Practical Use *of* Multiple Imputation

Timothy Peter Morris

UCL

DISSERTATION SUBMITTED FOR THE DEGREE OF DOCTOR *of* PHILOSOPHY

I, Timothy P. Morris, confirm that the work presented in this thesis is my own.

Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Multiple imputation is a flexible technique for handling missing data that is widely used in medical research. Its properties are understood well for some simple settings but less so for the complex settings in which it is typically applied. The three research topics considered in thesis consider incomplete continuous covariates when the analysis model involves nonlinear functions of one or more of these.

Chapters 2–4 evaluate two imputation techniques known as *predictive mean matching* and *local residual draws*, which may protect against bias when the imputation model is misspecified. Following a review of the literature, I focus on how to match, the appropriate size of donor pool, and whether transformation can improve imputation. Neither method performs as well as hoped when the imputation model is misspecified but both can offer some protection against imputation model misspecification.

Chapter 5 investigates strategies for imputing the ratio of two variables. Various 'active' and 'passive' strategies are critiqued, applied to two datasets and compared in a simulation study. ('Active' indicates the ratio is imputed directly within a model; 'passive' means it is calculated externally to the imputation model.) Without prior transformation, passive imputation after imputing the numerator and denominator should be avoided, but other methods require less caution.

Chapter 6 proposes techniques for combining multiple imputation with (multivariable) fractional polynomial methods. A new technique for imputing dimension-one fractional polynomials is developed and nested in a chained-equations procedure. Two candidate methods for estimating exponents in the fractional polynomial model, using Wald statistics and log-likelihoods, are assessed via simulation. Finally, the type I error and power are compared for model selection procedures based on Wald and likelihood-ratio type tests. Both methods can out-perform complete cases analysis, with the Wald method marginally better than likelihood-ratio tests.

Structure of thesis

1 · Introduction

2 · Introduction to partially parametric imputation

3 · Univariate simulation studies assessing predictive mean matching and local residual draws

4 · Multivariable simulation study: imputation transformations, predictive mean matching and local residual draws

5 · Multiple imputation for an incomplete covariate that is a ratio

6 · Combining multivariable fractional polynomial models with multiple imputation

7 · Discussion

¶ · Appendices

Contents

List of Figures

List of Tables

Acknowledgements

I am grateful to many people and groups who have made the three years of my PhD such a pleasure.

Ian White, my primary supervisor, has been superb in his guidance and input. He has been thorough, insightful, engaged, patient and encouraging throughout. It would be impossible to overstate how good Ian is a supervisor.

Patrick Royston, my secondary supervisor, has been generous with his time and advice, and helped with various programming issues and keeping the big picture. *Shaun Seaman*, *Angela Wood* and *Jonathan Bartlett* have also provided valuable ideas and advice.

I thank the *trauma group*, the *Aurum institute* and the *Epic-Norfolk* team for allowing me to use their datasets as examples. *Epic-Norfolk* is supported by grants from the Medical Research Council and Cancer Research UK.

The MRC Clinical Trials Unit at UCL and the MRC Biostatistics Unit have been excellent environments to work towards a PhD, for which I thank *colleagues* at both places, particularly *Rachel Jinks*, who played the role of PhD big sister to me for almost three years.

Except for my *Dad*, my friends and family have next to no idea what I do, but are supportive nonetheless. I would like to acknowledge the support of *Richard & Maggie*, *Ann-Marie*, *John & Hannah*, and *P-man & Jules* and *Joel & Marysia*, and obviously my wife and favourite person Becky.

I thank God for His lavish provision, especially of the above people.

1 Introduction

1.1 MISSING DATA

Missing data are any values that were intended to be recorded in a study but, for one reason or other, were not.

Missing data are a pervasive problem in medical research. In clinical trials, we fail to fully follow up all patients. In observational studies, we fail to record all the covariate data needed and then fail to fully follow up all patients. Ignoring observations with some missing items is wasteful of the resources invested to collect the observed items. Bias and inefficiency are likely consequences. There is no single answer to dealing with the issues arising from missing data and we are left with incomplete datasets that are difficult to analyse as intended.

Multiple imputation is a popular and flexible technique for handling missing values in partially observed datasets. Missing data are imputed in a way that fully reflects the uncertainty about them. To obtain valid estimates of variance, imputation must be performed $M > 1$ times. Each of the imputed data sets must be analysed identically and the results combined using a simple and general set of rules known as *Rubin's rules*[1].

In any analysis with missing data it is important to consider the way/s in which items of data might have become missing. There are three important assumptions in thinking about the process by which data go missing[2]:
Missing not at random (MNAR)
The probability of data being missing depends on unobserved information, such as the value of the missing datum itself, or an unmeasured variable.
Missing at random (MAR)
The probability of data being missing does not depend on unobserved information, such as an incomplete or unmeasured variable.
Missing completely at random (MCAR)
The probability of data being missing does not depend on any observed or unobserved information. (MCAR is a special case of MAR.)

MCAR is a special case of MAR, and it is possible to distinguish between MAR and MCAR in partially observed datasets, for example by fitting a logistic regression model for an incomplete variable's missingness indicator on complete variables. It is not possible to determine whether data are MNAR in a partially observed dataset without access to information external to the data at hand, other than by making untestable modelling assumptions[3]. Any analysis with missing data relies on untestable assumptions.

In practice the likely mechanism by which data might go missing should be proposed by research workers, ideally those involved 'on the ground' in collecting the data, such as nurses who conduct baseline interviews. It is also critical that any analysis based on the posited assumption is supplemented by further analyses under alternative plausible assumptions about the missing data. This provides an assessment of the sensitivity of results to the assumptions about missingness.

1.2 MULTIPLE IMPUTATION

1.2.1 *Notation*

A full list of the notation used in this thesis is given at this point, which readers are advised to refer back to. The list is arranged in order of Greek letters (alphabetically), Roman letters (alphabetically), analphabetic characters and diacritics. Bold face denotes a vector or matrix.

- α Parameter/s of the imputation model
- β Parameter/s of the analysis model
- γ Parameter/s of a logistic model used to simulate MAR data
- δ Matching distance used PMM and LRD
- ε Residual error
- ρ A correlation
- σ Standard deviation
- a_1 Numerator of a ratio
- a_2 Denominator of a ratio
- $\boldsymbol{B}$ Between-imputation variance: the squared standard deviation of $\boldsymbol{\beta}_m$
- c Indexes the covariates in the analysis model
- D Selected dimension of FP
- $D_{\max}$ Maximum dimension of FP considered
- d Indexes the FP dimensions $1, \ldots, D_{\max}$
- E The expectation
- $\boldsymbol{f}(\boldsymbol{x})$ Normalising transformation of covariates
- $\boldsymbol{g}(\boldsymbol{x})$ Transformation of covariates used in the analysis model
- h Indexes individuals with observed x
- i Indexes all individuals with observed or missing data
- j Indexes individuals with missing x
- k Size of donor pool in PMM and LRD
- l Indexes replicates in simulation studies
- M Number of imputed datasets
- m Indexes the M imputed datasets
- N A normal distribution
- n Number of individuals in the dataset

n_h Number of individuals with observed data

n_j Number of individuals with missing data

p [chapters 2 and 5:] Number of covariates in analysis model

[chapter 6:] 'Exponents' or 'powers'; parameter for transformation of x

q Number of variables with missing data

R^2 Proportion of explained variation

s Total number of simulation replicates

S The set of exponents considered for fractional polynomial transformation

U A uniform distribution

$\text{Var}(-)$ The variance (of $-$)

$\text{Var}(\hat{\boldsymbol{\beta}})$ Total variance of $\hat{\boldsymbol{\beta}}$, calculated from $(1 + (1/M))\boldsymbol{B} + \boldsymbol{W}$

$\boldsymbol{W}_m$ Variance in the mth imputed dataset

$\boldsymbol{W}$ Within-imputation variance: the mean of the $\boldsymbol{W}_m$

$\boldsymbol{w}$ Fully observed covariate(s), if a distinction between incomplete and complete $\boldsymbol{x}$ is being made

$\boldsymbol{x}$ Covariate(s) in the analysis model; $\boldsymbol{x} = (\boldsymbol{w}, \boldsymbol{z})$

y Outcome of the analysis model, assumed to be complete

$\boldsymbol{z}$ Partially observed covariate(s), if a distinction between incomplete and complete is being made

$'$ Used to indicate 'not'; $\boldsymbol{x}_{c'}$ denotes the variables in $\boldsymbol{x}$ that are not x_c

$*$ A draw from some distribution. A draw of the imputation model parameters is denoted $\boldsymbol{\alpha}^*$; an imputed value of x_j is denoted x_j^*

$\hat{}$ Denotes an estimate of the parameter it sits above

$f(.)g$ Increments of size . between f and g; 0(1)10 means 'from 1 to 10 in increments of 1'.

1.2.2 *Imputation*

As introduced by Rubin[1], missing values should be imputed by draws from the posterior predictive distribution of a Bayesian model. Schafer defines imputation as *Bayesianly proper* if imputed values are 'independent realisations of the posterior predictive distribution under some complete-data model and prior'[4]. This means allowing for all sources of uncertainty implicit in the model used to impute missing values, the 'imputation model'.

An example of proper imputation in a simple setting is as follows[5, 6]. Let y and x denote two continuous variables, and in truth $(y, x) \sim$ BVN (meaning their joint distribution is bivariate normal). Assume that with complete data the analysis of interest would be a linear regression of y on x, but some values of x are missing (assumed to be MCAR or MAR conditional on y).

To impute missing values x_j, a linear regression for $(x_h \mid y_h)$ is fitted to the individuals with observed data using noninformative prior distributions, returning posterior estimates $\hat{\boldsymbol{\alpha}}$ with covariance matrix $\widehat{\text{Var}}(\hat{\boldsymbol{\alpha}})$ and root mean squared error $\hat{\sigma}$. Values of $\boldsymbol{\alpha}^*$ and σ^* are drawn

from their posterior distribution as follows. A draw of σ is taken as

$$\sigma^* = \hat{\sigma}\sqrt{\frac{n_h - b}{e^*}} \tag{1.1}$$

where e^* is a random draw from a χ^2 distribution on $(n_h - 2)$ degrees of freedom. Next, $\boldsymbol{\alpha}^*$ is drawn from

$$\boldsymbol{\alpha}^* = \hat{\boldsymbol{\alpha}} + \left(\frac{\sigma^*}{\hat{\sigma}}\right)\boldsymbol{u}_1 \boldsymbol{V}^{\frac{1}{2}} \tag{1.2}$$

where $\boldsymbol{u}_1$ is a vector of 2 independent draws from a standard normal distribution. Independent realisations of x_j^* for individuals with missing values of x are given by

$$x_j^* = \boldsymbol{\alpha}^* \boldsymbol{z}_j + u_{2j}\sigma^*, \tag{1.3}$$

where u_{2j} is a draw from a standard normal distribution.

The linear regression of x on y is the appropriate conditional model for x derived from the bivariate normal joint model. The analysis of interest can also be derived from the bivariate joint normal model.

1.2.3 *Multiple imputation inference: Rubin's rules*

Having created M imputations Rubin's rules are applied as follows. Each imputed datasets is analysed identically using whatever model would have been used in the absence of missing data (the 'analysis model'); here, a linear regression of y on x. Parameter estimates $\hat{\boldsymbol{\beta}}_m$ for all M imputed datasets are obtained with corresponding covariance matrices $\widehat{\boldsymbol{W}}_m$. The overall estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\boldsymbol{\beta}}_m. \tag{1.4}$$

Its variance is estimated by

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \left(1 + \frac{1}{M}\right)\widehat{\boldsymbol{B}} + \widehat{\boldsymbol{W}}, \tag{1.5}$$

$$\text{where } \widehat{\boldsymbol{B}} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\boldsymbol{\beta}}_m - \hat{\boldsymbol{\beta}})^2 \tag{1.6}$$

$$\text{and } \widehat{\boldsymbol{W}} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{W}}_m. \tag{1.7}$$

$\hat{\boldsymbol{B}}$ is designated the *between-imputation variance* and $\widehat{\boldsymbol{W}}$ the *within-imputation variance*.

Hypothesis tests and confidence intervals are based on the t distribution[1]

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \sim t_v, \tag{1.8}$$

$$\text{where } v = (M-1)\left(1 + \frac{\widehat{\boldsymbol{W}}}{\left(1 + \frac{1}{M}\right)\widehat{\boldsymbol{B}}}\right)^2. \tag{1.9}$$

These rules are very general and can be used to combine estimates of population parameters, but not statistics that are not estimators, such as the log-likelihood. See White, Royston and Wood for examples of quantities that can and cannot be combined using Rubin's rules[6].

Despite imputations being drawn from a Bayesian model, the combining rules can provide valid frequentist inference[1, 7], meaning estimators are asymptotically unbiased with variance estimation such that the coverage rate of confidence intervals is equal to its stated value. There are a number of subtleties to be aware of that can thwart this in practice. For example, the following are required:

- The imputation model uses noninformative priors.
- The imputation model is specified correctly.
- The imputation and analysis models are 'compatible' (implying that the analysis model is correctly specified).

1.2.4 Multivariate imputation

Section 1.2.2 above outlined an approach to imputing a single incomplete variable. With multiple incomplete variables, if the missing data pattern is monotone (that is, if ordering variables from least to most missing data yields a dataset where the cth variable only contains missing values if variable $c + 1$ is missing), variables can be imputed using sequential univariate models, going from the least to the most missing, and conditioning on the complete and previously imputed variables at each stage. If the pattern of missing data is not monotone, the ideal approach would be to impute from a joint model for the incomplete covariates. If the incomplete variables are all continuous, a multivariate normal model may be a sensible choice; if they are all binary, a log-linear model might be used. However, in many real scenarios identification of an appropriate joint model to impute from may be impractical.

Van Buuren, Boshuizen and Knook introduced an alternative approach to multivariate imputation that is more practical and intuitive[8]. With q incomplete variables, it would usually be necessary to specify a q-variate imputation model. Instead, the suggestion of van Buuren et al. was to specify q univariate models, one for each of the incomplete variables. Imputation of x_c is performed on each variable in turn, conditioning on the previously imputed values of all other incomplete variables, and continuing for a few 'cycles' until the imputation models are thought to be stable in some sense.

The method is most commonly termed to as 'multiple imputation by chained equations' (*mice*), but is variously referred to as 'fully conditional specification' (FCS)[9], 'sequential regression'[10] or 'switching regression'[11].

The method has been met with enthusiasm by applied researchers because of its flexibility and the scope to specify relatively simple univariate models as desired: ordinal logistic regressions for an ordered categorical variable; linear regression for a continuous variable, and so on.

When mice was introduced there were some concerns about the theoretical lack of equivalence to any multivariate model. For example, with two incomplete variables, one continuous and the other binary, imputed by linear and logistic regression respectively, the two conditional distributions from which mice draws are incompatible; in this scenario mice is not Bayesianly proper in Schafer's sense[4]. While these concerns are not irrelevant, they appear to cause negligible problems in practice[12], and so mice will be considered to be a method of proper imputation.

1.3 CONTEXTS

Missing data cause problems in a vast number of studies and this thesis necessarily restricts focus to settings in which applied methodological research is thought to be required. The contexts considered in this thesis are:

1. *Item missingness*. This means that all individuals have at least some observed data. This is the setting to which MI is most suited. In medical research item missingness is the type of missing data most commonly seen.
2. *Incomplete covariates*. This is a common problem to which multiple imputation can be well suited and is commonly used.
3. *Missing at random*. While MNAR is interesting and may be realistic, it complicates exploration of the main issues addressed in this thesis. MI usually comes with a guarantee that inference will be valid under MAR. This claim will be untrue for some of the problems considered in this research if standard imputation methods are used. However, see section 7.4.3 for a note on extensions to MNAR.

1.4 RESEARCH TOPICS

The specific research topics I consider are outlined briefly in the subsections below. All three are concerned with imputation, while the third also considers methods for building models in multiply imputed datasets.

1.4.1 Research topic 1: Partially parametric techniques for imputation (chapters 2–4)

It is well known that when the assumptions underlying a parametric imputation model are incorrect, multiple imputation inference can suffer. Two techniques known as *predictive mean matching* (PMM) and *local residual draws* (LRD) relax some of the parametric assumptions of imputation. These are reviewed and further investigated. Both methods involve identifying, for an individual with a missing value, the k closest matches from individuals with observed values, and 'borrowing' some information from one of these donors.

Chapter 2 introduces the methods and reviews their development in the multiple imputation literature, with a focus on the method for defining 'closest matches' and on how large the pool of potential donors should be. Chapter 3 aims to assess how matching should be done and how large the pool of donors should be in simulation studies with a single incomplete covariate. The best versions of PMM and of LRD are taken forward to chapter 4, which compares the performance of the favoured versions in a simulation study with multiple covariates, some complete and some incomplete, with a complex missing data pattern.

1.4.2 Research topic 2: Multiple imputation for an incomplete covariate which is a ratio (chapter 5)

The use of ratios is common in medical research; examples are BMI, waist–hip ratio, and the ratio of total-to-HDL cholesterol. When the value of a ratio is missing, this may be because the numerator, the denominator or both are missing. If both are not missing it seems sensible to make use of the value that was observed. One such strategy, termed 'passive imputation', is

to choose an imputation model that imputes the numerator and denominator separately, and calculate the ratio from the two imputed values externally to the imputation model. Another possibility is to impute the ratio 'actively', explicitly within the imputation model, and ignore the observed component if its counterpart is missing. These strategies for imputing ratio covariates, and some that lie between, are critiqued and evaluated. Although attention focuses on the use of a ratio covariate, the same issues arise when the outcome is a ratio and it is assumed that the same results would apply.

1.4.3 *Research topic 3: Combining multivariable fractional polynomials with multiple imputation (chapter 6)*

Fractional polynomials (FP) and multivariable fractional polynomials (MFP) are commonly used in prognostic research. The MFP model selection algorithm aims to select the best-fitting model for multiple continuous covariates from a restricted set of simple but flexible power transformations[13]. With complete datasets the algorithm proceeds on the basis of likelihood ratio tests. The kind of dataset in which MFP model selection is typically applied will have missing covariate problems, and so it is important to adapt the algorithm to work with MI data.

In combining MI with MFP methods it is important to be able to impute missing values in a way that acknowledges uncertainty about the final analysis model, which is unknown at the point of imputation. Two existing imputation methods are therefore adapted in an attempt to perform this task. The next difficulty arises from the fact that likelihood ratio tests will be invalid: MI data do not have a likelihood that can be used for formal inference and model selection cannot therefore proceed on the basis of likelihood ratio tests. While Wald tests have proved useful for variable selection problems[14], they cannot be used for testing the significance of, say, a fractional polynomial vs. linear term. Two new strategies, based on a modified likelihood ratio tests[14] and on the difference in Wald statistics, are evaluated in simulation studies.

1.5 DATASETS

Three real datasets are analysed in this thesis. The first, a dataset based on trauma registries, motivated some of the research of chapters 4 and 6.

1.5.1 *Dataset 1: Trauma registry*

The trauma registry data come from one to two years of trauma admissions (5,693 in total) from the registries of five large trauma centres: Amsterdam (n = 649), Cologne (n = 1,705), London (n = 788), Oslo (n = 2,167) and San Francisco (n = 384). The primary publication[15] of these data had two aims:

1. To explore the association between death and the number of red cell packs received; particularly whether there was any evidence behind the notion of a transfusion threshold ('massive transfusion') at which point there is a leap in the odds of death.
2. To develop a multivariable prognostic model for 'massive transfusion', defined as >10 red cell transfusions, to improve the planning and delivery of red cells from blood banks.

Table 1.1: Summary of variables in the trauma dataset relevant to this work, $n = 5,693$

Variable	Frequency missing (%)	Mean (SD) or frequency (%) in observed data
Massive transfusion	0 (0%)	518 (9%)
Age (years)	0 (0%)	40 (20)
Sex: male	0 (0%)	4,161 (73%)
Injury type: penetrating	23 (0.4%)	580 (10%)
Time to emergency dept. (mins)	2,396 (42%)	65 (40)
Systolic blood pressure (mmHg)	425 (7%)	126 (29)
Base deficit	868 (16%)	3.4 (5.1)
Prothrombin time	1,648 (29%)	17 (8)

The second aim is of interest in this project. The analysis model was a logistic regression of massive transfusion on covariates identified as important.

The planned analysis was complicated by missing data in the covariates considered as (potentially) prognostic for massive transfusion. A summary of the variables involved and the extent of missing data is given in table 1.1.

Missing values were multiply imputed under missing at random using mice. Fifty imputed datasets were created after 100 cycles of chained equations. Each variable being imputed used a univariate model that included all other variables in table 1.1 as imputation covariates, including the binary outcome. Injury type was imputed using logistic regression. Due to highly skewed distributions in some variables, all other incomplete continuous variables were initially transformed towards normality as far as possible using shifted-log transformations and then imputed using PMM[6]. What will be referred to as *type 1* matching was used to identify the *three* closest matches for each individual with a missing value, of which one observation was selected at random and imputed. (Details on PMM will be clarified in chapter 2.)

In the absence of missing data, the aim would have been to build a prognostic model for massive transfusion using multivariable fractional polynomials (MFPs). However, at the time of publication there was no methodological work on how to impute data for MFP models, or how MFP models should be built in multiply imputed data. Due to concerns about compatibility of the imputation and analysis models, the prognostic model used the same shifted-log transformations of continuous covariates that were used for imputation.

This dataset provided motivation for the work in chapters 4 and 6. Niggling questions remained about whether the use of MFP methods could have improved the performance of the prognostic model. It was obvious that imputation would need to be tailored to allow covariates to have nonlinear effects in the imputed data, and also that the standard likelihood ratio tests could not be used in imputed data.

The trauma dataset is explicitly used to inform a simulation in chapter 4 and later as an illustrative analysis in chapter 6.

Table 1.2: Aurum summary of variables; $n = 1,348$

Variable	Frequency missing (%)	Mean (SD) or frequency (%) in observed data	Categories used in [17]
Death	0	185 (14%)	
Age (years)	0	37 (9)	18–29, 30–34, 35–39
			30–49, 50–70
Sex: male	0	542 (40%)	N/A
Hæmoglobin (g/mL)	143 (11%)	11.4 (2.3)	<8, 8.1–9.9,
			10–11.9 (–12.9 for men)
			>11.9 (>12.9 for men)
*Viral load (copies per mL)	162 (12%)	4.8 (0.8)†	<4, 4–5, >5
*CD4 count (cells per μL)	94 (7%)	8.9 (4.5)†	0–49, 50–99
			100–200, >200
BMI (kg/m^2)	381 (28%)	21.9 (4.9)	<18.5, 18.5–25, >25
‡Weight (kg)	376 (28%)	58 (12)	
‡Height (m^2)	275 (20%)	2.7 (0.3)†	

*Transformation used for viral load is $\log_{10}$(VL); transformation used for CD4 count is $\sqrt{\text{CD4}}$. These are standard transformations in HIV research.
† Summarised on transformed scale.
‡ Only enters into the analysis model via BMI.

1.5.2 *Dataset 2: Aurum cohort*

[Note – some of the below information is repeated in chapter 5, which was published by *Statistics in Medicine* online in August 2013[16].]

The Aurum institute in South Africa specialises in research and health systems management, focusing primarily on the prevention, treatment and care of TB and HIV.

Beginning in 2005, the Aurum Institute conducted a cohort study, recruiting 1,350 HIV-infected individuals beginning antiretroviral therapy. Participants were recruited from 27 centres in five provinces between February 2005 and June 2006, and followed up to March 2007. Information was recorded on several baseline characteristics and participants were followed up for death (time to death is the primary outcome). Of the recruited participants, 1,348 had a recorded time of death/censoring. One hundred and eighty five deaths occurred within the follow-up period. (Analysis is restricted to these 1,348 individuals for this project.) Table 1.2 summarises the variables involved.

The work by Russell et al. aimed to identify risk factors for mortality using a Cox proportional hazards model[17]. The authors showed that of the risk factors considered, CD4 count and hæmoglobin appeared to be associated with mortality, with the report focusing on the hæmoglobin result, the more novel of the two[17]. The multivariable model reported as the primary analysis was based on the individuals with complete covariate data. The model was selected using a process of backwards elimination with the significance level for rejection set at

0.1. Continuous covariates were categorised according to the groups given in table 1.2.

The article explains briefly how missing BMI measurements were imputed 10 times, and that this had no appreciable impact on the adjusted hazard ratios for BMI (the confidence intervals based on multiple imputation are not given).

Note that the variable with the largest proportion of missing data is BMI, a ratio, and over 100 participants had observed height but missing BMI. This dataset is used for topic 2 (chapter 5).

1.5.3 *Dataset 3: Epic-Norfolk cohort*

[Note – some of the below information is repeated in chapter 5, which was published online in August 2013[16].]

The Epic (European Prospective Investigation of Cancer) study is a European study that investigates associations between dietary factors and cancer. One centre of this study has recruited over 30,000 participants living in Norfolk (details can be found in [18]). The data used from Epic-Norfolk come from a subset of 22,754 of these individuals. The outcome in this particular analysis is time to death, and there are several non-dietary baseline covariates, summarised in table 1.3.

Part of the motivation for topic 2 was the controversial publication of the initial *Q-risk* cardiovascular risk score in 2007, where the imputation of the ratio of total-to-HDL cholesterol went very wrong[19]. The Epic-Norfolk dataset is of particular interest because it contains the same ratio, which is also incomplete. The proportion of missing data is far smaller than in Q-risk but the pattern of missingness within total cholesterol and HDL is similar.

The effect of cholesterol ratio has previously been estimated from a Cox proportional hazards model[20]. The variables considered in that publication were not all available in the dataset used for this work, and so a suitable proportional hazards model was selected from the available variables. Rather than categorising cholesterol ratio for analysis, it is included as linear in our analysis model.

1.6 THEMES

Difficulties with multiple imputation tend to arise if the imputation model is misspecified. (While problems can also occur when the analysis model is misspecified, this problem is not specific to multiple imputation.) Two approaches to imputation model misspecification, based on transformations prior to imputation, appear in all three topics. The first is to transform continuous covariates towards normality; the second is to transform to achieve compatibility.

1.6.1 *Normality*

Standard imputation for a continuous covariate draws missing values from a normal distribution. If the incomplete variable in fact follows a lognormal distribution, the imputed values will look fairly different to the observed. An obvious solution is to find a transformation $\boldsymbol{f}(\boldsymbol{x})$ of x such that the observed values of $\boldsymbol{f}(\boldsymbol{x})$ are normally distributed. Imputation can then be done on

Table 1.3: Epic-Norfolk summary of variables; $n = 22,754$

Variable	Frequency missing (%)	Mean (SD) or frequency (%) in observed data
Death	0	830 (4%)
Age (years)	0	59 (9)
Sex: male	0	10,145 (45%)
Smoking status: ever smoked	0	11,971 (53%)
Systolic blood pressure (mm Hg)	52 (<1%)	135 (18)
Diastolic blood pressure (mm Hg)	52 (<1%)	82 (11)
Cholesterol ratio	2,155 (9%)	4.7 (1.6)
[†]Total cholesterol (mg/dl)	1,514 (7%)	6.2 (1.2)
[†]HDL (mg/dl)	2,155 (9%)	1.4 (0.4)

[†] Only enters into the analysis model via cholesterol ratio

this scale and the inverse transformation performed before fitting the analysis model. This of course requires $\boldsymbol{f}(\boldsymbol{x})$ to be an invertible function.

This approach is noted as an option in [6]. Eddings and Marchenko see comparison of the observed and imputed values as the appropriate approach to imputation diagnostics, noting that 'problems with the imputation model can be corrected before the imputed data are analysed'[21].

When imputed values do not closely resemble the observed data it is not necessarily a cause for concern; this is to be expected somewhat under departures from MCAR. Schafer finds that imputation via a multivariate normal distribution can be remarkably robust to incorrect distributional assumptions[4]. With this in mind, the rationale for an entirely different transformation is given below.

1.6.2 Compatibility and congeniality

The term *compatibility* is used throughout this thesis to describe the relationship between the imputation and analysis models. The working definition is *a joint model exists that implies both the imputation and analysis models as conditionals*.

To put it another way, the imputation model and analysis model can be thought of as being embedded in a single (hypothetical) model. This hypothetical model does not need to be known or fitted, but its existence is important: Rubin's combining rules assume that the imputation and analysis models are correctly specified, and compatibility is a necessary condition for this.

Problems may arise when the models are incompatible, though it has been shown more than once that certain incompatible imputation models can improve inference[22, 23, 24]. (In these examples compatibility or incompatibility is related to parameter restrictions rather than covariate transformations.) Considering these examples, I define two departures from compatibility:

Semi-compatibility
There is a special case of the imputation model that is compatible with the analysis model.
Incompatibility
There is no special case of the imputation model that is compatible the analysis model.

Liu et al. came up with these definitions separately[25], and my designations follow theirs. Incompatibility implies that the imputation and analysis models cannot together represent any model from which data might have been generated. Semi-compatibility implies that the imputation model might represent a realistic data generating model, while the analysis model places restrictions on certain parameters of this model; the imputation model captures all the features in the analysis model, plus some more.

The concept of *congeniality*, introduced by Meng[22], is closely related to compatibility. Meng's formal definitions consider a Bayesian imputation model, a frequentist analysis procedure, and a Bayesian full probability model[22]. 'Congeniality' holds if:

1. The posterior predictive distribution for missing values is identical for the imputation model and the Bayesian model.

2. The posterior mean and variance for a parameter of interest from the Bayesian model are asymptotically equivalent to the estimate and variance from the frequentist analysis procedure (given complete or incomplete data).

Congeniality and compatibility are closely related concepts. Meng notes that *uncongenial* 'essentially means that the analysis procedure does not correspond to the imputation model'. Meng's Bayesian model is a hypothetical joint model in which the imputation and analysis models can be considered as embedded or not. However, Meng's definitions require that the researcher's incomplete and complete data analysis procedures be specified. Bartlett et al. also chose the term compatible rather than congenial, and state their rationale for this choice[26].

The above note highlights that there is difficulty in defining these terms and how they relate to each other. My preference is for the term compatibility, but readers who prefer to can read 'congenial' in place of 'compatible' in this thesis.

1.7 AIMS AND PRINCIPLES

1.7.1 Software

This project aims to provide applied statisticians with better approaches to dealing with MI when covariates have a nonlinear effect on outcome. It is important that methods suggested are practical to statisticians working under time pressures. Dembe et al. note that Stata and SAS are by far the most commonly used software in health services research[27]. For analysis with missing data, R also deserves a mention for its well-developed routines.

From a software perspective we regard Stata's `ice` and `mim` commands and the `mi` suite as 'practical'; all are in common use in medical research and are used for most of this thesis.

Alternative software, notably WinBUGS and MLwiN & Realcom, may in theory have the flexibility to provide superior results, but it is assumed that the majority of researchers use more

general purpose software the majority of the time, and that such users are unlikely to change software simply to deal with missing data. In any applied problem missing data are unlikely to be the only difficulty; if for example the analysis model involves multivariable fractional polynomials then Stata, R or SAS would be more feasible than WinBUGS or MLwiN & Realcom.

1.7.2 *The identities of 'the imputer' and 'the analyst'*

Rubin developed MI with the specific context of public use databases in mind[23]. Two distinct entities, working independently, are assumed to be involved: an imputer and an end-user. Rubin intended the burden of dealing with missing data to fall on the imputer, who is responsible for making realistic assumptions about the missing data and imputing values accordingly. The imputation model must be rich enough to ensure that it is not incompatible with the analysis model: it should allow for any reasonable model the analyst may be able to dream up. The analyst, who is assumed to have limited knowledge of missing data, is only required to fit his/her analysis model to each of the completed-and-released datasets and to combine parameter estimates using Rubin's rules (section 1.2.3)[23]; a fairly simple task.

The scenario Rubin invokes poses many difficult problems for the imputer but makes life simple for the analyst. While the approach is undoubtedly helpful for users of publicly available data, this is a more complex scenario than is usually seen in medical research, where the entities of 'imputer' and 'analyst' are likely to be the same person, or possibly two people working closely. It is assumed that one is not blind to what the other is doing, and the imputation model can be tailored to the analysis and further augmented if desired.

1.7.3 *Estimation*

In general, I focus on estimation. In assessing methods our key concerns about their frequentist properties are:

1. Consistency. As $n \to \infty$, $\hat{\beta} \to \beta$.
2. Coverage. A $(1-\alpha)$% confidence interval should have the property that it advertises, that $(1-\alpha)$% of intervals constructed identically from repeated samples will contain β. This is a function of consistency and of the discrepancy between the estimated and true variance of $\hat{\beta}$. Coverage that is greater than its stated value is sometimes seen in multiple imputation work. This is not necessarily a problem, and the property is termed 'confidence validity'[23]. Coverage levels lower than the stated value are more of a cause for concern.
3. Efficiency. Assuming two competing methods are unbiased and have nominal (or greater than nominal) coverage, the method yielding the shortest confidence intervals is preferable. This method is the most precise and thus powerful.

Note that some of the simulation work in chapter 6 departs from the focus on estimation, concentrating rather on the type I error rate and power of model selection procedures.

It is often suggested that after producing imputations, researchers inspect the imputed values and compare them to observed data; see for example White, Royston and Wood[6], or Eddings and Marchenko[21].

An important point about MI is that its aim is not to recreate the missing values but to provide a way of estimating the parameter/s of interest in a way that fully allows for missing

data uncertainty. This should be the main consideration in evaluating methods. As Rubin emphasises[23]:

> *Judging the quality of missing data procedures by their ability to recreate the individual missing values* (*according to hit rate, mean square error, etc.*) *does not lead to choosing procedures that result in valid inference, which is our objective.*

Referring to this quote is not intended to disparage the practice of inspecting imputed values. Doing so will be useful for flagging poor imputation models, but alone it is insufficient to demonstrate the adequacy of an imputation model. For example, a variable may be imputed with no covariates in the imputation model. The marginal distribution of imputed values may match the observed, but associations with other variables will be biased towards the null in imputed data.

2 *Introduction to partially parametric imputation*

2.1 FULLY PARAMETRIC IMPUTATION

For a partially observed normally distributed continuous covariate x the standard imputation method described by White, Royston and Wood[6] is to fit a normal errors model for observed x_h on covariates $\boldsymbol{w}_h$ and impute from this model. The method for drawing x_j^* has been described already in section 1.2.2.

This relies on distributional assumptions which can be problematic if incorrect. With complete x, specification of its probability distribution would not be required by the analysis model. However, when x is incomplete some specification is necessary; incorrect assumptions may introduce more problems than they solve when attempting to account for missing data.

The regression of $x_h|y_h$ for the imputation model should correctly specify the mean and variance structures. If the analysis model were $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, then there would be no obvious model for $x_j|y$ which respects (i) the shape of the association between y and x assumed by the analysis model, and (ii) the fact that x^2 is a deterministic function of x. Further, the regression of $x_h|y$ in the imputation model is assumed to correctly specify the variance structure, assuming by default homoscedasticity and $\varepsilon_i \sim \mathrm{N}(0, \sigma_\alpha^2)$.

For missing continuous x, an imputed item x_j^* drawn from the posterior predictive distribution could take any value if the distribution of x is assumed to be normal. Observable measurements of x may be be constrained to take certain values, for example $x > 0$, but posterior draws carries the risk of imputing unobservable values.

Some of the above problems can be overcome: if x is constrained to be greater than zero, $\log(x_j)$ can be imputed. 'Passive' imputation of $x_j^* = \exp[\log(x_j^*)]$ would then guarantee imputed values of x_j^* are greater than zero. However, this will not overcome all problems – in particular, the imputation model now assumes the $\log(x)$–y relationship to be linear, which is incompatible with an analysis model that assumes the y–x relationship is linear.

2.2 PARTIALLY PARAMETRIC IMPUTATION

I now consider three less parametric methods of imputation based on matching[28]: hot deck, predictive mean matching (PMM) and local residual draws (LRD), with focus on the imputation of a single incomplete covariate from multiple complete covariates. These methods usually impute values with greater face validity, but do not have the same theoretical basis as posterior draws and so application of Rubin's rules does not guarantee valid inference.

For the remainder of this chapter $\boldsymbol{w}$ is assumed to contain y.

2.2.1 *Hot deck*

Hot deck is a non-parametric method of imputation[29], traditionally used in survey research. Strata are defined from p complete variables $\boldsymbol{w}_i$, after continuous $\boldsymbol{w}_i$ have been categorised. For all j, an individual h in the same stratum is selected at random and x_j^* imputed as x_h.

2.2.2 *Predictive mean matching*

PMM mimics the first step of posterior draws by fitting an imputation model for $(x_h|\boldsymbol{w}_h, y)$ to estimate the mean for missing values. Missing observations are matched to $k \geq 1$ potential donors with observed values of x with the closest predictive mean, where the difference in predictive means is denoted δ_{hj} (some different ways to calculate δ_{hj} are described in section 2.4). One of these k individuals is randomly selected and its value imputed.

2.2.3 *Local residual draws*

As with PMM, missing observations are matched to $k \geq 1$ potential donors x_h who have the closest predicted mean δ_{hj}. LRD then imputes by adding an empirical *residual*, selected at random from the donor pool, to a draw from the predictive mean[28]. That is x_j^* for individual j with match h is imputed as

$$x_j^* = \boldsymbol{\alpha w}_j + (x_h - \boldsymbol{\alpha w}_h). \tag{2.1}$$

(See section 2.4 for definitions of $\boldsymbol{\alpha w}_h$ and $\boldsymbol{\alpha w}_j$).

2.3 CHOICE OF k IN PMM AND LRD

It is not obvious what values of k are sensible choices for imputation by PMM and LRD, but it is clear that some values are not sensible; useful choices will be a trade off between the obvious problems arising with very large or very small k. The reasoning for why these extremes will be problematic follows.

For PMM, $k = n_h$ effectively ignores the imputation model and implies that any x_h is as likely as any other to be close to the missing value, forcing x_j^*–$\boldsymbol{w}_j$ associations towards the null in the imputed data. Meanwhile, $k = 1$ imputes identical values (or at least very similar values; see section 2.4 below) for each of the m imputations, except when there are ties in δ_{hj} (see note in section 2.5). This can produce imputed datasets with $\hat{\boldsymbol{B}} = 0$ (for type 0 and type 2 matching; see section 2.4). This will lead to underestimation of standard errors, confidence intervals that are too short (i.e. coverage is less than nominal) and hypothesis tests that overstate statistical significance. In combining estimates, all $\hat{\boldsymbol{\beta}}_m$ will be equal, and $\hat{\boldsymbol{\beta}}$ will be too variable because it is the mean of one observation rather than M observations.

For LRD, $k = n_h$ might not be a poor choice and should in fact be comparable to posterior draws. However, this is of no real interest: any problems assumed for posterior draws will also exist for LRD. A theoretical basis exists for using Rubin's rules after imputation by posterior draws, but does not for LRD (this is equally true for PMM and hot deck). While there is equally

no guarantee of Rubin's rules working for $k < n_h$, there are other perceived advantages. For LRD, $k = 1$ is likely to suffer from the problems outlined above for PMM.

Between the extremes above lie the default values of $k = 3$ used by the `ice` command in Stata[30] and $k = 5$ used by Stata's `mi impute pmm` and the `mice` library in R[31] (the default in `mice` was $k = 3$ until 2014).

In preference to a fixed value of k, some authors have identified matches based on the match quality by defining donors as individuals for whom $\delta_{hj} < \delta_{\max}$[28, 32]. The value of $\delta_{\max}$ then needs to be specified for each variable to be imputed in any multiple imputation analysis, which is no small task. As with hot deck imputation, it is possible that some j with no h where $\delta_{hj} < \delta_{\max}$ can be identified and so $\delta_{\max}$ would then need to be increased, compromising the matching quality for other j. Setting k to a fixed value thus has real practical advantages, while it can still be altered from problem to problem and so remains reasonably flexible.

2.4 MATCHING METRIC IN PMM AND LRD

Matching is defined by a scalar prediction of the missing value for each individual j (usually the linear predictor where there is more than one independent variable in the imputation model). The method used to define a distance δ_{hj} has been inconsistent in the literature. While there has been no explicit *dis*agreement, authors have defined δ_{hj} in at least three ways, all of which were introduced by Little (in references [33] and [34]). The donor for individual j is chosen from the k individuals h with the smallest values of:

$$\text{Type 0} \qquad \delta_{hj} = |\hat{\boldsymbol{\alpha}}\boldsymbol{w}_j - \hat{\boldsymbol{\alpha}}\boldsymbol{w}_h| \tag{2.2}$$

$$\text{Type 1} \qquad \delta_{hj} = |\boldsymbol{\alpha}^*\boldsymbol{w}_j - \hat{\boldsymbol{\alpha}}\boldsymbol{w}_h| \tag{2.3}$$

$$\text{Type 2} \qquad \delta_{hj} = |\boldsymbol{\alpha}^*\boldsymbol{w}_j - \boldsymbol{\alpha}^*\boldsymbol{w}_h| \tag{2.4}$$

for all j, where $(j = 1, \ldots, n_j)$ are missing and $(h = n_j + 1, \ldots, n)$ are observed. (Note that this designation corresponds to the order in which they were introduced and to the number of * symbols appearing in the calculation; these names are used by the `ice` command in Stata and `aregimpute` in the R package `hmisc`.) δ_{hj} can be considered to be a measure of matching quality, where smaller values indicate a better match.

In the algorithm for defining the match pool it is possible to have tied δ_{hj} when selecting the donor pool. If the kth and $(k+1)$th closest matches (or more on either side) are tied this causes problems. Allowing k to be increased would reduce the expected match quality for individual j. Instead, one or more of the ties for the kth closest match are randomly selected until the donor pool equals k.

Type 0 matching was used by Rubin[35], David et al.[36] and Little[33]. Type 2 has been in common use in the literature, for example by Heitjan and Little[34] and Schenker and Taylor[28]. It was the default value used by R's `mice` package until 2010. Type 1 was initially proposed by Little[33] but disappeared from the literature until the thesis of Meinfelder in 2009[37] and a tutorial paper by White, Royston and Wood in 2011[6]. Type 1 matching has always been the default metric used by Stata's `ice` command for multiple imputation by chained equations[11]. Since 2010 it has also been the implementation in the R package `mice` and is the default for the `aregimpute` function of the R package `hmisc`.

For LRD the use of $\boldsymbol{\alpha}^*$ and/or $\hat{\boldsymbol{\alpha}}$ for h and j in equation (2.1) need not be the same as the chosen matching metric. A local residual analogous to type 2 matching would be defined as

$$x_j^* = \boldsymbol{\alpha}^* \boldsymbol{w}_j + (x_h - \boldsymbol{\alpha}^* \boldsymbol{w}_h). \tag{2.5}$$

However, it is not obvious that this is a sensible choice. Both PMM and LRD should aim to draw from a distribution centred at $\boldsymbol{\alpha}^* \boldsymbol{w}_j$, but the residuals $(x_h - \boldsymbol{\alpha}^* \boldsymbol{w}_h)$ do not have zero mean. Rather, the draws of x_j^* obtained by (2.5) are centred at $\hat{\boldsymbol{\alpha}} \boldsymbol{w}_j$. This problem can be solved by replacing the residual in (2.5) with $(x_h - \hat{\boldsymbol{\alpha}} \boldsymbol{w}_h)$, regardless of the type of matching. Both Schenker and Taylor[28], and Barnes, Lindborg and Seaman[38] use (2.5), without justification.

This type 1 *residual* is used throughout simulation studies presented in chapters 3 and 4. Simulations using of a type 2 residual for LRD were also run and in fact give very similar results, but the type 1 residual has a firmer basis.

When the variance of $\hat{\boldsymbol{\alpha}}$ is low, the three matching metrics will tend to select similar donor pools (that is, the donor pools are likely to overlap). Conversely, when the variance of $\hat{\boldsymbol{\alpha}}$ is high, donor pools will tend to be different.

This point is demonstrated in figure 2.1, where $\boldsymbol{\alpha}^* \boldsymbol{w}$ is plotted against $\hat{\boldsymbol{\alpha}} \boldsymbol{w}$ in a single simulated dataset. The selection of the two closest matches for type 0 (orange), 1 (purple) and 2 (blue) matching are shown. For each solid line, there is no other observation for which a parallel line using a different donor would be closer to the dashed line, according to that metric. The contrast in this example is stark: the three metrics select donors pools with *no* shared observations, and for each metric the donors selected by other metrics are poor matches.

To explore behaviour of PMM and LRD when the imputation model is poorly specified, I take an extreme example of imputation model misspecification (type 0 matching is not considered here because it is identical to type 2 in this scenario). Figure 2.2 plots $M = 30$ imputations of x for two individuals with missing values using posterior draws, PMM, and LRD, where y_h is linear in x_h^2 and the imputation model uses a linear regression of x_h on y_h.

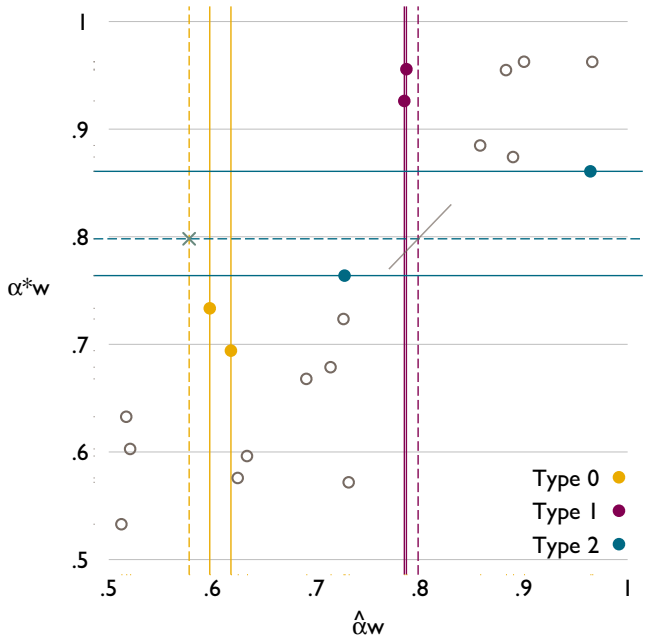
Posterior draws does an appalling job of preserving the shape of the association in imputed values. Both PMM and LRD manage to impute from a bimodal distribution, which is the correct imputation distribution, and preserve the bivariate relationship to some extent (with type 2 appearing to do so better than type 1). The example is very artificial in having such a strong association between x and y, but it illustrates the type of situation where PMM and LRD may be useful alternatives to fully parametric imputation. It remains to be seen whether they will be useful in less extreme cases.

2.5 Some situations where hot deck, PMM and LRD may break down

Below, I describe and contrast some of the advantages and issues with using posterior draws, hot deck, PMM and LRD.

Imputation using posterior draws can easily handle many covariates in the imputation model. When the imputation and analysis models are correctly specified Rubin's rules provides consistent parameter and variance estimation. However, posterior draws can impute unobservable values, and when the imputation model is misspecified the Rubin's rules estimator will fail to some degree.

