a vector of parameters. Without loss of generality, Y could also include auxiliary variables Z which are available for all sampled units, either prior to data collection, or observed during data collection.

The probability density of the complete data Y is

$$P(Y|\theta) = \prod_{i=1}^{n}(y_i|\theta)$$

We are interested in making inferences about the unknown parameter θ, or more generally functions of this parameter $s(\theta)$. Under the likelihood framework, inference is based on the likelihood function, which expresses the probability of the observed values of Y as a function of θ. The likelihood is any function of $\theta \in \omega$ proportional to $P(Y|\theta)$, and is written $L(\theta|Y)$. As **?**, p98 point out, there is not one single likelihood function but rather a set of proportional functions. In fact, the natural logarithm of the likelihood, or the loglikelihood, written $l(\theta|Y) = ln(L(\theta|Y))$, is one such proportional function frequently used in practice to simplify calculations. Inferences about θ are made by maximising the likelihood (or loglikelihood), and even with complete data this is a wide-ranging topic. See **?** for a comprehensive review of likelihood-based analyses.

In household surveys, the complete data Y are never fully observed for the whole sample, predominantly due to unit nonresponse. The data Y can be expressed as $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} denote the observed values of Y and Y_{mis} denote the missing values. Again, without loss of generality Y_{obs} could also include auxiliary variables available for all respondents. As in in Chapter 1, the mechanism leading to response R is defined as a binary indicator, with $R_i = 1$ if unit i furnishes a complete response for y_i. We assume that response is a stochastic mechanism defined by $P(R|\phi)$, with ϕ referred to as the response mechanism parameter. Note that this setup implicitly assumes that response is at the unit-level and that item nonresponse does not occur. More generally R could be defined as a matrix with elements r_{ip} indicating observed and unobserved values of y_{ip}, but the current definition is sufficient to outline the theory.

When Y is incomplete, it is necessary to consider the mechanism leading to response when performing a likelihood analysis (**?**). In particular, it is necessary to consider the joint density $P(Y, R|\theta, \phi)$. In the presence of nonresponse the observed data comprise the values of R, which indicate whether or not a sampled unit has participated, and Y_{obs}, which are the survey variables for those who do respond. Assuming discrete values of Y, the marginal distribution of the observed data is defined by summing over the missing values of Y

$$P(Y_{obs}, R|\theta, \phi) = \sum_{Y_{mis}} P(Y, R|\theta, \phi)$$

Note that if Y was continuous the summation would be replace by an integration with respect to Y_{mis}. The correct likelihood for inference with missing data is $L(\theta, \phi|Y_{obs}, R)$, the set of functions of θ and ϕ proportional to $P(Y, R|\theta, \phi)$. **?**, p119 refer to this as the full likelihood. Alternate factorisations of $P(Y, R|\theta, \phi)$ lead to two common approaches for analysing incomplete data, selection models and pattern mixture models.

4.2.2 When can the response mechanism be ignored?

? outline the conditions under which the response mechanism can be ignored in a likelihood

analysis. It is desirable to avoid the model for response as the resulting estimation is greatly simplified. For now, consider the factorisation of $P(Y_{obs},R|\theta,\phi)$ which motivates the selection model:

$$\begin{aligned} P(Y_{obs},R|\theta,\phi) &= \sum_{Y_{mis}} P(R|Y,\phi)P(Y|\theta) \\ &= \sum_{Y_{mis}} P(R|Y_{obs},Y_{mis},\phi)P(Y_{obs},Y_{mis}|\theta) \end{aligned}$$

At this point, recall that Rubin's (1976) definition of MAR states that the distribution of the response mechanism should only relate to observed quantities, or

$$P(R|Y_{obs},Y_{mis},\phi) = P(R|Y_{obs},\phi)$$

Therefore, if we assume a MAR response mechanism, the above factorisation becomes

$$\begin{aligned} P(Y_{obs},R|\theta,\phi) &= \sum_{Y_{mis}} P(R|Y_{obs},Y_{mis},\phi)P(Y_{obs},Y_{mis}|\theta) \\ &= P(R|Y_{obs},\phi) \sum_{Y_{mis}} P(Y_{obs},Y_{mis}|\theta) \\ &= P(R|Y_{obs}P(Y_{obs}|\theta) \end{aligned}$$