Figure 2.1: Types 0, 1 and 2 matching donor pools with $k = 2$. The × represents the missing observation; all other points are potential donors. The two closest matches are shown type 0 in orange, type 1 in purple and type 2 in blue.

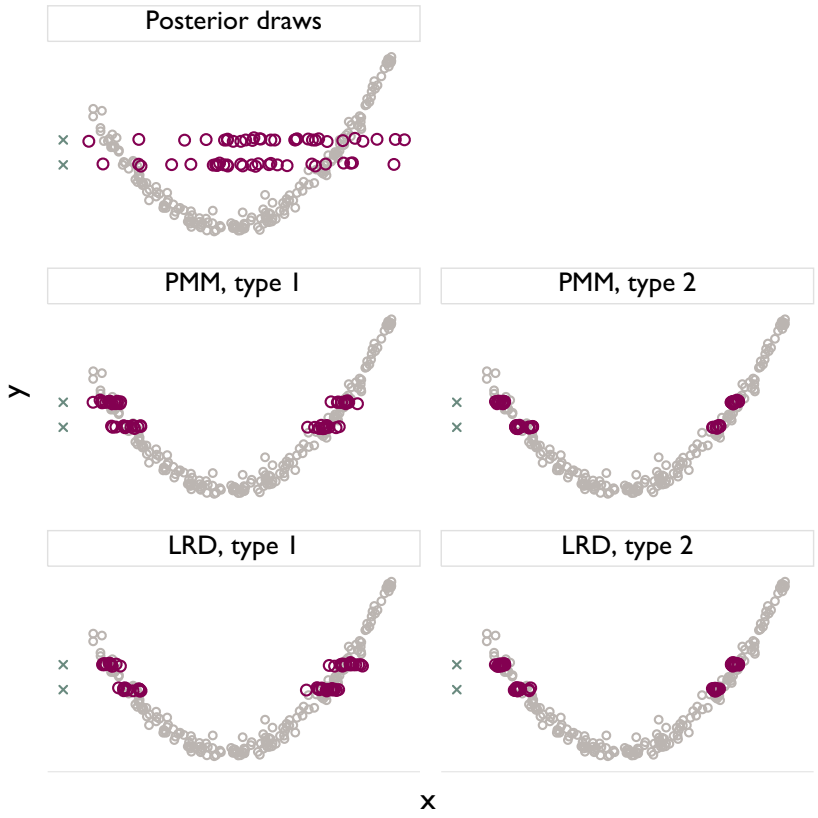


In hot deck imputation x_j^* are imputed from individuals with similar values of the imputation model covariates $\boldsymbol{w}$. (Technically there is no explicit 'model' for imputation, though David et al. note the hot-deck approach as being analogous to fitting a 'fully interactive' model to define a predictive mean before adding an empirical residual[36].) Hot deck solves some of the problems of posterior draws: an observable value is always imputed and the mean and variance of the imputed data should be similar for observed and imputed data under MAR.

Hot deck imputation not is without its own problems: it is not *proper* according to Schafer's definition (see 1.2.2)[4], meaning the Rubin's rules variance estimator may be invalid; continuous variables must be categorised before strata can be defined; once defined, strata will differ in the number of observed and missing values; strata may, in extreme cases, contain all missing values and no observed values, particularly under departures from MCAR; variables used to define strata may themselves be partly missing (though this issue might be solved by using a mice-type approach); and the number of variables that can be used to define strata (effectively imputation model covariates, despite the lack of a parametric model) is restricted because strata quickly become sparse.

PMM retains some advantages of hot deck, such as imputing observable values, but some key disadvantages are absent. The imputation model is used by PMM only to identify the k potential donors, and is not involved at any further point in producing imputations. However, it is not clear whether parameter and variance estimation based on Rubin's rules will work. Figure 2.2 leads us to expect that under imputation model misspecification PMM may reduce

Figure 2.2: Imputation of missing x_j, where $x \sim \mathrm{U}(-4.5, 5.5)$ and the true model is $y_i \mid x_i \sim N(x_i^2, 1)$. Grey dots are observed values; ×'s in the margins show the observed y_j values where x_j are missing; purple dots show $M = 30$ imputations of x^* for each j.



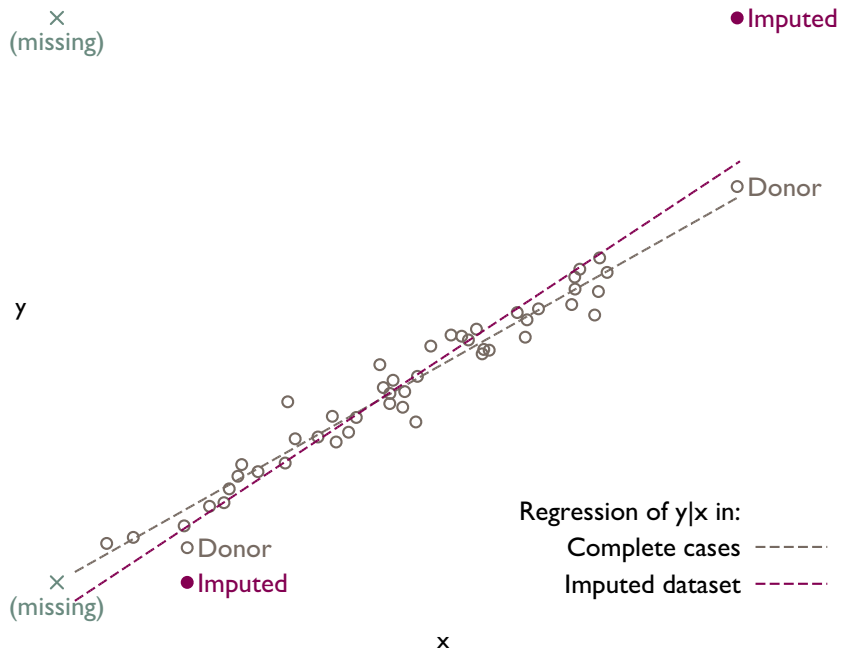
bias compared with posterior draws.

Assume δ_{hj} is a useful metric for defining matches. 'Poor quality' matches are then those with large values of δ_{hj}. For PMM we would expect poor quality matches to yield false relationships between x and y in the imputed data. LRD may be immune to this disadvantage.

LRD lacks the key cosmetic feature of PMM and hot deck: imputed values may not be observable and can lie outside the range of observed measurements. This perhaps explains its lack of popularity with applied statisticians in comparison with PMM. LRD will guard against heteroscedasticity in the same way as PMM because the local variance is approximated by using local residuals. LRD differs from PMM in its explicit use of the predictive mean to generate imputations. For PMM, poor quality matches will tend to attenuate x–$\boldsymbol{w}$ relationships (consider the extreme case above of $k = n_h$); meanwhile LRD may be forgiving of poor quality matching, particularly if the imputation model mean is only moderately misspecified. While LRD may appear at first glance to be more parametric and thus less robust than PMM, it may in fact be more robust to (i) departures from MCAR and (ii) misspecification of the imputation model.

A visualisation of the above point is given in figure 2.3. Forty x and y observations are

Figure 2.3: Potential bias in PMM when missing values are in the tails. The × symbols represent missing x values and observations with x observed are in grey. The imputation model is correctly specified but the sign of δ_{hj} is the same for all j, introducing bias.



simulated from a bivariate normal distribution. Given complete data, the analysis model would be $y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma_y^2)$. However, two x values are missing so this analysis is not an option. Instead the two values are imputed by PMM with $k = 1$ after fitting the imputation model $x_h \sim \mathrm{N}(\alpha_0 + \alpha_1 y_h, \sigma_x^2)$, which is correctly specified. Note that both the y_j lie beyond the observed y_h range, meaning the sign of δ_{hj} is the same for all possible h given an individual j. The regression of y on x is contrasted for complete data (dashed grey line) and the imputed dataset (dashed blue line). The problem with the slope comes from the imputed values lying beyond the range of y for which x was observed, giving them high leverage. PMM is unable to impute outside the x_h range and so some bias is introduced. This does not cause problems for LRD.

Figure 2.3 is exaggerated to make a point, and such a strong relationship along with such a strong and specific type of MAR is unlikely in practice. However, it raises a concern about the potential for upwardly biased regression coefficients after PMM. This is possible under MCAR but is more likely under departures from MCAR, where missing values in the tails of the $\hat{\boldsymbol{\alpha}}\boldsymbol{w}$ distribution would be more likely under certain MAR processes.

2.6 A REVIEW OF IMPUTATION BY PMM

The following is a review of the literature on PMM since the concept was introduced by Rubin in 1986[35]. A Web of Knowledge search for articles containing the phrases *predictive mean*

matching and *imputation* in title and/or keywords was initially run in October 2010 and updated in December 2011. All articles were read in full; those which developed, reviewed or evaluated some aspect of PMM are outlined below.

The existence of LRD was only noted during the course of this review and because of its similarity to PMM a further search was done on LRD, using the terms *local residual draws* or *local random residual* along with *imputation*.

Having read the articles returned by the above search, any work cited that appeared to be relevant were also obtained and, if relevant, added to the review. Figure 2.4 shows as a timeline the articles included.

The main focus of the review was to find

- The justification for PMM and LRD, and situations where authors advocate their use.
- Research on varying k in PMM and LRD.
- Comparisons of types 0, 1, 2 matching (and any other definition of δ_{hj}).

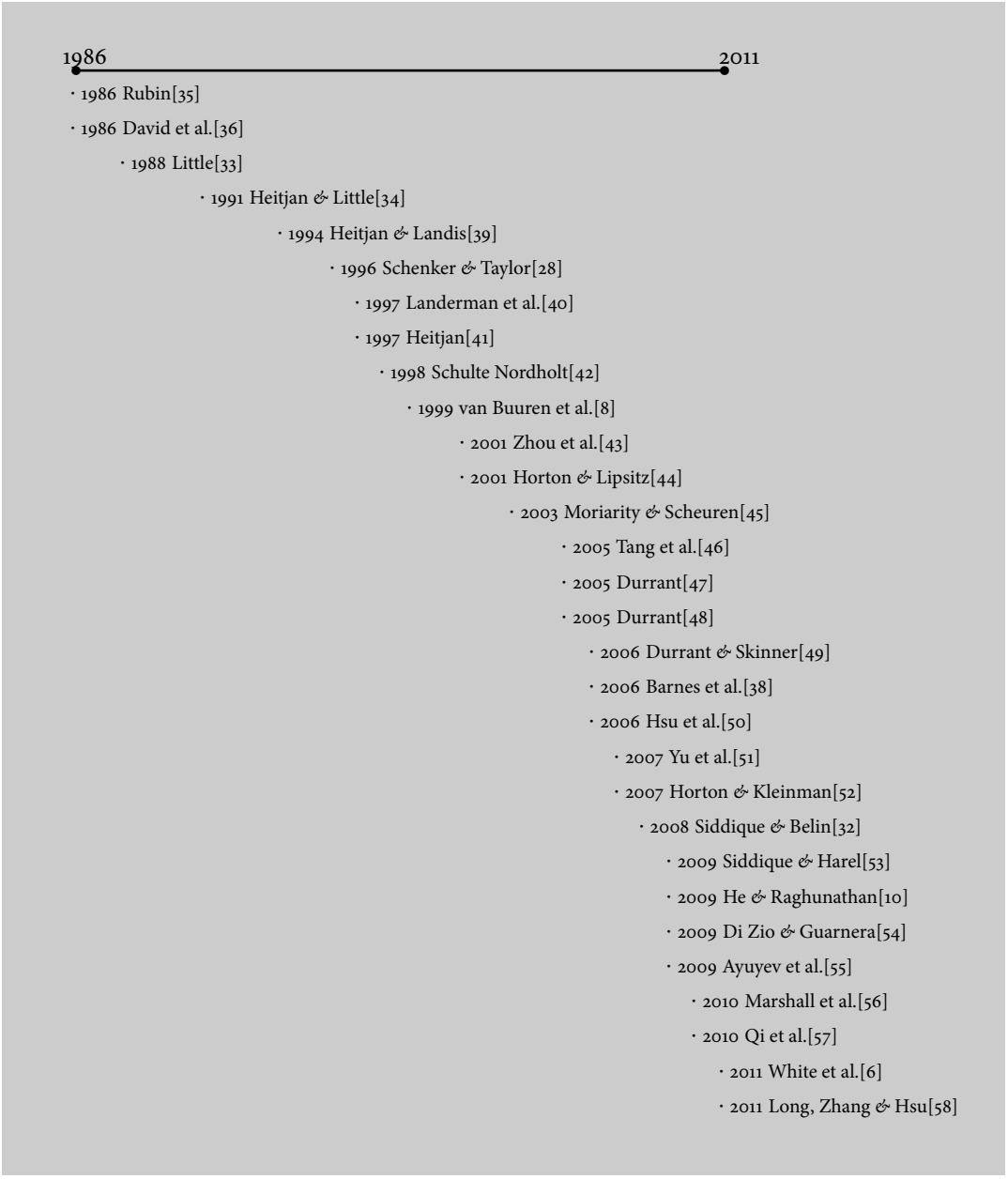
The review is divided up as follows: articles which give definitions of the metric for matching (that is, type 0, 1 or 2) are in section 2.6.1 along with relevant evaluations; those which outline a way of defining the donor pool are in section 2.6.2 along with relevant evaluations; articles which describe methods of sampling from the pool of potential donors are in section 2.6.3 along with relevant evaluations; general evaluations of PMM (not the specific aspects given above) are in section 2.6.4; case studies are in section 2.6.5 and the remaining articles are in section 2.6.6.

2.6.1 *Defining the matching metric δ_{hj}*

Rubin planted the seed for PMM by suggesting the use of MI as a way of improving matching[35]. His interest was in the situation where two variables, aiming to measure the same construct, are recorded in one cohort each, but are never simultaneously observed. The matching Rubin describes seems to be type 0. David et al[36] also used type 0 matching in what seems to be the origin of the LRD. Little then suggested the use of matching as a means of improving imputations[33], and wrote down the calculation of a matching metric δ_{hj}, which was type 0. In the same article it is noted – following Rubin's work on hot deck imputation[1] – that variances are underestimated if the parameters of the imputation model are treated as known. As an attempt to correct for this problem in PMM a modification to δ_{hj} was proposed, which was type 1 matching.

Heitjan and Little[34] suggested type 2 matching, using the argument Little[33] had used previously, that it would account for uncertainty about parameters in the imputation model. Draws of $\boldsymbol{\alpha}^*$ are obtained by drawing a bootstrap sample from the complete cases and fitting the imputation model to this sample (the *approximate Bayesian bootstrap*, ABB). Matching of h to j is then done on the linear predictor from this model. In a small simulation study (100 replications) based on sampling without replacement, type 2 matching gave arguably better coverage than type 0 matching (tending to err on the conservative side). The argument for type 2 does however seem to have been persuasive: since this paper authors describing PMM have outlined type 2 (although in most articles the description is not clear enough to identify the type of matching used).

Figure 2.4: A timeline of articles included in review of PMM and LRD



Hsu, Taylor, Murray and Commenges used imputation to try and recover information from censored patients in survival analysis[50] by imputing failure times. Type 2 matching was used to define a donor pool, termed the 'imputing risk set', taken from individuals still at risk (those who have not failed or been censored) at time t. Simulation studies focusing on a Kaplan–Meier estimator after bootstrap demonstrated better results for type 2 matching than type 0, with coverage being too low under the latter. Neither method was adequate with time-dependent covariates and dependent censoring, but PMM was the best of the methods investigated.

White, Royston and Wood's paper was a tutorial on mice and discussed PMM[6]. The authors' define their calculation of δ_{hj}, corresponding to type 1. Other calculations of δ_{hj} are not mentioned. A graphic justifying why PMM might be useful for protecting against misspecification of the imputation model is also given in [6]. Prior to this most articles had used δ_{hj} as type 0 or type 2 (for example [28, 34, 36, 38, 39, 43, 46, 57]), the only exceptions being Little[33] and Meinfelder[37].

2.6.2 *Defining the donor pool*

There are two broad approaches to defining the donor pool. The first is to use a fixed k; the second to use a fixed $\delta_{\max}$.

David et al.[36] initially imputed using global residual draws, but found this did not work particularly well. They replaced this with a set matching distance of \$2,000 ensuring $\mathrm{E}(\delta_{hj})$ was equal for all j.

The idea of selecting one from a pool of several potential donors in PMM was not apparently present in the initial work of Rubin[35] and Little[33], who both matched to the nearest donor only ($k = 1$). Heitjan and Little[34] introduced a pool of potential donors with $k = 5$, though no justification for this change was given. Since this paper authors have largely used fixed $k > 1$ (e.g. [39, 47]).

Schenker and Taylor suggested an adaptive method for choosing k based on 'the density of available donors in the vicinity of the incomplete case'[28]. Some $\delta_{\max}$ was defined within which potential donors lie and k is always greater than 1 and less than the number of observations with observed x ($\delta_{\max}$ is increased if $k < 2$).

2.6.3 *Sampling from the donor pool*

Heitjan and Little[34] only identified the pool of potential donors once and randomly sampled with replacement five times to obtain $m = 5$ imputations. Under the type 2 matching used this may be a computationally efficient way to generate imputations. However, for any imputation model with more than one covariate the imputed data sets cannot be regarded as independent under this sampling scheme (a single draw of $\boldsymbol{\alpha}^*$ is shared across all m). When there is only one covariate in the imputation model type 2 is identical to type 0 and therefore also fails to account for any uncertainty about $\hat{\boldsymbol{\alpha}}$. The different imputations are not independent given the model and so are improper.

Using type 2 matching, Schenker and Taylor drew $\boldsymbol{\alpha}^*$ and selected one of the potential donors at random separately for each imputed dataset[28]. This is more computationally intensive than Heitjan and Little's method[34] but draws of $\boldsymbol{\alpha}^*$ are independent across imputed

datasets, as they should be. Following this article, others seem to have followed Schenker and Taylor's method.

Moriarity and Scheuren[45] suggested the use of 'constrained matching' to build on Rubin's idea of imputation for statistical matching. While Rubin's method could impute the same individual h's value many times – unconstrained matching – constrained matching requires imputation of each observed value once, while 'slightly constrained' matching applies a penalty to any donor who has already donated (see section 7.4.1 for a remark on constrained matching). The description is unclear but this seems to apply within each imputed dataset, and the authors imply that not all missing values will be imputed under constrained matching. Their aim was to preserve the marginal distributions in the imputed data. In a simulation study they found MI with univariate- and with multivariate-constrained matching improved upon Rubin's method.

Durrant described various imputation techniques in a technical report[47]. PMM was compared to other imputation procedures via simulation and apparently tended to outperform the others provided multiple, not single, imputations were used. There is very little description of the simulation study in this report. It was however described more fully in a later article[49]. Interest was in imputing hourly pay, where the 'direct' variable (hourly pay measured *without* error) was partially observed but an indirect variable (hourly pay measured *with* error) was always available. A simulation study compared matching based methods including PMM with varying k, 'class-based' methods, and a version of PMM which imposes a penalty on a donor each time it is selected by reducing the probability of being selected as donor for another j ('slightly constrained' according to Moriarity and Scheuren's definition). Some small bias was observed for the class-based methods but not for the penalised method. PMM with $k = 1$ was seen to have the largest standard errors. PMM with $k = 10$ was then compared to propensity score weighting where the imputation model was correct, slightly misspecified and very misspecified. Both methods performed well under the correct imputation model, getting worse when the models were misspecified, but propensity score weighting deteriorated far faster than PMM. A brief simulation under MNAR was presented, which demonstrated increased error in both methods, but this was not enough to render them unusable.

Siddique and Belin imputed missing covariates and outcomes for a randomised trial using a form of PMM with fixed $k = 10$ but with the probability of donor selection inversely related to δ_{hj} [32]. A 'closeness parameter' ≥ 0 was used which could be altered to increase or reduce the probability of selecting the nearest donors. When imputing data from their trial it was observed that small values of the closeness parameter led to a larger estimate of the regression coefficient of interest. A simulation study motivated by the same trial demonstrated that smaller closeness parameters led to upwards bias, while larger closeness parameters led to more variable estimates. Imputation was done using something similar to a *forwards–backwards* type approach (rather than including measurements at all time points)[9], the reason for which was not given. They found that, more than the imputation method used, the order of imputation was important for the analysis model: imputing the least-missing variable first gave better results than imputing in a random order or chronological order, which were similar and tended to be biased away from the null. It seems unlikely this would have occurred if all other variables had been used when imputing each partially missing variable. Although not explicitly stated, the authors appear to

have used a mice approach (without chained equations the order for imputing variables would not have been noted).

2.6.4 *Other comparative evaluations of PMM or LRD*

David et al.[36] had access to a second database which contained 61% of the missing observations they had been imputing, and so were able to compare how well imputation methods performed. Unfortunately there was no comparison of inferences after different methods and instead the 'average deviations' from the true values were summarised by imputation methods. While hot deck was found to have the lowest prediction error – probably implying important interactions in the imputation model – Rubin's quote (1.7.3) tells us that this does not necessarily make it the best imputation method. For example, drawing imputations conditional on $\hat{\alpha}$ rather than α^* may reduce the average deviation but provide estimates of variance that fail to allow for missing data uncertainty.

Schenker and Taylor performed an extensive evaluation of PMM for imputing outcomes under MCAR[28]. In a fully factorial setup they varied the true model for the data, sample size, proportion of incomplete cases, the quantity to be estimated and the imputation method, with primary interest in which imputation methods provided valid inference. When the assumptions of posterior draws were correct, both PMM and LRD were passable, although their MSE was always higher than posterior draws and coverage tended to be slightly too low. When the assumptions of posterior draws were incorrect PMM and LRD tended to retain their performance even when posterior draws began to fail: coverage of 95% confidence intervals was around 92–94% for PMM and LRD even when the proportion missing was large and the sample size was small.

Zhou, Eckert and Tierney assessed the performance of PMM for imputing outcomes and then covariates, motivated by two real examples[43]. Type 2 matching was used with $\boldsymbol{\alpha}^*$ drawn via the approximate Bayesian bootstrap, following Heitjan and Little's recommendation, along with $k = 5$ and $m = 5$[34]. In a simulation study the authors aimed to compare the effect of matching on more vs. fewer covariates in the imputation model, but it appears the only variables included in the imputation model were those being matched, and so the comparison was effectively of larger vs. smaller imputation models, and matching was not compared to anything else. Two incompatible and two compatible imputation models were used: all gave reasonable results for bias, but coverage was lower than nominal, improving when more variables were included in the imputation models. A 'large', semi-compatible imputation model, using a total of 74 variables for imputation (many of which were auxiliary), consistently gave the best results.

Horton and Lipsitz investigated different software implementations of multiple imputation available in 2001, mainly discussing them from a usability point of view[44]. PMM was available in Solas, S-Plus and R but not SAS. In an example analysis of a single large ($n = 10,000$) simulated dataset PMM was seen to perform very similarly to multivariate normal imputation, arguably having slightly larger bias and slightly smaller variance. PMM in S-Plus and R performed well when x was MAR and its value predicted by other covariates excluding outcome y, but poorly when x was predicted by y.

Tang, Song, Belin and Unützer compared a PMM method to MI under a multivariate normal model, CC and LOCF for missing responses in the *Impact* clinical trial[46]. The method referred to as 'hot deck' was PMM with type 2 matching (with approximate Bayesian bootstrap used to draw α^*, as in [34]) with 'classes' formed based on the predicted mean. A simulation study was done based on Impact. This used sampling without replacement: allowing replacement could have favoured PMM by allowing missing items to be matched to their original observation. The sole focus of this simulation was coverage (which I regard as a very poor choice: as the number of observations sampled approaches the original sample size coverage must approach 100%, and so is meaningless under this simulation method). Using their coverage measure, PMM appeared to be the best technique (implying that its true coverage might have been too low), followed by multivariate normal imputation, complete cases and finally last observation carried forward.

Barnes, Lindborg and Seaman compared posterior draws, PMM and LRD for imputing outcomes with the variance structure of the imputation model misspecified[38]. In describing matching methods, $k = 1$ is described for PMM, while a donor pool is mentioned for LRD. In simulation studies $k = 1$ is used for both methods, along with type 2 matching. Two sets of simulations were suited to the assumptions of posterior draws (with different strengths of correlation). A third simulation used a mixture of the two correlation matrices from the first and second studies. This violates the posterior draws assumptions because the variance structure is misspecified. Unsurprisingly posterior draws were excellent in the first two studies. Surprisingly the standardised bias and coverage remained best for posterior draws even when the variance structure was misspecified. PMM and LRD were more biased than posterior draws but far less than LOCF. LRD gave coverage close to 95% in all cases, and the confidence intervals were generally shorter than posterior draws. PMM returned the shortest confidence intervals but gave coverage which was too low. The comparison between PMM and LRD is of course potentially confounded if a $k = 1$ was used for PMM but $k > 1$ was used for LRD.

Yu, Burton and Rivero-Arias mentioned PMM's availability in certain software[51]. A simulation study looked at how well various implementations retained the marginal distributions of variables, where these were slightly or very non-normal. Bias, standard errors and mean squared error are calculated compared to the complete dataset. R and Stata implementations of PMM performed equally well. Coverage was seen to be slightly poor for PMM implementations but this was still better than any of the others. The type of matching was not explicitly described but R's `mice` package was used (default type 2 with $k = 5$), while the `ice` command in Stata had type 1 matching and $k = 1$ as the only available option for PMM at this time (see figure 1 of reference [59]), so the coverage results for Stata may be surprising.

He and Raghunathan were interested in imputing missing values by chained equations when the assumption of normal errors was false[10], and so posterior draws were expected to break down. PMM, LRD and two other potentially robust imputation methods (global residual draws and Tukey's g and h distribution to model non-normal errors parametrically) were compared by simulation. LRD and PMM tended to have good performance with non-normal error variances. Some problems appeared with biased analysis model coefficients when the covariate being imputed followed a log-normal distribution. In some cases this produced under-coverage of

confidence intervals.

Di Zio and Guarnera described a PMM method that should be more robust to model misspecification than than the standard one, based on a Gaussian mixture model[54]. They had already suggested using Gaussian mixture models for MI by posterior draws, and this paper developed the idea for PMM. The justification was that standard PMM relies to some extent on a linearity assumption, and mixture models can relax this. In a small simulation study (100 replications due to the computationally intense model-fitting), non-normal data were generated along with a MAR pattern of nonresponse (the generation of MAR was not described). The preservation of the mean and variance structure for a multivariate dataset was investigated. Both seemed relatively well preserved by the new method, but it appeared to offer no great advantage over 'nearest neighbour donor' (an unclear and unreferenced method but described as a general form of PMM), predicted mean imputation based on a Gaussian mixture model or posterior draws based on a Gaussian mixture model. The latter is recommended by the authors for most situations, unless there is a requirement to imputed observable values.

Ayuyev, Jupin, Harris and Obradovic compared PMM to a new imputation method which they called 'dynamic clustering-based imputation'[55]. The implementation of PMM used was `WinMICE`; it is not clear what the matching metric or default value of k is in this package. In a simulation study they first assessed the quality of imputation methods on the proportion of correct imputations (categorical variables) and relative imputation accuracy (continuous variables). Both these measures are solely interested in how close multiple imputations come to the true value. It is a shame there was no focus on how close the resulting inferences were to those obtained on the dataset prior to deletion. However, by their hit-rate measure, dynamic clustering-based imputation was superior to posterior draws, PMM, linear regression and multilevel linear regression, with PMM coming a close second in terms of relative imputation accuracy for various proportions of missing data. Next, they looked at the proportion of correct classifications in a classification tree (a similar measure to the relative imputation accuracy); results were the same as for the previous study. This was all done under MCAR and the authors seem unaware of the existence of other missingness mechanisms. This paper is by computer scientists who seemed largely unaware of other literature on missing data.

Meinfelder compared several forms of PMM in simulation studies[37]. A method designated *parametric predictive mean matching* was described and noted to relate to Little's method, which would indicate type 1 matching, although the description of how the method was implemented points to type 2, highlighting the confusion that exists. A second method, designated *Bayesian bootstrap predictive mean matching* was also described. The description also corresponds to type 2 matching. In simulation studies, both methods were observed to have low bias but to underestimate variances and thus lead to low coverage. This is possibly due to the use of type 2 matching, or because $k = 1$ is used for both methods.

Marshall, Altman, Royston and Holder investigated the handling of missing data for prognostic modelling studies[56]. Survival time data were simulated with MAR missingness in the covariates. Complete cases and single imputation by PMM were compared to various MI methods, including MI by PMM with transformation towards marginal normality, MI by PMM without transformation and MI using flexible additive imputation models with PMM.

Simulations used R's `mice` package for the first two methods, so probably used $k = 5$ with type 2 matching, although this is not stated. For the flexible additive models, R's `aregimpute` package was used which by default uses $k = 1$ with unknown type 1 matching, but also has an option for using an adaptive technique based on Siddique and Belin[32]. Marshall et al. do not describe the option used. CC gave the least bias for regression coefficients under MAR and approximately nominal coverage, but was also inefficient. MI by PMM without covariate transformation was the second least biased and had generally acceptable coverage, certainly offering an improvement on the other imputation methods, and also had smaller bias than CC. PMM with covariate transformations tended to return similar results to imputation methods with stronger assumptions. Further simulations, where the missingness mechanism was MNAR, showed PMM without transformation to be the least biased of the methods investigated.

In a related paper, Marshall, Altman and Holder[60] performed a simulation study based on resampling with replacement. They compare CC, single imputation by PMM, MI by PMM, posterior draws using chained equations and flexible additive imputation models. Unlike the previous paper, CC was very poor compared to any imputation method. As in the previous paper, mice using PMM was the most useful.

Qi, Wang and He compared MI to their 'fully-augmented weighted estimator' (augmented inverse probability weighted) for handling a missing covariate in the Cox model[57]. This estimator is doubly robust and does not require a parametric model for the missingness probability or the distribution of missing covariates given the observed data. The PMM approach described suggests type 0 matching with $k = 1$. A simulation scenario involved one covariate which was MAR given two other covariates and outcome. PMM performed poorly compared to the fully-augmented weighted estimator when the imputation and analysis models were incompatible (some bias and under-coverage) but was acceptable otherwise. A second simulation study investigated problems when the censoring time depended on the missing covariate. PMM was biased in this situation (as were posterior draws) but coverage and efficiency were acceptable to the authors. In a third study the strength of correlation between missing and observed covariates was altered, showing bias barely changed as the correlation changed, while precision decreased as correlation increased.

Andridge and Little reviewed hot-deck related imputation methods and outline predictive mean matching, noting that matching on the linear predictor may be more fruitful than using the Mahalanobis distance, and that PMM is a more parsimonious method than the 'adjustment cell' method (standard hot deck)[61]. A simulation study compared various hot deck methods, including a 'proper' version of PMM based on the Bayesian bootstrap. Unfortunately it is not possible to tell what the type of matching is or the number of donors used, but the method had fair coverage properties and returned acceptable length confidence intervals, though it was not always the best method.

Long, Zhang and Hsu investigated MI of biomarker values which were MAR[58]. Interestingly, type 0 and type 2 matching were compared in two situations via simulation. While there was little difference in terms of bias, coverage tended to be closed to 95% under type 2 matching and they note that their estimators are 'further improved through a bootstrap set'[58].

2.6.5 *Non-comparative evaluations of PMM and case studies*

Landerman, Land and Pieper performed simulations motivated by problems with imputing income, a covariate in their analysis model[40]. Income was problematic because of its skew and the requirement that imputed values are non-negative. PMM was investigated with 5, 10 and 20% of observations missing under a MAR missingness pattern (although the MAR that was imposed is unclear). Bias in standard errors was small for all variables. Bias in point estimates and t statistics was slightly higher for one or two variables but essentially acceptable. PMM was not compared to any other method so it is unclear how alternatives would have fared. $k = 1$ was apparently used throughout. It was noted that 'weak' imputation models provided poor results when they were 'smaller' than the substantive model (incompatible); further, imputation models which were larger than the substantive model (semi-compatible) gave better results.

Schulte Nordholt describes hot deck in the context of longitudinal data as imputing a value observed in another individual from the same wave[42]. 'Cold deck' is described as imputing a value observed in the same individual on a different wave. In a simulation study PMM was not included but a short description of the method was given. In the examples of real data analysis the only comment was 'PMM gives similar results but is more computationally intensive'.

Van Buuren, Boshuizen and Knook first introduced the idea of MI by chained equations[8]. PMM was not previously feasible for non-monotone missing data problems and so this article opened up new possibilities for PMM. Although the article was not specifically about PMM the authors did use the method throughout (termed the 'closest predictor') apparently using $k = 1$ and type 0 matching.

2.6.6 *Miscellanies*

Heitjan and Landis[39] used PMM in the way described by Rubin[35]. They aimed to assess changes in underlying blood pressure over time. More recent participants had generally had high blood pressure treated to lower their blood pressure. *Treated blood pressure* and *untreated blood pressure* were to be imputed. Since these were never simultaneously observed Rubin's general approach is followed (performing sensitivity analyses to assumptions about the variables' partial correlation). There is no evaluation of the usefulness of PMM compared to any other imputation method, but it is the imputation method of choice for the sensitivity analysis in this paper. A nice explanation is given of potential advantages of PMM vs. hot deck, similar to that of David et al[36].

Heitjan[41] wrote to praise and criticise some authors who used MI in an applied paper[62]. MI by posterior draws is described in simple terms and it is noted briefly that PMM may be preferable if posterior draws provide cosmetically poor imputations, for example beyond the range of the observed data.

Durrant developed a method of imputation by data augmentation, which involved PMM using type 2 matching instead of posterior draws[48]. This was developed for univariate missing data problems where there is a complete surrogate version of the incomplete variable, but the surrogate is measured with error. The notation and description makes up the entire paper. There is no evaluation of the method or any mention of its implementation.

Horton and Kleinman were interested in missing covariates and describe the main methods for dealing with problems[52].Their main aim was to compare software available in 2007. There is little offered on PMM, the only comment noting that some of the software available at that time could impute by PMM as an alternative to posterior draws.

Siddique and Harel describe the `midas` SAS macro used for Siddique's earlier paper[53]. This was a flexible implementation of PMM which allowed for user specified imputation equations with PMM. However, it is unclear how sensible the combination of PMM with forwards–backwards imputation was (see [32]).

2.6.7 *What is already known about PMM and LRD?*

PMM and LRD seem to be considered as tools for dealing with imputation model misspecification. Under a correctly specified imputation model, posterior draws appear to provide superior inference; under a poorly specified imputation model, PMM and LRD may have a degree of robustness to misspecification that posterior draws does not.

Beyond these conclusions, the literature on PMM is fragmented. There has been no systematic exploration of how the number of donors k or different matching methods influence the analysis model estimates when one or more covariates are partially observed. Choices of k and δ_{hj} appear to have been largely chosen ad-hoc, or as software defaults. Some articles did vary k [28, 50], while others did something producing a similar effect[32]. Results seemed to show that smaller k produced more variable $\hat{\boldsymbol{\beta}}$, while larger k could lead to bias. One article contrasting different types of matching favoured type 2 over type 0 but a small simulation study did not demonstrate much difference[34]. While this did not seem to justify the general uptake of type 2, two more recent articles have noted a more pronounced difference in coverage for type 2 as compared to type 0[50, 58]. No articles have ever compared type 1 to types 0 or 2.

Intuitively, PMM should help protect against misspecification of the imputation model, as supported by figure 2.2. Chapter 3 describes and reports simulation studies investigating how k and the type of matching influences bias, efficiency and coverage of PMM and LRD. Chapter 4 takes the best form of each method forward to a multivariable simulation study based on the trauma dataset described in section 1.5.1.

DrawSuit

A study with bivariate normal y and x (section 3.1). The aim is to see how well various forms of PMM and LRD perform in a scenario which is ideally suited to posterior draws.

J-thwart & U-thwart

Two studies set up to thwart posterior draws (section 3.2). For both, the true model for y is linear in x^2. 'j' and 'u' describe the true shape of y–x association in the simulated datasets.

Trauma

A multivariable simulation study in which the transformation which best predicts outcome may not be the most appropriate transformation towards normality (section 4). The two most favoured forms of PMM and LRD from J-thwart (3.2.2) and U-thwart (3.2.1) are used.

3 *Univariable simulation studies assessing PMM & LRD*

3.1 SIMULATION STUDY DESIGNED TO SUIT POSTERIOR DRAWS: DRAWSUIT

This study uses simulation to evaluate PMM and LRD for missing x when the assumptions underlying posterior draws are correct. Data are generated from $x_i \sim \mathrm{N}(0, 1)$ and $y_i \mid x_i \sim N(\beta x_i, 10^2)$. The mean and variance structure of the imputation models are correctly specified and the missing data mechanism is MCAR or MAR. For this investigation, several imputation methods are investigated: posterior draws, PMM with type 1 and 2 matching and k = 1, 3, 5 and 10, and LRD with type 1 and 2 matching and k = 1, 3, 5, 10 and 20. It is of interest to know if these methods perform adequately in this scenario, and if not, what and how bad any problems are.

The factors to be varied are split into those which are a part of the analysis method to be compared (*) and those which are a part of the data generating model and may influence the comparison of analysis methods (○):

* *Imputation method*. The two matching-based methods are of central interest, while posterior draws are included for comparison (as well as the complete case and complete data analyses).
* *Type of matching*. Match types 1 and 2 will be investigated. Notice that type 0 is identical to type 2 matching when the imputation model includes only one covariate.
* *Size of potential-donor pool*. The number of donors is varied for imputation by PMM and LRD.
○ *Strength of association*. Different values of β will be used in the imputation model.
○ *Missingness mechanism*. It is expected that MAR will be a stronger test of PMM and LRD because of different expected matching distances for different values of observed y_i. Further, missingness under MAR will be in the tails of the distribution, which may introduce bias for PMM (see figure 2.3).
○ *Sample size*. It is plausible that a larger donor pool is less of a problem for larger sample sizes, where it represents a smaller proportion of the dataset.
○ *Proportion of missing values*. Any issues with PMM or LRD are likely to be magnified with a larger proportion of missing values, partly because the analysis relies more on the imputed data and partly because there are fewer close matches.

The study is not fully factorial. In particular, the factors not varied factorially with one another are sample size and proportion of missing values.

Table 3.1: Factors and levels to be varied in DrawSuit simulation study

Factor	Variations
Imputation method	Posterior draws, PMM and LRD
Match types	1 and 2
Donor pool	k = 1, 3, 5, 10
Strength of association	$\beta^{\dagger}$ = 0, 3.33 and 10
Missingness mechanism	MCAR, Weak MAR, strong MAR
Sample size	100, 500, 6000
Proportion of missing values	0.25, 0.5, 0.75

†Values of β are chosen to correspond to R^2 = 0, 0.1 and 0.5, termed 'zero', 'weak' and 'strong' association respectively.

3.1.1 *Simulation procedures*

Stata 11.2 is used for all aspects of these simulations. The `ice` command is used to impute missing data and `mim` to analyse imputed datasets and combine estimates using Rubin's rules.

The number of replications s used for each scenario, is 1000. For each replication, x_i are simulated from a normal(0,1) distribution. y_i are then simulated from

$$y_i|x_i \sim N(\beta x_i, 10^2), \tag{3.1}$$

implying a bivariate normal distribution for x and y.

Vectors R indicating whether x_i are observed are then simulated. For MCAR, $R_i = 1$ if $U_i < \pi$ where $U_i \sim$ uniform$(0, 1)$ and π is the desired proportion of missingness. For MAR,

$$\text{logit}[\text{P}(R_i = 1|y_i)] = \gamma_0 + \gamma_1 y_i. \tag{3.2}$$

For weak MAR, γ_1 is set to 0.05 (a log-odds ratio of 0.5 per standard deviation change in y); for strong MAR, γ_1 is set to 0.1. γ_0 is then altered to achieve the desired π. To achieve 25% missing x, -1.15 and -1.3 work well for weak and strong MAR respectively.

The imputation model will be

$$x_h|y_h \sim \text{N}(\alpha_0 + \alpha_1 y_h, \sigma^2) \tag{3.3}$$

for all imputation methods. Note that the normality of residuals for y means match quality will vary for different j.

The same analysis model, $y_i \sim \text{N}(\beta_0 + \beta x_i, \sigma^2)$, is fitted for each scenario (the intercept $\hat{\beta}_0$ is estimated even though $\text{E}(\hat{\beta}) = 0$).

Before imposing missingness, the parameters of the analysis model are estimated in the complete data. The response indicator, R, is then applied to x (deleting those values where $R = 0$) and the complete case analysis is fit. The various imputation methods are applied $m = 5$ times in turn to the partially observed dataset. Rubin's rules are used to obtain the estimate $\hat{\beta}$ and its standard error.

Table 3.2: Strength of MAR in univariable simulations

Missingness	Area under ROC curve
MCAR	0.50
Weak MAR	0.65
Strong MAR	0.75

The dependence of missingness mechanism on y was of interest and so upon generating R, the area under the ROC curve was calculated for each missingness mechanism in each dataset and summarised.

By applying different missingness mechanisms to one dataset, and different analysis strategies to each of the missingness patterns, there is a moderate dependency between the results of related simulation runs. This makes simulations particularly sensitive to differences between methods.

3.1.2 *Evaluating the performance of methods for different scenarios*

Bias in a point estimate is assessed by calculating

$$\frac{1}{s}\sum_{l=1}^{s}\hat{\beta}_l - \beta, \tag{3.4}$$

where β is the true parameter for the analysis model (used in the data generating model) and $\hat{\beta}_l$ are the estimated values of β from the lth replication.

Empirical standard errors are defined as

$$\mathrm{SE}_{\mathrm{emp}} = \frac{1}{s-1}\sum_{l=1}^{s}\sqrt{\left(\beta_l - \hat{\beta}\right)^2}, \tag{3.5}$$

the standard deviation of $\hat{\beta}$ over the s replications.

Coverage is assessed as the percentage of times the 95% confidence interval for $\hat{\beta}_l$ contains the true β.

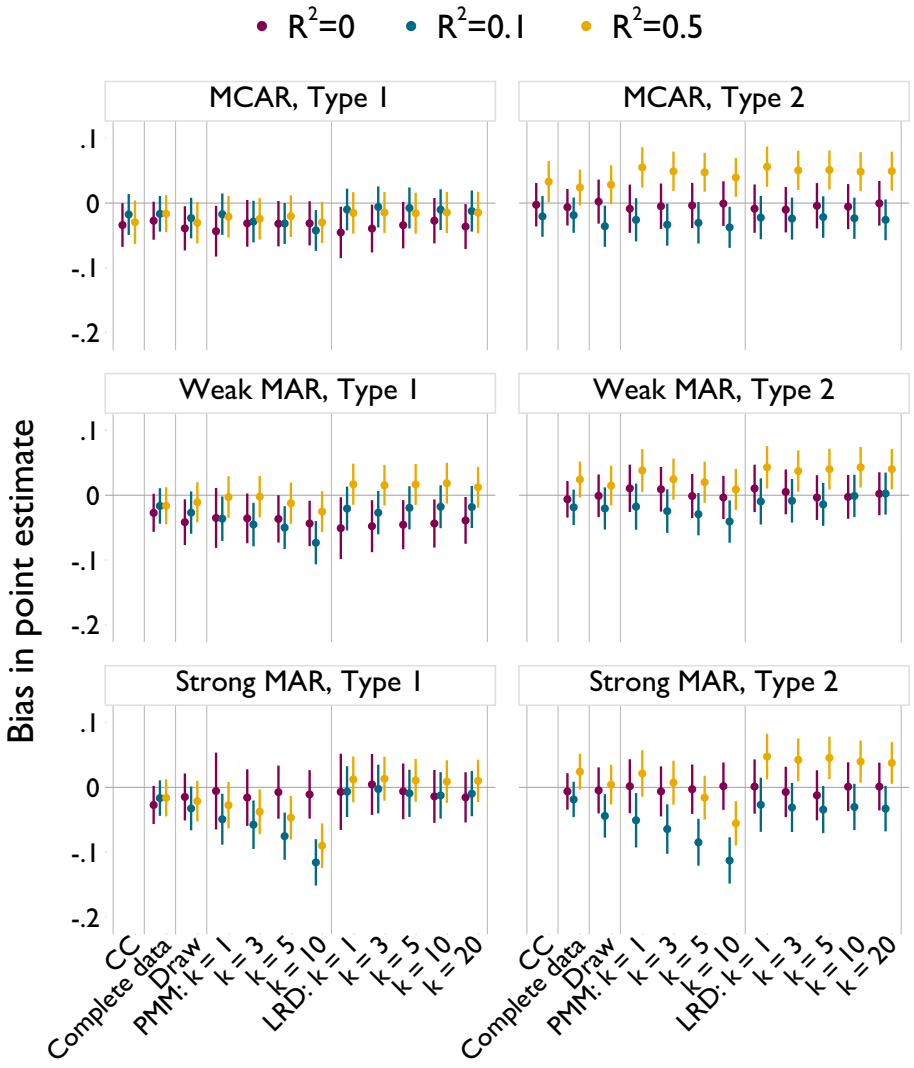
Results are presented graphically with point estimate for a simulation parameter, along with 95% Monte Carlo confidence intervals[63], on the vertical axis and the method of imputation on the horizontal axis.

3.1.3 *DrawSuit results: n = 500, 25% missing x*

Table 3.2 shows the area under the ROC curve for each of the three missingness mechanisms used. These values will be the same for other univariable simulation studies with larger and smaller sample sizes as well as more missing values and non-linear associations. Although these values on their own may lack meaning, this area under the ROC curve is given for simulation studies throughout this thesis so as to provide a standardised degree of MAR across simulation studies.

Posterior draws provides estimates which are unbiased with correct coverage in these cases (figure 3.1). These are provided for reference. None of the matching methods perform as well as

Figure 3.1: DrawSuit: Bias in point estimates, $n = 500$ (Error bars are ±2× Monte Carlo standard errors)



posterior draws, regardless of the number of donors or the type of matching. However, there are indications in these figures as to which forms of imputation by matching methods are worthy of further study.

PMM appears to give very similar results to complete data, complete cases and posterior draws when type 1 matching is used under MCAR. Under MAR a slight downward bias appears with larger values of k. This is expected: consider k = the number of observations with nonmissing x (global instead of local matching). This would not distinguish between good and bad donors, drawing at random any value of x from the observed data. Thus, increasing k for PMM may be viewed as introducing a degree of incompatibility to the imputation and analysis models. Although the results do demonstrate this bias, it is very small. Recall that the true values of β are 0 (black), 3.3 (blue) and 10 (green).

For type 2 matching, PMM is always unbiased when $R^2 = 0$. Under MCAR there is a slight downward bias for $R^2 = 0.1$ and a slight upward bias for $R^2 = 0.5$. This is unaffected by k. For MAR, both of these biases move slightly downwards overall with this shift more pronounced as k increases.

LRD appears to be unbiased when type 1 matching is used, regardless of the strength of association in the data or the strength of the missingness mechanism (clearly there is then no effect of changing k). With type 2 matching there appears to be a small upwards bias associated with LRD when $R^2 = 0.5$. There is also a hint of a downwards bias when $R^2 = 0$. Although present, it is worth noting that the magnitude of these biases are tiny. There is no effect of k on bias because LRD explicitly uses the predictive mean in imputation, unlike PMM which only uses it to identify suitable donors.

Empirical standard errors are similar for PMM and LRD and broadly similar for type 1 and type 2 matching (figure 3.2). Imputation methods with smaller k tend to have lower precision than counterparts with larger k. With the largest values of k precision is comparable to posterior draws. This is less true as the strength of MAR increases. For type 1 matching, standard errors are highest for $R^2 = 0$. Standard errors after type 2 matching are much less sensitive to the strength of association than after type 1.

Figure 3.3 shows nominal 95% confidence intervals under type 1 matching come close to giving nominal coverage under MCAR, except when $R^2 = 0.5$ with PMM, when it is around 93–94% and also $R^2 = 0$ for LRD. As the strength of missingness mechanism is increased, coverage is lower (between 90 and 95%), and consistently slightly worse for LRD than PMM. Larger k improves coverage somewhat for LRD and PMM.

For type 2 matching the coverage is generally far worse than type 1, especially with small k. The coverage increases, approaching 95% with larger values of k. In contrast to type 1, larger values of R^2 always have coverage closer to 95% than the smaller values. As with type 1 matching, coverage is pushed downwards slightly by stronger missingness mechanisms.

Taking all these results together, PMM with $k = 10$ and LRD with $k \geq 10$ appear to provide the best alternative to posterior draws for $n = 500$. Type 1 matching provides slightly superior results in the case explored here. However, neither method is as good as posterior draws in terms of its minimal bias, small standard error and nominal coverage of confidence intervals.

Figure 3.2: Drawsuit: Empirical standard error, n = 500 (Error bars are ±2× Monte Carlo standard errors)

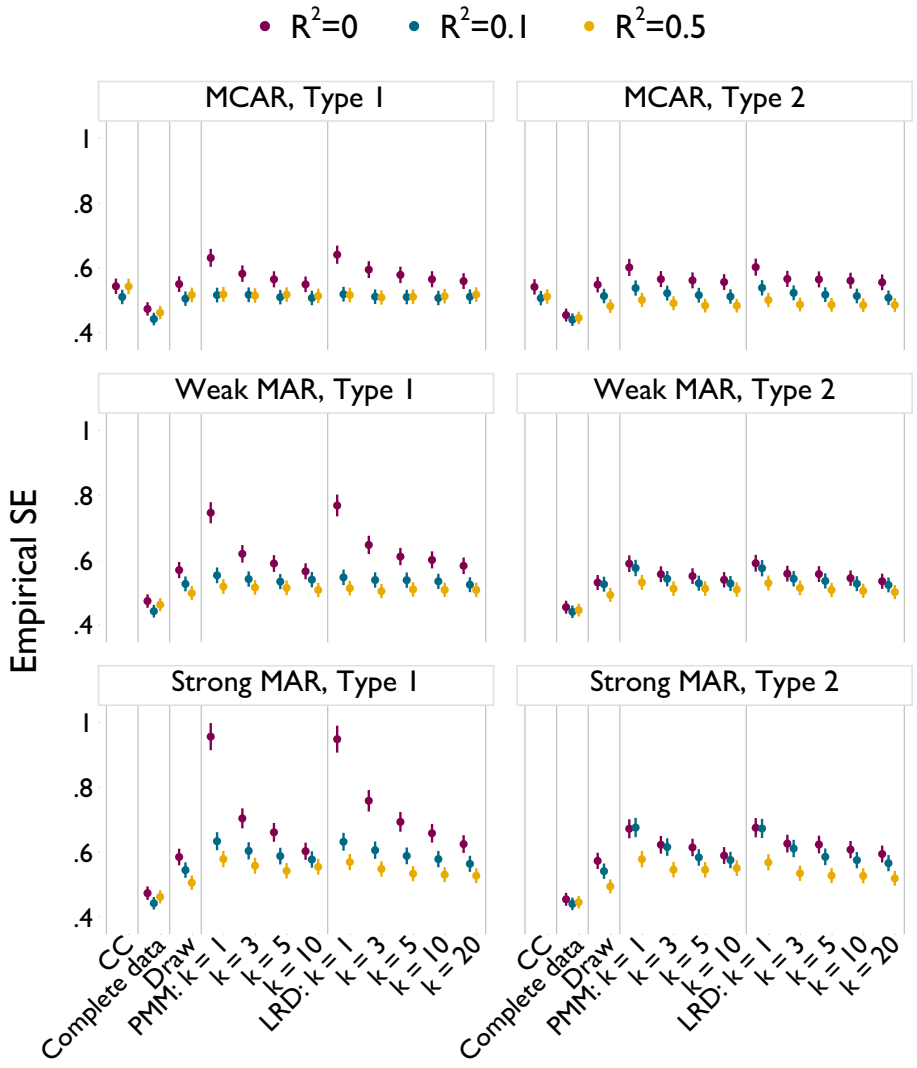


Figure 3.3: DrawSuit: Coverage of nominal 95% CI, n = 500 (Error bars are ±2× Monte Carlo standard errors)

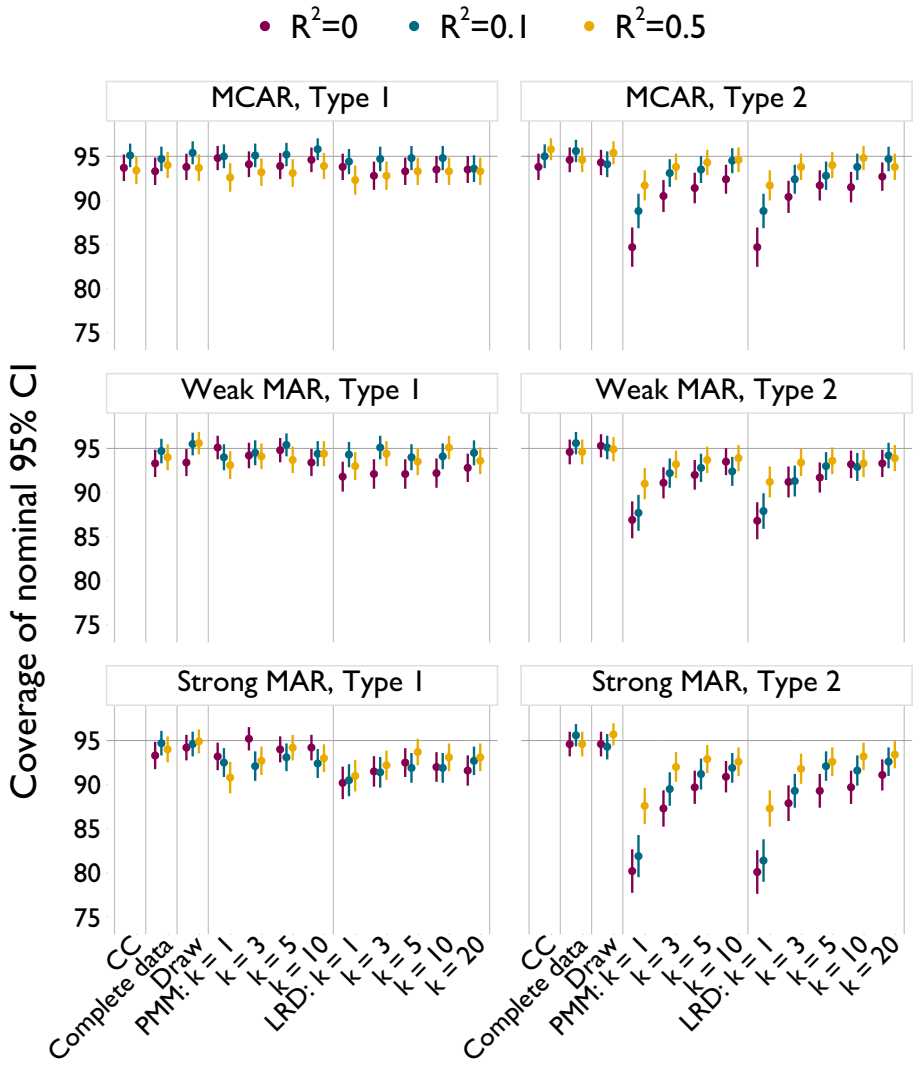
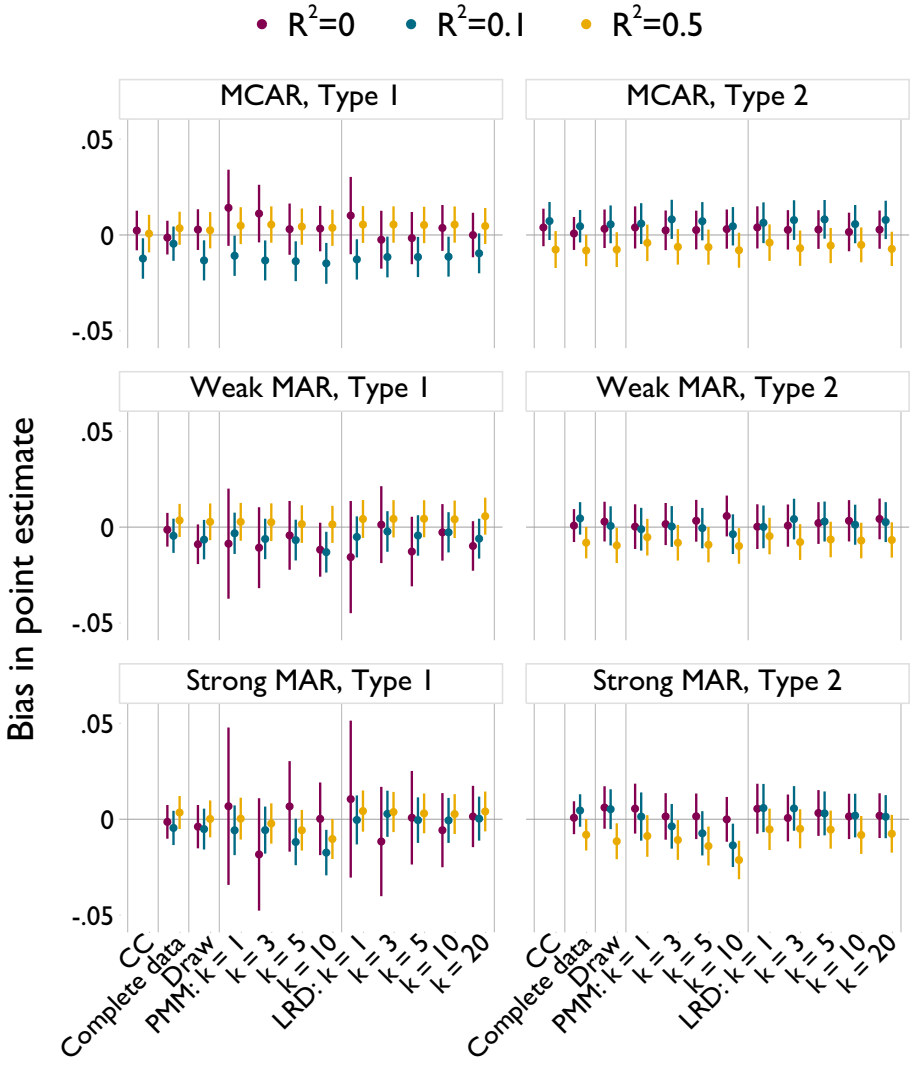


Figure 3.4: DrawSuit: Bias in point estimates, $n = 5000$ (Error bars are ±2× Monte Carlo standard errors)

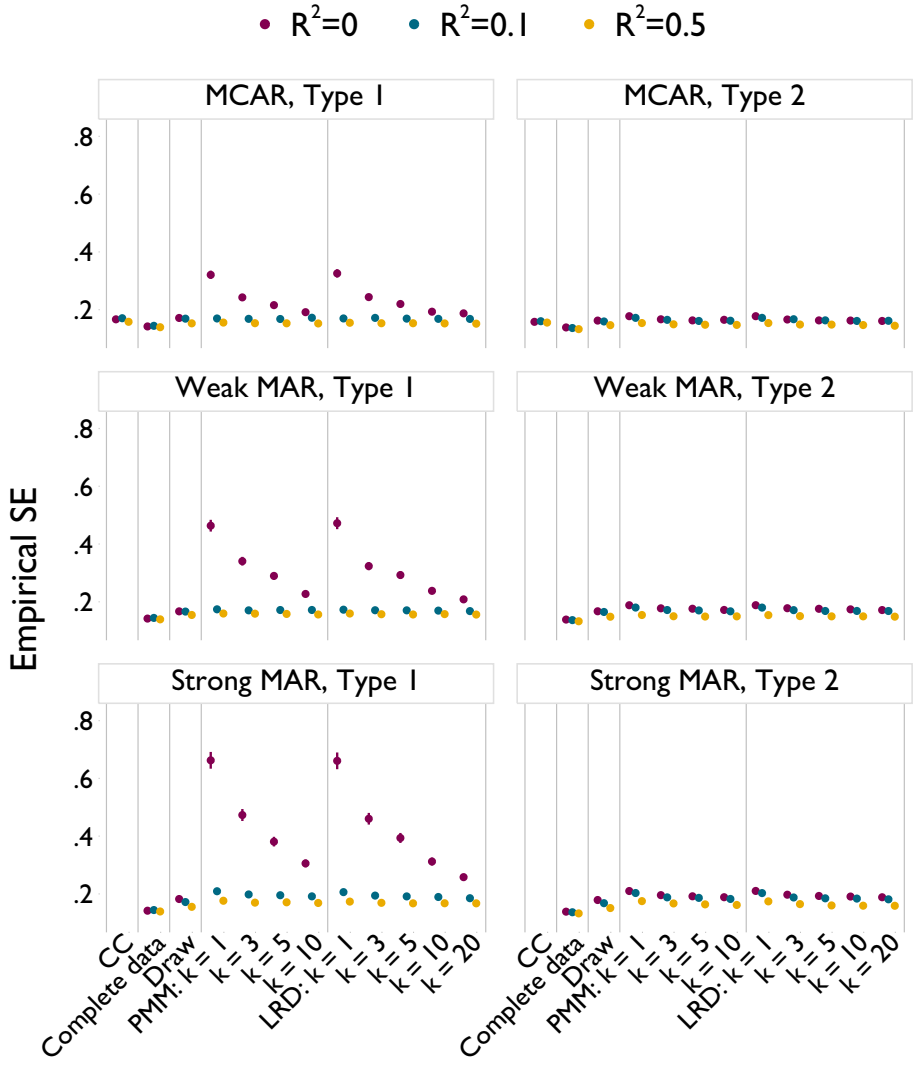


3.1.4 *DrawSuit results: n = 5000, 25% missing x*

It was of interest to see whether the same patterns occur for much larger datasets, where it is possible that the performance of PMM and LRD would improve due to improved matching quality on average. Further simulations were run for $n = 5000$, with all other factors identical to the above scenarios. Plots of the results are again presented side by side for type 1 and 2 matching. In general these results showed similar patterns to $n = 500$ and $n = 100$, with minor differences in places.

Bias in point estimates is very small and never estimated as greater than ±0.02 (figure 3.4). Although sometimes the Monte Carlo error bars do not cross the line of zero bias, the magnitude

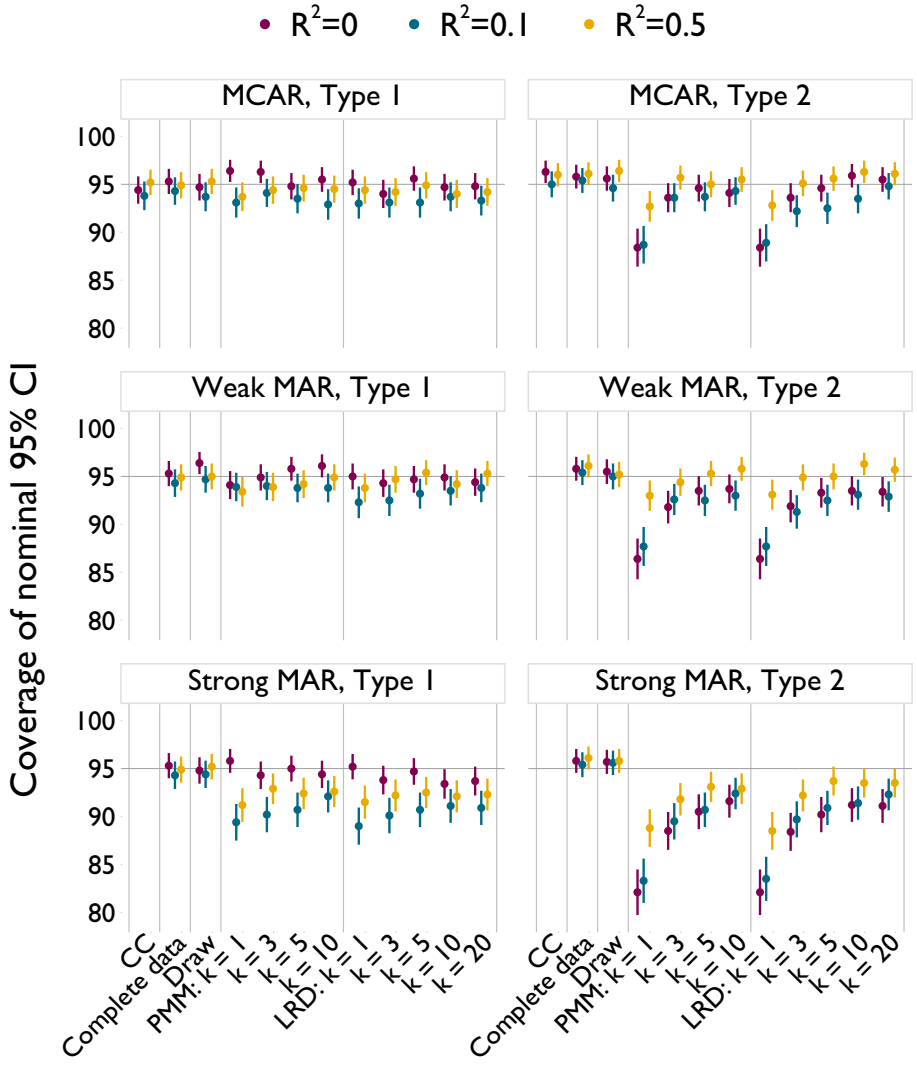
Figure 3.5: DrawSuit: Empirical standard error, $n = 5000$ (Error bars are ±2× Monte Carlo standard errors)



is once again minuscule relative to the values of β used. Perhaps surprisingly, PMM still has increasing downwards bias for larger k. This is again a larger problem for type 1 matching than type 2.

Standard errors (figure 3.5) are relatively similar, contrasting methods, to the $n = 500$ case, with one important difference: under type 1 matching and $R^2 = 0$ the empirical standard errors are markedly larger than in any similar cases. With type 2 matching and other values of R^2 empirical standard errors are unsurprising. Standard errors decrease with increasing k but even in the best cases are still around 20% larger than those for posterior draws. Standard errors are higher and become less comparable to posterior draws as the strength of MAR increases, regardless of the matching type or k. As with the two smaller sample sizes, type 1 matching has

Figure 3.6: DrawSuit: Coverage of nominal 95% CI, n = 5000 (Error bars are ±2× Monte Carlo standard errors)



larger standard errors than type 2 in the worst situations.

Coverage results shown in figure 3.6 are very similar to the n = 500 case. In the best cases, coverage for PMM and LRD sometimes reaches 95%. Again, type 1 matching seems to give more appropriate coverage than type 2 for smaller values of k. As R^2 and strength of MAR increase the coverage of both methods is pulled down below 95%.

3.1.5 *DrawSuit results: n = 100, 25% missing x*

For n = 100 with 25% missing x, results regarding PMM, LRD, the type of matching and k are broadly similar to results for n = 500 and n = 5000. The main difference is that any problems

are exaggerated. These results are presented in appendix B.

3.1.6 *DrawSuit results: Increasing the proportion missing*

Two variations away from the base case of 25% missing were investigated: 50% and 75% missing, with a sample size of 500 only. Results are not shown for these simulations as they are a simple extension of the results with 25% missing. Problems observed with 25% missing follow similar patterns but are exaggerated (since a larger proportion of the data in each analysis depends on the imputations used and δ_{hj} will generally be larger). No substantial differences are present that would alter conclusions about choice of k or the type of matching.

3.1.7 *DrawSuit: Conclusions regarding matching metric and size of donor pool*

The results for this simulation study based on a very simple model are complicated. While the factors chosen for an analysis – type of matching and size of donor pool – can be chosen, others cannot and may be unknown. This makes it difficult to recommend a specific method when faced with a dataset. In this situation, some of the factors that were varied in this simulation will be known, such as sample size. To some extent, the expected strength of association in a dataset could be assumed 'known' from the observed data, which may help to choose an imputation method, but this could be affected by the missing data and may in turn affect which imputation method is chosen. It is also possible to verify the strength of MAR from the observed data, but ideally any method would be robust to different missingness mechanisms.

With respect to PMM and LRD:

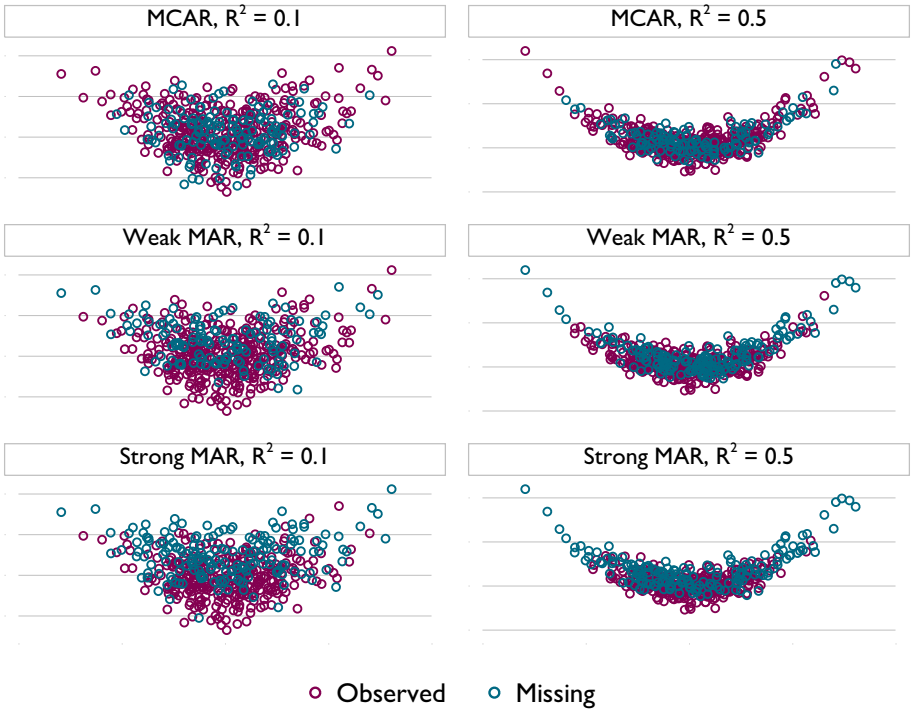
- Larger k is almost always sensible. For PMM increasing k does result in a bias towards the null but this was tiny for the case investigated here. For LRD the bias is not present and the value of 20 is uniformly better than any others considered. In the situation considered here a global residual draw would be very similar to posterior draws and in the presence of heteroscedasticity larger k may not provide adequate imputations. Since this sort of situation is one reason to investigate alternatives to posterior draws, global residual draws will not be considered here, and k is restricted to 20 for LRD.
- Type 2 matching seems to be better than type 1 in terms of point estimates – these are less biased and more precise. The model standard errors after type 2 matching are also more accurate. However, coverage is poor compared to type 1, particularly when k is small. For $k \geq 10$, coverage for type 1 and 2 matching is very similar.

It remains unclear is how these methods will perform in more complex settings – specifically where posterior draws is expected to perform poorly.

3.2 Simulation studies designed to thwart posterior draws

The following two simulation studies are very similar in set up to 3.1 except that the data generating model for y is linear in x^2 instead of x. The imputation model is (3.3) as in section 3.1.1, imputing missing x (not x^2) in exactly the same ways as before, making the naïve and false assumption that x is linear in y. It is then expected that posterior draws will produce poor imputations, as demonstrated in the example of figure 2.2. Issues around the choice of analysis

Figure 3.7: Typical simulated datasets in U-thwart



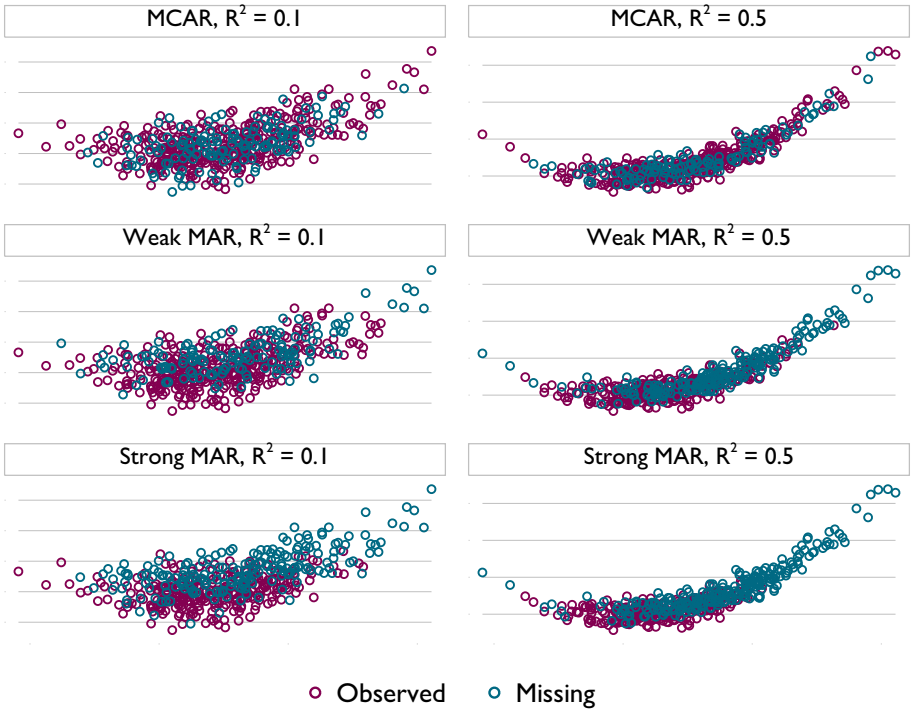
model are not considered here. Throughout, $x \sim \mathrm{N}(0, 1)$ and the data generating model and analysis model is $y_i \sim \mathrm{N}(\beta_0 + \beta_1 x + \beta x^2, \sigma_y^2)$. Although $\mathrm{E}(\beta_0) = \mathrm{E}(\beta_1) = 0$ both parameters are estimated. Interest is in the estimation of β.

Two different non-linear scenarios are investigated: in the first, the expected minimum value of y is set to the mean of x, while in the second the expected minimum of y is set to one standard deviation below the mean of x, giving 'U' and 'J' shaped curves (styled *U-thwart* and *J-thwart* respectively). The same strengths of association and missingness mechanisms are simulated as in DrawSuit.

Figures 3.7 and 3.8 show typical simulated data over six different settings. Observed values are represented by purple dots, and the true value for missing values by blue dots. The $R^2 = 0.1$ setting is shown to the left and $R^2 = 0.5$ to the right; $R^2 = 0$ is omitted because $(y, x) \sim$ BVN with correlation 0, which does not require visualisation. The complete datasets in the left-side panels are identical, as they are in the right-side panels. The top panels show MCAR, the middle panels 'weak' MAR and the bottom panels 'strong' MAR. Notice that under MAR x is more likely to be missing for higher values of y, meaning missing values tend to occur in the tails. This provides a difficult test for matching methods.

U-thwart is assumed to be less realistic than J-thwart: it seems unlikely that in real datasets the estimated maximum or minimum of one variable would be at the exact mean of a covariate. In U-thwart, the imputation model, a linear regression of x on y, is null in expectation regardless of the strength of association. This provides a particularly tough test for type 1 matching,

Figure 3.8: Typical simulated datasets in J-thwart



explained below.

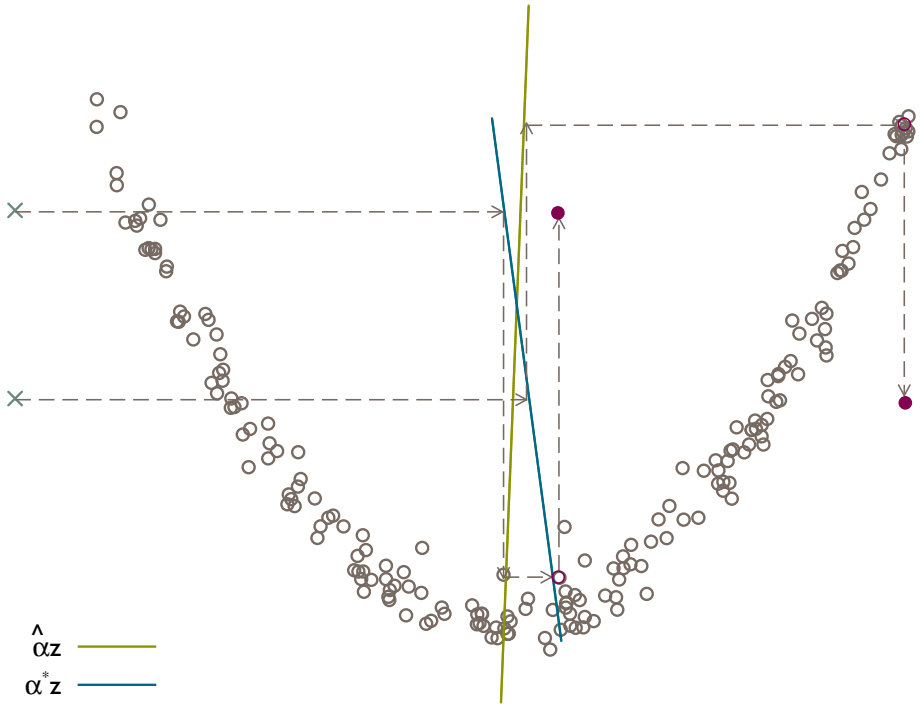
When there is a non-null association in J-thwart, $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^*$ will tend to have the same sign because the regression of x on y is non null. However the symmetry of U-thwart means that even if y_i is deterministically equal to x_i^2, the regression of x on y is null. There is then a strong possibility that the signs of $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^*$ will differ, implying matches may be extremely poor quality even with very small δ_{hj}. Figure 3.9 demonstrates this point, with arrows indicating the type 1 matching algorithm. In general, type 1 matching will provide poor quality imputations whenever the sign of $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^*$ differ.

The DrawSuit simulation study above (section 3.1) showed consistency across different sample sizes and proportions of missing values. For the following simulations these factors are not varied and it is assumed results would remain consistent: results are presented only for $n = 500$ and 25% missing x throughout.

3.2.1 *U-thwart results*

As expected, imputation by posterior draws proves very poor and introduces bias in this scenario when $R^2 \neq 0$ (figure 3.10). PMM and LRD also appears to have flaws, which was less expected. Bias observed in point estimates is large under type 1 matching as was anticipated (see figure 3.9) and no method provides anything close to unbiased estimation of β. The magnitude of bias is roughly proportional to the size of the true parameter. Bias is slightly reduced by increasing

Figure 3.9: Issue with type 1 matching for U-thwart. Grey dots are fully observed data; grey crosses are observations with x missing and y observed; dashed lines show how $\hat{\boldsymbol{\alpha}}\boldsymbol{w}$ is matched to $\boldsymbol{\alpha}^*\boldsymbol{w}$ to select a donee (hollow blue circle) and impute that value (filled blue circle)



k for MCAR and weak MAR, but even then parameters are underestimated by around 20%. For strong MAR, increasing k seems to have no effect, and increases bias if anything. Posterior draws are always more severely biased than PMM and LRD.

Type 2 matching is unbiased under MCAR. Under both MAR scenarios there is an *upwards* bias for $R^2 = 0.5$, no bias for $R^2 = 0$ and a slight downwards bias for $R^2 = 0.1$. The upwards bias for $R^2 = 0.5$ is understandable: both PMM and LRD are imputing values in the tails donated from observed individuals at or near to $x_{h\min}$ and $x_{h\max}$, which will induce upwards bias in $\hat{\beta}$ (similar to that seen in the left panel of figure 3.14, coming later).

Empirical standard errors are larger when type 1 matching is used than type 2 (figure 3.11). This is most pronounced when $R^2 = 0.5$, but is also larger under MAR than MCAR. Type 2 matching always gives less variable point estimates than type 1. As k increases the standard errors reduce.

Despite the severe bias, low precision, and poorly estimated standard errors, coverage of confidence intervals obtained by type 1 matching is not as far from 95% as might be expected (figure 3.12). Coverage is too high under MCAR but is often closer to 95% under MAR. Interestingly, confidence intervals for $R^2 = 0.1$ always have better coverage than for $R^2 = 0$ and $R^2 = 0.5$. Coverage is always lower for $R^2 = 0.1$ than for other values of R^2.

Confidence intervals after type 2 matching tend to have worse coverage than type 1, the

Figure 3.10: U-thwart: Bias in point estimates (Error bars are ±2× Monte Carlo standard errors)

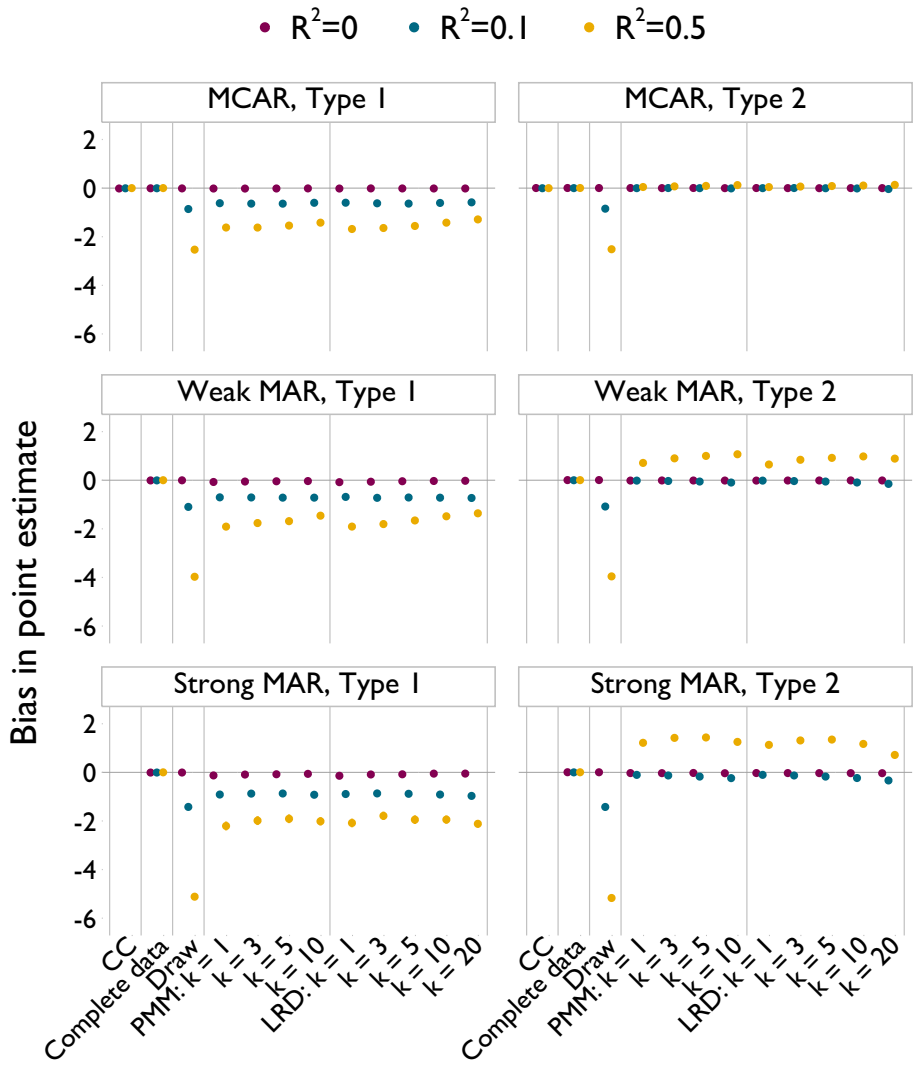


Figure 3.11: U-thwart: Empirical standard errors (Error bars are ±2× Monte Carlo standard errors)

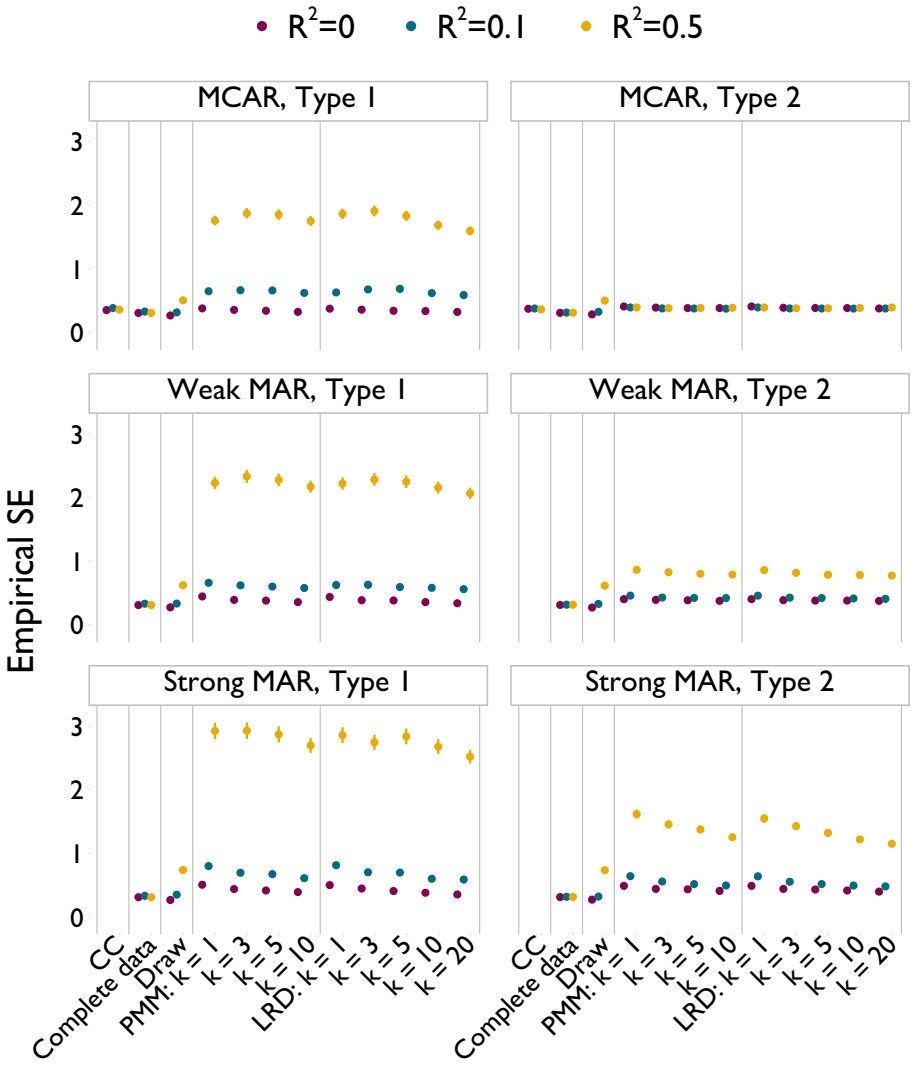


Figure 3.12: U-thwart: Coverage of nominal 95% CI (Error bars are ±2× Monte Carlo standard errors)

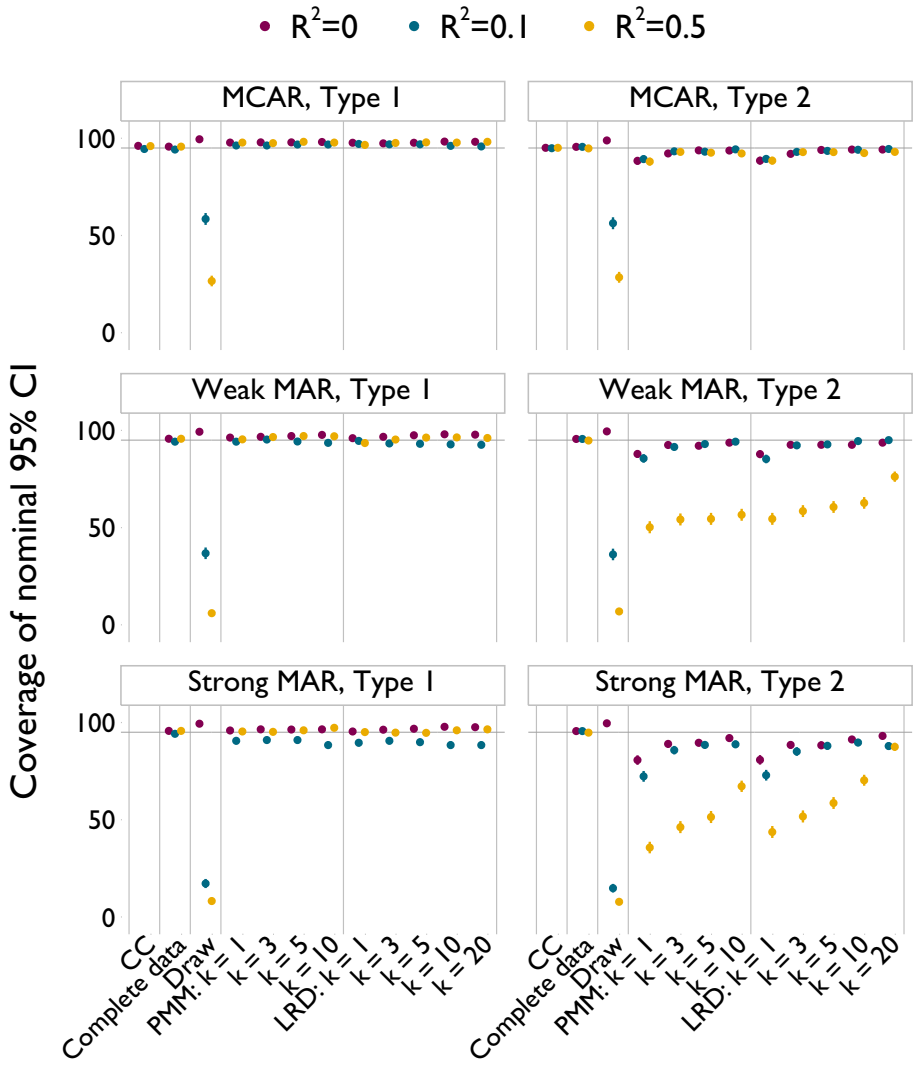
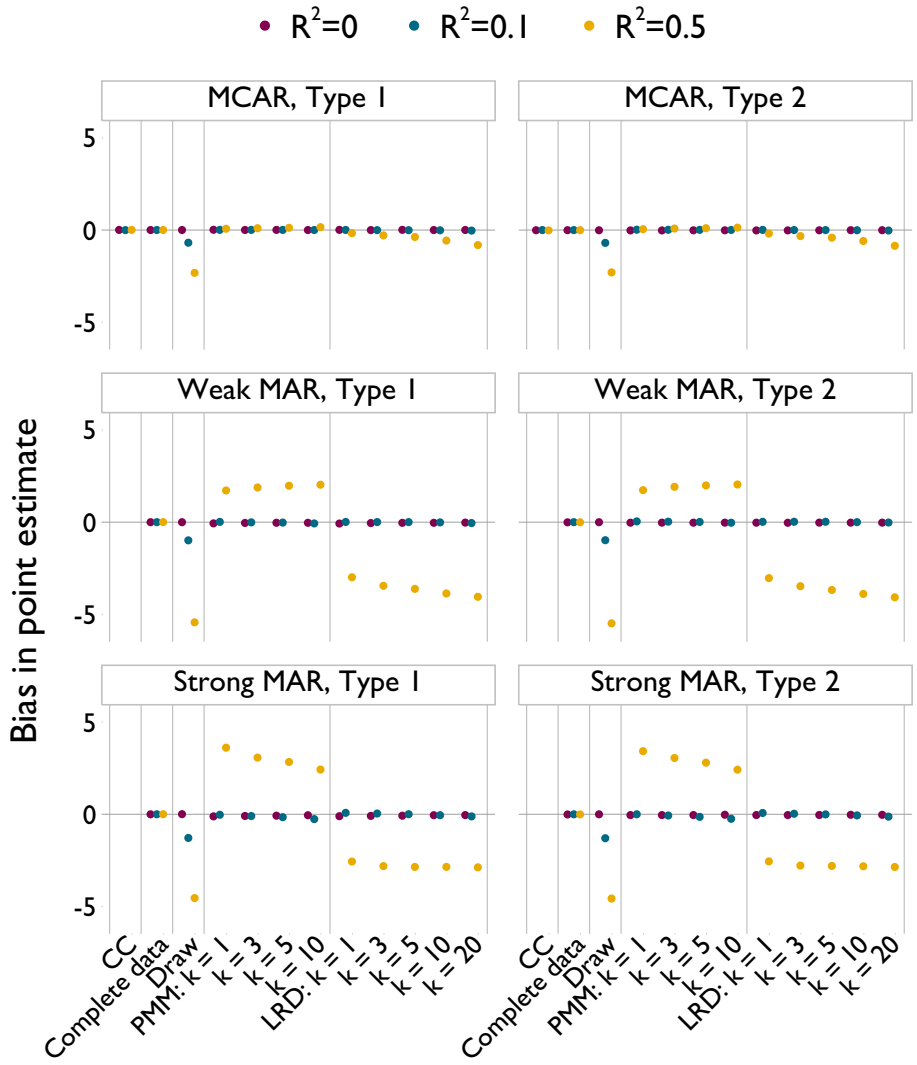


Figure 3.13: J-thwart: Bias in point estimates (Error bars are ±2× Monte Carlo standard errors)

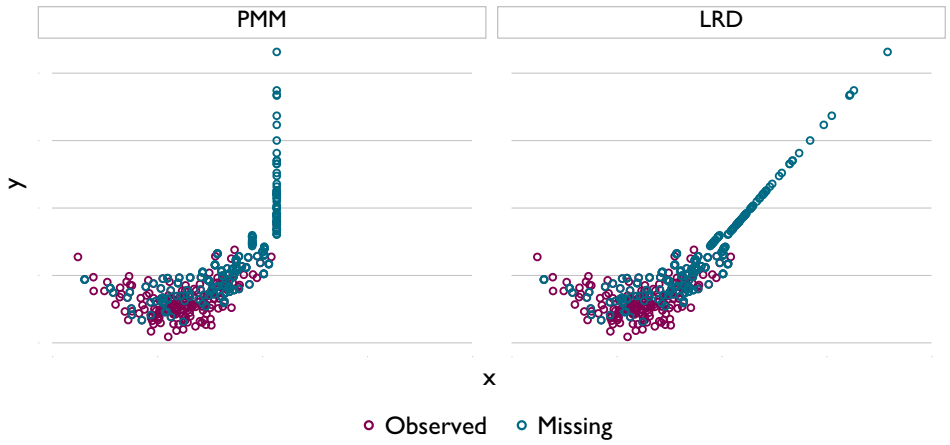


only exception being under MCAR. Under MAR, coverage is almost always below 95%. Again, the poorest performance occurs when $R^2 = 0.5$ and this is mainly due to the bias. $R^2 = 0$ and 0.1 give coverage which is relatively much higher but often still well below 95%. LRD gives consistently slightly better coverage than PMM with a corresponding k.