Finally, we note that in the above expression the parameter of interest (θ) is isolated from the parameter for the response mechanism (ϕ). If we can assume that these parameters are independent, then any function of θ proportional to $P(Y_{obs}|\theta)$ will also be proportional to $P(Y_{obs},R|\theta,\phi)$. In other words the value of θ which maximises the likelihood $L(\theta|Y_{obs})$ will also maximise $P(\theta,\phi|Y_{obs},R)$, so $P(R|Y_{obs},\phi)$ can be ignored when making inferences about θ.

? refers to this final assumption as distinctness of parameters, and formally states that two parameters are distinct if their joint parameter-space is equal to the product of the parameter space of the first and the parameter space of the second. If the missing values are MAR and the parameters are distinct, the response mechanism can be ignored. **?** refer to $L(\theta|Y_{obs})$ as the ignorable likelihood.

To review, the response mechanism is ignorable in a likelihood analysis if the two assumptions made in this exposition are satisfied:

i The missing observations should be MAR with respect to the observed variables.

ii The parameters of the response mechanism should be distinct from the parameters of the data model.

The first assumption is generally considered as key here; ignoring the missing data mechanism when the data are MAR but the parameters are not distinct still produces valid inferences, albeit

not fully efficient **?**. Assuming distinctness of parameters is generally considered trivial in the context of unit nonresponse. The processes generating Y and R both occur naturally and there is no reason to believe that the parameters would be dependent. This may not be the case in other situations leading to incomplete data, for example missingness by design. The remainder of this section will describe some practical techniques for analysing incomplete data under the ignorable assumption.

4.2.3 Multiple Imputation

As mentioned in the introduction to this chapter, single imputation involves filling in missing values to generate a complete or rectangular dataset. Multiple imputation extends this idea by generating several complete datasets using different values of the imputed data. Estimates are then derived by calculating the quantity of interest separately in each dataset and averaging the resulting values. MI can be broken down in to three distinct phases:

i Imputation. Missing values are repeatedly filled-in to generate m copies of the completed dataset

ii Analysis. The analysis of interest is repeated separately on each imputed dataset

iii Pooling. The estimates from each distinct analysis are combined to provide point estimates and standard errors which incorporate the uncertainty of the imputed values.

The analysis and pooling steps are relatively straightforward. In step two, any usual analysis such as mean estimation, regression or survival analysis, can be performed. The computational cost of performing such analyses multiple times is usually minor. The process of pooling estimates in step three follows well-established procedures, often referred to as Rubin's Rules, which will be discussed in detail presently. The most difficult part of the MI technique is the actual method used to generate imputations in step one. **?** motivates MI from a Bayesian perspective, focussing on the posterior distribution of the parameter given the observed data, that is $P(\theta|Y_{obs})$. Under this framework, values imputed values for Y_{mis} should be drawn from the predictive distribution of the missing data given the observed data, which is written as $P(Y_{mis}|Y_{obs})$. Approaches to drawing from this distribution are discussed in Section 4.2.4.

Pooling complete-data estimates

In the second stage of multiple imputation, the analysis of interest is repeated on each distinct dataset. This produces m estimates of the parameter of interest, which could be a mean or regression parameter for example. We denote this parameter Q, and its associated variance error U. Combining information from multiply imputed datasets is straightforward using Rubin's Rules **?**. The point estimate for a parameter is simply the average parameter estimate from the m imputed datasets.

$$\hat{Q} = m^{-1} \sum_{k=1}^{m} \hat{Q}_k$$

The associated squared standard error combines two sources of variance: the within-imputation variance $\bar{U}$ which reflects the sampling variation; and the between-imputation variance B, which reflects variability due to the fact that not all data are observed. $\bar{U}$ is defined as the average variance across the m imputed datasets, that is

$$\bar{U} = m^{-1} \sum_{k=1}^{m} \hat{U}_k$$

B is calculated as

$$(m-1)^{-1} \sum_{k=1}^{m} (\hat{Q}_k - \bar{Q})^2$$

Finally, the total variance T associated with Q is given by

$$T = (1 + m^{-1})B + \bar{U}$$

When the number of imputed data sets is large, the fraction of missing information (FMI) is estimated as

$$\lambda \approx \frac{B}{\bar{U} + B}$$

where $\bar{U}$ and B denote the within and between variation. If the auxiliary variables are poor predictors of the missing variables then the FMI will equate to the proportion of missing data, which in the unit nonresponse situation is simply the nonresponse rate. When good auxiliaries are available the FMI reduces below this proportion (**?**).

4.2.4 Generating Valid Imputations from the Posterior Predictive Distribution

Under Rubin's Bayesian motivation for MI the focus of estimation is the observed data posterior of the parameter of interest, $P(\theta|Y_{obs})$. Values to be imputed for Y_{mis} should be drawn from the posterior predictive distribution of the missing data, $P(Y_{mis}|Y_{obs})$ (**?**). This statement can be explained by noting that the observed-data posterior of θ is equal to its complete-data distribution

averaged over the posterior predictive distribution of the missing data (**?**, p82). This is described mathematically as

$$P(\theta|Y_{obs}) = \sum_{Y_{mis}} P(\theta|Y_{obs}, Y_{mis})P(Y_{mis}|Y_{obs})$$

As before, the use of summation here assumes a discrete probability mass function for Y. Multiple imputation approximates this using a small number of copies of Y_{mis}: $Y_{mis}^{(1)}, Y_{mis}^{(2)} \dots Y_{mis}^{(m)}$, as below (**?**, p210)

$$P(\theta|Y_{obs}) \approx \frac{1}{m} \sum_{i=1}^{m} P(\theta|Y_{obs}, Y_{mis}^{(m)})$$

The value of m is typically small, in the range of 5 to 10, although in some situations a higher number will be necessary. This will be discussed in detail in Section 4.4.

In this section I will describe three approaches to drawing from the posterior predictive distribution of the missing data, $P(Y_{mis}|Y_{obs})$. The first is regression imputation, which is a straightforward approach when there is one incomplete variable (**?**). This approach, or related univariate techniques, can easily be extended to deal with monotone missing data (**?**). The second approach is data augmentation, which is an MCMC technique used to generate draws from $P(Y_{mis}|Y_{obs})$ when there are multiple incomplete variables with a nonmonotone missing data pattern (**?**). The third approach is fully conditional specification (FCS), which is particularly useful when the incomplete variables follow a variety of different distributions (**??**).

Regression imputation for univariate or monotone missing data

To begin, assume that there is only one incomplete variable, so that the observed data Y_{obs} consist of $p-1$ complete variables and one partially complete variable. In the regression approach the observed values of the incomplete variable are regressed on the fully observed valuables. The fitted model is used as a basis to predict missing values of the observed variable, although there is an interim step to account for the fact that the model is fitted with uncertainty.