3.2.2 *J-thwart results*

Bias in J-thwart is shown in figure 3.13. Interestingly, this is not just a slightly attenuated version of U-thwart as might be expected. The bias in posterior draws is very similar, but there are few similarities to U-thwart when it comes to the matching methods. There seems to be no difference between type 1 and 2 matching, which may be because there are fewer cases than

Figure 3.14: J-thwart: understanding the biases for PMM and LRD



U-thwart where the sign of $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}^*$ differ (see figure 3.9). The main contrast is now between LRD and PMM. Bias is seen to be close to 0 for $R^2 = 0$ and 0.1 but with $R^2 = 0.5$ point estimates are biased upwards for PMM by up to four (strong MAR), and downwards for LRD by as much as five (weak MAR). Under weak MAR, this bias increases with k for both PMM and LRD, as well as for LRD under strong MAR.

Figure 3.14 helps to understand the biases observed in figure 3.13. (The simulation scenario corresponds to the bottom right panel of figure 3.13.) Observed (purple) values are plotted beneath imputed (blue) values for PMM (left panel) using $k = 1$ and type 2 matching. PMM cannot impute outside the range of observed data, and the censoring of imputed values at $(x_{h\max})$ means an upwards bias is introduced. For LRD (right panel), imputed values lie parallel to $\hat{\alpha}_1$, the slope of the imputation model. Beyond the range of observed values, this linear function leads to attenuation of the curve, creating downward bias in the coefficient for x^2. Results are not shown for type 1 matching because of the close similarity.

Type 1 and 2 matching again give very similar results in terms of empirical standard errors (figure 3.15). The main differences are again for cases where $R^2 = 0.5$. LRD has larger standard errors than PMM under MCAR, and increasing k causes the standard errors to increase further. Under weak MAR, the standard errors for PMM and LRD increase dramatically for $R^2 = 0.5$ but PMM still returns smaller SEs than LRD. Under strong MAR, standard errors for PMM become larger than those for LRD. Under MAR larger values of k reduces the empirical standard error.

Although there are few apparent differences between type 1 and 2 matching in terms of bias and standard errors, there are more pronounced differences in coverage rates (figure 3.16). In particular, type 2 matching seems to be a better choice for LRD, while type 1 matching performs better for PMM. It is worrying that when $R^2 = 0.5$ the coverage is consistently well below 95% and in any given set up there may be no method which gives coverage close to 95%.

Figure 3.15: J-thwart: Empirical standard errors (Error bars are ±2× Monte Carlo standard errors)

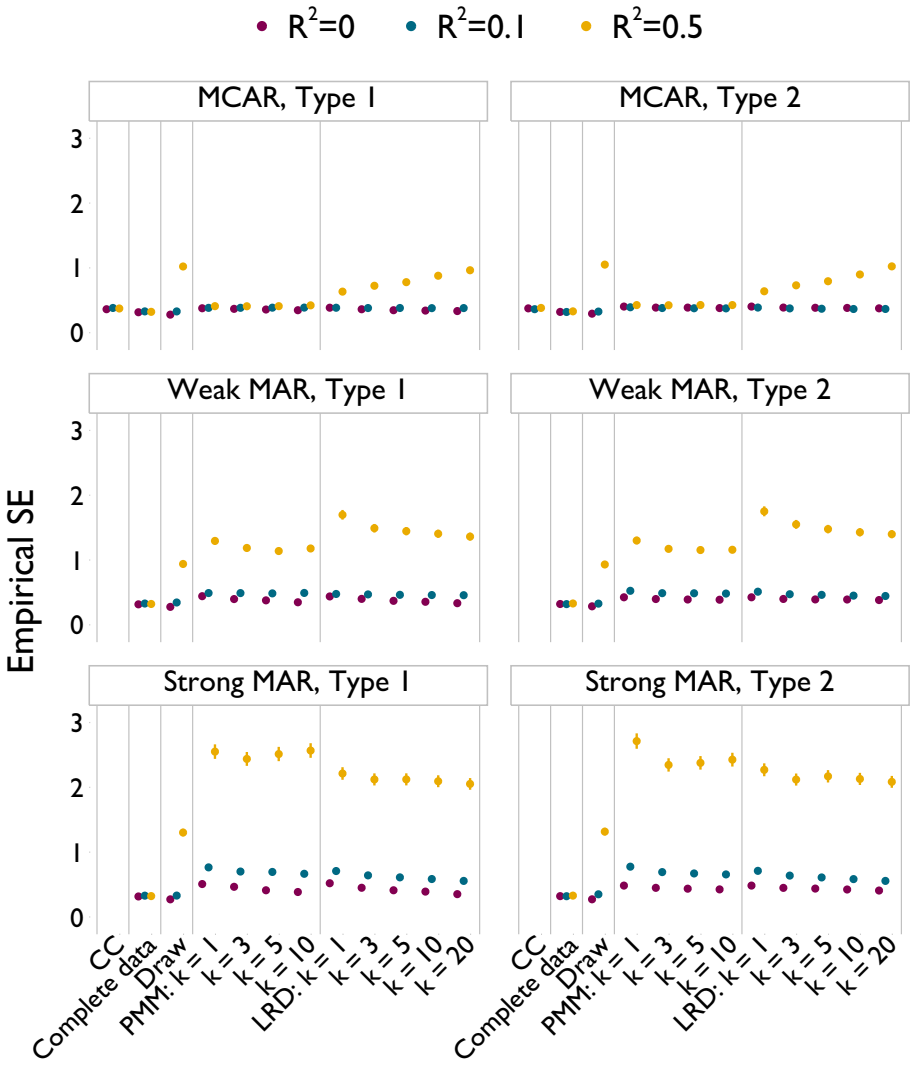
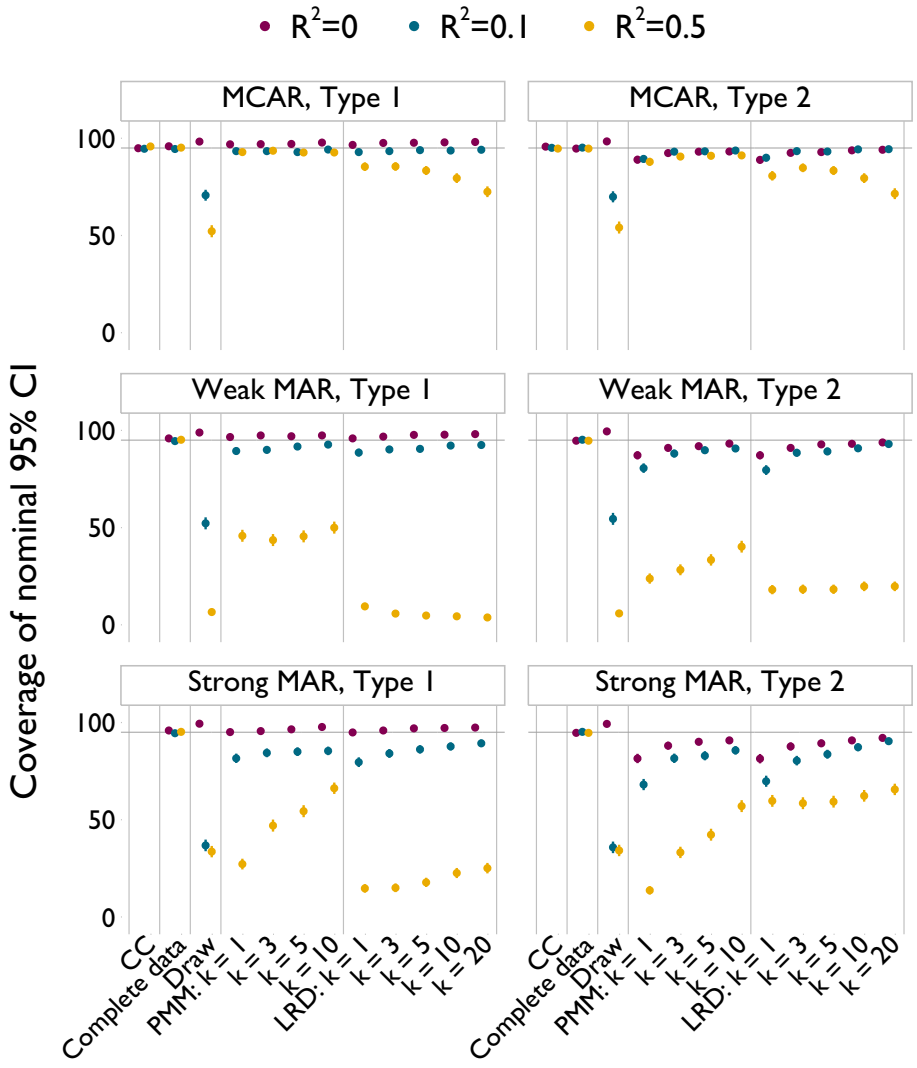


Figure 3.16: J-thwart: Coverage of nominal 95% CI (Error bars are ±2× Monte Carlo standard errors)



3.2.3 *U-thwart and J-thwart conclusions*

The results of these simulation studies are awkward: it is not easy to say that a certain value of k, a certain type of matching and one of PMM or LRD are the best imputation strategies to use in practice.

Choice of a method depends heavily on:

- The strength of association in the complete data and
- The missingness mechanism.

Both are unknown, but researchers can make educated guesses through knowledge of the study design and subject area, and inspection of the observed data.

On balance, my preference is for PMM with $k = 10$ and type 1 matching.

4 *Multivariable simulation: imputation transformations, PMM and LRD*

4.1 MOTIVATION

The above studies are useful but the simple designs do not reflect situations typically dealt with by applied statisticians. The following simulation study is designed with the aim of investigating imputation by PMM and LRD in a more realistic scenario.

A further complication considered considered here is that an analysis model may not involve fitting untransformed continuous covariates $\boldsymbol{x}$ but use a transformation of the covariates $\boldsymbol{x}$, denoted $\boldsymbol{g}(\boldsymbol{x})$. There may also be different transformations $\boldsymbol{f}(\boldsymbol{x})$ towards marginal multivariate normality.

The simulation procedure involves generating $\boldsymbol{f}(\boldsymbol{x})$ as (marginal) multivariate normal, and then simulating binary outcome as linear-logistic in $\boldsymbol{g}(\boldsymbol{x})$. Three potential imputation strategies which may be considered are[6]:

1. Impute untransformed $\boldsymbol{x}$ using a linear imputation model and passively impute $\boldsymbol{g}(\boldsymbol{x})$.
2. Impute $\boldsymbol{f}(\boldsymbol{x})$ using a linear imputation model and passively impute $\boldsymbol{g}(\boldsymbol{x})$. This strategy aims to impute from the correct marginal distribution[6].
3. Impute $\boldsymbol{g}(\boldsymbol{x})$ directly using a linear imputation model. This may mean the marginal distributions are poorly approximated, but aims for compatibility with the analysis model[64, 65, 22].

The study is designed such that all three imputation approaches are misspecified. The aim is to compare the covariate-transformation strategies, and to investigate whether the use of posterior draws, PMM or LRD is most robust.

4.2 DESCRIPTION OF TRAUMA DATA

The motivating example comes from the analysis of a trauma registry dataset reported by Stanworth et al[15]; the data were introduced in chapter 1. The authors aimed to produce a prognostic model to predict the probability of a patients requiring 'massive transfusion' ($\geq$ 10 units of packed red blood cells). Data were collected from five geographical locations: London, Oslo, Germany, Amsterdam and San Francisco, resulting in a dataset of 5,693 patients. Candidate predictors were those available on, or shortly after, arrival to emergency department. A good model would enable trauma departments to notify blood banks early if a patient is likely to

require large amounts of blood. The dataset includes patient age, sex, type of injury (blunt or penetrating), time from injury to arrival at emergency department, systolic blood pressure at admission, base deficit, prothrombin time and injury severity score. Of these variables, only sex and type of injury are categorical, while the remainder are continuous. Most variables have some values missing and the missingness pattern is non-monotone (see section 1.2.4). Table 4.1 summarises these variables.

Although missingness is not strictly monotone in the dataset, some of the commonest patterns are monotone. For the variables with the largest proportion of missing data – time to emergency dept., prothrombin time, base deficit and systolic blood pressure – the most common missingness patterns are represented in table 4.2. Although most variables had a small amount of missing data, this study only involves missingness in these four variables.

In the publication of this study[15], the $\boldsymbol{f}(\boldsymbol{x})$ transformations of covariates were taken for the imputation model, the aim being to produce imputed values with approximately the same marginal distributions as observed values. These same transformations were used in the analysis model because of concerns about incompatibility under any other transformation. However, the $\boldsymbol{f}(\boldsymbol{x})$ transformations would not necessarily be the strongest predictors of response and so may reduce the prognostic ability of the model.

4.3 SIMULATION PROCEDURES

Imputation by posterior draws, PMM with type 1 matching and $k = 10$, and LRD with type 2 matching and $k = 20$ are compared. The values of k and matching types chosen are used because on balance these seem to give the best results in J-thwart. Note that any differences between methods cannot then be ascribed to PMM or LRD because they are confounded by k and the type of matching.

The parameters from a complete data analysis of the trauma dataset are unknown. The first imputed dataset used in the published analysis is therefore used as one possible representation of the complete data, taken as the basis of this simulation study, providing what are regarded as the complete data parameters for simulating data. For the covariates listed in the top section of table 4.1, a transformation towards marginal normality was taken (using Stata's `lnskew0` command) giving $\boldsymbol{f}(\boldsymbol{x})$. The correlation matrix $\hat{\boldsymbol{\Sigma}}$ for $\boldsymbol{f}(\boldsymbol{x})$ was then estimated (table 4.3).

A multivariable fractional polynomial (MFP) logistic regression model was fit to the complete dataset[13], allowing a single function of each covariate, giving the transformations $\boldsymbol{g}(\boldsymbol{x})$ which best predict massive transfusion. Table 4.1 shows the specific transformations used in $\boldsymbol{f}(\boldsymbol{x})$ and the powers used in $\boldsymbol{g}(\boldsymbol{x})$. Prothrombin time and base deficit have powers for $\boldsymbol{g}(\boldsymbol{x})$ that are different to both $\boldsymbol{x}$ and $\boldsymbol{f}(\boldsymbol{x})$, along with non-trivial proportions of missing data (see table 4.1). Estimated analysis-model coefficients for these variables are of primary interest in this study. Prothrombin time will be particularly interesting as the two transformations are nothing like each other (in that if $\ln(x)$ is normal then x^{-2} will be severely skewed), while the logarithmic and square-root transformations for used base deficit are less dissimilar. Figure 4.1 shows the distributions of three transformations for prothrombin time and base deficit. Note that when variables have a large coefficient of variation all transformations would be essentially the same.

Table 4.1: Description of variables in the trauma registry data. Recall that $\boldsymbol{f}(\boldsymbol{x})$ denotes the normalising transformation and $\boldsymbol{g}(\boldsymbol{x})$ denotes the fractional polynomial transformation used by the analysis model.

Variable, x	Type	Frequency missing *	$f(x)$	$g(x)$ exponent	Mean (SD)	$\hat{\beta}$ (SÊ)
Age in years	Continuous	24 (0.4%)	$\ln(x + 36.9)$	1	40 (20)	0.006 (0.003)
Sex	Categorical	0	–	–	–	−0.04 (0.12)
Injury type	"	23 (0.4%)	–	–	–	0.81 (0.17)
Time to emergency dept.	Continuous	2396 (42%)	$\ln(x + 11.5)$	1	66 (39)	0.001 (0.001)
Systolic blood pressure	"	425 (7%)	–	1	126 (29)	−0.018 (0.002)
Base deficit	"	868 (15%)	$\ln(x + 7.6)$	0.5	3.2 (4.9)	2.09 (0.24)
Prothrombin time	"	1648 (29%)	$\ln(x - 9.0)$	−2	16.2 (7.7)	−0.032 (0.003)
Injury severity score	"	86 (2%)	$\ln(x + 12.0)$	0.5	19.8 (15.4)	1.56 (0.12)
Massive transfusion	Categorical	0				
Death	Categorical	0				

* Total number of incomplete cases = 3196 (56%)

Table 4.2: Six most common patterns of missing values in trauma registry data

Systolic blood pressure	Base deficit	Prothrom-bin time	Time to emergency dept.	Frequency
✓	✓	✓	✓	2,451 (43%)
✓	✓	✓	.	1,133 (20%)
✓	✓	.	✓	562 (10%)
✓	✓	.	.	461 (8%)
✓	.	.	.	323 (6%)
.	.	.	.	152 (3%)

Figure 4.1: Kernel densities for base deficit and prothrombin time, shown on the x, $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{g}(\boldsymbol{x})$ scales, in a single simulated dataset.

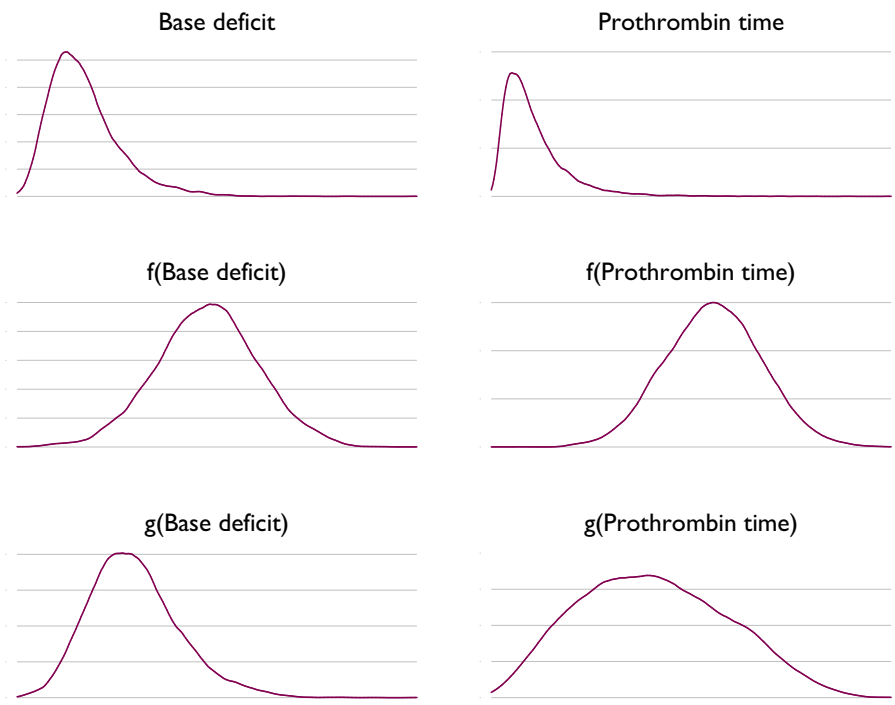


Table 4.3: Correlation matrix $\mathbf{\Sigma}$ for trauma registry covariates

	f(Age)	Sex	f(Time to ED)	Penetrating injury	f(ISS)	f(SBP)	f(Base deficit)	f(Prothrombin time)
f(Age)	1.00							
Sex	-0.07	1.00						
f(Time to ED)	-0.00	0.01	1.00					
Penetrating injury	-0.09	0.10	-0.18	1.00				
f(ISS)	0.19	-0.01	0.13	-0.12	1.00			
f(SBP)	0.11	0.07	-0.01	-0.08	-0.28	1.00		
f(Base deficit)	0.00	-0.03	0.01	0.11	0.38	-0.33	1.00	
f(Prothrombin time)	0.02	-0.04	0.10	-0.02	0.34	-0.28	0.33	1.00

In generating the covariates, resampling and simulation approaches were both considered. Simulation was favoured due to the potential for ties in δ_h for PMM and LRD under the resampling approach. Resampling with replacement might also have given PMM and LRD with type 2 matching an unfair advantage since observations to be imputed would often be matched to their 'original' observations, as noted by Tang et al.[46]. Simulation has an added advantage that the data generating model for the covariates is known and so the distribution of $\boldsymbol{f}(\boldsymbol{x})$ is multivariate normal.

For each replication, $\boldsymbol{f}(\boldsymbol{x})$ are drawn from a multivariate normal distribution with correlation $\boldsymbol{\Sigma}$ (see table 4.3). $\boldsymbol{x}$ and $\boldsymbol{g}(\boldsymbol{x})$ are then calculated. Binary outcome, y_i, is simulated as 1 if expit($\boldsymbol{\beta}[\boldsymbol{g}(\boldsymbol{x}_i)]$) > U and 0 otherwise. U is a uniform (0,1) random variable and $\boldsymbol{\beta}$ are parameters from the MFP logistic regression model fit to the complete cases, which was used to determine $\boldsymbol{g}(\boldsymbol{x})$.

The correctly specified imputation model for covariates might appear to be to impute $\boldsymbol{f}(\boldsymbol{x})$ from a multivariate normal model, because covariates were generated on the $\boldsymbol{f}(\boldsymbol{x})$ scale. However, note that because the MFP model includes non-zero regression coefficients imputation of covariates must condition on y[66]. Although $\boldsymbol{f}(\boldsymbol{x})$ is generated from a multivariate normal marginal distribution this does not imply multivariate normality conditional on y (this would not even be true if be if y were generated including $\boldsymbol{f}(\boldsymbol{x})$ linearly as covariates). The correct specification of the imputation model is therefore not standard.

In keeping with the earlier simulation studies, three different strengths of association are involved: all coefficients equal to the 'truth' (table 4.1); all coefficients half the magnitude; and all double the magnitude. For the 'halved' and 'doubled' scenarios the intercept term is iterated until the correct proportion of $y = 1$ is achieved.

As with the earlier simulation studies, three different missingness mechanisms are considered: MCAR, the observed strength of MAR and a stronger version of MAR, where the coefficients predicting missingness are all doubled. The dataset had a complex missing data pattern which the simulations aim to reproduce. MCAR is therefore incorporated into simulations by merging the observed matrix of missingness indicators R with the simulated data and deleting any values where the indicator is 0. This preserves the observed *pattern* of missingness and does not depend on any observed or unobserved variables, as required by MCAR.

As shown by table 4.2, some of the commonest missingness patterns in the trauma dataset are monotone. The imputed dataset used as a possible representation of the complete data was merged with a dataset containing indicators for whether a variable was actually observed. These indicators were fit as the response in separate logistic regression models and estimates saved. To simulate missing at random, response for each incomplete variable was simulated using parameters estimated from its logistic regression on outcome and fully observed covariates. Although there are alternative ways to simulate MAR (see for example [67] and [68]), these methods are not used here because in calibrating this simulation study the method used gave an adequate representation of the missingness patterns in the observed data.

Stata version 11.2 is used for all aspects of these simulations. The `ice` command is used to impute missing values and `mim` to analyse the m completed datasets. For each simulation scenario 1,000 replications are used.

The simulation process is as follows. For each of the 1,000 datasets of covariates simulated, response is simulated according to the three strengths of association. Complete data analysis is run on each. Three missingness mechanisms are then imposed on the dataset with the observed strength of association, and the 'observed' MAR mechanism on the doubled- and halved-association datasets. Complete cases analysis is run on these five datasets. For each of these incomplete datasets, MI is performed in nine ways: three transformations $\boldsymbol{f}(\boldsymbol{x})$, x and $\boldsymbol{g}(\boldsymbol{x})$ are then performed, and for each, imputation by posterior draws, type 1 PMM with $k = 10$, and type 2 LRD with $k = 20$. Each imputation involves 10 cycles of chained equations and $m = 5$ imputations. Passive imputation of $\boldsymbol{g}(\boldsymbol{x})$ is used where the active imputation is on $\boldsymbol{f}(\boldsymbol{x})$ or $\boldsymbol{x}$. The analysis model is fitted to each imputed dataset and estimates are combined using Rubin's rules[1].

As with the above simulations, point estimates ±2×Monte Carlo error are plotted graphically across the imputation strategies and methods.

In a change from the various univariable studies, per cent bias is now presented. This is favoured because it gives a common scale for the two variables, which have different true coefficients. In the univariable studies this was not possible because a zero association case was included, for which per cent bias would be infinite for all methods and settings.

4.4 TRAUMA STUDY RESULTS

Analysis of complete data and of complete cases is unbiased for both prothrombin time and base deficit. With multiple imputation there is some bias, of varying magnitude, for the two variables: MI introduces more bias for prothrombin time than for base deficit. The magnitude of bias is proportional to the strength of association, with stronger associations displaying greater relative bias. Bias is worst under MCAR, and shrinks under observed MAR and again under strong MAR. This seems to be because missing values occur more in the tails and the imputed values overstate the curvature of the x^{-2} function.

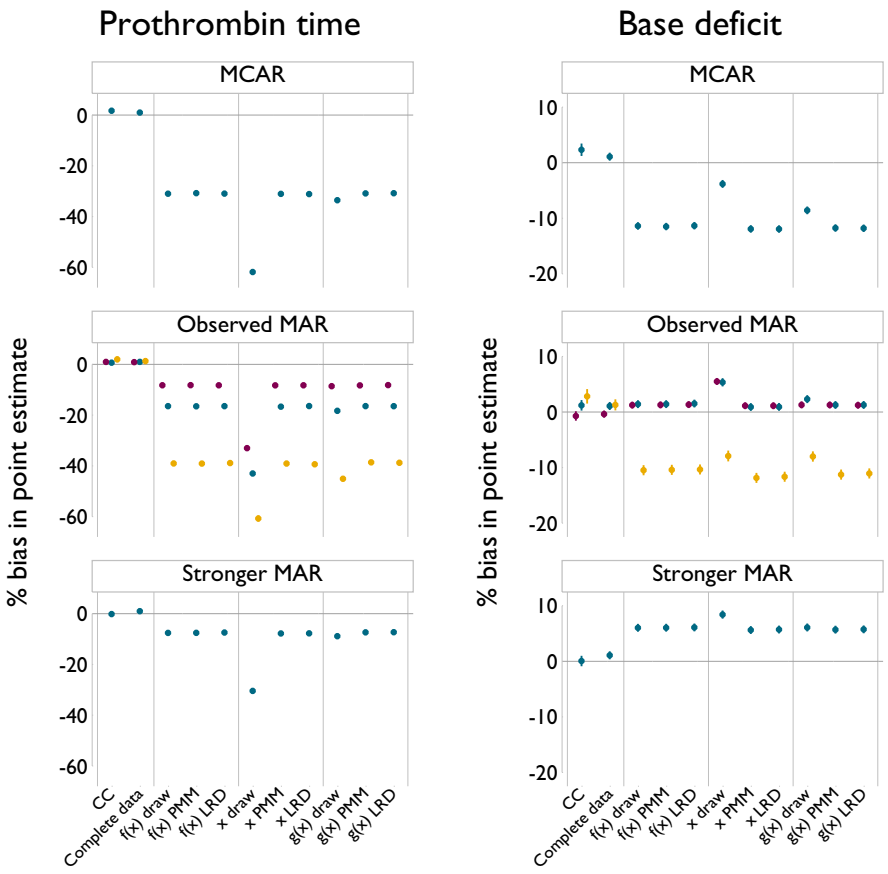
There are surprisingly few differences between the various transformations or imputation methods, with the exception of posterior draws on $\boldsymbol{x}$. This strategy is erratic: for prothrombin time this is the most biased imputation strategy; for base deficit it returns lower bias than other methods under MCAR and observed MAR, but higher under strong MAR. There are no important differences between PMM and LRD. As with previous results, there is a large difference between weaker and stronger associations, where the relative bias is proportional to the true strength of association.

The complete cases standard error is in most cases larger than any other method, as might be expected. Complete data would be expected to have the smallest empirical standard errors, but for both variables most of the imputation methods are seen to have lower empirical standard errors. This is due to their bias towards the null seen in figure 4.2.

Again, there is little to choose between the three transformations and imputation methods. For prothrombin time, posterior draws on $\boldsymbol{x}$ are inefficient, but this effect is much smaller for base deficit.

Empirical standard errors are very similar across missing data mechanisms. Standard errors across associations are not compared; a stronger association will return a more variable

Figure 4.2: Trauma study: Per cent bias in point estimates. Observed association results are in blue, halved associations in purple, and doubled associations in orange. Error bars are ±2× Monte Carlo standard errors.



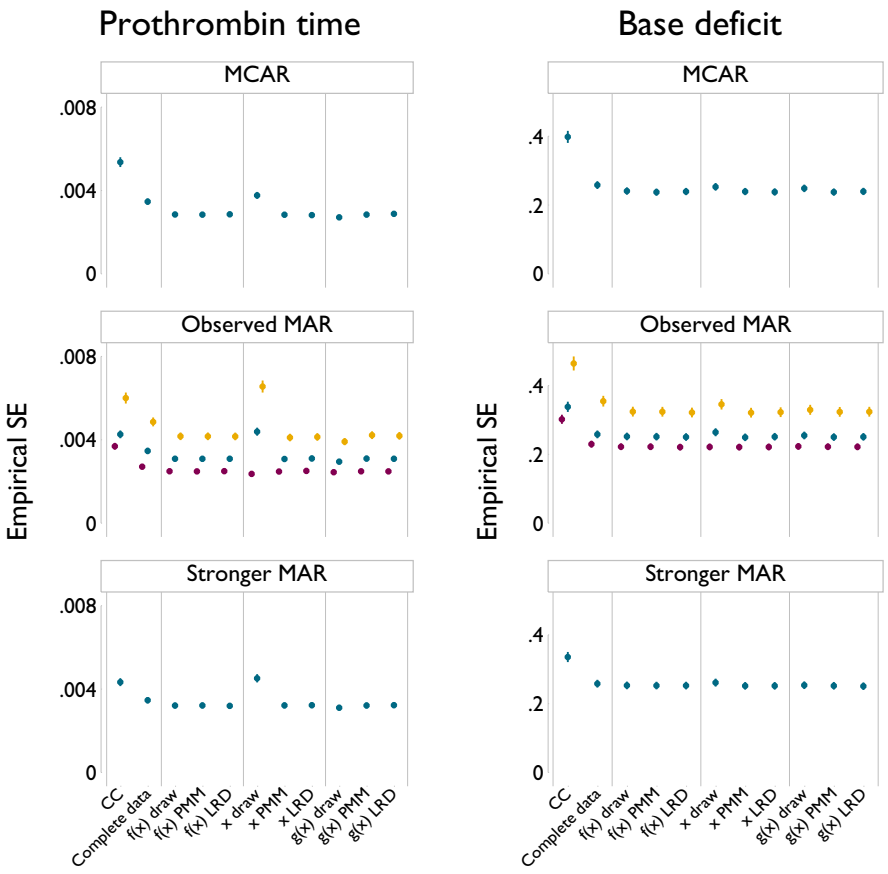
parameter estimate.

Coverage of confidence intervals is nominal for both complete cases and complete data for both variables across all settings.

Coverage after MI is less impressive. Due to the large biases for prothrombin time, coverage of confidence intervals is generally lower than 95%; under MCAR it is pitifully low for all imputation strategies. Imputation of $\boldsymbol{x}$ by posterior draws remains the most erratic method – note how the half and double strengths of association have worse coverage than the observed strength. For other methods, coverage tends to be slightly low at best. Again, coverage performance is similar across all transformations and imputation methods except for posterior draws of $\boldsymbol{x}$.

For base deficit the biases were smaller than for prothrombin time, and results for coverage after MI tend to be better. Coverage is a little too low for the various imputation approaches under some settings. Posterior draws on $\boldsymbol{x}$, always gives nominal or high coverage; the method is biased and so to achieve this coverage the Rubin's rules variance has to be overestimated. Posterior draws on $\boldsymbol{g}(\boldsymbol{x})$ also achieves slightly better coverage tha other methods. Coverage

Figure 4.3: Trauma study: Empirical standard errors. Observed association results are in blue, halved associations in purple, and doubled associations in orange. Error bars are ±2× Monte Carlo standard errors.



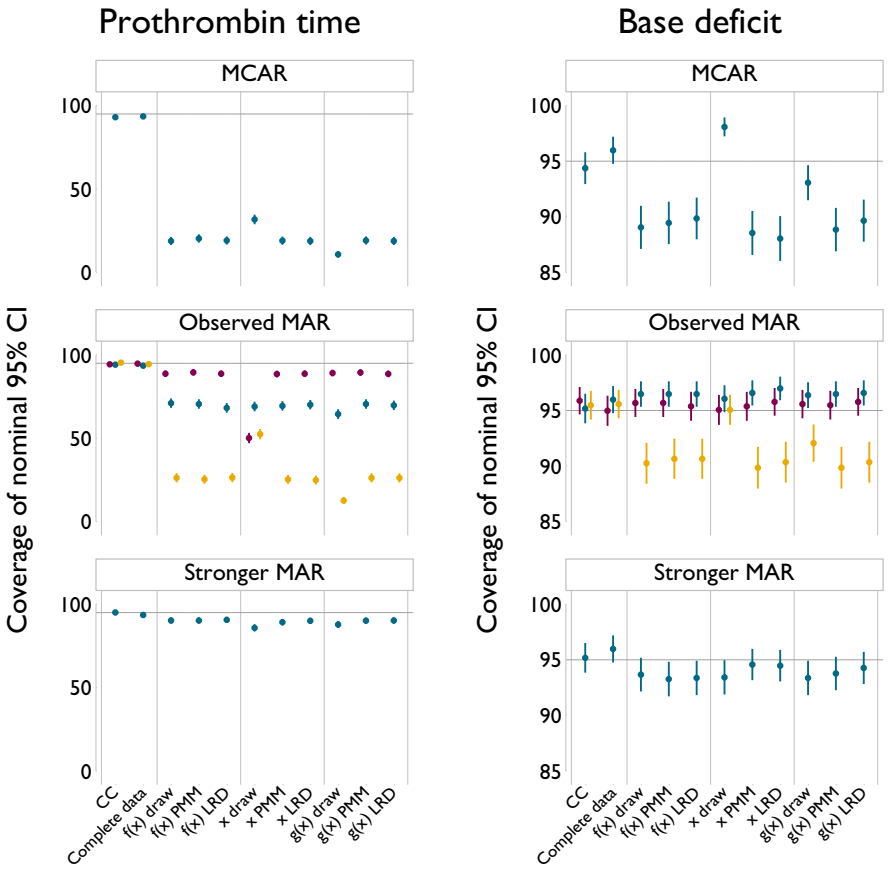
is at its worst under (1) double-strength association, and (2) MCAR. The poorest coverage corresponds to situations with most bias. For base deficit the safest imputation approach in terms of coverage seems to be to impute $\boldsymbol{x}$ or $\boldsymbol{g}(\boldsymbol{x})$ by posterior draws.

4.4.1 *Note on complete-case analysis*

The results for complete cases are initially surprising and confusing: with a complex missing data pattern and a large proportion of missing values, it is unbiased with nominal coverage rates for all settings (but with low efficiency).

This result is in fact not by chance, but is specific to the general simulation scenario: the probability of being a complete case depends on outcome, and the analysis model is logistic regression. This mimics case–control sampling. It is well known that the intercept for logistic regression is biased in case–control studies, but the coefficient for the exposure of interest is not, meaning the disease–outcome association can be estimated consistently even though the

Figure 4.4: Trauma study: Coverage. Observed association results are in blue, halved associations in purple, and doubled associations in orange. Error bars are ±2× Monte Carlo standard errors.



incidence of disease cannot. In this study an additional complexity exists because the probability of being a complete case further depends on fully observed covariates. This is likely to induce bias in the complete-cases coefficients relating to the fully observed covariates, but not those relating to partially observed covariates.

Although this is a quirk relating to logistic regression, it is worth noting for any studies similar in design, assuming only the incomplete covariates are of substantive interest. Because estimates from complete cases are consistent, an MI-based estimate that is different may rationally be taken as somewhat biased. The imputation model may then need calibrating to provide a result close to complete cases.

4.5 TRAUMA STUDY CONCLUSIONS

The differences between results for base deficit and prothrombin time are very interesting. There are large disparities which are presumably related to the specific g-transformation (the relationship between $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{x}$ was similar for both). For base deficit this is $\sqrt{x}$, which is not

dissimilar to its f-transformation, $\ln(x)$. However for prothrombin time the g-transformation is x^{-2}, which is an extremely skewed distribution if $\ln(x)$ follows a normal distribution.

Despite the arguments for PMM and LRD being potentially useful in this sort of scenario, the results of this study seem to go against intuition: posterior draws are rarely much worse than an equivalent PMM or LRD method and can sometimes provide better inference. This is a surprising contrast to the earlier simulation studies described in section 3.2.2, where both methods appeared to be more robust than posterior draws when the imputation model was misspecified and incompatible with the analysis model.

5 *Multiple imputation for an incomplete covariate which is a ratio*

The following chapter has been published as a research article by *Statistics in Medicine*[16]. The authors are myself, my supervisors and my advisors. This chapter is the same as the accepted manuscript and so should stand alone. The notation is similar to the rest of the thesis but note that p here denotes the number of covariates in the analysis model; this is different to its use in chapter 6.

5.1 ABSTRACT

We are concerned with multiple imputation of the ratio of two variables, which is to be used as a covariate in a regression analysis. If the numerator and denominator are not missing simultaneously it seems sensible to make use of the observed variable in the imputation model. One such strategy is to impute missing values for the numerator and denominator, or the log-transformed numerator and denominator, and then calculate the ratio of interest; we call this 'passive' imputation. Alternatively, missing ratio values might be imputed directly, with or without the numerator and/or the denominator in the imputation model; we call this 'active' imputation. In two motivating datasets, one involving body mass index as a covariate and the other involving the ratio of total to high density lipoprotein (HDL) cholesterol, we assess the sensitivity of results to the choice of imputation model, and as an alternative explore fully Bayesian joint models for the outcome and incomplete ratio. Fully Bayesian approaches were unusable in both datasets due to computational problems when estimating them in WinBUGS. In our first dataset, multiple imputation results are similar regardless of the imputation model; in the second, results are sensitive to the choice of imputation model. Sensitivity depends strongly on the coefficient of variation of the ratio's denominator. A simulation study demonstrates that passive imputation without transformation is risky because it can lead to downward bias when the coefficient of variation of the ratio's denominator is larger than about 0.1. Active imputation or passive imputation after log-transformation are preferable.

5.2 INTRODUCTION

Missing values of covariates are a common problem in regression analyses. Missing data are classified as being *missing completely at random* (MCAR) if missingness does not depend on observed or unobserved data, *missing at random* (MAR) if missingness does not depend on unobserved data given observed data, or *missing not at random* (MNAR) if missingness depends

on missing data even given the observed data[2]. Amongst methods that attempt to deal with missing data, rather than discarding them, multiple imputation (MI) can provide valid inference under MAR, and has become popular in practice since its inception over 30 years ago[69].

Briefly, MI works as follows. Missing values are replaced with imputed values, drawn from their posterior predictive distribution under a model given the observed data. We term this model the *imputation model*. The process is repeated $M > 1$ times, giving M imputed datasets with no missing values. Each imputed dataset is analysed using the model that would have been used had the missing values been observed. We call this model the *analysis model*. The M estimates of each parameter of interest are then combined using 'Rubin's rules'[1]. When the imputation model is correctly specified Rubin's rules can provide standard errors and confidence intervals that fully incorporate uncertainty due to missing data.

MI is an attractive tool for analyses with missing data: the nuisance issue of modelling missing data is neatly separated from the analyses of substantive interest; the imputation model can make use of auxiliary variables that it would be undesirable to include as covariates in the analysis model (such as post-baseline measurements in a randomised controlled trial); the same M imputed datasets can be used for a variety of substantive analyses; and the imputation model can be tailored to reflect possible departures from MAR, which is helpful for sensitivity analysis.

Ratios are commonly used as covariates in regression based analyses; examples are body mass index (BMI = Weight in kg ÷ (Height in m)^2)[17], waist–hip ratio[70], urinary albumin-to-creatinine ratio (Albumin concentration in mg/g ÷ Creatinine concentration in mg/g)[71], and what we refer to as 'cholesterol ratio' (Total cholesterol in mg/dL ÷ HDL in mg/dL)[20].

An individual's ratio measurement may be missing for one of three reasons:

1. The denominator is missing
2. The numerator is missing
3. Both components are missing

For both 1 and 2 the ratio is semi missing rather than fully missing; that is, one of the two components is observed. Ratio missingness due to more than one of these reasons for different observations in the same dataset means it is not obvious how best to impute the ratio. A mixture of reasons 1 and 2 is particularly awkward.

One reasonable question at this stage is, 'Why use a ratio covariate?' There are mathematical arguments against their use[72]. Senn and Julious claim ratios are always poor candidates for parametric analysis unless the components, and therefore the ratio, follow a lognormal distribution, or the ratio's coefficient of variation is small[73]. We make three points. First, applied researchers *do* use ratios and we are unlikely to persuade them to stop, especially since the use of certain ratios is well established; we should be pragmatic and try to guide practitioners on how to analyse datasets involving incomplete ratio covariates. Second, arguments against ratios assume that a ratio is not the correct functional form for a covariate, but it may be. Third, ratios are not used by accident: a ratio may be of genuine substantive interest when its separate components are not. For example, BMI is widely used because it measures weight-for-height

and as such is regarded as a proxy measure of body fat. Substantive interest is in the influence of body fat on outcome, not weight or height. Weight alone may be considered a measure of body fat but BMI is measured with less error since it aims to remove the effect of height (although it may not do so completely or accurately). It is our opinion that when researchers propose a relationship they believe, such as the influence of a ratio on outcome, this should not be cast aside lightly. The substantive question should not be altered for statistical convenience unless we have little choice.

We assume the aims of analysis are unbiased estimation of a parameter describing the association between a ratio and some outcome, confidence intervals with the ascribed coverage and fully efficient parameter estimation. There may be other covariates in the analysis model, and primary interest may be in one of these, but the properties of the ratio parameter estimator are important nonetheless. There has been no previous methodological work on MI for a ratio covariate, although [6] and [26] allude to the issue, but practitioners are imputing ratio covariates nonetheless[19]. We aim to highlight issues with imputing an incomplete ratio covariate and to identify imputation strategies that are sensible practicable for applied statisticians.

Despite the positive features listed above, MI is not the only approach to dealing with missing covariates, nor is it necessarily the best approach for any given analysis. Joint models for the outcome and covariates may be superior because they make use of the full likelihood in a coherent way. In this paper, we also investigate results for fully Bayesian joint models.

The remainder of this paper is as follows. In section 5.3 we introduce and describe our two motivating datasets; in section 5.4 we consider candidate models for imputing incomplete ratios; section 5.5 presents two case studies, contrasting the different imputation models (for the datasets introduced in section 5.3); section 5.6 presents a simulation study in a simpler setting than our case studies; section 5.7 is a discussion.

5.3 DATASETS: AURUM AND EPIC-NORFOLK

For both of our datasets, regression analyses involving a ratio as a covariate have previously been published[17, 20]. The analysis models used in our example analyses are not the same as the original articles because (i) we want to keep the analysis models and imputation models relatively simple and (ii) we do not wish to make any substantive claims about these data. Therefore we have chosen to use analysis models resembling but not matching those used in the earlier publications[17, 20].

For both datasets the analysis model is the Cox model,

$$h_i(t \mid \boldsymbol{x}_i) = h_0(t)\exp\left(\sum_{c=1}^{p} \beta_c x_{ci}\right), \tag{5.1}$$

where $h_0(t)$ is the nonparametric baseline hazard function at time t, $h_i(t \mid \boldsymbol{x}_i)$ is the hazard for the ith individual and x_{ci} is the value of the cth covariate in the ith individual. Survival (or censoring) times are assumed to be fully observed.

Table 5.1: Aurum summary of covariates and of the analysis model and components of BMI; $n = 1,348$

	Covariate	Frequency missing (%)	Mean (SD) or frequency (%)
x_1	Age (years)	0 (0%)	37 (9)
x_2	Sex: male	0 (0%)	542 (40%)
x_3	Hæmoglobin (g/mL)	143 (11%)	11.4 (2.3)
x_4	*Viral load (copies per mL)	162 (12%)	4.8 (0.8)†
x_5	*CD4 count (cells per μL)	94 (7%)	8.9 (4.5)†
$x_6 = a_1/a_2$	BMI (kg/m^2)	381 (28%)	21.9 (4.9)
a_1	‡Weight (kg)	376 (28%)	58 (12)
a_2	‡Height (m^2)	275 (20%)	2.7 (0.3)†

*Transformation used for viral load is $\log_{10}(x_4)$; transformation used for CD4 count is $\sqrt{x_5}$. These are standard transformations in HIV research, and we use them in the imputation models and the analysis models.
† Summarised on transformed scale.
‡ Only enters into the analysis model via BMI

5.3.1 *The Aurum cohort*

The Aurum dataset comes from a South African cohort study of 1,350 HIV infected participants starting antiretroviral therapy. Participants were recruited from 27 centres in five provinces between February 2005 and June 2006, and followed to March 2007. Information was recorded on a range of baseline characteristics and participants were followed up for death. The aim of the work by Russell et al.[17] was to estimate the influence of hæmoglobin on mortality using a Cox model. 1,348 of the participants had a recorded time of death/censoring, with 185 deaths occurring within the follow-up time. We restrict our analysis to these 1,348 individuals.

The analysis model is (5.1) with $p = 6$, where $x_1, \ldots, x_6$ are age in years, sex, hæmoglobin in g/mL, viral load in copies per mL, CD4 count in cells per μL and BMI. Table 5.1 provides a summary of these covariates and of weight and height. Any transformation of the covariate used in the analysis model is given, and the transformed measure is summarised in the final column. Note that 381 (28%) patients are missing a weight and/or height measurement, but only five of these have height missing when weight is observed. Five of the covariates are continuous and one (sex, which is complete) is categorical. Hæmoglobin, weight, height^2 and BMI appear to be approximately normal on the transformed scale, while (log) viral load and (square root of) CD4 count do not. We focus on the estimation of β_3 and β_6, the log hazard ratios for hæmoglobin and BMI respectively (hæmoglobin was the focus of the original publication[17]).