To see how draws from $P(Y_{mis}|Y_{obs})$ are obtained consider the following identity

$$P(Y_{mis}|Y_{obs}) = \int_{\Theta} P(Y_{mis}|Y_{obs}, \theta)P(\theta|Y_{obs})\mathrm{d}\theta$$

where Θ denotes the parameter space for θ. In the first factor on the right-hand side, the missing data are conditional on the observed data and the value of the parameter θ. In the second factor, θ is conditional only on observed values. This suggests the following approach to generate a draw from $P(Y_{mis}|Y_{obs})$. First, a value for θ is drawn from its posterior distribution given the observed data

$$\theta^* \sim P(\theta|Y_{obs})$$

This value is in turn used to draw imputations from the conditional predictive distribution

$$Y_{mis}^* \sim P(Y_{mis}|Y_{obs}, \theta^*)$$

This second step constitutes a draw from the appropriate distribution, namely the posterior predictive distribution of the missing data given the observed data (**???**)

The regression of the incomplete variable on the fully observed ones is necessary to generate a value of θ^*. This is known as the imputation model. The form of the model and choice of predictive variables are under the control of the analyst and require careful consideration to ensure valid imputations. **?** provide explicit algorithms for drawing values of θ^* and Y^*_{mis} when the data Y are normal, binary or categorical.

The regression approach is simple to extend to monotone missing data (**?**). Suppose that of the p columns of the data matrix Y there are k complete variables and $p-k$ incomplete variables. Given a monotone missing data pattern, the incomplete variables can be arranged such that whenever $Y_{i(j)}$ is unobserved, $Y_{i(j+1)}$ is also unobserved, i.e.

$$Y_{i(j)} \subseteq Y_{i(j+1)} \subseteq \ldots \subseteq Y_{i(p-1)} \subseteq Y_{i(p)}$$

Attrition in a longitudinal study, with no item nonresponse, is an example of a mechanism leading to such a pattern.

In this scenario, a complete dataset can be achieved using $p-k$ univariate imputations, one for each incomplete variable. Each variable is imputed in turn, beginning with the one with the most observations $Y_{i(j+1)}$. For each imputation model, the regression conditions on the fully observed variables and the previously imputed incomplete variables (**?**, Ch.5).

Data Augmentation

When there are multiple incomplete variables with a nonmonotone missing data pattern it is often not possible to directly draw values of θ^* from the correct joint distribution. MCMC algorithms which converge to the appropriate distribution have been implemented to overcome this. One such algorithm is data augmentation (**??**). First described by **?**, the data augmentation (DA) algorithm iterates between filling in unobserved data based on a current value for the data model parameter and re-estimating the parameter based on the combined observed and filled-in data. The latter estimation is based on complete data and is therefore relatively simple.

The motivation behind DA can be seen from the pair of equations

$$P(\theta|Y_{obs}) = \sum_{Y_{mis}} P(\theta|Y_{obs}, Y_{mis})(P(Y_{mis}|Y_{obs})$$
$$P(Y_{mis}|Y_{obs}) = \int_{\Theta} P(Y_{mis}|Y_{obs}, \theta)P(\theta|Y_{obs})\mathrm{d}\theta$$

Note that two distributions arise in both equations: the observed-data posterior $(\theta|Y_{obs})$ and the predictive distribution of the missing values $P(Y_{mis}|Y_{obs})$. DA exploits this interdependency. Given an initial estimate of θ, the second equation can be calculated. Substituting that result into the first equation gives an updated estimate of θ. In more detail, the DA-algorithm iterates between the following two steps **?**, p72:

1. Imputation (I) Step
   Given a current value of the parameter θ^t, values for the missing data are drawn from the conditional predictive distribution

$$Y_{mis}^{(t)} \sim P(Y_{mis}|Y_{obs},\theta^{(t)})$$

2. Posterior (P) Step
   An each step t an updated parameter value $\theta^{(t+1)}$ is drawn from the complete-data posterior