5.3.2 *The Epic-Norfolk cohort*

The Epic (European Prospective Investigation Into Cancer and nutrition)-Norfolk study is a large cohort study designed to investigate the link between dietary factors and cancer. Dietary and non-dietary factors were collected at baseline and participants were followed up for cancer

Table 5.2: Epic-Norfolk summary of covariates of the analysis model and of components of cholesterol ratio; $n = 22,754$

	Covariate	Frequency missing (%)	Mean (SD) or frequency (%)
x_1	Age (years)	0 (0%)	59 (9)
x_2	Sex: male	0 (0%)	10,145 (45%)
x_3	Smoking status: ever smoked	0 (0%)	11,971 (53%)
x_4	Systolic blood pressure (mm Hg)	52 (<1%)	135 (18)
x_5	Diastolic blood pressure (mm Hg)	52 (<1%)	82 (11)
$x_6 = a_1/a_2$	Cholesterol ratio	2,155 (9%)	4.7 (1.6)
a_1	†Total cholesterol (mg/dl)	1,514 (7%)	6.2 (1.2)
a_2	†HDL (mg/dl)	2,155 (9%)	1.4 (0.4)

† Only enters into the analysis model via cholesterol ratio

and non-cancer outcomes. We use some of the non-dietary characteristics as covariates and time to death as the outcome.

The analysis model is (5.1) with $p = 6$, where $x_1, \ldots, x_6$ are age, sex, smoking status, systolic blood pressure, diastolic blood pressure and cholesterol ratio. These six covariates and total cholesterol and HDL are summarised in table 5.2; no transformations are used. In total, 2,155 (9%) participants are missing a total cholesterol and/or HDL measurement. Total cholesterol is always missing when HDL is missing. Incomplete covariates are all continuous and appear approximately normal, except for HDL, which is positively skewed. We focus on the estimation of β_6, the log hazard ratio for cholesterol ratio.

5.4 METHODS AND MODELS

5.4.1 *Model for analysis*

The analysis model is the Cox model (5.1) with p covariates $(x_1, \ldots, x_p)$ made up of the ratio $x_p = a_1/a_2$ and $p - 1$ other covariates $(x_1, \ldots, x_{p-1})$, which we denote $(\boldsymbol{z}, \boldsymbol{w})$ where $\boldsymbol{z}$ are incomplete and $\boldsymbol{w}$ are complete (in both example datasets we have $\boldsymbol{z}$ and $\boldsymbol{w}$).

5.4.2 *Models for missing data*

Candidate models for the covariates are listed in table 5.3 (note the *Label* column, which we henceforth use to refer to models). For MI the outcome must be explicitly included as a covariate in the imputation model[66]. In table 5.3 we denote outcome by $f(y_i)$. For the Cox model $f(y_i)$ involves a censoring indicator and the Nelson–Aalen estimate of the cumulative hazard function to the survival time (an approximation to the cumulative baseline hazard function $H_0(t)$[74]), included as separate covariates in the imputation model. When the analysis model is linear or logistic regression $f(y_i) = y_i$.

Table 5.3: Candidate imputation models for $\boldsymbol{x}_i$

Imputation model	Label	Relationship to analysis model
$(\boldsymbol{z}_i, x_{pi} \mid f(y_i), \boldsymbol{w}_i) \sim \text{MVN}$	M1	Compatible
$(\boldsymbol{z}_i, x_{pi}, a_{1i} \mid f(y_i), \boldsymbol{w}_i) \sim \text{MVN}$	M2	Semi-compatible
$(\boldsymbol{z}_i, x_{pi}, a_{2i} \mid f(y_i), \boldsymbol{w}_i) \sim \text{MVN}$	M3	Semi-compatible
$(\boldsymbol{z}_i, x_{pi}, a_{1i}, a_{2i} \mid f(y_i), \boldsymbol{w}_i) \sim \text{MVN}$	M4	Semi-compatible
$^{\dagger}(\boldsymbol{z}_i, a_{1i}, a_{2i} \mid f(y_i), \boldsymbol{w}_i) \sim \text{MVN}$	M5	Incompatible
$^{\ddagger}(\boldsymbol{z}_i, \ln(a_{1i}), \ln(a_{2i}) \mid f(y_i), \boldsymbol{w}_i) \sim \text{MVN}$	M6	Incompatible

† Passive imputation of $x_{pi} = \frac{a_{1i}}{a_{2i}}$ is required
‡ Passive imputation of $x_{pi} = \exp[\ln(a_{1i}) - \ln(a_{2i})]$ is required

5.4.3 *Compatibility in relation to active and passive imputation*

Multiple imputation can provide an approximation to fitting a joint model if the models for imputation and analysis are compatible[75], where a joint model may be either maximum likelihood or Bayesian (if the joint model is Bayesian, compatibility also requires that priors are non-zero over the entire parameter space). Considering whether or not the models M1–M6 are compatible with the analysis model helps us to formulate hypotheses and understand future results.

By 'compatible', we mean a joint model exists that implies both the imputation model and the analysis model as conditional models. This does not mean the joint model is correct, but that the analysis model and imputation model are both implied by it and so the multiple imputation procedure is coherent. Appendix C.1 describes how to tell if models are compatible, and works through two examples of imputation models where one is compatible and the other is not (C.1.1 and C.1.2 respectively).

Non-compatibility of models is not always problematic; Meng[22] and Rubin[23] have both shown that there can be some *benefit* to using imputation models that correctly draw on information not used by the analysis model. Collins, Schafer and Kam demonstrate via simulation that auxiliary variables (i.e. variables that are in the imputation model but not the analysis model) are unlikely to be harmful, and may be of benefit by making the MAR assumption more plausible, while 'restrictive' imputation strategies can lead to problems[24].

We therefore distinguish between two types of non-compatibility: if there is a special case of the imputation model that is compatible with the analysis model, as when it includes auxiliary variables, then the imputation model is termed 'semi-compatible' (following Liu et al[25]); otherwise the imputation model is simply termed 'incompatible'. In previous work, imputation models that are compatible or semi-compatible appear to perform well even when misspecified[65, 64], but this is not necessarily true for imputation models that are incompatible[64, 24]. We hypothesise that imputation models that are compatible or semi-compatible will be more robust to modest degrees of misspecification than models that are incompatible.

Imputation of a ratio is done either actively or passively. Of the imputation models listed

in table 5.3, only M1 is compatible with the analysis model. Of the remaining models M2–M4, which use active imputation, are semi-compatible because they include a_1 and/or a_2, which do not appear in the analysis model, as auxiliary variables in the imputation model; models M5 and M6, which use passive imputation, are incompatible with the analysis model because x_p is present in the analysis model but not in the imputation model, while a_1 and a_2 are present in the imputation model but not in the analysis model. We expect models M5 and M6 to be prone to bias and poor coverage, despite making use of all the observed data when imputing the ratio.

5.4.4 *Motivation for missing data models*

The choice of a model listed in table 5.3 might be motivated by the way it makes use of observed information in a_1, a_2, which will depend on the pattern of missingness. Model M1 may be a good approach when a_1, a_2 are missing simultaneously. If a_1 is only missing when a_2 is missing, M2 may be used, because model M2 makes use of observed a_1 values when imputing the ratio, and there is no information in a_2 about missing values of a_1 that might be used to improve imputation of x_p. (Conversely, if a_2 is only missing when a_1 is missing, M3 may be attractive.) Note that M2 and M3 do not respect the deterministic relationship $x_p = a_1/a_2$.

Model M4 makes use of information on a_1, a_2 by imputing both alongside x_p; this may be motivated by having a_1, a_2 or both missing. This is similar to the approach advocated by von Hippel[65], which has been termed *just another variable* (JAV)[6, 64]. As with M2 and M3, the model ignores the deterministic relationship $x_p = a_1/a_2$ and assumes multivariate normality. This will appear a bizarre assumption; it is clearly wrong because the distributions of two of these variables must define the distribution of the third, yet software does not know this and will sample without complaint. If the assumption made by M4 is uncomfortable, we may be attracted to M5 or M6.

Model M5 is incompatible with the analysis model (see appendix C.1.1), and requires x_p to be imputed passively from imputed values of a_1/a_2. The components a_1, a_2 are not auxiliary but completely determine the values of x_p. The ratio of a_1 and a_2, which are both normal, is expected to be heavy tailed.

M6 alters the problem by transforming x_p into a linear function of its logged components and passively imputing it. Model M6 guarantees that imputed values of a_1, a_2 are positive, as with all observed ratios. While this may be desirable it is important to remember that our primary goal is valid inference, and we are not trying to recreate the missing values[23, 76]. The cosmetics of this model should therefore be a secondary consideration.

We have omitted from table 5.3 the imputation model $(\boldsymbol{z}_i, \ln(x_{pi}) \mid f(y_i)) \sim \text{MVN}$. We do not consider this because $\ln(x_p) = \ln(a_1) - \ln(a_2)$ where $\ln(a_1)$ and $\ln(a_2)$ are normal, and the sum of two normal distributions is normal. Model M6 is therefore equivalent to imputing $\ln(x_p)$, but makes more use of the observed data when components are not simultaneously missing. The only setting where modelling $\ln(x_p)$ alone is appropriate is if (a_1, a_2) are always either both observed or both missing. In this case the model would then be equivalent to M6.

To summarise our discussion of the models in table 5.3, there are conceptual problems with each one: Model M1 is compatible with the analysis model, but does not use information on observed a_1 or a_2 when the other component is missing; M2–M4 are likely to be misspecified;

and M5 and M6, the two models which make use of all the observed information on a_1 and a_2 and respect the relationship $x_p = a_1/a_2$, are incompatible with the analysis model.

5.4.5 *Software and details of imputation*

We used Stata 12's `mi` suite for MI in our case studies and simulations in section 5.6[77, 78]. Multiple imputations were produced using `mi impute mvn` and Rubin's rules were implemented using `mi estimate`.

Advice on the number of imputations typically suggests a small number (fewer than 10) is sufficient[79]. This idea comes from comparing the length of confidence intervals based on M imputations to intervals based on ∞ imputations. Our view on choosing the number of imputations, described in White, Royston and Wood[6], is slightly different, being based on the reproducibility of analyses. To achieve negligible Monte Carlo error from our MI analyses, we use $M = 300$ imputations for the Aurum case study, and $M = 100$ for EPIC-Norfolk. Note that we are not advocating such large values of M in general.

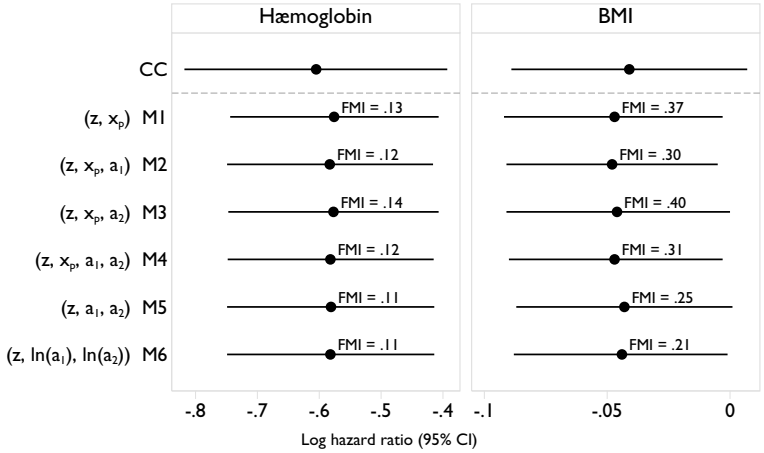
Our imputation models, all of which are based on a multivariate normal model, used a burn-in of 1,000 iterations of the MCMC chain. Thereafter, imputed datasets were stored at every 10th iteration of the chain.

5.5 CASE STUDIES

This section presents the results for multiple imputation. However, in analyses with missing data, Bayesian models are widely regarded as a sensible alternative if there is reason to be suspicious of MI results. Bayesian analyses of the Aurum and Epic datasets, corresponding to the MI approaches presented in this section, are outlined and presented in appendix C.2.

5.5.1 *Imputing body mass index in the Aurum cohort*

Figure 5.1: Results from analyses of Aurum data under different models for imputing BMI. The estimated fraction of missing information (FMI) is given next to MI analyses.

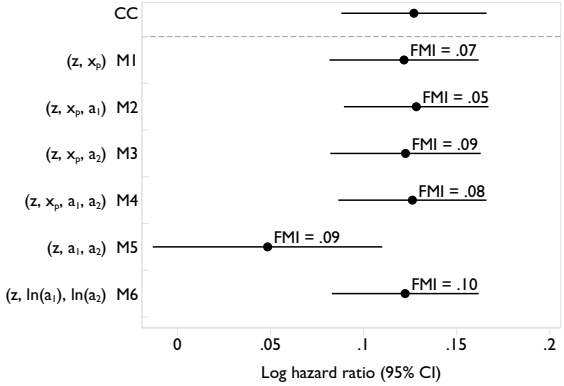


The MI procedures took between 2min 7sec (M1) and 2min 44sec (M6) to impute 300 times, fit the analysis model in each imputed dataset and use Rubin's rules to combine estimates.

Figure 5.1 shows estimates resulting from different imputation models. There is very little difference in the point estimates or width of confidence intervals; all returned essentially the same result. The number of imputations meant Monte Carlo error was negligible, at a maximum reaching 1/50th of the estimated standard error. The relative efficiency vs. infinite M was > 0.999 for all models. For both hæmoglobin and BMI, the MI estimates gave a slight change in the point estimate and a small reduction in the width of confidence intervals as compared to complete cases.

5.5.2 *Imputing cholesterol ratio in the Epic-Norfolk cohort*

Figure 5.2: Results from analyses of Epic-Norfolk data under different models for cholesterol ratio. The estimated fraction of missing information (FMI) is given next to MI analyses.

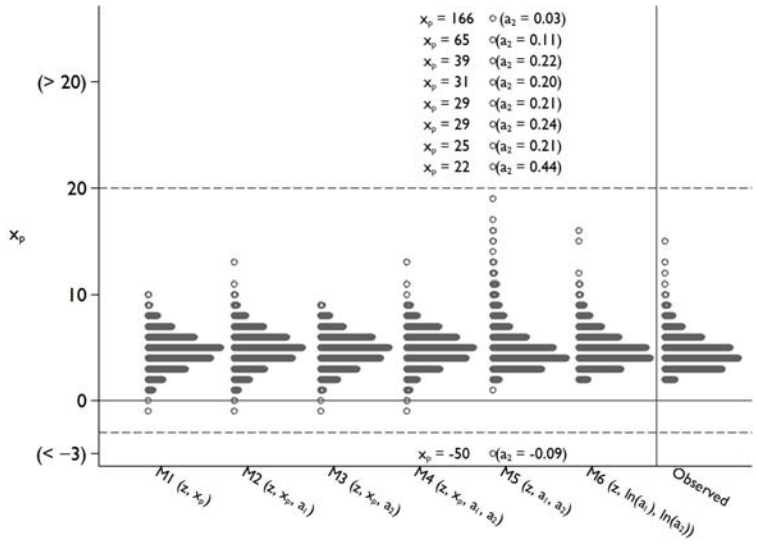


For MI of the Epic-Norfolk data, $M = 100$ was used. We used a smaller number of imputations than in Aurum because only 9% of individuals were missing cholesterol ratio. MI took between 19min 2sec (M1) and 21min 0sec (M5) to impute 100 times, analyse each imputed dataset and combine estimates using Rubin's rules. The relative efficiency vs. infinite M was > .999 for all models except M5, where relative efficiency was .991.

There was consistency between estimates from models that impute cholesterol ratio directly (figure 5.2). Monte Carlo error for point estimates was negligible (around 0.0005, less than 1/50th of the standard error) for all models except M5 where it was 0.003. MI models are less consistent than in the Aurum MI analyses but would in 5 of 6 cases give similar substantive conclusions. These estimates are also very similar to complete-cases analysis and, interestingly, the imputation model that passively imputes cholesterol ratio through log-total-cholesterol and log-HDL. However, the estimate after the standard passive imputation approach (M5) is much closer to the null, with wider confidence intervals.

Figure 5.3 demonstrates the problem with model M5 in the Epic-Norfolk data, plotting imputed values of cholesterol ratio from a single, typical, imputed dataset under models M1–M6 alongside 2,155 randomly selected observed values. The largest observed value of cholesterol ratio was 15.7. Note that for model M5 some imputed values were very large or very small; plotting these extreme values distorted the y-axis and so we have censored the y-axis below

Figure 5.3: Dotplot of imputed cholesterol ratio for single (typical) imputed datasets in EpicNorfolk under models M1–M6. Imputed values of $x_p < 3$ or $x_p > 20$ are not plotted but represented according to rank; imputed values of (x_p, a_1) are listed.



−3 and above +20, ranking and listing the values of imputed HDL and cholesterol ratio values outside of this range.

The problem with M5 arises because the mean and SD of HDL are 1.42 and 0.42 respectively, meaning its coefficient of variation (CV) is 0.30, resulting in a danger of a_2 being imputed close to zero or even negatively. This CV is far larger than in the Aurum data, where CV(height^2) = 0.11 and imputed values are never close to zero (data not shown).

Figure 5.3 also highlights the difference between the other imputation models. Imputation on the log scale (M6) is the only model to guarantee a_1, a_2 and x_p are positive. Further, the imputed values closely resemble the observed. M1–M4 can and did impute some $x_p < 0$; these models all assume x_pN and so the distribution of imputed values is symmetrical about its mean. By looking at figure 5.3, model M6 appears to be appealing, while from a statistical inference perspective (figure 5.2), there appears to be little to choose between M6 and M1–M4. From all perspectives M5 is a poor choice.

5.5.2.1 Predictive mean matching

A natural question about model M5 that arises from figure 5.3 is whether removing the high-leverage points could reduce the bias. For example, a truncated normal imputation model could be used to invoke the constraint $x_p > 0$, which would remove the negative outliers of model M5.

A better alternative, which can also remove the positive outliers, is *predictive mean matching* (PMM)[33, 28, 6]. Briefly, the imputation model is fitted and, for each individual with a missing value, the k individuals ('donors') with observed values with the closest predicted mean are identified. One of these is selected at random and their value 'donated' as the imputed value. This ensures that imputed values are within the range of observed values.

To improve model M5, PMM is most easily implemented in a chained equations procedure[6].

Imputation of a_1 and a_2 uses PMM, and x_p is passively imputed. The largest possible imputed value of x_p is then the ratio of the largest observed value of a_1 to the smallest observed value of a_2 (and vice versa for the smallest imputed value of x_p).

We used this imputation model on the Epic-Norfolk data, using $k = 10$ and storing imputed values after 10 cycles of chained equations. This reduced the bias of model M5, giving an estimated log-hazard ratio of .119 (95% CI .079–.159). See appendix C.3 for the full results.

5.6 SIMULATION STUDY

5.6.1 *Design*

We performed a simulation study designed to investigate models M1–M6 in a simpler setting than the two case studies. With x_p as the only covariate and a continuous outcome y, we investigated the performance of the imputation models and how this varied with the strength of x_p–y association and the coefficient of variation of the ratio's denominator, $\mathrm{CV}(a_2)$. This affects the distribution of x_p and we hypothesise that when $\mathrm{CV}(a_2)$ is large model M5 will be biased. An imputed value of a_{2i} may be very small, meaning the corresponding value of x_{pi} will be large, and possibly outside the range of observed x_p. The x_{pi} will thus have high leverage. For such values, there are unlikely to be appropriately large or small y to preserve the true x_p–y relationship, which leads us to expect bias towards no association.

Scenarios investigated include two values of $\mathrm{CV}(a_2)$: 0.1, taken from height^2 in the Aurum data, and 0.3, taken from HDL in the Epic-Norfolk data; these are varied factorially with R^2 values of 0.1 and 0.3. All simulations were performed using Stata 12[77]. Our simulation procedures were as follows:

1. Simulate $n = 500$ complete values of $\ln(a_1), \ln(a_2)$ to follow a bivariate normal distribution. In our first scenario the mean, standard deviation and correlation are taken from ln(weight) and ln(height^2) in the Aurum data: $\ln(a_1)$ has mean 4 and SD 0.21, $\ln(a_2)$ has mean 0.97 and SD 0.11, and $\mathrm{Corr}(\ln(a_1), \ln(a_2)) = 0.22$. This gives $\mathrm{CV}(a_2) = 0.1$.

2. Generate complete $x_p = \exp(\ln(a_1) - \ln(a_2))$, meaning that x_p follows a lognormal distribution. For the ratios and components in our two example datasets the lognormal distribution seems to be a suitable choice.

3. Simulate $y \sim \mathrm{N}(\beta_0 + \beta_1 x_p,\ \sigma^2)$. We used the same value of β_1 (arbitrarily 2) throughout to make bias comparable across all simulation settings. To vary the strength of association we altered σ^2 to achieve the desired R^2.

4. Simulate binary indicators of response, R_1 and R_2 for a_1 and a_2 respectively. Each R is generated independently from the model $\mathrm{logit}\{\mathrm{P}(R = 1)\} = \gamma_0 + \gamma_1 y$. Under MCAR $\gamma_1 = 0$. Under MAR, γ_1 is chosen so that ROC analysis of y versus an indicator of response R produces a mean area under the curve of 0.65. This is to achieve the same degree of MAR across scenarios. We then alter γ_0 so that $\mathrm{P}(R_1 = 1) = \mathrm{P}(R_2 = 1) = 0.75$. Because γ_1 has the same sign for both R_1 and R_2 and both depend on y, the probability of a_1, a_2 being missing simultaneously is slightly larger under MAR than MCAR. This means

the overall proportion of observations missing x_p is slightly smaller under MAR (42% missing x_p) than MCAR (44% missing x_p).

5. Set a_{1i} to missing if $R_{1i} = 0$ and a_{2i} to missing if $R_{2i} = 0$ and x_{pi} to missing if $R_{1i} = 0$ or $R_{2i} = 0$.

6. Impute x_p five times using each of the models M1–M6 (table 5.3).

7. Fit the correct analysis model to each imputed dataset and combine results using Rubin's rules.

We used 5000 replicates of this process under each combination of simulation settings. Interest is in β_1. Bias, coverage of 95% confidence intervals and efficiency of $\hat{\beta}_1$ (expressed by the empirical standard error, SD($\hat{\beta}_1$) over all replications[63]) were calculated under models M1–M6, with analysis of complete data (i.e. before any data are set to missing) and complete cases (dropping observations with missing x_p) also provided for reference.

5.6.2 *Results*

Table 5.4 summarises the results of our simulation study. Results of the complete data and complete cases analyses are both as expected. Complete data is always unbiased with 95% coverage and the smallest empirical standard error of all methods. Complete cases is unbiased under MCAR but biased under MAR. Coverage is correspondingly low and efficiency is lower than complete data.

M1 is mainly unbiased, but there is a small upward bias under MAR and $R^2 = 0.3$ and coverage is slightly low when data are MAR. This is perhaps because it assumes normality for x when it is actually lognormal. M1 also tends to be inefficient compared to other imputation models, as would be expected, regardless of the missingness mechanism.

With this general pattern of missingness, M3 is usually more biased than M2, though coverage tends to be similar (except where CV(a_2) = 0.3 and $R^2 = 0.3$). Efficiency of the M2 and M3 seems to depend on CV(a_2) and R^2. Model M4 has similar bias to M2 and M3; at worst this reaches about 4% with both large CV(a_2) and R^2. Empirical standard errors for M4 are at least as small as M2 and M3, while coverage tends to be good except when both CV(a_2) and R^2 are 0.3.

Model M5 performs well in the two scenarios when CV(a_2) = 0.1. There is a small downward bias but efficiency and coverage are both good compared with other methods. However, when CV(a_2) = 0.3 we observe unacceptable bias towards the null and lower efficiency than other methods, although coverage is still over 90%. When considered alongside bias this coverage implies that while the empirical standard error is large, the estimated standard errors are even larger, reducing the effect of the large bias on coverage and implying low power.

M6 is more biased than M5 when CV(a_2) = 0.1, but much less so when CV(a_2) = 0.3. Across all of our settings, it is more efficient than M1–M5 and with coverage close to 95%. If the small bias seems acceptable then this is the best imputation model.

Table 5.4: Simulation results: bias, coverage and efficiency of different imputation models

R^2	CV(a_2)	Imputation model		Bias (β_1 = 2)		Empirical SE		Coverage	
				MCAR	MAR	MCAR	MAR	MCAR	MAR
0.1	0.1	Complete data		0.000		0.273		95.2	
		Complete cases		0.003	−0.172	0.366	0.352	95.1	92.6
		x	M1	−0.005	−0.004	0.368	0.386	93.8	94.9
		x, a_1	M2	−0.001	0.002	0.333	0.345	94.6	94.7
		x, a_2	M3	−0.009	−0.003	0.363	0.383	94.6	94.9
		x, a_1, a_2	M4	−0.005	0.005	0.330	0.342	94.7	95.0
		a_1, a_2	M5	−0.017	−0.016	0.328	0.337	94.8	95.0
		$\ln(a_1), \ln(a_2)$	M6	−0.016	−0.034	0.329	0.332	94.9	95.1
0.1	0.3	Complete data		0.006		0.267		95.3	
		Complete cases		0.001	−0.168	0.359	0.351	95.3	92.9
		x	M1	−0.009	0.005	0.358	0.385	94.7	94.9
		x, a_1	M2	−0.007	0.014	0.348	0.372	94.9	94.9
		x, a_2	M3	−0.001	0.031	0.334	0.362	95.4	95.0
		x, a_1, a_2	M4	−0.001	0.038	0.325	0.346	95.0	94.7
		a_1, a_2	M5	−0.562	−0.665	0.350	0.334	94.3	92.6
		$\ln(a_1), \ln(a_2)$	M6	−0.038	−0.064	0.313	0.318	95.8	95.4
0.3	0.1	Complete data		0.003		0.137		95.2	
		Complete cases		0.001	−0.139	0.183	0.188	95.5	88.5
		x	M1	−0.005	0.031	0.171	0.187	95.3	94.0
		x, a_1	M2	−0.003	0.026	0.159	0.171	95.8	95.0
		x, a_2	M3	−0.007	0.029	0.170	0.188	95.2	93.8
		x, a_1, a_2	M4	−0.003	0.026	0.159	0.171	95.9	94.6
		a_1, a_2	M5	−0.016	0.000	0.158	0.168	96.1	95.3
		$\ln(a_1), \ln(a_2)$	M6	−0.016	−0.031	0.158	0.163	96.2	95.6
0.3	0.3	Complete data		−0.002		0.137		95.0	
		Complete cases		−0.006	−0.143	0.184	0.192	94.9	88.5
		x	M1	−0.009	0.054	0.174	0.196	94.2	93.0
		x, a_1	M2	−0.012	0.057	0.172	0.193	94.8	93.3
		x, a_2	M3	−0.010	0.076	0.170	0.191	94.3	91.5
		x, a_1, a_2	M4	−0.009	0.080	0.167	0.187	94.2	91.8
		a_1, a_2	M5	−0.580	−0.814	0.287	0.300	94.3	93.3
		$\ln(a_1), \ln(a_2)$	M6	−0.051	−0.070	0.162	0.164	95.1	94.6

We have presented the results of two case studies involving commonly used ratios and a simulation study based in part on these datasets. A key message is the caution against passive imputation of a_1 and a_2 without prior transformation. Superficially, the approach appears to make more use of the available data, however it is often inefficient and can suffer from large bias. Our analysis of the Epic-Norfolk data demonstrated this problem in practice. However, in our Aurum case study, the use of passive imputation appeared to make little difference to the substantive results compared to active imputation. Our simulation study confirmed that problems arise when $\mathrm{CV}(a_2)$ is large. Note that a ratio with very small $\mathrm{CV}(a_2)$ is unlikely to be used in applied work (unless $\mathrm{CV}(a_1)$ is also very small) because as $\mathrm{CV}(a_2) \to 0$, x_p becomes a function of a_1 divided by a constant. We therefore recommend that incomplete ratios be imputed actively, or passively after log transformation as in model M6.

In considering models for missing data, joint models for the covariates and outcome are attractive because they use the full data likelihood in a coherent way. In our two case studies we attempted to fit fully Bayesian joint models and summarise posterior distributions for parameters of interest. Computational problems prevented this approach from being useful. In one dataset some of the models did not appear to converge to any true posterior distribution (or if they did, results were extraordinarily sensitive to the choice of model for the ratio). In the other dataset, it was not possible to load the observed data into WinBUGS and so the attempt was abandoned.

Compatibility is a useful concept for considering whether various imputation models are sensible. We hypothesised that models M1 and M2–M4 would perform well due to being compatible and semi-compatible respectively, while models M5 and M6 would perform poorly because of being incompatible. In our simulations, M1–M4 did tend to perform well despite being misspecified, and model M5 did often perform poorly. In our Epic-Norfolk example, where model M5 gave nonsense results, problems could be identified by inspecting the imputed values of x_p.

Model M6 was surprisingly as good as any other model considered throughout. Despite being more robust than M5, we know it is not completely 'safe'. In our simulation study, the imputation model assumed $(\log(a_1), \log(a_2) \mid y)\mathrm{N}$, and since $\log(x_p) = \log(a_1) - \log(a_2)$, this implies $(\log(x_p) \mid y)\mathrm{N}$. The imputation model therefore has mean function $\log(x_p) = \alpha_0 + \alpha_1 y$, while the analysis model has mean function $y = \beta_0 + \beta_1 x_p$. In further simulations, we noted that M6 was still robust when $R^2 = .5$ and $\mathrm{CV}(a_2) = 0.3$ (results not shown). We can provide no guarantee for greater values other than that this model will eventually fall apart. However, it is our experience that associations stronger than $R^2 = 0.5$ are rare in medical applications.

The imputation models considered in this work were all based on the multivariate normal distribution. This facilitated understanding of the relationship between the imputation and analysis models. However, it would have been feasible to use mice-based approaches to customise the imputation models. An example of such an approach might be to use the following

equations as a cycle of chained equations:

$$
\begin{aligned}
a_1 &\sim \mathrm{N}(\alpha_0^{(1)} + \alpha_1^{(1)} a_2 + \alpha_2^{(1)} x_1 + \ldots + \alpha_{(p-1)}^{(1)} x_{(p-1)}, \sigma^{2(1)}) \\
a_2 &\sim \mathrm{N}(\alpha_0^{(2)} + \alpha_1^{(2)} a_1 + \alpha_2^{(2)} x_1 + \ldots + \alpha_{(p-1)}^{(2)} x_{(p-1)}, \sigma^{2(2)}) \\
&\quad [\text{passively impute } x_p^* = a_1^*/a_2^*] \\
x_1 &\sim \mathrm{N}(\alpha_0^{(3)} + \alpha_1^{(3)} x_2) + \ldots + \alpha_p^{(3)} x_p, \sigma^{2(3)})
\end{aligned}
$$

While this does not represent a well defined imputation model, it may be superior inferentially to some of the models that were used, particularly M5.

Some of the issues with model M5 could have been alleviated by using partly parametric imputation techniques such as predictive mean matching (PMM)[33] or local residual draws[28]. In practice, this requires a switch to the chained equations approach rather than a multivariate imputation model. Since a parametric model is used only to identify suitable donors, this makes it impossible to think about compatibility. We investigated PMM in the problematic Epic-Norfolk dataset and found model M5 much improved. PMM may therefore be a useful adjunct to a suitably chosen imputation model.

In evaluating methods we have focused on bias, coverage, and efficiency. For those interested in accurate prediction, efficiency may be more important and coverage less so or even unimportant[80]. It is worth noting that precision is also lower for model M5. Therefore if passive imputation is to be used for a ratio in prediction settings it should be done on the log scale.

We have considered the imputation of ratio covariates. Some similar issues arise when the analysis model contains any nonlinear function, for example interactions and squares. The difference is that in both cases the main effects and their interaction, or the variable and its square, are included in the analysis model. In the case of squares, a measurement and its square will also be observed or missing simultaneously. Imputation is then complicated by the fact that the analysis model contains both the untransformed variable and a nonlinear function as covariates, rather than just the nonlinear function, as in the case of ratios. This makes issues around compatibility somewhat more complicated. See von Hippel[65], Seaman, Bartlett and White[64] and Bartlett et al[26] for recent work on imputation of squares and interactions.

Bartlett et al proposed the use of rejection sampling when producing imputations and showed it to be useful for imputing squares and interactions; this may therefore be a good approach for imputing ratios. By explicitly involving the analysis model in the specification of the imputation model, each imputation model used in the chained equations is compatible with the imputation model[26]. However, the method is more time-intensive than any imputation models investigated here and it is yet to become available in standard software packages. It also sacrifices one of the advantages of multiple imputation: separation of missing data issues from substantive analyses. However, this may be necessary, and has already been partly conceded when we tailor imputation models that to be compatible with the analysis model.

6 Combining multivariable fractional polynomial models with multiple imputation

Until now, this thesis has dealt with situations where the correct analysis model is fixed and known, and evaluated alternative approaches to multiple imputation. This chapter investigates a very different scenario, where the analysis model is (at least partly) unknown and to be chosen using a semi-automatic model selection procedure. The focus is on fractional polynomials but combining MI with model selection using splines is also largely unresearched; I believe some of the concepts presented in this chapter would be relevant to any such future research.

6.1 BACKGROUND

Fractional polynomial (FP) and multivariable fractional polynomial (MFP) models are a flexible but relatively simple method for modelling nonlinear effects of one or more continuous covariates in regression analyses. The type of datasets to which fractional polynomial models are applied commonly have missing covariate data. There is currently no satisfactory approach to handling missing data when using fractional polynomial methods. This chapter aims to extend MI to accommodate MFP and vice versa so that they can be combined effectively.

6.1.1 *Single-variable fractional polynomials*

For a regression model involving a single continuous explanatory variable x, a fractional polynomial model of dimension D, termed 'FPD', has D terms in x and linear predictor

$$\beta_0 + \sum_{d=1}^{D} \beta_d x^{p_d}. \tag{6.1}$$

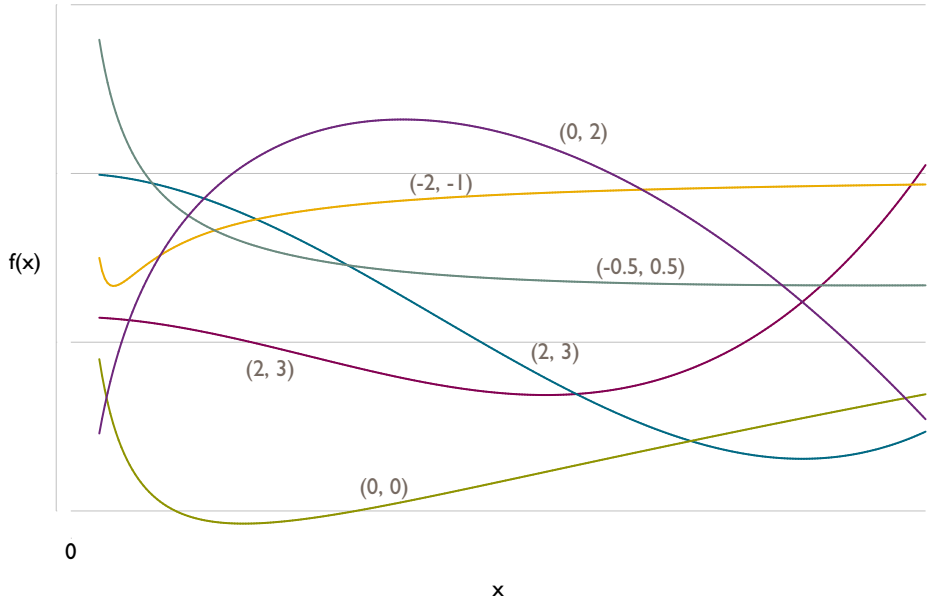
Possible values of the exponents $p_1 \leq \ldots \leq p_D$ are typically restricted to the set

$$S \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\},$$

which is a suggestion rather than a rule. By convention $x^0 = \log(x)$. With $D > 1$ it is possible that $p_d = p_{d'}$; the dth term is $x^{p_d}\log(x)$ rather than x^{p_d}. For $D = 1$, there are eight possible models for x; for $D = 2$ there are 36, and so on. In practice, $D > 2$ is rarely considered.

Note that equation (6.1) is inclusive of conventional polynomials such as quadratics or cubics, but is considerably more general and therefore flexible. Conversely, fractional polynomial functions are Box–Tidwell transformations[81], where the parameter space of $\boldsymbol{p}$ is restricted to the set S, and so FP models are comparatively less flexible. Figure 6.1 plots a selection of FP functions with $D = 2$, demonstrating the flexibility available.

Figure 6.1: Example FP2 functions. Exponents p used to plot the curves are given in parentheses.



6.1.2 *Multivariable fractional polynomials*

Multivariable fractional polynomials are the natural extension of fractional polynomials to the setting with multiple $\boldsymbol{x}$. There are C continuous explanatory variables with linear predictor

$$\beta_0 + \sum_{c=1}^{C} \sum_{d=1}^{D_c} \beta_{cd} x_c^{p_{cd}}. \tag{6.2}$$

The D_c indicates that the complexity D of FP function is allowed to differ for different c. A variable suspected to have a u-shaped relationship with outcome would need at least $D = 2$. Meanwhile $D = 1$ may be desirable for certain variables because it forces outcome to be a monotonic function of x, which may be biologically plausible.

6.1.3 *Model building with a single x and fully observed data*

Assume there is a single covariate x. If the scalar D and vector $\boldsymbol{p}$ are known for a given dataset the main task for an analyst is to estimate the vector $\boldsymbol{\beta}$. Aside from the context of model validation, such a scenario is very unlikely. It is usually necessary to estimate $\boldsymbol{p}$. Further, D_c will often be uncertain and so we require a method for selecting between models of different dimension.

Ambler and Royston describe a procedure for selecting D from $d = 1, \ldots, D_{\max}$ and estimating $p \mid D$[82]:

1. For $d = 0, \ldots, D_{\max}$ (where $d = 0$ is a model omitting x) fit the models of dimension d for all combinations of $p \in S$ (6.2). For each d, the candidate models are of identical complexity, and so the best fit is the model which maximises the log-likelihood, $\log(L)$.

2. For the dimensions under consideration the best models are compared as follows using likelihood ratio tests (this assumes that a null model is the least complex form considered for x; in practice the least complex model considered might be a linear model for x):

   a) Test the best FP$D_{\max}$ model at the chosen α level against the null model using $2D_{\max}$ d.f. If the test is not significant the null model is chosen; if the test is significant continue.

   b) Test the best FP$D_{\max}$ model at the chosen α level against the linear model on $2D_{\max} - 1$ d.f. If the test is not significant the linear model is chosen; if the test is significant continue.

   c) Test the best FP$D_{\max}$ model for x at the chosen α level against the best fitting FP1 model on $2D_{\max} - 2$ d.f. If the test is not significant the FP1 model is chosen; if the test is significant continue.

   d) ...

   e) Test the best FP$D_{\max}$ model for x at the chosen α level against the best fitting FP$(D_{\max} - 1)$ model on $2D_{\max} - (2D_{\max} - 2)$ df. If the test is not significant the FP$(D_{\max} - 1)$ model is chosen.

   f) If all tests are significant, the FP$D_{\max}$ model is chosen ($D = D_{\max}$).

3. Estimate $\boldsymbol{\beta} \mid D, \hat{\boldsymbol{p}}$.

With complete covariate data, the above algorithm is the standard for fractional polynomial model selection. It has the characteristic of a closed testing procedure[83], meaning the overall type I error rate is maintained at the chosen significance level (in this case slightly lower because the df allocated to each p_d is conservative; see 6.1.5). Having pre-edited data in preparation for FP model building, for example deciding on how to deal with extreme values, important tasks for the analyst are deciding on α, D and the values in S, although standard choices are implemented as the defaults in Stata.

6.1.4 *Model building with multiple $\boldsymbol{x}$ and fully observed data*

In the context of *multiple* continuous covariates to be considered for FP functions, Royston and Sauerbrei extend the model selection procedure as follows[13]:

1. The $c = 1, \ldots, C$ covariates are ordered in terms of decreasing significance in a normal errors model with all x_c's included linearly.

2. x_1 is subjected to the function selection procedure described in section 6.1.3, holding $x_2, \ldots, x_C$ as linear.

3. x_2 is subjected to the function selection procedure, fixing the function(s) of x_1 selected at step (2) and holding $x_3, \ldots, x_C$ as linear.

4. ...

5. After running the function selection procedure for x_C the first 'cycle' is complete. Note that correlations between $x_1, \ldots, x_C$ will mean the best fitting functions for x_c can depend on the chosen functions for $x_{c'}$, so the process continues.

6. Function selection is repeated for x_1, fixing the selected functions of $x_2, \ldots, x_C$ chosen in cycle 1.

7. ...

8. The process continues until all functions remain stable for a full cycle, indicating the model has stabilised.

The algorithms outlined above depend entirely on likelihood ratio testing to select models, which poses problems in the multiple imputation setting.

6.1.5 *Points to note on $\boldsymbol{p}$ and D*

Even with fully observed datasets, some points related to the function selection procedure lead to slight miscalibration of error rates. The first two are related to the role of the parameters $\boldsymbol{p}$, and the lack of any estimate of $\mathrm{Var}(\hat{\boldsymbol{p}})$. One point is concerned with testing procedures, and the other with estimation of $\boldsymbol{\beta}$.

1. The parameters $\boldsymbol{\beta}$ are estimated conditional on $\hat{\boldsymbol{p}}$, thereby treating $\hat{\boldsymbol{p}}$ as fixed and known. This leads to precision of $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ that is overstated for *estimation*.

2. The parameter space for $\hat{\boldsymbol{p}}$ is discrete. In the testing procedure outlined above, the allowance of 1df for each $\hat{\boldsymbol{p}}$ would be correct if the parameter space were continuous in $(-\infty, \infty)$. However, since this is not the case, the 1df used for each term in $\boldsymbol{p}$ is too generous, leading to conservatism in the *testing* procedures.

A more obscure point is that in testing between models of different dimension, D is estimated, also with a discrete parameter space (typically 0, 1, 2). The notion of D as a parameter has not been considered in any previous literature on fractional polynomials or, to my knowledge, variable selection. The current work aims to integrate multiple imputation with current multivariable fractional polynomial methods, and so I do not consider this further.

6.2 Considerations for combining fractional polynomials with multiple imputation

This initial motivation for this work arose from the problem of model selection. The standard procedures outlined above depend entirely on likelihood ratio tests, but MI data do not yield a meaningful likelihood for the purposes of inference.

It is clear that there is a second problem in combining MI with MFP: the statistical characteristics of any model selection procedure will depend not only on the tests used but on the method of imputation. It is critically important that imputation allows for the uncertain nature of the analysis model, and leads to consistent and efficient estimation of $\boldsymbol{p}$.

In the following sections, problems with extending MFP methodology to MI are separated into three parts, described in sections 6.2.1, 6.2.2 and 6.2.3. Each is discussed and addressed in turn in sections 6.3 to 6.8, and good methods from one step are carried into the subsequent work.

6.2.1 *Imputation allowing for the analysis model to include (unknown) FP functions*

The issue with imputation is ensuring (semi-)compatibility of the models for imputation and analysis. An appropriate imputation method must allow for the uncertainty about both $\boldsymbol{p}$ and $\boldsymbol{\beta}$. Imputing x using a linear regression of x^1 on y would be wrong: this is compatible with an analysis model where $p = 1$, but incompatible with other values of p considered by the analysis model (unless $p = 1$ in truth, in which case there would be no sense in exploring FP transformations). This approach would bias estimation of p towards 1. This issue is explored in section 6.3, where I propose one proper imputation method and propose an extension to a second.

6.2.2 *Estimation of $\boldsymbol{p}$ from candidate FPd models*

With complete data, $\boldsymbol{p}$ is estimated by comparing the fit of models based on the log-likelihood. It is well known that the log-likelihood cannot be used for formal inference in multiply imputed data. However, comparing models of identical complexity does not make reference to any distribution, and so the use of log-likelihoods is acceptable for estimation of $\boldsymbol{p}$. However, this does not mean other methods are not possible, and estimating $\boldsymbol{p}$ by finding the value which maximises the Wald statistic is a feasible, and plausibly superior, alternative. This issue is addressed in section 6.4, where the performance of methods based on log-likelihoods and Wald statistics are compared to log-likelihoods based on complete data and on complete cases.

6.2.3 *Tests in the function selection procedure*

Hypothesis testing based on log-likelihoods *does* constitute formal inference, meaning that standard likelihood ratio tests cannot be used. With multiply imputed data, it is usual to based hypothesis tests on Wald statistics constructed using Rubin's rules. However, the testing involved in selecting functions is non-standard: comparing a model including x as linear to an FP2 model (which does not contain x^1) is a difficult problem for Wald tests because the models are not nested.

6.2.4 *Estimation of parameters from the selected model*

Following model selection, inference will always be based on Rubin's rules for $\boldsymbol{\beta}$. There is nothing controversial about this step so it is not considered further. The only caution is related to the point in 6.1.5, that by treating $\hat{\boldsymbol{p}}$ as fixed and known, $\text{Var}(\hat{\boldsymbol{\beta}})$ will be underestimated. This is a problem inherent to MFP methods and so is not of specific concern in combining MI with MFP.

An imputation method appropriate for (M)FP models must:

1. Allow for the uncertainty about $\boldsymbol{p}$ implied by the analysis procedure.
2. Impute positive values of x_j.

The methods of imputation below have potential.

6.3.1 *Predictive mean matching and local residual draws*

Predictive mean matching has previously been proposed as a method for imputing when the analysis model contains nonlinear functions of covariates[6], and local residual draws have similar potential. The rationale is that we wish to impute with minimal assumptions about the functional form. PMM and LRD assume a functional form to identify donors but not to impute missing values, and so there is less dependence on the functional form than in using posterior draws.

The results of chapters 3.2 and 4 showed that despite the potential, neither method was able to impute nonlinear functions without introducing bias, particularly under MAR, although their performance was fair when the functional form was approximately correct. Seaman, Bartlett and White also demonstrated bias associated with PMM when the analysis model contained nonlinear functions of incomplete covariates[64]. PMM and LRD without transformation are not pursued further here.

One attractive feature of PMM (but not LRD) is that imputed values are always taken from observed values, and so if observed values are positive, imputed values will also be positive. Fractional polynomial transformations require that covariates only take positive values and so PMM may prevent strange imputation, particularly for variables with unusual distributions.

In chapter 5 it was noted and demonstrated that PMM can be a useful adjunct to a suitably chosen imputation model, and it is used in this context below.

6.3.2 *Imputing x for FP1 models via the approximate Bayesian bootstrap*

Multiple imputation aims to draw missing values from their posterior predictive distribution, which requires uncertainty in parameter estimates to be fully acknowledged. A method of proper imputation in preparation for (M)FP1 analysis models is outlined below.