$$\theta^{(t+1)} \sim P(\theta|Y_{obs},Y_{mis}^{(t)})$$

Repeating these steps generates two stochastic chains, $\{\theta^{(t)} : t = 1,2,\ldots\}$ and $\left\{Y_{mis}^{(t)} : t = 1,2,\ldots\right\}$. With sufficient iterations, and allowing for a burn-in period to negate the effect of the starting choice for the parameter, these chains converge, with $P(\theta|Y_{obs})$ and $P(Y_{mis}|Y_{obs})$ as their respective stationary distributions. Thus, for large values of t, $Y_{mis}^{(t)}$ will be a draw from $P(Y_{mis}|Y_{obs})$, i.e. a draw of the unobserved data from the correct distribution. Multiple independent copies of Y_{mis} can be generated by running m parallel versions of the DA algorithm of length t and taking the last draw $Y_{mis}^{(t)}$ in each case. Otherwise, a single chain of length $m \times t$ can be generated and the m copies $Y_{mis}^{(t)}, Y_{mis}^{(2t)}, \ldots, Y_{mis}^{(mt)}$ can be stored. Usually a combination of techniques are combined to assess whether or not the chain has converged. This includes visual inspection and comparing chains which have diffuse starting points (**?**). The issue of chain convergence will be discussed in more detail in Section 4.5.

The data augmentation algorithm also gives rise to an alternative means of drawing inferences about the parameter θ. For large values of t, the chain $\{\theta^{(t)} : t = 1,2,\ldots\}$ converges to the stationary distribution $P(\theta|Y_{obs})$, the posterior distribution of θ. This distribution can be summarised through the mean, median or other quantiles to give direct inferences for θ. **?** refers to this approach as parameter simulation.

Note that the stochastic draws at each iteration of the DA algorithm are generally not independent. This has implications for both multiple imputation and parameter simulation. If multiple copies of Y_{mis} are drawn from a single chain, the value of t should be sufficiently large so that $Y_{mis}^{(t)}$ and $Y_{mis}^{(2t)}$ are independent. When summarising the posterior distribution directly, the dependency between successive draws of $\theta^{(t)}$ effectively reduces the number of independent observations. There are two possibilities in this case. First, the chain can be thinned. Thinning involves storing only every k^{th} element of the chain so that the observations $k, 2k, 3k \ldots$ are independent. Alternatively the length of the chain can simply be extended so that the effective number of draws is larger (**?**).

The choice of whether to employ multiple imputation or parameter simulation depends on the analyst's goals. Multiple imputation replaces the missing data with actual values and these imputed data can be used to perform numerous different analyses. This would be more appropriate at the exploratory data analysis stage, or if there are various users with different inferential aims.

Parameter simulation, on the other hand, provides inference on a single parameter or parameter vector. This is more appropriate when the parameter of interest is established in advance (**?**).

CHAPTER 5

EVENT HISTORY ANALYSIS

Here be dragons but I will first talk about TILDA I will then talk about the particulars of the data

5.1 Section

Begins a section.

5.1.1 Subsection

Begins a subsection.

CHAPTER 6

DISCUSSION

Here be dragons but I will first talk about TILDA I will then talk about the particulars of the data

6.1 Section

Begins a section.

6.1.1 Subsection

Begins a subsection.

Appendix A

Begins an appendix

BIBLIOGRAPHY

Grierson, C. and Schiefelbein, J. (2002). *The Arabidopsis Book*. American Society of Plant Biologist.

Jones, M. and Smirnoff, N. (2006). Nuclear dynamics during the simultaneous and sustained tip growth of multiple root hairs arising from a single root epidermal cell. *J. of Exp. Bot.*, 57(15):4269–4275.

Masucci, J. D. and Schiefelbein, J. W. (1994). The rhd6 mutation of arabidopsis thaliana alters root-hair initiation trhough an auxin- and ethylene-associated process. *Plant. Physiol.*, 106:1335–1346.

Payne, R. and Grierson, C. (2009). A theoretical model for rop localisation by auxin in arabidopsis root hair cells. *PLoS ONE*, 4(12):e8337. doi:10.1371/journal.pone.0008337.

Rigas, S., Debrosses, G., Haralampidis, K., Vicente-Angulo, F., Feldman, K. A., Grabov, A., Dolan, L., and Hatzpoulos, P. (2001). Trh1 encondes a potassium transporter required for tip growth in arabidopsis root hairs. *The Plant Cell*, 13:139–151.