To draw values $\boldsymbol{p}^*$ from the posterior, a method based on the approximate Bayesian bootstrap (ABB) can be used[34]. The Bayesian bootstrap is operationally similar to the bootstrap, but rather than sampling observations with probability $1/n_h$, the probability of sampling any h is random[84]. The probabilities are calculated by drawing $(n_h - 1)$ values from a uniform$(0, 1)$ distribution; these are ordered and the absolute differences $(0 - u_1), (u_1 - u_2), \ldots, (u_n - 1)$ calculated. The vector of absolute differences are the vector of probabilities for sampling with replacement, and form an improper non-informative Dirichlet prior[61]. Rubin shows that this simulates a draw from the posterior distribution[84]. The approximate Bayesian bootstrap approximates the draws from a Dirichlet posterior distribution by drawing from a scaled multinomial distribution, reducing the computational burden, but leading to a very slight loss of efficiency[61]. The Bayesian bootstrap and approximate Bayesian bootstrap lead to similar

inference to the frequentist bootstrap, particularly with large n_h, greater than say 50[61]. In the implementation below, what is referred to as 'ABB' is in fact based on a frequentist bootstrap, but simulation results would be practically identical for the sample sizes considered.

Considering a single x, the following imputation procedure is compatible with FP1 functions of x:

1. Draw an ABB sample of n_h observations from the n_h individuals with observed values of x.

2. For $p = -2(.)3$, fit a linear regression of x^p on y. This is compatible with the assumption that the analysis model is a regression model for y on x^p for unknown p. Values in (.) should at a minimum include the powers considered by the analysis, but could be far less coarse. Increments of 0.2 are used in the present chapter.

3. Record the value of p which returns the largest value of $\log(L) + J$, where J is the Jacobian for the transformation from x to x^p (required in order to make the log-likelihoods comparable). Denote this value p^*. (As the maximum from a bootstrap sample, p^* is an approximate nonparametric draw from the posterior of p.)

4. Restore the partially observed dataset.

5. Impute $(x_j)^{p^*}$ once using the appropriate linear regression from step 2.

6. Passively impute x_j^* by taking the p^*th root of $(x_j^*)^{p^*}$.

7. Repeat steps 1–6 until M imputed datasets exist.

As noted previously, it is important that x_j^* are positive, so that the standard fractional polynomial transformations can be calculated for all x_j^*. Two options for imputation have been implemented:

1. Impute using a truncated regression imputation model. Specify a (lower) truncation bound for x^1 at some value > 0. This is transformed to a bound for x^{p^*} in step 5 (a lower bound for $p^* \geq 0$ and an upper bound for $p^* < 0$).

2. Perform the imputation in step 5 using PMM. Provided the observed values of x are positive, the imputed values will be.

A Stata command implementing this method of imputation for a single FP1 variable has been written (`tuni`), with options to impute by truncated regression or PMM. It is relatively straightforward to see how the method could be generalised to multiple incomplete covariates by nesting it within a chained-equations type procedure to handle $C > 1$. The only subtlety to note in the extension is that steps 2 and 5 should condition on the current draw of $p_{c'}^*$ in imputing x_c. This has also been implemented in Stata as `icet`.

However, the approach is partially limited by its incompatibility with FP models involving $D > 1$. The $D = 1$ case is justified by the fact that when the analysis model is a linear regression of y on x^p, linear regression of x^p on y is a compatible imputation model. For $D > 1$ the model for $x|y$ is not a linear regression.

6.3.3 *Imputing x for MFP models via 'substantive-model-compatible fully-conditional-specification'*

Bartlett et al. developed and evaluated a method of proper imputation when the analysis mode contains nonlinear functions of covariates: *substantive-model-compatible fully-conditional-specification* (SMC FCS)[26]. The simulation studies presented by the authors demonstrated that the method tends to have good properties in terms of bias, efficiency and coverage.

Bartlett et al. write the analysis model as $f(y \mid \boldsymbol{x}, \boldsymbol{\beta})$. Their method is motivated by the fact that for a partially observed covariate x, the conditional distribution $f(\boldsymbol{x} \mid y)$ can be expressed as

$$\frac{f(y, \boldsymbol{x})}{f(y)} \propto f(y \mid \boldsymbol{x}) f(\boldsymbol{x}) \tag{6.3}$$

by Bayes' theorem. In SMC FCS a model for x_c is specified as $f(x_c \mid \boldsymbol{x}_{\boldsymbol{c}'}, \boldsymbol{\alpha})$ – whereas mice would also condition on y. The imputation model that is implied by this x_c model together with the analysis model then involves densities proportional to

$$f(y \mid \boldsymbol{x}, \boldsymbol{\beta}) \times f(x_c \mid \boldsymbol{x}_{\boldsymbol{c}'}, \boldsymbol{\alpha}). \tag{6.4}$$

This imputation model does not often belong to a standard parametric family of distributions. However, Bartlett et al. show that if it is easy to draw from $f(x_c \mid \boldsymbol{x}_{-\boldsymbol{c}}, \boldsymbol{\alpha})$, it is possible to use rejection sampling to draw from the distribution specified in (6.4). This involves repeatedly drawing from a 'proposal distribution' $f(x_c \mid \boldsymbol{x}_{\boldsymbol{c}'}, \boldsymbol{\alpha})$ and rejecting proposed draws of x_c^* unless a certain criterion is satisfied, where the acceptance probability is proportional to $f(y \mid \boldsymbol{x}, \boldsymbol{\beta})$. The criteria for rejection / acceptance of a proposal draw are described for linear regression, discrete outcomes and proportional hazards models in [26]; for full details, see Bartlett et al[26].

In joint work with Jonathan Bartlett, this method has been coded in Stata as `smcfcs`, and an article describing it submitted to the Stata Journal.

SMC FCS was developed with the aim of imputing incomplete covariates $\boldsymbol{x}$ where the analysis model contains nonlinear transformations of $\boldsymbol{x}$. The scenarios considered restricted attention to cases where the analysis model was known and correctly specified. For fractional polynomials, the analysis model is not known, and specifying the imputation model is more difficult.

The natural way to extend SMC FCS for FP models is via a model that is *at least* as general as any analysis model that might be selected. For a single x, imputation could therefore include all eight FP transformations of x as analysis model covariates for rejection sampling. Following imputation, whatever FP1 model is actually selected will be semi-compatible with the analysis model specified by SMC FCS.

Technically $x^p \log(x)$ should also be imputed for each p in S if the analysis model considers FP2 functions, since these repeated powers may be selected for the final model. This implies that for FP2 each variable x should specify, for purposes of imputation, an analysis model with 16 transformations of x. This could lead to hugely complex imputation models that sometimes fail to converge, and the suggestion is unlikely to be met with enthusiasm in practice. In the interests of pragmatism we suggest omitting the repeated-power transformations.

6.3.4 *Choice of imputation method*

Despite the promise of SMC FCS, in early runs of the simulations in section 6.5, the method sometimes performed extremely poorly. In particular, problems occurred when the coefficient of variation of the variable being imputed was high and/or when the marginal distribution of a covariate was misspecified. Appendix D gives a brief graphical overview of some issues that arise with fractional polynomials.

The difficulties with SMC FCS led to the incorporation of the univariate `tuni` command described in section 6.3.2 in a mice-type procedure (`icet`).

The simulation studies presented in this chapter are entirely based on data imputed by `tuni` or `icet`. However, SMC FCS remains a promising method that may be superior to `icet` when certain issues have been resolved.

6.4 ESTIMATION OF p: SIMULATION STUDY

The fractional polynomial function-selection procedure which considers maximum dimension $D_{\max}$ requires null, linear, and the best-fitting FPd models for $d = 1, \ldots, D$. With complete data, 'best-fitting' for any given d is defined by the model returning the largest value of the log-likelihood. This section considers methods for estimation of the best-fitting models in MI data. The two methods below are considered.

6.4.1 *Candidate methods*

In a simulation study, two methods of estimating p are compared.

Log-likelihoods. The M imputed datasets are stacked and each FPd model is fitted, treating the data as a single dataset of $n \times M$ observations, and p is selected to maximise the log-likelihood.

Wald statistics. $\hat{\beta}_p$ and $\widehat{\mathrm{Var}}(\hat{\beta}_p)$ are estimated using Rubin's rules and the square of the z_p statistic is calculated:

$$z_p^2 = \left(\frac{\hat{\beta}_p}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_p)}} \right)^2, \tag{6.5}$$

with p selected to maximise z_p^2.

Although it is unusual to use likelihoods in MI data, problems with using the likelihood arise from referring the quantity to distributions: twice the difference in the log-likelihoods for two models does not follow a χ^2 distribution. However, when attempting to determine the optimum exponents for $\boldsymbol{x}$, the use of log-likelihoods does not refer the statistics obtained to any distribution. Here, the likelihood is used to compare the fit of models of identical complexity, and so the method is uncontroversial.

Wald statistics have not previously been used for MFP in complete data, but it is not clear that the method would estimate p poorly. Further, Wald statistics based on Rubin's rules have been shown to be the ideal basis for variable selection methods in MI data[14], which provides motivation for their use with MFP models.

If log-likelihoods and Wald statistics are unbiased, as expected, the method which estimates p with greatest precision would be favoured. If bias and precision of both methods are comparable then both will be carried forward to the stage of testing between models of different complexity (sections 6.5–6.8). As usual, analysis of the complete data will be included as a gold-standard. Analysis of complete cases is also of interest as a benchmark; if the MI methods cannot outperform complete cases then they are not worth using in practice because CC is practically the more convenient method.

6.4.2 *Simulation design*

To compare our candidate methods, a simulation study based on FP1 is used. The true model involves linear regression of a continuous outcome y on an FP1 function of a single continuous covariate x. Because we aim to compare bias and precision of log-likelihoods vs. Wald statistics for estimating p, for estimation we use a set of Box–Tidwell transformations rather larger than the usual eight transformations in the set S usually used in fractional polynomials. This does not impact on the methods themselves, but provides a finer picture of bias and precision for the purpose of comparing methods.

The general simulation procedure is as follows.

1. Complete data are simulated on $n = 300$ observations under the bivariate normal distribution with parameters

$$(y, x^p) \sim \text{BVN}\left(\begin{bmatrix} 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}\right). \tag{6.6}$$

   This implies the true analysis model is a linear regression of y on x^p. It is important to produce a strong association between x and y, such that power for the true analysis model is close to 100%. If $\text{Corr}(y, x^p) \approx 0$ in any simulated dataset the profile for $\hat{p}$ will be flat regardless of true p, and it becomes impossible to distinguish between good and bad methods.

2. 40% of values of x are set to missing under a missing at random mechanism*: the probability of x being missing is 0.2 when $y \leq 0$ and 0.6 when $y > 0$.

3. Missing values are multiply imputed using the method outlined in section 6.3.2 (`tuni`); see section 6.4.2.1 below.

4. For $p' = -2(.2)3$, the linear regression analysis model for $(y|x^{p'})$ is fitted, and the log-likelihood and Wald statistics based on MI data recorded. The log-likelihood for complete data and complete cases analysis are also recorded.

5. $\hat{p}$ is estimated as the value of p' maximising the log-likelihood or Wald statistic.

(*The step-MAR process breaks from the smoother linear-logistic MAR used in other chapters. This simulation study was originally run for two analysis models: normal errors with continuous outcomes and logistic regression with a binary outcome. Results were very similar and so

are not shown for logistic regression. The step-MAR process was used in order to make the interpretation of the two sets of results comparable.)

Four values of p are considered: 0, .5, 1 and 2. This process is replicated a total of 10,000 times. The mean and variance of $\hat{p}$ over the replications are of interest, and results summarised graphically.

6.4.2.1 Use of `tuni` in favour of `smcfcs`

For this simulation study it is useful to consider a parameter space for p that is closer to continuous, since this provides a finer view of the distribution of $\hat{p}$ according to different methods. It would not be possible to impute using SMC FCS because using only the eight standard FP1 transformations would favour these values of p, and it would be impossible to fit imputation models that include all x^p for $p = -2(0.2)3$. Further, since the data generating model involves dimension-1 FP transformations, `tuni` can used for imputation (see section 6.3.2).

6.4.3 Simulation results

The simulation results are displayed as a spikeplot in figure 6.2. The columns represent different true values of p; from left to right $p = 0, 0.5, 1$ and 2. Rows represent different methods for estimating p; from top to bottom complete data using the log-likelihood (CD-ll), complete cases using the log-likelihood (CC-ll), Wald statistics based on MI data (MI-Wald), and log-likelihoods based on MI data (MI-ll). The horizontal axes represent different values of $\hat{p}$, and are labelled with the powers used in S. The vertical axes display the frequency of a value being selected over 10,000 replications. The vertical axes all originate at 0 but the maxima are scaled individually to make each sub-plot as clear as possible.

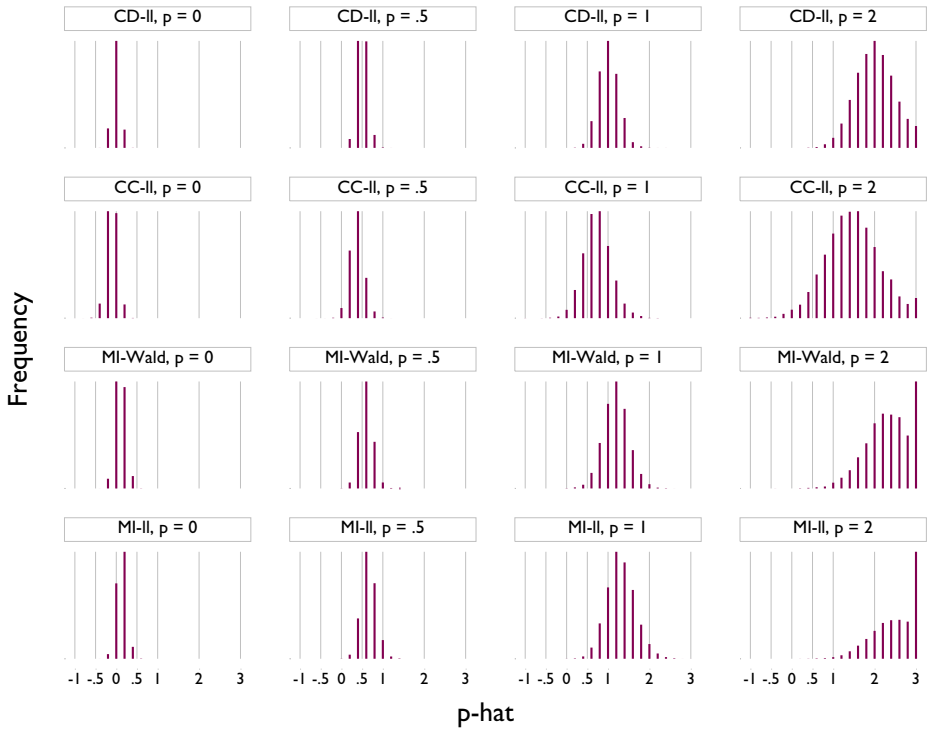
Note that across all methods it is apparent that the variance of $\hat{p}$ increases with p. This is not important in of itself but is noticeable in figure 6.2. To understand why, consider the following: if $p = 2$ in truth, $\hat{p} = 3$ is closer to the true model than $\hat{p} = 2$ is when $p = 1$ in truth. That is, a cubic is closer to a quadratic than a quadratic is to a straight line. It is thus expected that $\text{Var}(\hat{p})$ will be different for different values of p.

With complete data, use of log-likelihoods is unbiased and efficient, as expected. Data are missing at random, and so there is some bias associated with complete-case analysis, as well as low precision because the mean sample size with complete cases is 0.6×300.

The MI-Wald method exhibits a slight upwards bias for p. This bias is lowest for $p = 0$, increasing slightly for each larger value of p. The Wald method is also less precise than using complete-data log-likelihoods, but slightly more precise than complete-cases log-likelihoods.

The MI log-likelihood method also exhibits a small upwards bias, and this is slightly greater than the bias in the MI-Wald method. Again, precision is lower than for complete data and higher than for complete cases.

Figure 6.2: Simulation results: estimation of $\hat{p}$ according to method (10,000 replicates)



6.4.4 *Conclusions*

For the estimation of p the gold standard would be to have full data and estimate p based on log-likelihoods. Since this is infeasible with incomplete covariate data, Wald statistics and log-likelihoods based on multiply imputed data both offer an improvement over analysis of the complete cases. With imputed data, Wald statistics appear to do slightly better than log-likelihoods in terms of both bias and precision. However, the differences are small, particularly in relation to the set of powers in S typically used in fractional polynomial models. In this example, complete cases was the worst method, although often only slightly worse. It is worth noting that its performance will degrade further with multiple incomplete covariates.

Both the log-likelihood and Wald methods will be carried forward to the following section, which focuses on hypothesis testing. This is because the two methods for MI data under consideration in section 6.5 are again based on Wald statistics and log-likelihoods. The negligible differences between the two methods in this section mean that the two methods considered in the next section can be coherent: Wald statistics for estimating p and Wald statistics for model selection, or log-likelihoods for estimation of p and log-likelihoods for model selection.

Note that it is possible that bias in estimating $\boldsymbol{p}$ will be more problematic for multivariable fractional polynomials, where each covariate is addressed in turn. Selecting the wrong values of $\boldsymbol{p}_c$ for x_c may have knock on effects on exponents for $\boldsymbol{x}_{c'}$, magnifying problems.

6.5 METHODS FOR (M)FP MODEL SELECTION IN MI DATA

The candidate methods we consider for selecting between fractional polynomial models of different dimension are outlined below. The method which most closely resembles analysis with complete data in simulations will be recommended, provided it is superior to analysis of complete cases.

6.5.1 *Likelihood ratio tests based on 'stacked' data*

Wood, White and Royston[14] proposed new methods for hypothesis testing in multiply imputed data based on log-likelihoods, which extend naturally to multivariable fractional polynomial models. The methods, designated 'stacking', involved treating the M imputed datasets as one dataset of $n \times M$ observations, as in section 6.4. The best stacking method explored, which the authors designated W3, involved weighting the observations by $w_c = (1 - f_c)/M$, where f_c is the fraction of missing data for the cth covariate[14]. Equal weights are assigned to all observations for each test, but the weight changes according to the covariate being tested.

The use of the fraction of missing data for calculating weights is an attempt to weight each variable back to the correct amount of information; f_c is an approximation to the fraction of missing *information*. When the approximation holds, stacking will work well. This requires a fully observed outcome, missing values to be MCAR, and the covariate with missing values to be uncorrelated with other covariates. These conditions are extremely unlikely to be met in practice. When they are not met, stacking will perform less well, but it is of interest to investigate how quickly it degrades under departures from these conditions.

6.5.2 *Wald and ΔWald tests*

Wald tests based on Rubin's rules are a valid and powerful test for the significance of regression coefficients[14]. There are two subtleties related to fractional polynomial model selection that will stretch this validity.

Consider a single covariate x. For each model considered that includes x in some form, the Wald test vs. a null model for the parameters $\beta_1 \dots \beta_D$ is calculated following Rubin's rules. This is a standard Wald test.

It is not possible to calculate a Wald statistic to test between non-nested models. It is instead proposed to use the *difference* between two models' Wald statistics; the method is therefore designated 'ΔWald'. This is motivated by the similarity to calculating the difference between two models' χ^2 based on log-likelihoods. With complete data the two methods are asymptotically equivalent.

There is no guarantee that the ΔWald statistic will be positive. In one sense this is not a problem for testing – a negative Wald Statistic is not significant at any level – but this behaviour in one tail of the distribution could flag unusual behaviour in the tail we are interested in.

It is proposed that model selection proceeds on the basis of Wald tests where possible and ΔWald otherwise. The χ^2 reference distributions and their degrees of freedom are the same as those used in the function selection procedure with complete data.

Consider the test of FP1 vs. a null model. The Wald statistic is calculated is from β_{c1} and tested using χ^2_2 as the reference distribution. The df come from the two extra parameters, β_{c1} and p_{c1}, as compared with the null model, but the Wald statistic is calculated only from β_{c1}, a single parameter. There is thus reason to expect tests to be conservative. Conversely, recall from section 6.1.5 that $\widehat{\mathrm{Var}}(\hat{\beta})$ will be underestimated because it assumes $\hat{p} = p$. This results in the Wald statistic for β being too large. It is possible that these two wrongs will cancel out to some extent. This holds for the test of any model vs. the null; that is, the first test of the standard model selection procedure outlined in section 6.1.4.

For the remainder of this chapter, Wald tests with a genuine null and tests calculated from the difference in Wald statistics will be designated ΔWald

6.5.3 Other methods

Two other approaches to this problem might have been evaluated but were not after consideration. A brief description and justification of their omission is given below.

The first approach is Meng and Rubin's likelihood-ratio test for multiply imputed data[85]. This is derived from the asymptotic equivalence of Wald and likelihood-ratio tests, and was developed as a convenience tool to avoid calculation and inversion of M variance–covariance matrices in high-dimensional datasets. By aiming to approximate Wald tests, it will at best perform as well as Wald tests. In unpublished work, Patrick Royston has found the test to have extremely low type I error rates for fractional polynomials (Patrick Royston, personal communication). I do not therefore consider this approach further.

The second approach is that of Robins and Wang[7]. Readers are referred to [7] for details. While their approach is theoretically strong, there are several practical difficulties. A short description and the reasons it was deemed impractical in the context of MFP model building is given below.

Robins and Wang take a different approach to imputation: imputed values are drawn conditional on the observed data and the observed-data MLE $\hat{\boldsymbol{\alpha}}$ rather than drawing $\boldsymbol{\alpha}^*$ from the posterior. The imputer must save datasets containing the score function of the imputation model and the derivation of the score function with respect to the parameters of the imputation model. The analysis model is then applied to the M stacked imputed datasets assuming observations are independent. The analyst must then save a dataset and matrix containing the estimating equations of the analysis model and the derivative of these equations with respect to the parameters of the analysis model. The approach provides consistent variance estimation when the imputation and analysis models are incompatible, although it is unimpressive at small sample sizes.

While the Robins and Wang method has been implemented in some simple cases involving monotone missingness, the demands are too great to attempt to apply to (M)FP problems. For MFP even 'standard' imputation and analysis models tend to be complex. It is assumed that in practice the above would be too much to ask of researchers looking to use MFP models with incomplete data.

The aim of the following simulation study is to investigate the error rates of model selection by complete cases, ΔWald and stacking in the relatively simple context of univariable FP1 models, as usual comparing these to complete data analysis as the gold standard. All scenarios involve a continuous outcome and two covariates: x_1, which is incomplete, and x_2, which is complete. The outcome, y has regression function based on x_2 and an FP1 function of x_1. Throughout this section it is assumed that x_2 is known to be included in linear form.

6.6.1 *Design*

The following general setup of simulation study is replicated 5,000 times. Two sample sizes are used for all settings: $n = 200$ and $n = 500$.

Two continuous covariates are simulated from the model

$$(x_1^{-0.5}, x_2) \sim \text{BVN}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1 & \rho \\ \rho & \sigma_2 \end{bmatrix}\right). \tag{6.7}$$

The parameters μ_1 and σ_1 must be chosen with care as FP transformations will have more or less effect depending on the CV of the variable being transformed. FP transformations provide a degree of nonlinearity for a variable with mean 5 and variance 1, in that fitting all FP1 models may give fairly different log-likelihoods. If the mean is increased but the variance remains the same, FP transformations of the new variable will be closer to linear, in that the log-likelihoods for the FP1 models will be less different. This is why the `fracpoly`, `fp` and `mfp` Stata commands perform a preliminary scaling of FP covariates unless instructed otherwise. The parameter values used are $\mu_1 = 0.6$ and $\sigma_1 = 0.2$, meaning x_1 has mean 3 and variance 1 (approximately), and $\mu_2 = 3$ and $\sigma_2 = 1$. This affords the opportunity for certain transformations to provide better or worse fit than others. The value of ρ is set to 0 or 0.5 depending on the scenario.

The outcome y is simulated from

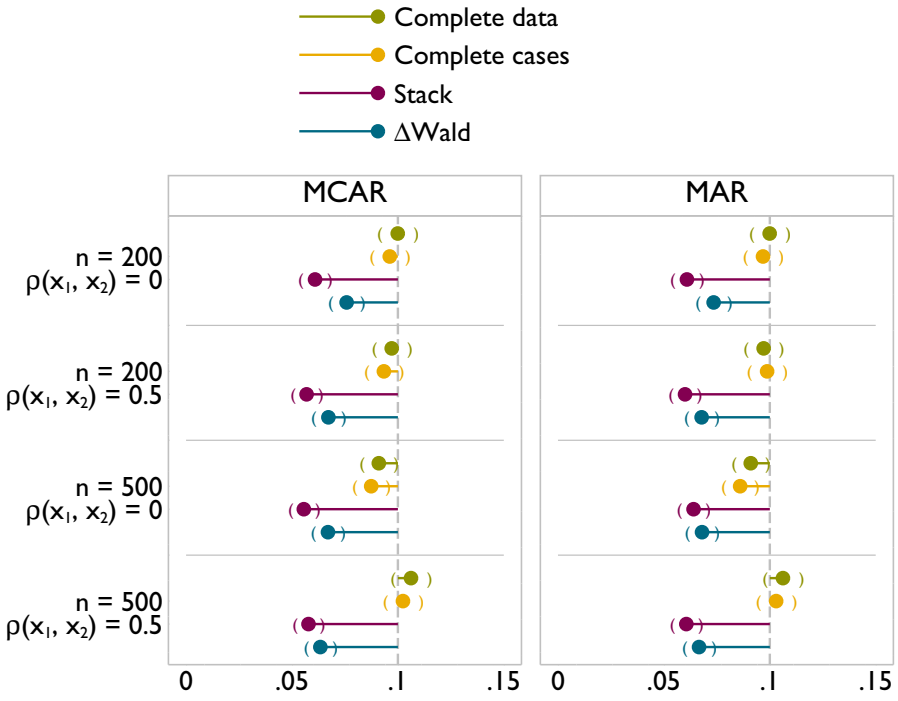
$$y_i \sim \text{N}(\beta_0 + \beta_1 x_{1i}^{-0.5} + \beta_2 x_{2i}, \sigma_y^2). \tag{6.8}$$

The linear predictor includes an FP1 function of x_1 and a linear function of x_2. The same value of p_1 was used in (6.7) and (6.8) so that the joint distribution of the complete data is $(x_1^{-0.5}, x_2, y) \sim$ MVN. For investigations of type I error, β_1 is set to 0. For investigations of power, β_1 is chosen such that, with complete data, the test for inclusion of x_1 has 90% power. Note that this means β_1 changes for different values of ρ and n. The true value of p_1 was chosen as -0.5 because this is relatively far from 1, meaning the test has a good degree of power. When complete data analysis had 90% power for a test of FP1 vs. null, the test of FP1 vs. linear had ≈80% power.

The parameters associated with x_2 are not of particular importance.For consistency with later simulation studies and with the x_1 parameters, these are set to to $\mu_2 = 3$ and $\sigma_2 = 1$, with β_2 chosen such that the likelihood-ratio test for inclusion of x_2 would have 90% power with fully observed data.

For the present study missingness is in x_1, while x_2 and y are complete. Two missing data scenarios are investigated. The first involves 30% of x_1 values being set to missing completely at

Figure 6.3: Type I error for test of nominal size 0.1 for FP1 vs. null on a single incomplete covariate



random. The second involves 30% being set to missing at random conditional on y, according to a linear-logistic model (equation 3.2, as used in sections 3.1.1 and 5.6). The association of R_{x_1} with y is similar to that used for simulations in other chapters: the degree of MAR is set so that the expected area under a ROC curve relating R_{x_1} to y is 0.65.

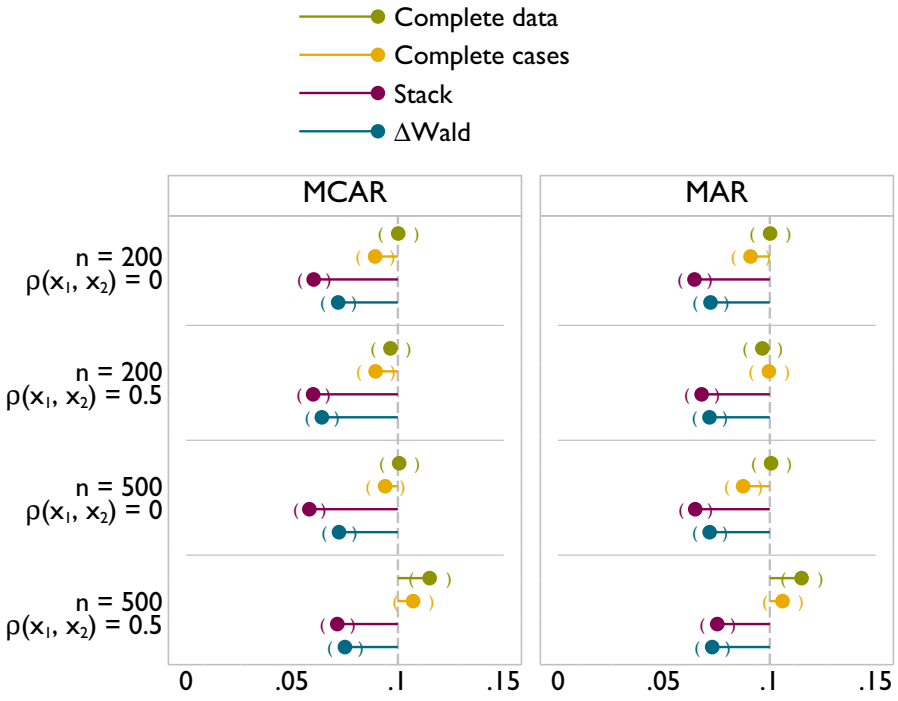
Missing x_1 values are imputed using `tuni` with 10 cycles of chained equations, and $M = 10$ imputations are used.

The function selection procedure is set up and run for complete data, complete cases, and MI data using stacking and ΔWald. The nominal size of tests used is $\alpha = 0.1$ throughout, following Ambler and Royston[86]. $D_{1\max} = 1$ and x_2 is always correctly included as a linear term. The most complex function considered is FP1. This is first tested against the null model and then against a model including x_1 as linear. The simulation summary of interest is the rejection rate for each method. When $\beta_1 = 0$, this should be as close to α as possible. When $\beta_1 \neq 0$, this should be as close to 1 as possible.

The scenario expected to best suit stacking is $\rho = 0$ with x_1 MCAR, because the 'FMI = FMD' approximation will hold in this case. Data being missing at random and $\rho = 0.5$ will provide a sterner test for stacking.

The test against a null model is based on a true Wald statistic. The test of FP1 vs. linear will provide a tougher test because it is based on ΔWald.

Figure 6.4: Type I error for test of nominal size 0.1 for FP1 vs. linear on a single incomplete covariate



6.6.2 *Results: type I error*

Figure 6.4 summarises the type I error of the test of the best-fitting FP1 model vs. a model that excludes it, with a test of size $\alpha = 0.1$.

Complete data analysis is seen to be close to this nominal value for both $n = 200$ and $n = 500$, and both $\rho = 0$ and $\rho = 0.5$. Complete cases analysis appears to have very similar type I error rates to CD, both under MCAR and MAR.

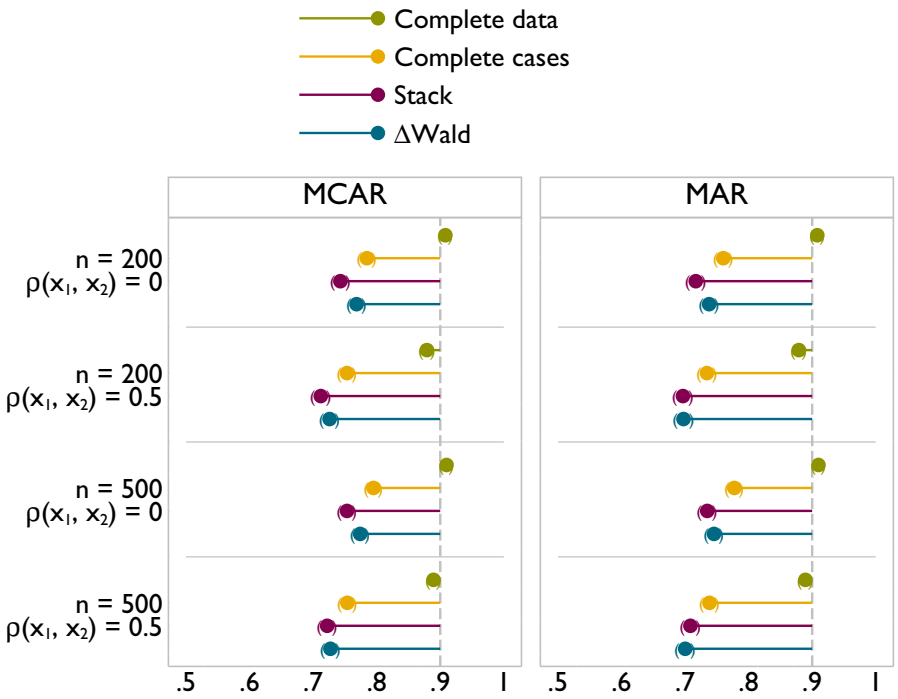
The type I error rates for stacking are lower, at around 0.05–0.06. This does not vary according to n or ρ, but is slightly worse under MCAR than MAR.

ΔWald also has low type I error rates, around 0.06–0.07. Again, there appears to be little effect of n or ρ. There is also slightly less contrast between MCAR and MAR scenarios.

Figure 6.4 gives results for the test of FP1 vs. a straight line, for the same simulated datasets as shown in figure 6.3. This scenario tests the ΔWald method. Note that results are based on the same simulated datasets as used in producing figure 6.3, but using a different test.

Results are similar to those for the test of FP1 vs. null: type I error is close to the nominal 0.1 for complete data and complete cases, while it is lower for stacking and ΔWald. For stacking, the small effect of the missing data mechanism is again seen. ΔWald performs surprisingly well, with results close to stacking, but usually slightly closer to the nominal 0.1 level.

Figure 6.5: Power of FP1 vs. null test of nominal size 0.1 with a single incomplete covariate



6.6.3 *Results: power*

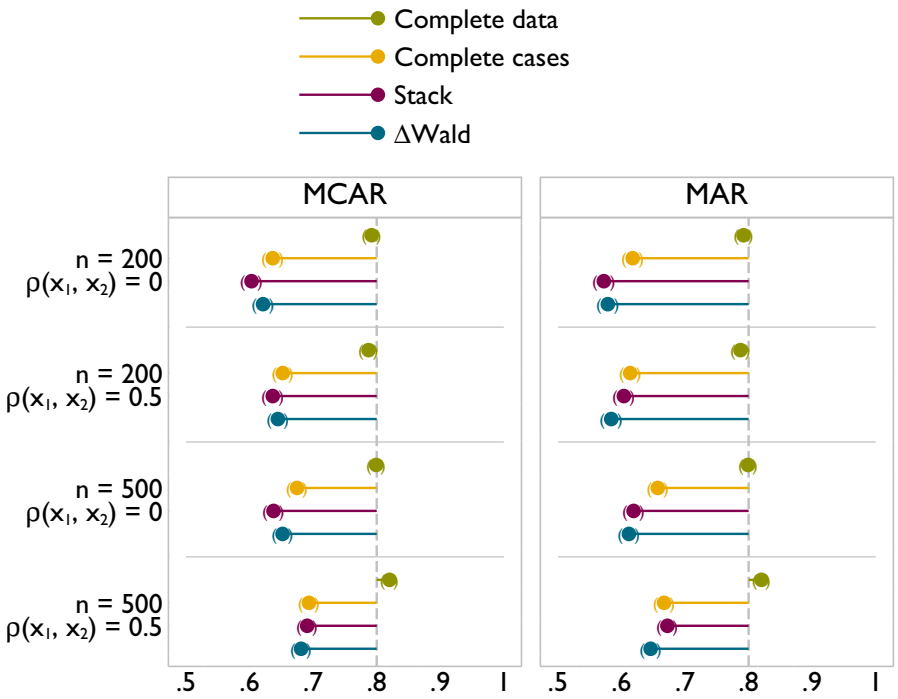
Figure 6.5 gives rejection rates for the FP1 vs. null test when there is an association between x_1 and y. Outcome was simulated such that with complete data this test had 90% power. This is seen throughout, subject to slight Monte Carlo error.

Complete cases analysis has lower power, with rejection rates of between 0.7 and 0.75. This is lower under MAR than MCAR, and also appears to be slightly lower for $\rho = 0.5$ than $\rho = 0$. Stacking and ΔWald have very similar behaviour, although the rejection rates are even lower than for complete cases. In this scenario they would be expected to have very similar power to complete cases; the lower power comes because the type I error is also low, making both tests conservative.

Figure 6.6 shows rejection rates for the FP1 vs. linear test for the simulations depicted in figure 6.5. The choice of $p_1 = -0.5$ for the true exponent meant that with complete data this test had power of around 80%, although this was not calibrated specially. Results are based on the same simulated datasets as used in producing figure 6.5, but using a different test.

Results are largely similar: complete cases has power of 10–20% lower than complete data analysis, and this tends to be slightly worse under MAR than MCAR. Stacking and ΔWald again have lower power than complete cases, but not substantially lower.

Figure 6.6: Power of FP1 vs. linear test of nominal size 0.1 on a single incomplete covariate



6.6.4 *Conclusions*

This simulation study has investigated rejection rates of two different tests for x_1, first under the null and then under the alternative. Type I error is well calibrated with complete data and with complete cases. For both stacking and ΔWald the type I error is too low. This serves to make the tests conservative, which has a knock-on effect on power. If a single covariate is incomplete, and this covariate is of interest, it may be desirable to use complete cases (with the caution that this may lead to bias in the estimation of $\boldsymbol{p}$). In other settings complete cases is likely to lose power. I explore these settings in sections 6.7 and 6.8.

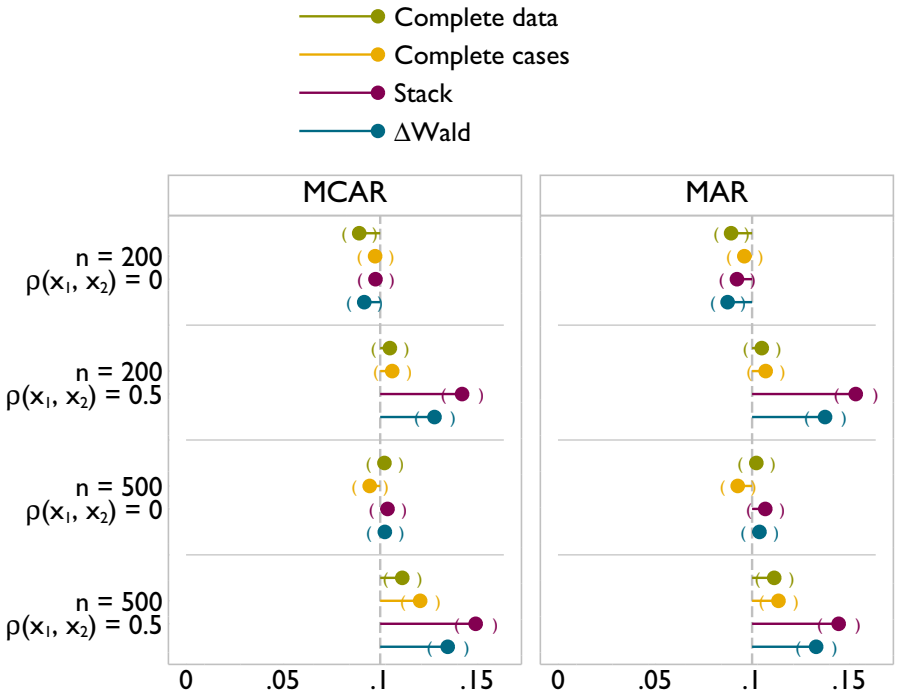
6.7 MFP model selection on two variables where a confounder is incomplete: simulation study

The following simulation study investigates the performance of methods when *M*FP model selection is used. In a change from section 6.6, x_1 is complete and x_2 is incomplete. If the association between y and x_1 is of interest but we wish to adjust for x_2, we may lose power despite observing all data on the variables that are of substantive interest.

6.7.1 *Design*

The design is similar to the study presented in section 6.6, which was set up with the aim of being consistent with the present and subsequent sections.

Figure 6.7: Type I error of FP1 vs. null test of nominal size 0.1 with an incomplete confounder



The first difference in the data generating model is that x_2 is subject to missingness rather than x_1. However, the proportion of missingness and MAR mechanism used are the same as those for x_1 in section 6.6.

The second difference is in the analysis: methods for model selection are run on both x_1 and x_2; this is full MFP, rather than FP. The parameter relating x_2 to y is again set such that in complete data, the test of FP1 vs. linear has 90% power.

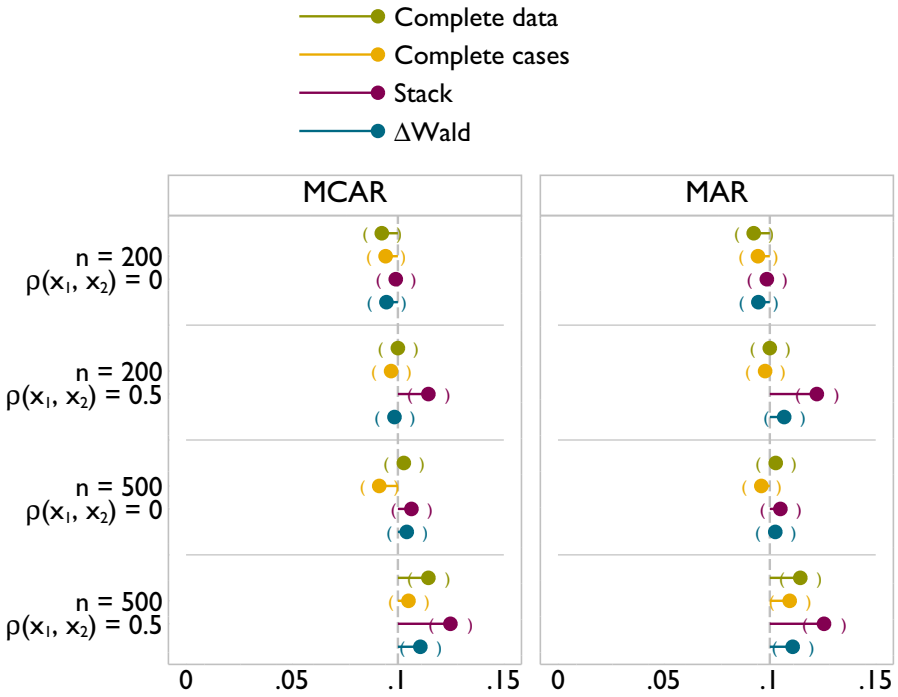
These changes mean that x_2 needs to be imputed rather than x_1. Even though the true regression function is linear in x_2, the MFP model selection means imputation accommodates the potential selection of an FP function. Further, because this involves FP terms for both x_1 and x_2, the draw of p_2^* is taken from the maximum of ll + J over possible values of (p_1, p_2). To reduce the computational burden, the candidate values of p_1 are rather coarser than p_2, using the standard set S.

Although model selection is done on both x_1 and x_2, only the rejection rate for x_1 is summarised. It may be obvious, but under certain settings the rejection rates for x_1 and x_2 are correlated, for example when $\rho \neq 0$.

6.7.2 *Results: type I error*

Figure 6.7 gives results for the test of FP1 vs. null. For complete data the rejection rate is close to 0.1, although appears to be slightly higher with $\rho = 0.5$ than with $\rho = 0$. This is because there is a true relationship between x_2 and y, but this test has 90% power and so x_2 will be incorrectly

Figure 6.8: Type I error of FP1 vs. linear test of nominal size 0.1 with an incomplete confounder



omitted 10% of the time. When $\rho = 0.5$ this relationship will sometimes be mopped up by the test on x_1. Complete cases analysis gives results very close to complete data.

Stacking and ΔWald have similar rejection rates across different simulation settings, although this is different to complete data and complete cases analysis. With $\rho = 0$ the rejection rate is close to 0.1, but it is higher with $\rho = 0.5$. The effect is a little more pronounced under MAR than MCAR, and more so for stacking than ΔWald.

For stacking this happens because under MAR and/or $\rho = 0.5$ the FMI ≈ FMD assumption is incorrect. Interestingly the effect of the value of ρ used here appears to be greater than the effect of the MAR mechanism. It is less clear why this happens for ΔWald tests.

Results for a test of FP1 vs. a straight line are given in figure 6.8. For complete data and complete cases analysis the increased type I error with $\rho = 0.5$ disappears, because now x_2 is always included, and its correct functional form is linear.

Stacking still has high type I error rates with $\rho = 0.5$. Meanwhile ΔWald has better type I error rates, comparable to complete data analysis, and the highest type I error rate is 0.11 and the lowest 0.09.

It is interesting that the small problems with ΔWald seen in figure 6.7 disappear in figure 6.8. This demonstrates that using ΔWald rather than a true Wald test is not actually a problem.

Figure 6.9: Power of FP1 vs. null test of nominal size 0.1 with an incomplete confounder

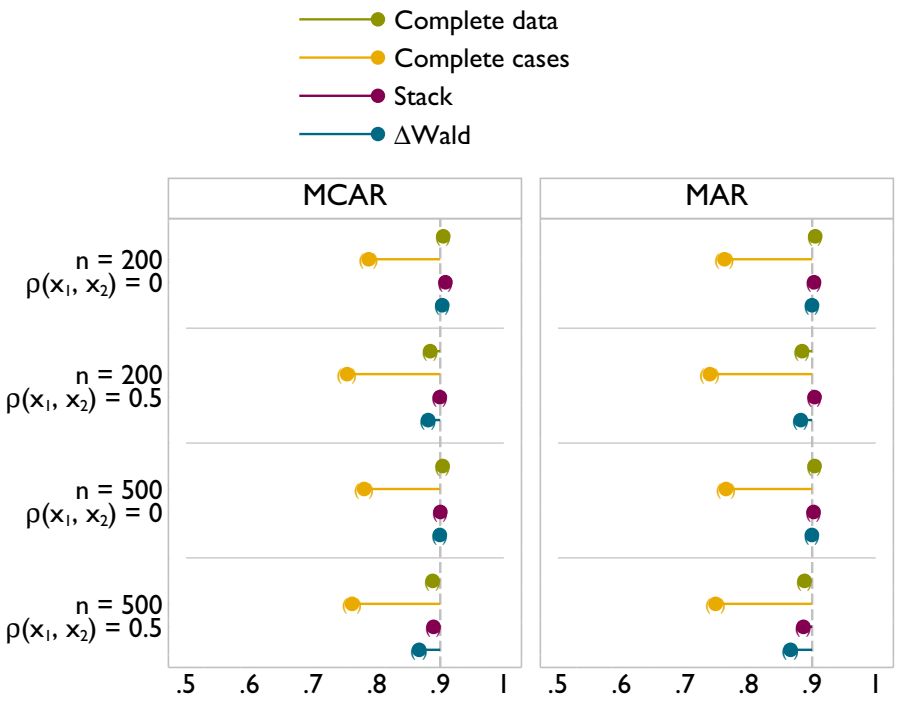
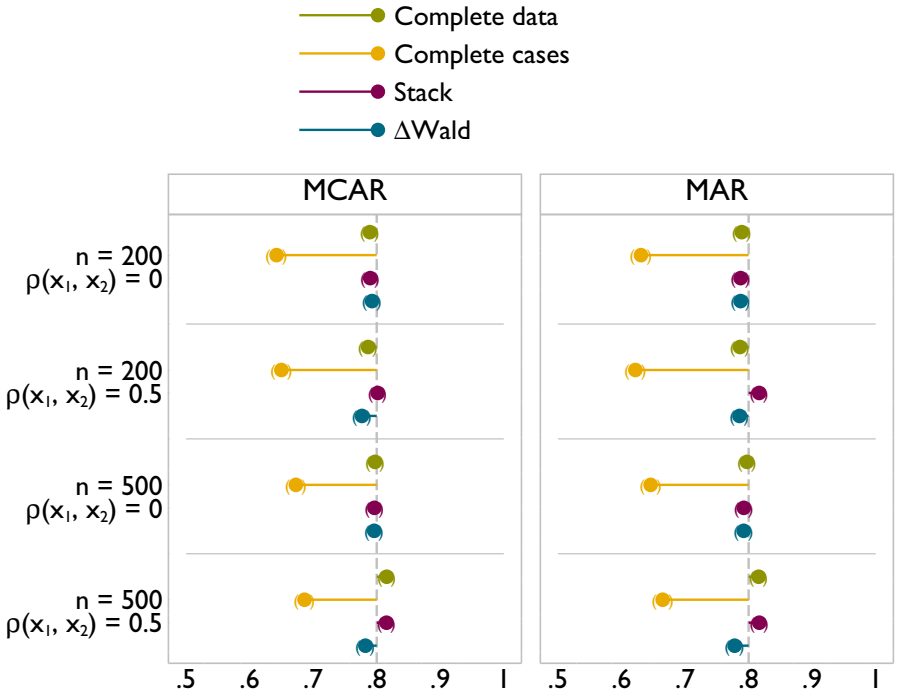


Figure 6.10: Power of FP1 vs. linear test of nominal size 0.1 with an incomplete confounder



6.7.3 *Results: power*

Results for the test of FP1 vs. null are given in figure 6.9. Complete data analysis has power of around 0.8 for all settings. There is again a dramatic loss in power for complete cases. This is 0.62 in the worst case and 0.69 in the best. Again, stacking has higher power comparable to complete data. In some settings the power of stacking is higher than complete data; this is again due to the high type I error rate. ΔWald also has power that is comparable to complete data, but never higher, due to the correctly calibrated type I error rates.

Results for the test of FP1 vs. a straight line are given in figure 6.10. These are extremely similar to results shown in figure 6.9.

6.7.4 *Conclusions*

This simulation study has investigated the rejection rates of two different tests for x_1 when x_2 is incomplete, first under the null and then under the alternative. Throughout, the methods have been compared to analysis of complete data – a gold-standard – and complete cases. The type I error is well calibrated for complete data and complete cases. Stacking and ΔWald can have high type I error when the correlation between x_1 and x_2 is not zero, although the correlation of 0.5 used here may be higher than would be observed in many medical datasets. When the type I error rates are high, ΔWald is closer to the nominal 0.1. Power is encouragingly substantially higher for both tests than for complete cases, and comparable to complete data.

This simulation study has demonstrated that in principle both methods can be used to test hypotheses with fairly good calibration of type I error and high power. Practically, the test of FP1 vs. null is fairly unrealistic. A confounder would not usually be included in a model because of its statistical significance, but due to prior knowledge of the disease under study. However, the test of FP1 vs. linear may be used for both variables in such a setting. It is encouraging that the ΔWald method performs as well for this test as for FP1 vs. null and has power close to complete data for this test.

The simulation scenarios presented in this section are further extended in section 6.8.

6.8 MFP model selection with two incomplete variables: simulation study

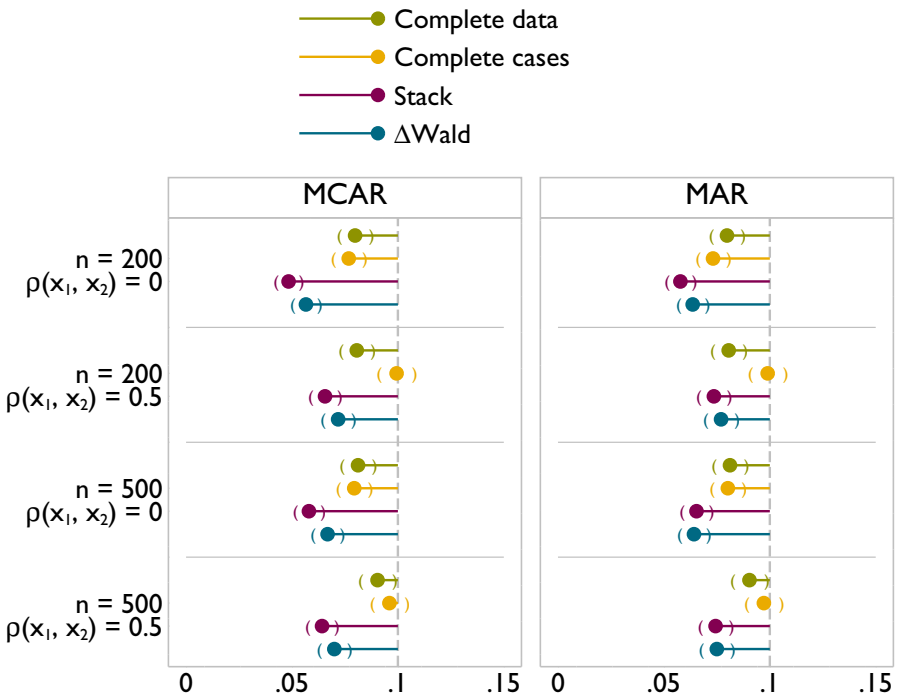
The following simulation study aims to further extend the scenarios presented in sections 6.6 and 6.7. Missingness will be in both x_1 and x_2. Testing procedures are again full MFP with $D_{\max} = 1$ for both variables.

6.8.1 *Design*

The design is largely the same as in sections 6.6 and 6.7, which were designed to be consistent with the present section.

The first difference in the data generating model is that x_1 and x_2 are both subject to missingness. The proportion of missing values and degree of MAR used are the same as in the earlier sections. Indicators of missingness, R_{x_1} and R_{x_2} are simulated independently. Under MAR, these are associated since both depend on y, but they are conditionally independent

Figure 6.11: Type I error of FP1 vs. null test of nominal size 0.1 on x_1 with both covariates incomplete



given y. The correlation means that the number of complete cases is slightly larger under MAR than under MCAR.

This change means that x_2 needs to be imputed, and in the same way as x_1 to allow for FP functions of x_2, even though the true regression function is linear in x_2. This means the imputation now uses the mice-based extension of `tuni`: `icet`.

Again, full MFP is used. Interest is again in the rejection rate for x_1 only.

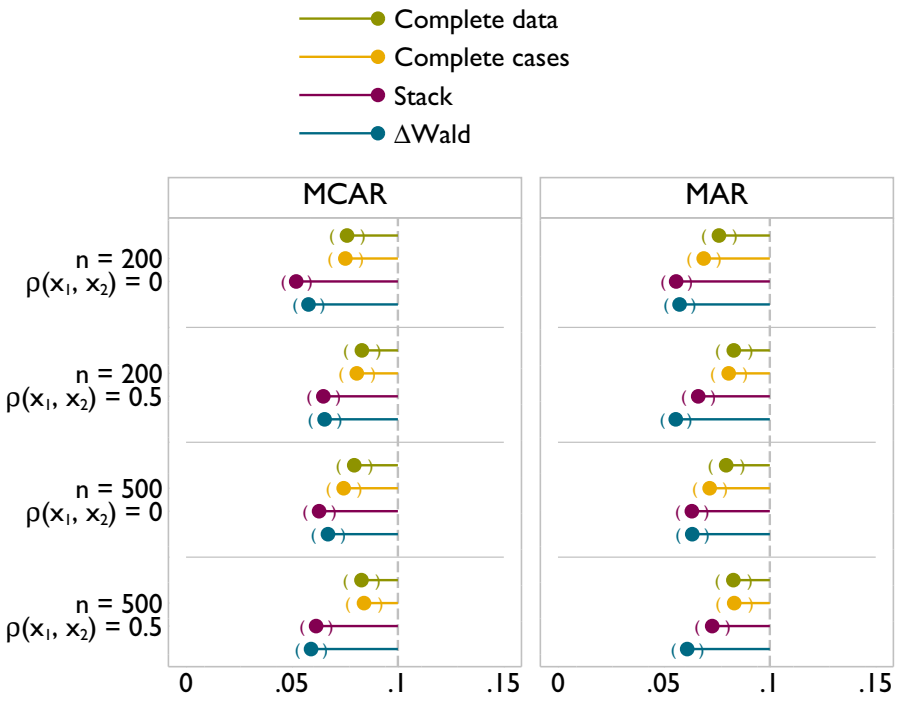
6.8.2 *Results: type I error*

The results of simulations investigating the type I error rate for tests of FP1 vs. null are given in figure 6.11.

The type I error rate is slightly lower than 0.1 for complete data analysis in all scenarios. With $\rho = 0$ and $n = 500$ type I error is closer to 0.1 than other settings. Complete cases analysis has very similar type I error to complete data with $\rho = 0$ and slightly higher with $\rho = 0.5$ (closer to 0.1). Stacking and Wald tests have marginally lower type I error than complete data or complete cases in all settings.

The effect on stack and ΔWald of varying n and MCAR/MAR is negligible. The rejection rate for ΔWald is slightly closer to 0.1 than stack under MCAR but under MAR they are more similar. For $\rho = 0.5$ the rejection rate is closer to 0.1 than for $\rho = 0$, and MAR results are similarly better than MCAR.

Figure 6.12: Type I error of FP1 vs. linear test of nominal size 0.1 on x_1 with both covariates incomplete



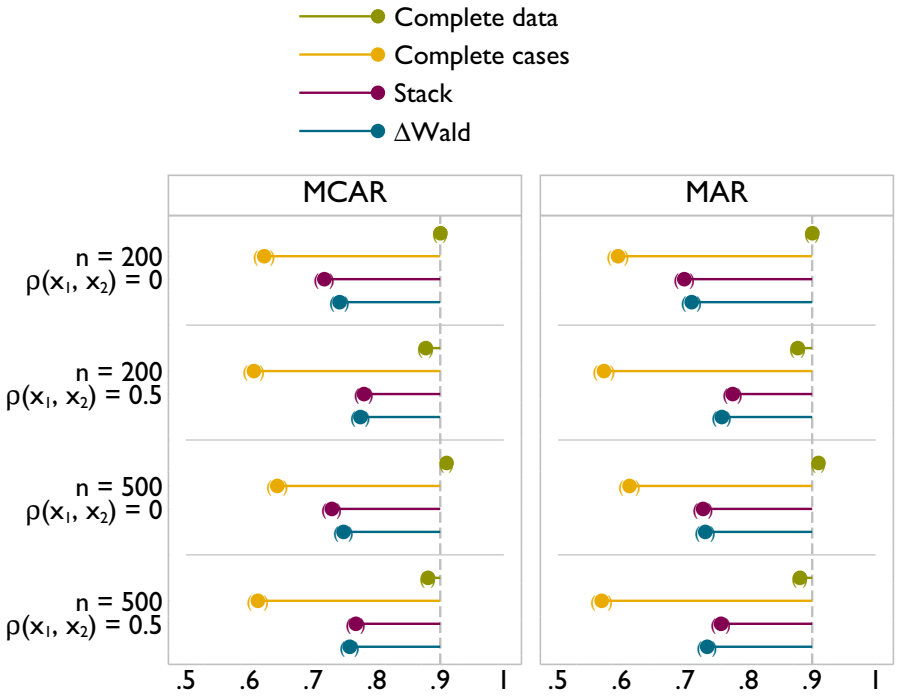
Results for the test of FP1 vs. a straight line are given in figure 6.12. Again the purpose of considering this test is to bring the ΔWald method under scrutiny.

Rejection rates are largely similar to those seen for the test of FP1 vs. null, and for all methods are consistently slightly low, at around 0.08–0.09. Complete data analysis again tends to be closest to the correct rejection rate of 0.1. Complete cases tends to have very similar rejection rates to complete data. Stack and ΔWald again have lower type I error rate, but are very similar. Under MAR with $\rho = 0.5$ stacking comes slightly closer to the nominal 0.1 level. The effects of sample size and of the type of missingness are small.

Overall These results demonstrate that the use of stack or ΔWald give us no cause for concern about the type I error rate for a test of FP1 vs. null or of FP1 vs. linear. There is slight miscalibration of type I error rates with both stacking and ΔWald beyond that observed for complete data and complete cases. There is little to choose between stacking and ΔWald. In absolute terms the performance of methods is generally good: in the worst of the scenarios considered the rejection rate is as low as 0.05, which should not cause concern, but is never higher than the nominal value of 0.1.

Having demonstrated that the type I error rate is well controlled, I now consider whether stack and ΔWald can offer and improvement in power over complete cases.

Figure 6.13: Power of FP1 vs. null test of nominal size 0.1 on x_1 with both covariates incomplete



6.8.3 *Results: power*

Results of simulations investigating the power of tests of FP1 vs. null are given in figure 6.13. Parameters of the data-generating model were chosen so that analysis of complete data would have 90% power for this test on x_1, and 0.9 is thus given as the reference line.

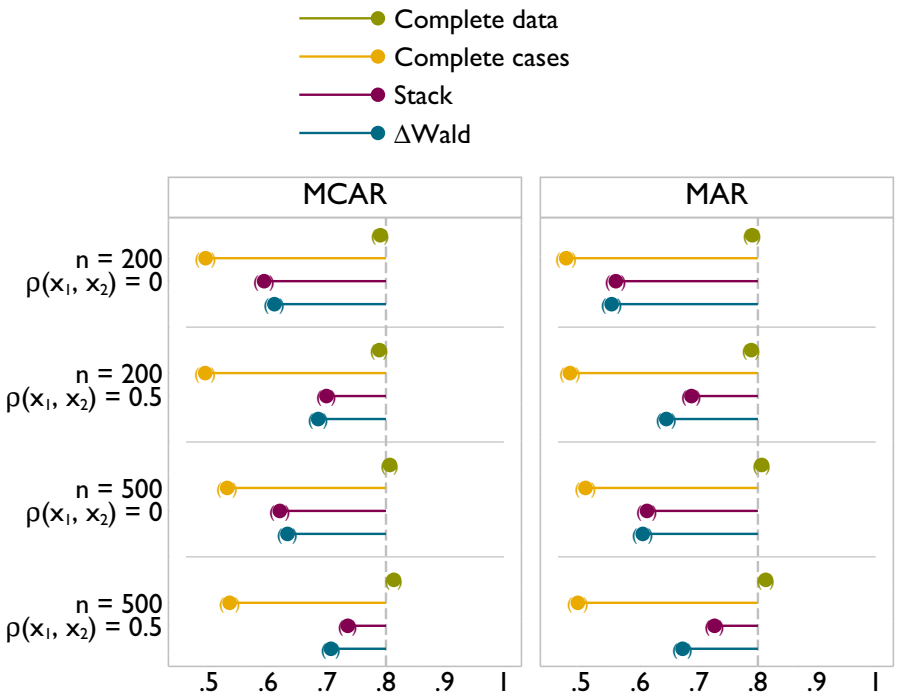
Complete data analysis has the highest power, as would be expected, and complete cases the lowest. Again, differences between stacking and ΔWald are small, with both offering improved power over complete cases. Gains in power are between 9% and up to 20%. Compared to complete data analysis MI always involves a loss in power of 10–20%.

There is little to choose between stack and ΔWald, and the differences are predictable from the results on type I error for this test.

Results for the power of the FP1 vs. null test are given in figure 6.14. As in section 6.7 the data-generating model was not calibrated to give any specified power, but the test was applied to the same simulated datasets summarised in figure 6.13. On average, this was approximately 80%.

The patterns that emerge in figure 6.14 are similar to those seen in figure 6.13. Again, complete data is the most powerful method, and power can be woefully low for for complete cases, sometimes below 0.5. Stacking and ΔWald tests offer increased power over complete cases, but the relative advantages/disadvantages are more variable than for the test of FP1 vs. null. Stacking seems to have an appreciable advantage over ΔWald under MAR and $\rho = 0.5$,

Figure 6.14: Power of FP1 vs. linear test of nominal size 0.1 on x_1 with both covariates incomplete



which is related to the different type I error rates. With MCAR and $\rho = 0$ ΔWald is slightly more powerful. In the worst cases around 7% power is gained over complete cases, and in the best cases the gain is over 20%.

6.8.4 *Conclusions on model selection*

The above simulation studies have demonstrated that both the stacking and ΔWald methods can be used to build multivariable fractional polynomial models in multiply imputed datasets.

The type I error is controlled to some extent by both methods. In the above simulation studies the type I error rates are 0.05 at the lowest and 0.14 at the highest for a test of nominal size 0.1. When a covariate of interest is incomplete but the outcome and confounder/s are complete there may be little gain from using MI instead of complete cases analysis: the type I error rates are lower and power is very similar (although under MAR complete cases will lead to biased estimation of p; see section 6.4).

When a confounder is partially observed but the variable of interest is complete the gains from using MI can be large. Type I error rates are higher than nominal in this setting, but generally not by enough to cause concern. The power gains of stack and ΔWald over complete cases can be large here, coming close to the power of complete data analysis in the best scenarios (although given that type I error is too high, power is strictly not comparable.)

When both the covariate of interest and a confounder are incomplete, results are the average of the two settings on their own. Again, stacking and ΔWald have type I error rates that are too

low – lower than complete data or complete cases. Power can be gained for one variable when the other is subject to missingness.

The simulation study described in section 6.8 is perhaps closest to the way MFP methods are most typically used, which is for building prognostic models. In such settings there will typically be several covariates with a complex missing data pattern. The results of section 6.8 demonstrate that in such a setting, the use of MI and stack or ΔWald will be beneficial for variables with smaller proportions of missing data, where complete cases may be useless. This will lead to the development of more accurate models of greater benefit to the medical community.

6.9 ILLUSTRATIVE EXAMPLE

This section compares MFP model selection using complete cases, stacking and ΔWald methods in the trauma data described in section 5.3. This is not intended to provide a comprehensive or clinically meaningful analysis, but is a demonstration that the methods do work in a large dataset containing variables of different types with awkward distributions and varying degrees of missing data, and that different methods can and do give somewhat different results.

Although the imputation approach described in section 6.3.4 has been implemented for some simple settings, the implementation is as yet unsatisfactory for applying to these data. The focus of this chapter was initially on model building, rather than imputation, and so the imputed data used in the analysis of reference [15] is used. This comes with a caution about the potential for bias towards the exponents used in the imputation model.

6.9.1 Analysis with complete data

Given complete data for all individuals, the analysis would involve a multivariable fractional polynomial logistic regression of massive transfusion on the covariates sex (binary), age (continuous), time to emergency department (continuous), blunt/penetrating injury (binary), systolic blood pressure (continuous), prothrombin time (continuous) and base deficit (continuous).

For the continuous covariates, $D_{\max} = 2$, except for time to emergency department, where $D_{\max} = 1$ (it is implausible that particularly short or long waiting times would have more similar probabilities of massive transfusion than medium times).

Because the aim of the analysis is to derive a prognostic model, and the number of candidate variables is small, the analysis errs on the side of caution by performing the test of FP$D_{c\max}$ vs. null with nominal significance set at $\alpha = 0.5$, meaning weak prognostic variables can be included, but they will be excluded if significance is extremely low. For the remaining tests, the significance level is set at $\alpha = 0.1$.

Due to missing data on covariates the above analysis is not possible. The dataset contains 2,456 complete cases (45%) of 5,693 individuals in total. This will potentially lead to bias and the tests losing power for all variables, particularly complete cases.

6.9.2 Selected models

Table 6.1 shows the variables and exponents selected by complete cases, stacking and ΔWald. For all three methods convergence was achieved after two cycles through the MFP algorithm.

Table 6.1: Models selected in trauma data. The numbers give the exponents selected for each variable in the final model.

	Complete cases	Stack	ΔWald
Age	−2	0.5, 1	1, 1
Time to emergency dept.	1	1	1
Systolic blood pressure	1	1	−2, 0.5
Base deficit	1	−1	−0.5
Prothrombin time	−0.5, −0.5	−0.5, −0.5	−0.5, −0.5
Injury† (blunt/penetrating)	1	1	1
Sex†	–	1	1

†For binary variables an exponent of 1 indicates inclusion in the final model.

The three methods all selected different final models. Only time to emergency dept. prothrombin time and injury type were included in the same form in all models. Complete cases selected the simplest model overall, omitting sex, selecting linear functions for time to emergency dept., systolic blood pressure and base deficit, an FP1 for age and FP2 for prothrombin time. Stacking included sex in the model and further included base deficit as FP1 and age as FP2. The ΔWald model was even more complex, selecting an FP2 for systolic blood pressure, which was linear in both other models. Note that the FP model selected in complete cases differs from that used in chapter 4 because there $D_{c\max} = 1$ for all c.

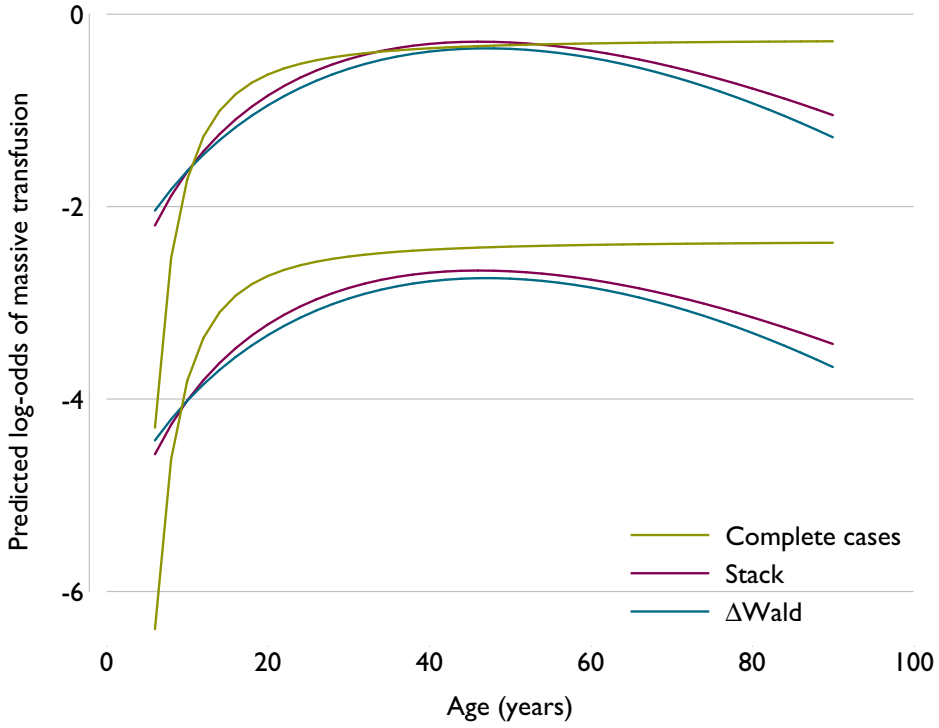
The values of $\hat{p}_c$ selected by the models were also sometimes different even when D was the same. For age, FP2 functions were chosen by both stacking and ΔWald, $\boldsymbol{p}_c = (0.5, 1)$ for stacking and $(1, 1)$ for ΔWald. For base deficit, $p_c = (-1)$ for stacking and (-0.5) for ΔWald. This is unlikely to be due to differences in power between the methods, since these are very similar. It is more likely that to be related to the result in section 6.4, where stacking was shown to estimate $\boldsymbol{p}$ with slightly more bias than ΔWald. With MFP, this can occur at any step of a cycle, and if the wrong form is selected for one variable then this will have a knock-on effect on the form for the next variables, unless these are uncorrelated.

The three methods selected different $\hat{\boldsymbol{p}}$ and D; because $\hat{\boldsymbol{\beta}}$ are only comparable conditional on $\hat{\boldsymbol{p}}$ and D, comparing the values of $\hat{\boldsymbol{\beta}}$ from the three selected models would be meaningless.

Instead, fitted FP functions are compared for age and base deficit from each of the three models for two plausible individuals. The data used are made up but are also realistic representations of individuals within the dataset. The covariate values used are given in table 6.2.

Figure 6.15 shows the comparison of fitted functions for these individuals across a range of values of age (from 6–90 years) and base deficit (from −5 to 20), both of which span most of the observed range of the covariates. Stack and ΔWald return very similar fitted functions within the range considered, despite selecting slightly different $\hat{\boldsymbol{p}}$. For both variables the fitted functions for complete cases are a completely different shape; in particular the effect below age 10 seems fairly extreme. The curves for the two individuals are also closer together for complete cases, possibly indicating that the other variables provide a lower degree of prognostic separation than for the models selected by stack and ΔWald.

Figure 6.15: Fitted functions for age and base deficit according to method of model selection



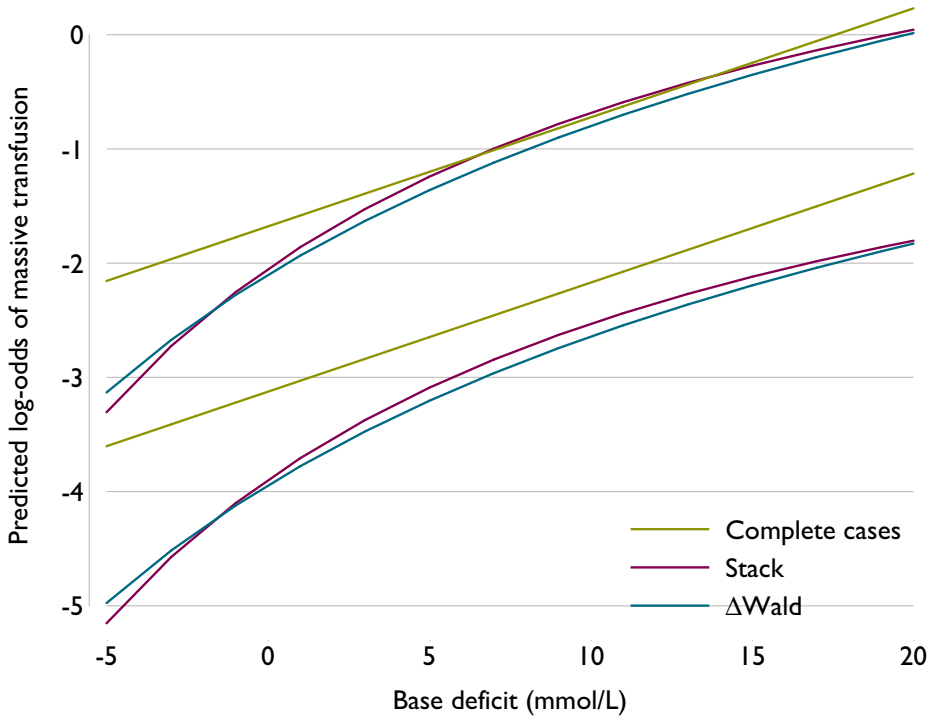


Table 6.2: Table of covariate values for two imaginary individuals.

Individual	A	B
Age in years	*34	*24
Time to emergency dept. in minutes	63	73
Systolic blood pressure in mmHg	91	130
Base deficit in mM	*13.5	*5.4
Prothrombin time in seconds	16.8	14.4
Injury type	Blunt	Blunt
Sex	Female	Male

*Values of age are fixed when base deficit is varied in figure 6.15 and vice versa

6.10 DISCUSSION

This chapter has attempted an attack on the problem of combining MI with MFP methodology, splitting the problem into three components: imputation, estimation of exponents, and model selection. The results of each component have been utilised and carried forward to subsequent components of work.

6.10.1 *Imputation for fractional polynomials*

Two approaches to imputation have been described that can considered proper for (M)FP model building. The first, based on the approximate Bayesian bootstrap, was useful for the investigations presented in this chapter, though strictly it is limited to imputing for $D_{\max} = 1$. The Stata implementation could be made more general to allow for other variable types. The second method is a more general approach that should in principle work for larger values of $D_{\max}$. The method failed to perform adequately in this work, but tuning the implementation to work with FP models is an area of planned development for `smcfcs` (see appendix D).

Neither method is controversial; both adapt existing methods to the task of fractional polynomial model building. As such, neither was assessed formally for the purpose. It is possible that there are other approaches to imputation that may be appropriate and could improve on the approaches used here.

6.10.2 *Model selection in multiply imputed data*

There are two distinct component parts to building fractional polynomial models: estimating the best exponents for a covariate, and selecting the appropriate complexity of FP function for that covariate.

The results of the simulations presented in section 6.4 indicate that for estimation of p, using log-likelihoods or Wald statistics on MI data are both superior to using log-likelihoods based on complete cases. This was a single missingness mechanism and the performance of complete cases could degrade further with different mechanisms, but would be unbiased under MCAR. Wald statistics appear to be slightly preferable, although differences appeared to be small.

Despite the slight advantage to using Wald statistics for estimation of exponents, both methods were carried forward to model selection work, which assessed testing procedures based on stacking and ΔWald. It was judged to be advantageous to have a coherent method for estimation of $\boldsymbol{p}$ and for variable selection: log-likelihoods before stacking, and Wald statistics before for ΔWald testing.

These methods were used in sections 6.6–6.8. Overall the type I error rates for ΔWald and stacking were less well calibrated than for complete cases, however, power could be higher even with lower type I error rates. The missing data mechanism and patterns in simulations were relatively simple, but complete cases can become extremely inefficient with complex missing data patterns, so whenever the proportion of complete cases is low it is preferable to base the analysis on MI. Stacking and ΔWald based on proper multiple imputation represented an improvement.

There is little reason to favour stacking over ΔWald method, or vice versa. For estimation of p, ΔWald seems to be very slightly less biased. In some scenarios stacking is more powerful while ΔWald is in others. For analysis with incomplete covariates either method should be used in preference to analysing complete cases.

Below is a practical suggestion for combining MI with MFP methods (6.10.3), and another for MFP, which unfortunately does not generalise to multiple imputation (6.10.4).

6.10.3 *Improve estimation by re-imputing and re-fitting the selected model*

In using SMC FCS and the approximate Bayesian bootstrap method for use with FP, it was noted that both methods make the imputation model/s semi-compatible with the analysis model. For reasons of efficiency, it may be preferable to use a smaller imputation model, and to produce imputations that are fully- rather than semi-compatible with the analysis model.

Consider fractional polynomial models in complete data. Although $\hat{\boldsymbol{p}}$ is 'estimated', it is subsequently treated as fixed and known. In the same spirit, it is possible to initially impute as outlined in sections 6.3.3, select the model, then impute a second time, this time using a more restricted imputation model.

For SMC FCS the restricted model would specify the selected analysis model at the second imputation step, and finally fit the selected model to the second set of imputations without any further model selection. This assumes in the imputation that $\boldsymbol{p}$ is fixed and known, which is exactly how $\boldsymbol{p}$ is treated by fractional polynomials with complete data, and may increase efficiency.

With `icet`, the draws of p^* might be replaced by imputing $x_j^{\hat{p}}$, where $\hat{p}$ is the exponent selected by FP.

Use of either method may have advantages for the analysis: if the selected exponents are accurate, the restricted imputation strategies will result in superefficient imputations[23]. Conversely, if the selected exponents are inaccurate the estimates after restricted imputation will be misleading. It is up to researchers to decide whether they are willing to take this risk.

6.10.4 *Improving the reference distribution for test statistics*

[Note – the below idea comes from discussions with Professor James Carpenter of the MRC Clinical Trials Unit at UCL and the London School of Hygiene & Tropical Medicine.]

Section 6.8.2 demonstrated the type I error of MFP model selection procedures is not necessarily equal to the significance level ascribed to the tests. While the absolute difference tended to be small, it may be possible to improve the error rate.

The problem with the stack and ΔWald methods is that the reference distributions used are not quite correct under the null. However, it is possible to use the bootstrap to simulate the distribution of the test statistics under the null. Two similar techniques are outlined below.

One possibility is to use the parametric bootstrap. For any parameter(s) considered, outcomes can be simulated stochastically from a parametric model in which the parameter(s) of interest is set to zero but all others are set at their observed value. Repeating this B times leads to an empirical reference distribution obtained under the null, to which the observed value of the test statistic can be referred.

A second possibility is to bootstrap residuals. The null model is fitted, and the expected outcome $\hat{y}$ is calculated for all individuals. The vector of residuals $y_i - \hat{y}_i$ is also stored, and a bootstrap sample of this vector taken. This sample is merged with the vector $\hat{y}_i$ and the test statistic of interest recorded. Repeating the process B times again leads to an empirical reference distribution. This method may be slightly better than simulating outcomes because it uses the empirical distribution of residuals rather than making parametric assumptions about the distribution.

With multiply imputed data neither bootstrap approach could be done prior to imputation because both rely on fully observed covariates to be able to bootstrap under the null, and so a bootstrap-then-impute approach is not possible. It is not clear how an impute-then-bootstrap approach would work. Bootstrapping MI datasets independently would return a reference distribution for the complete, rather than incomplete, datasets.

This approach has not previously been explored for fractional polynomials with complete data but is uncontroversial and so can be used in that context.

7 Discussion

7.1 SUMMARY OF THESIS

The problem of missing data is widespread in medical research, potentially leading to analyses which are inefficient and sometimes biased. While methods are available for dealing with missing data, almost all such analyses rely inherently on untestable assumptions. The best way to deal with missing data is to avoid them, and efforts to do so should be invested during data collection. This will minimise reliance on untestable assumptions.

Despite all efforts, missing data will inevitably continue to plague medical research. When missing data do occur it is important to approach analyses by considering plausible mechanisms by which data might have gone missing, and to perform analyses which are valid under the stated assumptions. The 'missing at random' assumption is the most general for which analyses can be performed without explicit modelling of the missing data mechanism, but this does not make it plausible. Any important analysis based on the MAR assumption should be supplemented with sensitivity analyses that make alternative relevant assumptions about departures from MAR[87].

This thesis has investigated statistical methods for dealing with missing data when they do occur, with covariates either missing at random or missing completely at random. The three investigations have centred around incomplete covariates that are assumed to have a nonlinear effect on outcome. In the absence of missing data (assuming covariates are measured without error), it is unnecessary to specify a probability model for the covariates. When the analysis model contains one or more nonlinear functions of covariates it can be difficult to specify a sensible model for the covariates, and inference may be invalid under missing at random.

Researchers are assumed to put less careful thought into specifying the imputation model/s than analysis model, the missing data being a nuisance to the analysis of substantive interest. This implies that the imputation model is more likely to be misspecified, or to have a greater degree of misspecification. In this work attention has been paid trying to find imputation models that are compatible with the analysis model. The MI approach used in Stanworth et al[15], where the normalising transformations of covariates for imputation models determined their form in the analysis model, seems to be an anomaly in the literature. Chapters 2–5 assume rightly that the analysis model is correctly specified, and consider different approaches to imputation.

7.1.1 Predictive mean matching and local residual draws

In chapters 2, 3 and 4, predictive mean matching and local residual draws were introduced, reviewed and investigated via simulation.

In chapter 2, PMM and LRD were introduced and the statistical literature developing and using either method was reviewed. The review initially focused only on PMM, but LRD was discovered while reviewing articles and, due to largely similar promise, the scope of the review was extended to LRD. LRD was noted to have more potential than PMM under (strong) MAR mechanisms. The specific MAR mechanism used was designed to stretch PMM and LRD to their limits and so always involved missingness in the tails of the distribution. An alternative MAR process could lead to more missingness at the centre of the distribution, which would not cause the same problems for PMM and LRD.

The review considered what was already known about PMM and LRD, and what had not been investigated. Interest was particularly focused on methods for defining the matching metric and on the choice of k (the size of the donor pool). Little's article introducing PMM had described two methods for defining the donor pool[33], designated type 0 and 1 in this thesis. Heitjan and Little introduced a third method, designated type 2[34]. A small number of authors, notably Schenker and Taylor[28], had considered whether donors should be selected from a fixed value of k or according to whether they lie within a defined distance of the donee. Few authors had compared the effect of varying k. Basing donor pools on a fixed donor–donee distance was deemed impractical here, and simulation work instead considered different fixed values of k.

Univariable simulation studies in chapter 3 demonstrated that when the imputation model is correctly specified, both PMM and LRD could be biased upwards or downwards, but the bias was miniscule and so of no practical concern. Increasing the value of k reduced the variance and also improved coverage. With misspecified imputation models posterior draws exhibited large biases towards the null, which were usually attenuated but not resolved by PMM or LRD. In certain cases both PMM and LRD could be upwardly biased. In univariable simulations it was clear that larger values of k improved the statistical properties: $k = 10$ for PMM and $k = 20$ for LRD are recommended. For PMM, type 1 matching appeared to be better than type 2, while type 2 appeared to be slightly better for LRD.

A multivariable simulation study with multivariate missingness in the covariates (chapter 4) used the recommended versions of PMM and LRD from the univariable simulation studies. Simulation parameters and missing data were based on the trauma dataset described in section 1.5.1. Different transformations of covariates prior to imputation were also investigated. Of the imputation methods considered, all led to downwards bias, while analysis of the complete cases was unbiased, despite data being simulated under MAR. This was due to the missing data mimicking case–control sampling for the covariates of interest. Further, while the variance of imputation methods was lower than complete cases, coverage was also too low, while coverage for complete cases was close to the nominal level.

It was concluded that while PMM and LRD were promising, they performed best in settings where posterior draws is a superior alternative. In other settings both methods usually outperformed posterior draws but were not necessarily superior to the analysis of complete cases. Neither method should therefore be treated as a general solution to the problems associated with misspecified imputation. I advise making serious efforts to specify the imputation model correctly, although this may be extremely difficult.

7.1.2 *Multiple imputation for a ratio covariate*

The ratio of two variables is a mathematically strange choice of covariate (see e.g. [72, 88]), but its use is common in medical research, and one or both of the variables making up a ratio are often incomplete. This work was motivated by an infamous analysis where missing values of a ratio were handled by imputing the two components separately and then calculating their ratio passively[19]. The association between the ratio and outcome was close to null with very high precision. However, a non-null association of consistent magnitude had been repeatedly observed in multiple previous studies. Following some rapid-responses to the article, the authors revised their imputation and the resulting estimate was in agreement with what was expected.

Six approaches to imputation were investigated, including the approach used by Hippisley-Cox et al[19]. Each was based on a fairly simple imputation model that could be easily implemented in general-purpose statistical software. The six approaches are recapped briefly below:

M1. Treat the ratio as missing if either of its components are missing, and impute it as normal (compatible; active imputation; may not use all observed data).

M2. Impute ratio and the numerator as bivariate normal (semi-compatible; active; may not use all observed data; ignores the deterministic relationship between numerator and denominator and their ratio).

M3. Impute ratio and the denominator as bivariate normal (semi-compatible; active; may not use all observed data; ignores the deterministic relationship between numerator and denominator and their ratio).

M4. Impute the ratio, denominator and numerator as trivariate normal (semi-compatible; active; makes use of all observed data; ignores the deterministic relationship between numerator and denominator and their ratio).

M5. Impute the numerator and denominator as bivariate normal, and calculate the ratio (incompatible; passive; makes use of all observed data; respects the deterministic relationship between the numerator and denominator and their ratio).

M6. Log-transform the numerator and denominator, impute them as bivariate normal, and exponentiate the difference to calculate the ratio (incompatible; passive; makes use of all observed data; respects deterministic relationship between the numerator and denominator and their ratio).

In the Aurum cohort, all methods for imputing the ratio led to similar substantive conclusions about its effect on outcome. In the Epic-Norfolk cohort, approach M5 led to very different conclusions to the other five.

It was hypothesised that the differences were due to the coefficient of variation of the denominator, which was around 0.1 in the Aurum dataset and 0.3 in Epic-Norfolk. A simulation study confirmed that when the CV of the denominator was 0.3 the approach led to substantial downwards bias and was grossly inefficient. A key message was thus against passive imputation without transformation. The simulation study also showed that of the models considered, approach M6 will tend to provide the best inference, though the active imputation approaches are generally not bad.

7.1.3 *Combining multiple imputation with multivariable fractional polynomials*

Chapter 6 considered the problem of using multivariable fractional polynomials when one or more covariates are incomplete. The difficulties in combining these techniques were broken down into three parts:

1. Imputing covariates when their role and form in the analysis model is unknown.
2. Estimating the exponents $\boldsymbol{p}$ in multiply imputed data.
3. Selecting (multivariable) fractional polynomial models in multiply imputed data.

The exponents are estimated in the analysis, and so multiple imputation must allow for this uncertainty. A method for imputing (potentially multiple) FP1 functions based on the approximate Bayesian bootstrap was proposed, coded in Stata and used in simulations. Bartlett et al.[26] have developed a different method – 'substantive-model-compatible fully-conditional-specification' – that was promising for imputing (possibly multivariable) fractional polynomials with dimension > 1, but on applying `smcfcs` in simulations it returned very low type I error and power, usually due to imputing outliers with high leverage. The first method was thus extended in Stata to impute multiple continuous covariates in a chained equations procedure.

For estimation of exponents two methods were compared to complete cases and complete data. The first used likelihoods in multiply imputed data and the second used Wald statistics. Both methods were an improvement on complete cases, which estimated the exponents with some bias. Wald statistics had lower bias and was more efficient than stacking, though the differences were minimal.

For model selection, two methods were investigated, and compared as before to complete data and complete cases.

Stacking
selects exponents using likelihoods. It then treats the M imputed datasets as one large dataset and, for each covariate, fits the best model of dimension d with weights based on M and the fraction of missing data for that covariate. Comparison of the best models of different dimension proceeds by referring likelihood ratio tests to a χ^2 distribution, as would be done in complete data.

Δ*Wald*
selects exponents by maximising the Wald statistic for each complexity of model considered. Model selection then proceeds on the basis of Wald tests, when comparing the FPD model to the null, and the difference between Wald statistics when comparing two non-null models.

In simulation studies the rejection rates of both methods were found to be worse (and of inconsistent direction) than complete cases when only the variable of interest was incomplete. However, both performed well when only a confounder was incomplete, giving type I error rates close to the nominal level and power close to that achieved with complete data. When the variable of interest and a confounder were both incomplete both methods had low type I error but large gains in power over complete cases.

7.1.4 *Themes*

The three topics above were approached as separate pieces of research, but there are some clearly linked ideas and themes. The introduction noted that multiple imputation is a well developed methodology for some simple settings, but less so for the sort of complex settings applied researchers typically face. All three investigations considered specific, seemingly simple settings with complex issues; for all, the analysis model included nonlinear functions of covariates where one or more was incomplete.

¶*Transformation towards marginal normality* Chapters 4, 5 and 6 all used passive imputation methods, sometimes after some initial transformation. In chapter 4 this approach worked as well as any other when a marginally-normalising transformation was used before imputation. Passive imputation without any transformation was the worst approach.

The two passive imputation approaches used in chapter 5 reflected these results. Taking logs before imputing the denominator and numerator and passively imputing the ratio gave good results. Passive imputation without prior transformation was the worst approach, as in chapter 4.

In chapter 6 the `tuni` approach relied on automated passive imputation. A draw of p, p^* was taken from a bootstrap sample, then in the original dataset $x_j^{p^*}$ was actively imputed and x_j^* passively imputed. This approach performed well. (This may be because in simulations $x^p \sim \text{N}$ and so draws of p^* would be centred about the tansformation that was simultaneously normalising and compatibilising.)

¶*Transformation to improve compatibility* A theme running through all three projects in this thesis has been the compatibility of the models for imputation and for analysis. Rubin's combining rules for multiple imputation inference assume the imputation and analysis models are correctly specified[1]. Compatibility is a necessary (but insufficient) condition for the models to be correctly specified, while incompatibility guarantees that at least one of the models is misspecified (without reference to any data), placing Rubin's rules on shaky ground. This can lead to bias, over- or under-estimation of variance or *superefficiency*, and over- or under-coverage of confidence intervals. In analysing a dataset with missing data, the impact of model incompatibility is not usually clear. For the purposes of this thesis, compatibility or semi-compatibility of the imputation and analysis models has been assumed to be desirable.

In chapters 4 and 5 compatibility appeared not to be the most important consideration in developing an imputation model, and was often in tension with a normalising transformation. However, compatibility is nonetheless desirable for imputation, provided it does not sacrifice all other desirable features of the imputation model.

An initial run of the simulation study in section 6.4 gave results that appeared to be a mistake: log-likelihoods and Wald statistics estimated exponents with greater precision than the complete data analysis! This was not a coding error: it occurred because imputation initially calculated x^p using the true value of p and then imputed it.

This is similar to the scenario described by Meng[22] and later Rubin[23], where the imputation model makes a correct assumption about the value of a parameter, resulting in the phenomenon of *superefficiency*. Rather than imputing with the correct amount of uncertainty,

each imputed dataset is drawn conditional on the true parameter value p rather than conditioning on independent draws p^*_m for each imputed dataset (the imputation model is incompatible and improper for the analysis model, but the imputing assumption is correct). This means the between-imputation variance is low, implying high precision, but coverage that is higher than stated. These are desirable properties, but they rely on the imputing assumption being correct. In Rubin's scenario with an 'imputer' and 'analyst', the imputer may have access to information that the analyst does not or cannot hold, so it is plausible that the imputer can make a correct imputing assumption that the analyst does not know. For the initial simulation on estimation of exponents the imputing assumption was correct, but outside of a simulation study p would be unknown. The imputation method based on the approximate Bayesian bootstrap was developed as a practical solution when p is unknown.

7.2 CAN WE DO BETTER THAN MULTIPLE IMPUTATION?

Some of the work presented in this project is encouraging, but it is important to note that multiple imputation is not the best solution in all scenarios and is not perfect for most scenarios. For some work complete cases analysis did not perform too badly, or all methods performed badly. It is important to acknowledge that multiple imputation is not always the best approach to handling missing data.

Assume an outcome y is fully observed and a covariate x is partially observed. Let R denote an indicator of response for x. The correct analysis model is a linear regression of y on x. If R_x depends on x but is conditionally independent of y given x, complete cases analysis will be unbiased, while multiple imputation of $x|y$ will lead to biased (though more efficient) inference[89]. However, if R_x is conditionally independent of x given y, this is MAR and the opposite will be true. If R_x depends on both x and y then neither method is appropriate. It is thus important that multiple imputation is used only after an assessment that it is likely to offer an advantage over, say, complete case analysis under the assumed missing data mechanism.

The likely utility of multiple imputation will also depend on the variable/s in which missing data occur. For example, if missing data are in the outcome alone, multiple imputation is not worthwhile unless the imputation model uses information external to that used by the analysis model. If missing data are in the covariate of interest, it is also unlikely that much information will be recovered (again assuming the imputation model does not include auxiliary variables). If a confounder is incomplete, or the variable of interest is only partly observed, as in the cases of ratios and interactions, multiple imputation can offer advantages over complete cases.

An alternative to multiple imputation is to model the missing data and outcome using a full probability model such as maximum likelihood or fully Bayesian methods. If the models used are correctly specified and will fit, these joint models should give equivalent inference to correctly specified multiple imputation with $M = \infty$. In chapter 5 it was demonstrated that fitting fully Bayesian joint models is not trivial. Although specifying full probability models is easier than correctly specifying the (corresponding) imputation model, fitting it is far more computationally complex, and will be inaccessible to many applied researchers.

While multiple imputation will not always be the best approach, with adequate consideration of the assumptions and careful choice of imputation model it will often be a better approach

than complete cases and more practical than full probability models.

7.3 Implications for the practical use of multiple imputation

PMM and LRD are less useful for dealing with nonlinear relationships than originally expected. Both are slightly inferior to standard posterior draws when the imputation model is correctly specified, however, when the imputation model is misspecified they can be superior. This implies it is necessary to think carefully about correct specification of the imputation model, but either method may be a helpful adjunct to this choice when there is uncertainty about its adequacy. In using either method, I *strongly* caution against the combination of type 0 or 2 matching with $k = 1$ (at the time of writing $k = 1$ with type 2 matching is the default option for Stata's `mi impute pmm`). For PMM, I advise type 1 matching with $k = 10$ and for LRD, type 2 matching with $k = 20$. For datasets with a particularly small number of available donors (for example $n_h = 30$), smaller values of k will be necessary; for particularly large datasets it may be preferable to use larger k. Note that the only software either of these methods is available in is `ice`[30]. In R, SAS and Stata's `mi` suite, PMM with type 2 matching is the only option.

For analysis with incomplete ratios, I offer a further strong caution: *avoid the use of passive imputation without prior transformation*. Bayesian full probability models do not appear to be viable for this problem, despite theoretical promise. Of the various imputation methods explored, passive imputation after log-transformation of the numerator and denominator was perfectly adequate. This was unexpected because the imputation model is incompatible with the analysis model, but the result held even for relatively strong associations.

In combining multivariable fractional polynomial models, the methods explored were, necessarily, less simple to implement than the sort of methods considered in previous chapters. First, special approaches to imputation were required to ensure the variable and function selection procedures were not biased by the method of imputation. Compared with more standard imputation, the methods I advocate are computationally intensive. Further, `smcfcs` can fail and `icet` lacks generality. In this setting it is more important than ever to inspect imputed values, comparing them graphically alongside observed values where possible. Where imputation is unsatisfactory, researchers should proceed by making careful restrictions or additions to the imputation model. Both methods described have been coded in Stata (`smcfcs` and `icet`).

For MFP model selection the ΔWald method appears to be slightly preferable to stacking for estimation of exponents, though there is little to choose between them for variable selection. Both have have been coded as Stata commands with similar generality to the commands for building MFP models in complete data, and so these are simple to apply once data have been imputed satisfactorily. I hope that these methods, particularly ΔWald, will begin to be used in practice now that their validity and value has been established.

7.4 Limitations and extensions

As with any research, these conclusions do not cover all settings. Some remarks on the limitations of my investigations are given below, and potential extensions.

7.4.1 *Diagnostics for PMM and LRD*

In Kazuo Ishiguro's novel *Never Let Me Go*[90], the protagonists are given life and grown to adults for the purpose of donating their vital organs to individuals from whom they were cloned. Each donates two or three times before dying. This bizarre idea suggests a diagnostic for PMM and LRD: a variable detailing, for each individual with observed data, the number of times they donate within each imputed dataset. Presumably a small number of individuals donating a lot of the time would be a cause for concern, although it is possible that it simply indicates a strong degree of MAR. An example of the use of such a diagnostic is in J-thwart; figure 3.8 shows clearly that one or two individuals have donated many times. While it is not this fact that leads to the problems with PMM and LRD, it does indicate the strong MAR mechanism that leads to them. If results of such a diagnostic were a cause of genuine concern, something like Moriarity & Scheuren's 'constrained' matching might be sensible, which makes repeated donation harder.

A second useful diagnostic could be envisaged for individuals with imputed data: the number of donors that have been used across the M imputations (for each imputed variable). If many individuals have a small number of donors providing data for most imputations this may be a cause for concern. Again, some sort of constraint could be invoked that makes it harder for a donee to receive repeatedly from any donor.

Both of these diagnostics may be useful additions to current implementations of PMM and LRD. However, there are two cautions to note before deciding on an implementation of either diagnostic.

1. Computationally, any implementation would increase the time taken to create imputations, perhaps considerably. Further, an implementation of constrained matching would add to this even further, since models would be required to record the number of donations per donor and augment δ_{hj} accordingly.
2. Further thought is required on the interpretation of these diagnostics. Although they may be interesting, it does not follow that the imputation model is wrong, or that a constraint on matching will improve inference. Consider the data shown in figure 3.8. If a constrained matching method were used for PMM, the problematic imputed values of x would lie on the vertical line at the right or to the left of it, which would increase the bias further. Even so, the diagnostic recording the number of times the observed value of x donates would be reduced.

These are strong reasons against the implementation of either of the above diagnostics

7.4.2 *Generalisability*

Some work has focused on concepts and how methods work, for example chapter 3 and sections 6.4 and 5.4.3; other work has focused more on practical application of methods, for example chapter 4 and sections 5.5 and 6.9. The work on concepts demonstrates strengths and weaknesses of different methods in different scenarios, but does not mean such scenarios will occur in practice. It was therefore important to motivate work with real data examples. However, this is in itself limited to the structure of the example datasets in use. For example, the outcome of using imputation model M5 is very different for the two datasets presented in section 5.5.

With just the first dataset we may have concluded that M5 was adequate. It is possible that with different datasets very different approaches would be necessary.

Continuous, normally distributed outcomes have primarily been used in simulation work. In imputing missing covariate values it is unlikely that using different outcomes would have substantially changed conclusions about the different methods, but it is possible that the contrast between methods may be different. It may be a useful extension to repeat some of the simulation studies to consider binary, ordinal, count and survival outcomes. (This does not apply to chapter 3; with a binary outcome there would only be two values of $\hat{\alpha}\, y_h$.)

7.4.3 Missing data and missing at random

The introduction stated that this thesis would deal with data missing at random and missing completely at random. Multiple imputation is usually assumed to provide valid inference under these assumptions. It does not mean that the assumptions are plausible, and in any given analysis the likely process by which data go missing must be considered. If this process seems to be other than missing at random then imputation models will need augmenting (this remark sounds innocuous but may be extremely difficult to do in practice and imputation under missing not at random is still a subject of active research).

Even where missing at random seems to be plausible it is critically important to assess the sensitivity of conclusions to departures from the assumption. One approach that may be fruitful is that of Carpenter, Kenward and White[91]. Briefly, imputations are drawn under MAR, but when Rubin's rules are applied, imputed datasets are weighted by the magnitude of a parameter estimate of interest. The approach was developed assuming a single parameter of interest, but research is ongoing for multiple parameters.

There are two cautions from the authors when using this approach[91]:

1. The true MNAR inference must lie within the range of the various MAR estimates.

2. Although an attraction is that it is not necessary to produce imputations under MNAR, a drawback is that estimates are often completely dominated by one imputed dataset, even for a reasonable number of imputations. Hundreds or even thousands of imputations may thus be required to tackle this problem. A high computational burden is then added.

7.4.4 Multi-level data

Many datasets in medical research involve some form of clustering, where an observed outcome is 'clustered' with others within some larger unit[92]. A 'cluster' is some common measurement shared across multiple units of observation. Common examples are studies in a meta-analysis, multiple individuals in a longitudinal study, and clusters in a cluster randomised trial.

In certain settings it is important to allow for the clustering in the analysis, while in other settings it will be desirable as a means of increasing precision. Indeed, individuals are clustered within recruiting centres in both the trauma and Aurum datasets.

When the analysis model allows for clustering, for example using random treatment-by-study interactions in a random-effects meta-analysis, it is important that the imputation allows for this. This is an area of active research. Two-level imputation models have been implemented

as `2L.norm` in R's `mice` package. WinBUGS and Realcom can also impute from general multilevel models.

In a recent paper, Goldstein, Carpenter and Browne have worked on developing methods for fitting multilevel models which may include interactions and nonlinear terms for covariates, although their approach is completely different to ours and it is unclear whether allowance for nonlinear terms would extend to fractional polynomial models[93].

7.4.5 Closeness of the selected MFP model to the true model

In chapter 6 a simulation study considered bias and variance in the estimation of exponents for fractional polynomial models according to different methods. Following this, the type I error and power of model selection procedures were considered. The power work is a step beyond much previous work on multivariable fractional polynomials. However, controlling the error rates is not the only care of frequentist statistics.

It would be of interest to have some measure of closeness to the true model. It is not sensible to summarise $\hat{\boldsymbol{\beta}}$ over repeated simulations because $\hat{\boldsymbol{\beta}}$ are only interpretable conditional on $\hat{\boldsymbol{p}}$, making 'closeness' to the correct model difficult to quantify. Royston and Sauerbrei advocate summarising the results of MFP models graphically, or numerically at pertinent covariate levels[13], asa in figure 6.15. The latter approach in simulations might be useful future work to facilitate an exploration of closeness to the true model. However, there is no reason to believe the conclusions would change.

7.4.6 Remark on simulation

This thesis has relied heavily on simulation studies, some relatively complex, to evaluate the performance of methods.

If a researcher were given any one of the simulated datasets, they may decide on a different approach to the analysis on inspection of the data. For example, with the simulation study performed in chapter 3, if the association between observed values and outcome appears to be approximately linear and $x|y$ appears to be approximately normal, posterior draws may be used in preference to PMM. The procedures blindly followed by simulation programs do not involve any representation of the decisions that are used in analysing data.

This is a strength in that simulation considers the statistical properties of a procedure regardless of the data, but a weakness in that it does not mimic the operating characteristics of the researcher's approach. This is true of the majority of Monte Carlo simulation studies but is worth noting. In principle simulation studies could be designed to follow certain decision-making procedures. Such a simulation may be interesting but focuses less on the method and more on the researcher's procedures.

Bibliography

[1] Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons: New York, 1987.

[2] Rubin DB. Inference and missing data. *Biometrika* 1976; **63**:581–592.

[3] Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society, series C* 1994; **43**:49–93.

[4] Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman and Hall: London, 1997.

[5] Kenward MG, Carpenter JR. Multiple imputation: current perspectives. *Statistical Methods in Medical Research* 2007; **16**(3):199–218. URL `http://smm.sagepub.com/cgi/content/abstract/16/3/199`.

[6] White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* 2011; **30**(4):377–399, doi:10.1002/sim.4067. URL `http://dx.doi.org/10.1002/sim.4067`.

[7] Robins JM, Wang N. Inference for imputation estimators. *Biometrika* 2000; **87**(1):113–124, doi:10.1093/biomet/87.1.113. URL `http://biomet.oxfordjournals.org/cgi/content/abstract/87/1/113`.

[8] van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 1999; **18**:681–694.

[9] Nevalainen J, Kenward MG, Virtanen SM. Missing values in longitudinal dietary data: A multiple imputation approach based on a fully conditional specification. *Statistics in Medicine* Dec 2009; **28**(29):3657–3669, doi:10.1002/sim.3731. URL `http://dx.doi.org/10.1002/sim.3731`.

[10] He Y, Raghunathan TE. On the performance of sequential regression multiple imputation methods with non normal error distributions. *Communications in Statistics - Simulation and Computation* 2009; **38**(4):856–883, doi:10.1080/03610910802677191. URL `http://dx.doi.org/10.1080/03610910802677191`.

[11] Royston P. Multiple imputation of missing values. *The Stata Journal* 2004; **4**:227–241.

[12] van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* 2006; **76**:1049–1064.

[13] Royston P, Sauerbrei W. *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley: Chichester, 2008. URL `http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470028424.html`.

[14] Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? *Statistics in Medicine* 2008; **27**(17):3227–3246, doi:10.1002/sim.3177. URL `http://dx.doi.org/10.1002/sim.3177`.

[15] Stanworth S, Morris T, Gaarder C, Goslings JC, Maegele M, Cohen M, Konig T, Davenport R, Francois Pittet JF, Johansson P, *et al.*. Reappraising the concept of massive transfusion in trauma. *Critical Care* 2010; **14**(6):R239+, doi:10.1186/cc9394. URL `http://dx.doi.org/10.1186/cc9394`.

[16] Morris TP, White IR, Royston P, Seaman SR, Wood AM. Multiple imputation for an incomplete covariate that is a ratio. *Statistics in Medicine* Jan 2014; **33**(1):88–104, doi:10.1002/sim.5935. URL `http://dx.doi.org/10.1002/sim.5935`.

[17] Russell E, Charalambous S, Pemba L, Churchyard G, Grant A, Fielding K. Low haemoglobin predicts early mortality among adults starting antiretroviral therapy in an HIV care programme in south africa: a cohort study. *BMC Public Health* 2010; **10**(1):433+, doi:10.1186/1471-2458-10-433. URL `http://dx.doi.org/10.1186/1471-2458-10-433`.

[18] Day N, Oakes S, Luben R, Khaw KT, Bingham S, Welch A, Wareham N. EPIC-norfolk: study design and characteristics of the cohort. european prospective investigation of cancer. *British journal of cancer* Jul 1999; **80 Suppl 1**:95–103. URL http://view.ncbi.nlm.nih.gov/pubmed/10466767.

[19] Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the united kingdom: prospective open cohort study. *BMJ* 2007; **335**(7611):136+, doi:10.1136/bmj.39261.471806.55. URL http://dx.doi.org/10.1136/bmj.39261.471806.55.

[20] Arsenault BJ, Rana JS, Stroes ESG, Despres JP, Shah PK, Kastelein JJP, Wareham NJ, Boekholdt SM, Khaw KT. Beyond low-density lipoprotein cholesterol: respective contributions of non-high-density lipoprotein cholesterol levels, triglycerides, and the total cholesterol/high-density lipoprotein cholesterol ratio to coronary heart disease risk in apparently healthy men and women. *Journal of the American College of Cardiology* 2010; **55**(1):35–41, doi:10.1016/j.jacc.2009.07.057. URL http://dx.doi.org/10.1016/j.jacc.2009.07.057.

[21] Eddings W, Marchenko Y. Diagnostics for multiple imputation in Stata. *The Stata Journal* 2012; **12**(3):353–367. URL http://www.stata-journal.com/article.html?article=st0263.

[22] Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 1994; **9**:538–558.

[23] Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91**:473–489.

[24] Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods* 2001; **6**(4):330–351, doi:10.1037//1082-989X.6.4.330-351. URL http://dx.doi.org/10.1037//1082-989X.6.4.330-351.

[25] Liu J, Gelman A, Hill J, Su YS. On the stationary distribution of iterative imputations 2012. URL http://arxiv.org/abs/1012.2902.

[26] Bartlett JW, Seaman SR, White IR, Carpenter JR, for the Alzheimer's Disease Neuroimaging Initiative*. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research* Feb 2014; :0962280214521348+doi:10.1177/0962280214521348. URL http://dx.doi.org/10.1177/0962280214521348.

[27] Dembe A, Partridge J, Geist L. Statistical software applications used in health services research: analysis of published studies in the U.S. *BMC Health Services Research* Oct 2011; **11**(1):252+, doi:10.1186/1472-6963-11-252. URL http://dx.doi.org/10.1186/1472-6963-11-252.

[28] Schenker N, Taylor JMG. Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis* 1996; **22**(4):425–446, doi:10.1016/0167-9473(95)00057-7. URL http://dx.doi.org/10.1016/0167-9473(95)00057-7.

[29] Reilly M. Data analysis using hot deck multiple imputation. *The Statistician* 1993; **42**:307–313. URL http://www.jstor.org/stable/2348810.

[30] Royston P. Multiple imputation of missing values: update. *The Stata Journal* 2005; **5**:527–536.

[31] van Buuren S, Oudshoorn CGM. *Multivariate Imputation by Chained Equations: MICE V1.0 User's manual*. TNO Report PG/VGZ/00.038. Leiden: TNO Preventie en Gezondheid 2000.

[32] Siddique J, Belin TR. Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine* 2008; **27**(1):83–102, doi:10.1002/sim.3001. URL http://dx.doi.org/10.1002/sim.3001.

[33] Little RJA. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics* 1988; **6**:287–296.

[34] Heitjan DF, Little RJA. Multiple imputation for the fatal accident reporting system. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 1991; **40**(1):13–29, doi:10.2307/2347902. URL http://dx.doi.org/10.2307/2347902.

[35] Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics* 1986; **4**(1), doi:10.2307/1391390. URL http://dx.doi.org/10.2307/1391390.

[36] David M, Little RJA, Samuhel ME, Triest RK. Alternative methods for CPS income imputation. *Journal of the American Statistical Association* 1986; **81**(393), doi:10.2307/2287965. URL http://dx.doi.org/10.2307/2287965.

[37] Koller-Meinfelder F. Analysis of incomplete survey data – multiple imputation via bayesian bootstrap predictive mean matching. PhD Thesis, Otto-Friedrich-Universität Bamberg, Bamberg Aug 2009.

[38] Barnes SA, Lindborg SR, Seaman JW. Multiple imputation techniques in small sample clinical trials. *Statistics in Medicine* 2006; **25**(2):233–245, doi:10.1002/sim.2231. URL http://dx.doi.org/10.1002/sim.2231.

[39] Heitjan DF, Landis RJ. Assessing secular trends in blood pressure: A multiple-imputation approach. *Journal of the American Statistical Association* 1994; **89**(427), doi:10.2307/2290900. URL http://dx.doi.org/10.2307/2290900.

[40] Landerman LR, Land KC, Pieper CF. An empirical evaluation of the predictive mean matching method for imputing missing values. *Sociological Methods & Research* Aug 1997; **26**(1):3–33, doi:10.1177/0049124197026001001. URL http://dx.doi.org/10.1177/0049124197026001001.

[41] Heitjan DF. Annotation: What can be done about missing data? approaches to imputation. *American Journal of Public Health* Apr 1997; **87**(4):548–550.

[42] Schulte Nordholt E. Imputation: methods, simulation experiments and practical examples. *International Statistical Review* 1998; **66**(2):157–180, doi:10.1111/j.1751-5823.1998.tb00412.x. URL http://dx.doi.org/10.1111/j.1751-5823.1998.tb00412.x.

[43] Zhou XH, Eckert GJ, Tierney WM. Multiple imputation in public health research. *Statistics in Medicine* 2001; **20**(9-10):1541–1549, doi:10.1002/sim.689. URL http://dx.doi.org/10.1002/sim.689.

[44] Horton NJ, Lipsitz SR. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician* 2001; **55**:244–254.

[45] Moriarity C, Scheuren F. A note on rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics* Jan 2003; **21**(1):65–73, doi:10.1198/073500102288618766. URL http://dx.doi.org/10.1198/073500102288618766.

[46] Tang L, Song J, Belin TR, Unützer J. A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine* 2005; **24**(14):2111–2128, doi:10.1002/sim.2099. URL http://dx.doi.org/10.1002/sim.2099.

[47] Durrant GB. Imputation methods for handling item nonresponse in the social sciences: A methodological review. *Technical Report*, University of Southampton, ESRC National Centre for Rsearch Methods and Southampton Statistical Sciences Research Institute (S3RI) 2005. URL http://www.ncrm.ac.uk/research/outputs/publications/methodsreview/MethodsReviewPaperNCRM-002.pdf.

[48] Durrant GB. A semi-parametric multiple imputation data augmentation procedure. *Proceedings of the Section on Survey Research Methods, American Statistical Association* 2005; :1943–1950.

[49] Durrant GB, Skinner C. Using missing data methods to correct for measurement error in a distribution function. *Survey methodology* Jun 2006; **32**(1):25–36. URL http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-001-X20060019260&lang=eng.

[50] Hsu CH, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via non-parametric multiple imputation. *Statistics in Medicine* 2006; **25**(20):3503–3517, doi:10.1002/sim.2452. URL http://dx.doi.org/10.1002/sim.2452.

[51] Yu LM, Burton A, Rivero-Arias O. Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research* 2007; **16**(3):243–258. URL http://www.swetswise.com.libproxy.ucl.ac.uk/eAccess/titleDetail.do?titleID=192257.

[52] Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician* 2007; **61**:79–90.

[53] Siddique J, Harel O. MIDAS: A SAS macro for multiple imputation using distance-aided selection of donors. *Journal of Statistical Software* Feb 2009; **29**(9):1–18. URL http://www.jstatsoft.org/v29/i09.

[54] Di Zio M, Guarnera U. Semiparametric predictive mean matching. *AStA Advances in Statistical Analysis* Jun 2009; **93**(2):175–186, doi:10.1007/s10182-008-0081-2. URL http://dx.doi.org/10.1007/s10182-008-0081-2.

[55] Ayuyev V, Jupin J, Harris P, Obradovic Z. Dynamic clustering-based estimation of missing values in mixed type data. *Data Warehousing and Knowledge Discovery*, *Lecture Notes in Computer Science*, vol. 5691, Pedersen T, Mohania M, Tjoa A (eds.). chap. 29, Springer Berlin / Heidelberg: Berlin, Heidelberg, 2009; 366–377, doi: 10.1007/978-3-642-03730-6_29. URL http://dx.doi.org/10.1007/978-3-642-03730-6_29.

[56] Marshall A, Altman D, Royston P, Holder R. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC medical research methodology* Jan 2010; **10**(1):7+, doi:10.1186/1471-2288-10-7. URL http://dx.doi.org/10.1186/1471-2288-10-7.

[57] Qi L, Wang YFF, He Y. A comparison of multiple imputation and fully augmented weighted estimators for cox regression with missing covariates. *Statistics in medicine* Nov 2010; **29**(25):2592–2604, doi:10.1002/sim.4016. URL http://dx.doi.org/10.1002/sim.4016.

[58] Long Q, Zhang X, Hsu CH. Nonparametric multiple imputation for receiver operating characteristics analysis when some biomarker values are missing at random. *Statistics in Medicine* Nov 2011; **30**(26):3149–3161, doi: 10.1002/sim.4338. URL http://dx.doi.org/10.1002/sim.4338.

[59] Royston P. Multiple imputation of missing values: update. *The Stata Journal* 2005; **5**:188–201.

[60] Marshall A, Altman D, Holder R. Comparison of imputation methods for handling missing covariate data when fitting a cox proportional hazards model: a resampling study. *BMC medical research methodology* Dec 2010; **10**(1):112+, doi:10.1186/1471-2288-10-112. URL http://dx.doi.org/10.1186/1471-2288-10-112.

[61] Andridge RR, Little RJA. A review of hot deck imputation for survey non-response. *International Statistical Review* Apr 2010; **78**(1):40–64, doi:10.1111/j.1751-5823.2010.00103.x. URL http://dx.doi.org/10.1111/j.1751-5823.2010.00103.x.

[62] Gomel MK, Oldenburg B, Simpson JM, Chilvers M, Owen N. Composite cardiovascular risk outcomes of a work-site intervention trial. *American Journal of Public Health* Apr 1997; **87**(4):673–676, doi:10.2105/ajph.87.4.673. URL http://dx.doi.org/10.2105/ajph.87.4.673.

[63] White IR. simsum: Analyses of simulation studies including monte carlo error. *Stata Journal* 2010; **10**(3):369–385.

[64] Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC medical research methodology* 2012; **12**(1):46+, doi: 10.1186/1471-2288-12-46. URL http://dx.doi.org/10.1186/1471-2288-12-46.

[65] Von Hippel PT. How to impute squares, interactions, and other transformed variables. *Sociological Methodology* 2009; **39**:265–291, doi:10.1111/j.1467-9531.2009.01215.x. URL http://dx.doi.org/10.1111/j.1467-9531.2009.01215.x.

[66] Moons K, Donders R, Stijnen T, Harrel F. Using the outcome for imputation of missing predictor values was preferred. *Journal of clinical epidemiology* 2006; **59**(10):1092–1101, doi:10.1016/j.jclinepi.2006.01.009. URL http://dx.doi.org/10.1016/j.jclinepi.2006.01.009.

[67] Robins JM, Gill RD. Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine* 1997; **16**:39–56.

[68] Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical Methods in Medical Research* 2007; **16**(3):277–298. URL http://smm.sagepub.com/cgi/content/abstract/16/3/277.

[69] Rubin DB. Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section*, 1978; 20–28.

[70] Yusuf S, Hawken S, Ôunpuu S, Bautista L, Franzosi MG, Commerford P, Lang CC, Rumboldt Z, Onen CL, Lisheng L. Obesity and the risk of myocardial infarction in 27,000 participants from 52 countries: a case-control study. *The Lancet* 2005; **366**(9497):1640–1649, doi:10.1016/s0140-6736(05)67663-5. URL http://dx.doi.org/10.1016/s0140-6736(05)67663-5.

[71] Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, Jacques PF, Rifai N, Selhub J, Robins SJ, *et al.*. Multiple biomarkers for the prediction of first major cardiovascular events and death. *New England Journal of Medicine* 2006; **355**(25):2631–2639, doi:10.1056/nejmoa055373. URL http://dx.doi.org/10.1056/nejmoa055373.

[72] Allison DB, Paultre F, Goran MI, Poehlman ET, Heymsfield SB. Statistical considerations regarding the use of ratios to adjust data. *International journal of obesity and related metabolic disorders: Journal of the International Association for the Study of Obesity* 1995; **19**(9):644–652. URL http://view.ncbi.nlm.nih.gov/pubmed/8574275.

[73] Senn S, Julious S. Measurement in clinical trials: a neglected issue for statisticians? *Statistics in Medicine* 2009; **28**(26):3189–3209, doi:10.1002/sim.3603. URL http://dx.doi.org/10.1002/sim.3603.

[74] White IR, Royston P. Imputing missing covariate values for the cox model. *Statistics in Medicine* 2009; **28**(15):1982–1998, doi:10.1002/sim.3618. URL http://dx.doi.org/10.1002/sim.3618.

[75] Schafer JL. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica* 2003; **57**(1):19–35, doi:10.1111/1467-9574.00218. URL http://dx.doi.org/10.1111/1467-9574.00218.

[76] Schafer JL, Graham JW. Missing data: Our view of the state of the art. *Psychological Methods* 2002; **7**(2):147–177, doi:10.1037//1082-989x.7.2.147. URL http://dx.doi.org/10.1037//1082-989x.7.2.147.

[77] StataCorp. *Stata Statistical Software: Release 12*. Stata Press: College Station, TX, 2011.

[78] StataCorp LP, 4905 Lakeway Drive, College Station, Texas 77845. *Stata multiple-imputation reference manual release 12* 2011.

[79] Schafer JL. Multiple imputation: a primer. *Statistical Methods in Medical Research* Feb 1999; **8**(1):3–15, doi:10.1177/096228029900800102. URL http://dx.doi.org/10.1177/096228029900800102.

[80] Copas JB. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)* 1983; **45**(3):311–354, doi:10.2307/2345402. URL http://dx.doi.org/10.2307/2345402.

[81] Box GEP, Tidwell PW. Transformation of the independent variables. *Technometrics* 1962; **4**(4):531–550, doi:10.1080/00401706.1962.10490038. URL http://dx.doi.org/10.1080/00401706.1962.10490038.

[82] Ambler G, Royston P. Fractional polynomial model selection procedures: investigation of type i error rate. *Journal of Statistical Computation and Simulation* 2001; **69**(1):89–108, doi:10.1080/00949650108812083. URL http://dx.doi.org/10.1080/00949650108812083.

[83] Marcus R, Eric P, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**(3):655–660, doi:10.1093/biomet/63.3.655. URL http://dx.doi.org/10.1093/biomet/63.3.655.

[84] Rubin DB. The Bayesian bootstrap. *The Annals of Statistics* 1981; **9**:130–134.

[85] Meng XL, Rubin DB. Performing likelihood ratio tests with Multiply-Imputed data sets. *Biometrika* 1992; **79**(1):103+, doi:10.2307/2337151. URL http://dx.doi.org/10.2307/2337151.

[86] Ambler G, Royston P. Fractional polynomial model selection procedures: Investigation of type I error rate. *Journal of Statistical Computation and Simulation* 2001; **69**.

[87] White IR, Horton NJ, Carpenter J, Pocock SJ. Strategy for intention to treat analysis in randomised trials with missing outcome data. *BMJ* Jan 2011; **342**, doi:10.1136/bmj.d40. URL http://dx.doi.org/10.1136/bmj.d40.

[88] Kronmal RA. Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 1993; **156**(3):379–392, doi:10.2307/2983064. URL http://dx.doi.org/10.2307/2983064.

[89] White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine* Dec 2010; **29**(28):2920–2931, doi:10.1002/sim.3944. URL http://dx.doi.org/10.1002/sim.3944.

[90] Ishiguro K. *Never Let Me Go*. Vintage, 2006.

[91] Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research* 2007; **16**(3):259–275. URL http://smm.sagepub.com/cgi/content/abstract/16/3/259.

[92] Kahan BC, Morris TP. Assessing potential sources of clustering in individually randomised trials. *BMC Medical Research Methodology* Apr 2013; **13**(1):58+, doi:10.1186/1471-2288-13-58. URL http://dx.doi.org/10.1186/1471-2288-13-58.

[93] Goldstein H, Carpenter JR, Browne WJ. Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *J. R. Stat. Soc. A* Aug 2013; :n/adoi:10.1111/rssa.12022. URL http://dx.doi.org/10.1111/rssa.12022.

[94] Arnold BC, Castillo E, Sarabia JM. Conditionally specified distributions: An introduction. *Statistical Science* 2001; **16**(3):249–265. URL http://www.jstor.org/stable/2676688.

[95] Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - a bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**(4):325–337, doi:10.1023/a:1008929526011. URL http://dx.doi.org/10.1023/a:1008929526011.

[96] Clayton D. Bayesian analysis of frailty models. *Technical Report*, MRC Biostatistics Unit, Cambridge 1994.

A INITIALISMS

The below section is a reference for the initialisms used in this thesis. These are listed alphabetically.

ABB	Approximate Bayesian Bootstrap
BVN	BiVariate Normal
CD	Complete Data
CC	Complete Cases
CV	Coefficient of Variation; the standard deviation divided by the mean
df	Degrees of Freedom
Epic	European Prospective Investigation of Cancer*
FMI	Fraction of Missing Information
FP	Fractional polynomial/s
FCS	Fully Conditional Specification (synonym for mice)
LRD	Local Residual Draw
MAR	Missing At Random
MCAR	Missing Completely At Random
MFP	Multivariable Fractional Polynomial/s
MI	Multiple Imputation
mice	Multiple Imputation by Chained Equations*
MNAR	Missing Not At Random
MVN	MultiVariate Normal
PMM	Predictive Mean Matching
ROC	Receiver Operating Characteristic
SD	Standard deviation
SE	Standard error
SMC	Substantive-Model-Compatible

*Initialisms that are pronounced as a word rather than initials are set in lowercase.

B DRAWSUIT RESULTS: $n = 100$, 25% MISSING x

It is hypothesised that the performance of PMM and LRD will degrade. In particular, increasing k may provide worse results since finding a good match is harder and k represents a larger proportion of the observed data. Therefore, the simulations in 3.1.3 are repeated with $n = 100$.

Biases are very similar to $n = 500$, but are generally around double the magnitude of the corresponding $n = 500$ cases. In relative terms, this is still very small and unlikely to matter in practice. Again, PMM is increasingly biased with larger k, while LRD is unaffected. As with $n = 500$, this bias is slightly stronger for type 1 matching than type 2. All biases present increase with the strength of MAR.

Standard errors from PMM and LRD tend to be slightly better relative to posterior draws than in the $n = 500$ case. Again larger k gives lower standard errors. Interestingly for PMM with type 1 matching, $R^2 = 0$ is more precise than posterior draws when $k > 3$. Again, the relative precision of both PMM and LRD is adversely affected by stronger missingness mechanisms. Greater precision would be expected for methods which are biased towards the null, but interestingly here the method least biased towards the null also has the lowest standard error.

Coverage showed similar patterns to $n = 500$, again slightly accentuated. As previously, type 1 matching gives better coverage than type 2 and LRD tends to give slightly low coverage, with PMM giving superior results under like-for-like matching and k. The largest values of k reduced problems for type 2 matching to a negligible size. In these cases there were no coverage issues for $R^2 = 0.5$; for $R^2 = 0$ and $R^2 = 0.1$ issues were small.

Figure B.1: DrawSuit: Bias in point estimates, $n = 100$ (Error bars are ±2× Monte Carlo standard errors)

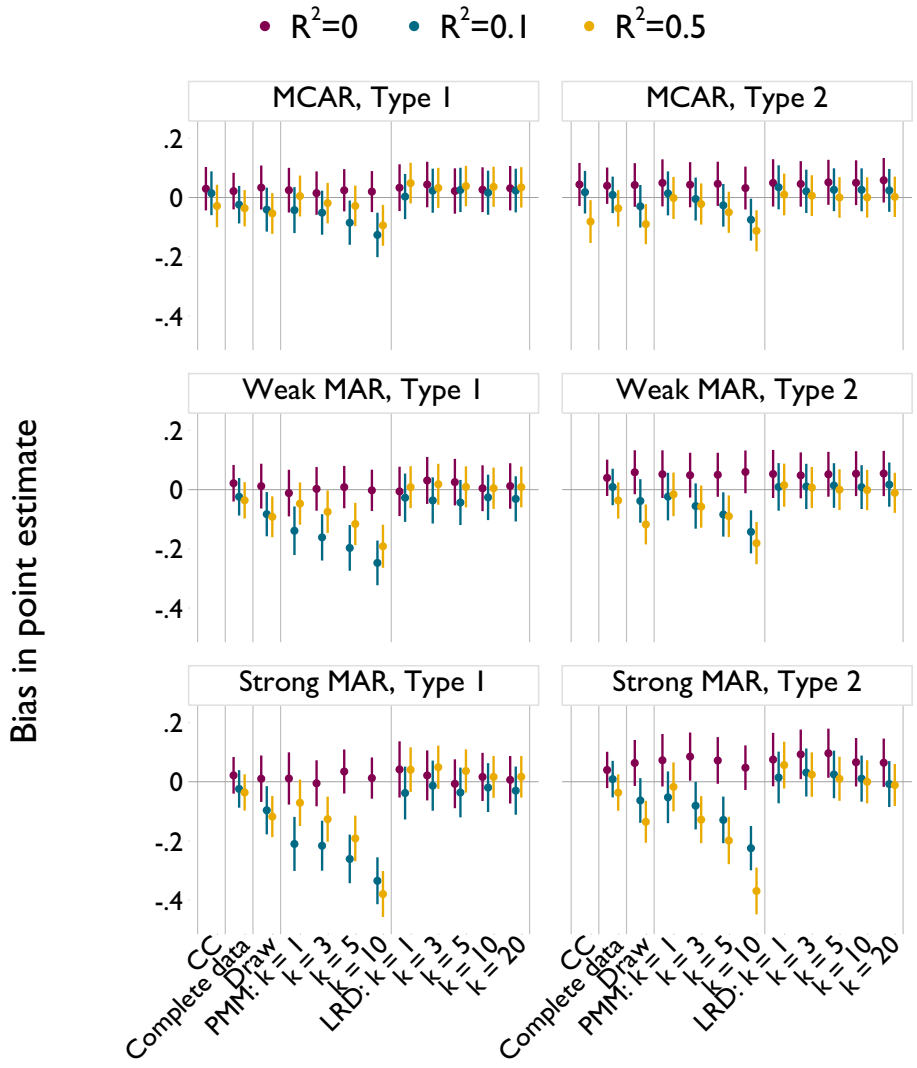


Figure B.2: DrawSuit: Empirical standard errors, $n = 100$ (Error bars are ±2× Monte Carlo standard errors)

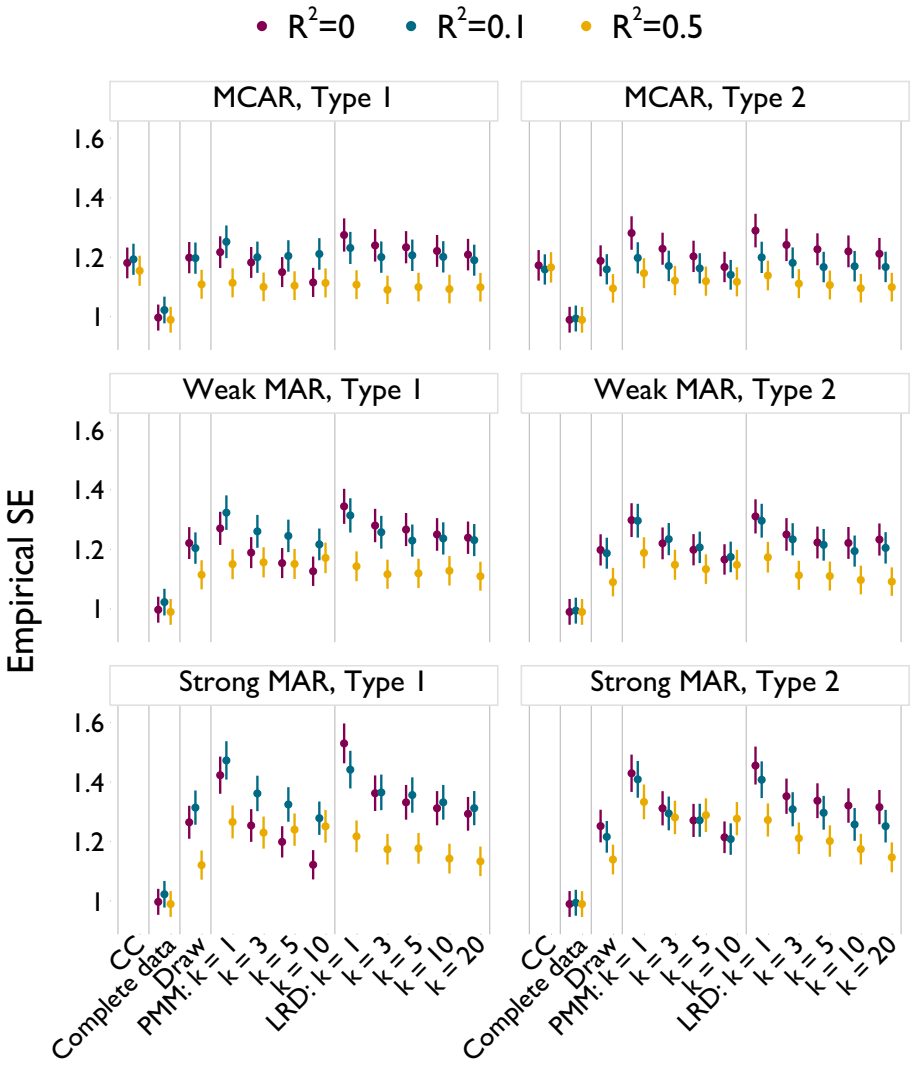
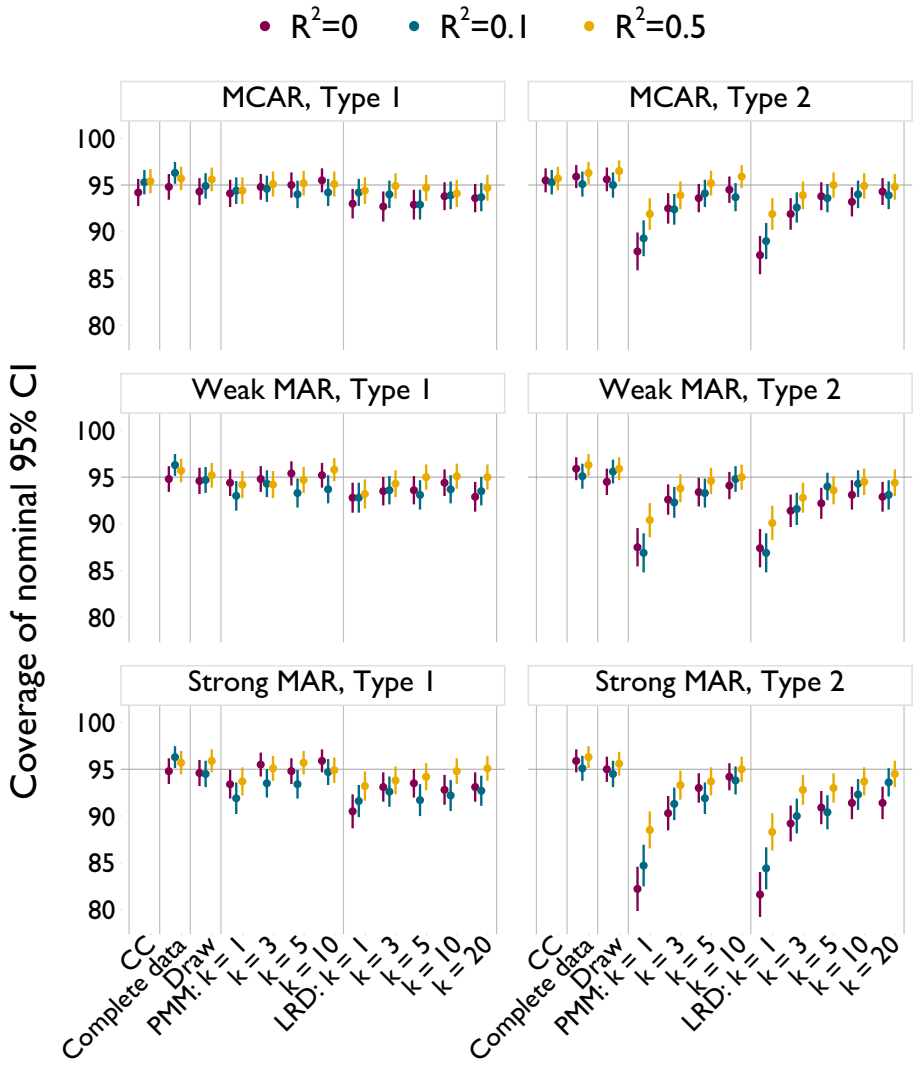


Figure B.3: DrawSuit: Coverage of nominal 95% CI, $n = 100$ (Error bars are ±2× Monte Carlo standard errors)



C APPENDICES RELATING TO MI FOR RATIOS (CHAPTER 5)

C.1 COMPATIBILITY FOR A RATIO COVARIATE

[Note: as in chapter 5, the appendix below matches the appendix of the published article[16].]

Section 5.4.3 says that models are compatible if a joint model exists that implies both as conditionals. How can we tell whether there is a joint model underpinning both the imputation model and analysis model? Arnold, Castillo and Sarabia give a theorem that is restated here for clarity[94].

Theorem 1 *Given two conditional densities $f(x|y)$ and $g(y|x)$, a joint density exists if and only if $\{(x, y) : f(x|y) > 0\} = \{(x, y) : g(y|x) > 0\}$ and there exist functions $u(x)$ and $v(y)$ such that, first,*

$$\frac{f(x|y)}{g(y|x)} = u(x)v(y), \tag{C.1}$$

and second, $u(x)$ is integrable.

Here $u(x)$ is a marginal density for x and $v(y)$ is a marginal density for y. Below, we posit an analysis model and check compatibility against two different imputation models using (C.1).

We distinguish between two kinds of non-compatibility:

SEMI-COMPATIBILITY *There is a special case of the imputation model that is compatible with the analysis model.*

INCOMPATIBILITY *There is no case of the imputation model that is compatible with the analysis model.*

That is, if setting certain parameters of the imputation model to 0 yields a compatible model, the imputation model is drawing on more information than the analysis model, and so is richer. If parameters of the imputation model cannot be set to 0 to identify a compatible model, the imputation model is using different information to, or less information than, the analysis model. Previous work has shown that incompatibility can be harmless or beneficial[22, 23, 24]. When the analysis model is correctly specified, these are examples of using semi-compatible imputation models, while incompatible imputation models can be harmful.

Appendices C.1.1 and C.1.2 work through two simple examples. For both, the analysis model involves only the ratio as a covariate. C.1.1 uses model M5 and is shown to be incompatible; C.1.2 uses model M1 and is shown to be compatible.

Instead of dividing the densities we subtract the log-densities. For clarity we omit the intercept terms α_0 and β_0 from the imputation model and the analysis model respectively, assuming both equal zero. Note that since neither parameter involves a_1, a_2 or y this does not impact on compatibility.

C.1.1 *Imputation model incompatible with the analysis model*

Suppose the proposed analysis model is a linear regression of y on the ratio a_1/a_2. The log-density for this is

$$-\ln(\sigma_y\sqrt{2\pi}) - \frac{\left(y - \beta\frac{a_1}{a_2}\right)^2}{2\sigma_y^2}. \tag{C.2}$$

The proposed imputation model is a bivariate normal model for a_1, a_2 given y:

$$(a_1, a_2 \mid y) \sim \text{BVN}\left(\begin{bmatrix} \alpha_1 y \\ \alpha_2 y \end{bmatrix}, \begin{bmatrix} \sigma_1 & \rho \\ \rho & \sigma_2 \end{bmatrix}\right),$$

which has log-density

$$-\ln(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}) - \frac{1}{2(1-\rho^2)} \tag{C.3}$$

$$\left[\left(\frac{a_1 - \alpha_1 y}{\sigma_1}\right)^2 + \left(\frac{a_2 - \alpha_2 y}{\sigma_2}\right)^2 - 2\rho\left(\frac{a_1 - \alpha_1 y}{\sigma_1}\right)\left(\frac{a_2 - \alpha_2 y}{\sigma_2}\right)\right] \tag{C.4}$$

The imputation model (C.4) is of the form $b(a_1, a_2) + c(y) + d_1(a_1 y) + d_2(a_2 y)$ and the analysis model (C.2) is of the form $b'(a_1, a_2) + c'(y) + d_3\left(\frac{a_1}{a_2} y\right)$. Subtracting one from the other, we cannot express the result as $u(a_1, a_2) - v(y)$, indicating that they are incompatible.

C.1.2 *Imputation model compatible with the analysis model*

The proposed analysis model is as in (C.1.1), and so the log-density is given by (C.2). However, the imputation model involves a linear regression of $\frac{a_1}{a_2}$ on y. The log-density is:

$$\ln f\left(\frac{a_1}{a_2}\right) = -\ln(\sigma_a\sqrt{2\pi}) - \frac{\left(\frac{a_1}{a_2} - \alpha y\right)^2}{2\sigma_a^2}. \tag{C.5}$$

Subtracting (C.5) from (C.2), we get

$$\ln(\sigma_a\sqrt{2\pi}) + \frac{\left(\frac{a_1}{a_2}\right)^2}{2\sigma_a^2} + \frac{\alpha^2 y^2}{2\sigma_a^2} - \frac{2\alpha\frac{a_1}{a_2} y}{2\sigma_a^2} - \ln(\sigma_y\sqrt{2\pi}) - \frac{y^2}{2\sigma_y^2} - \frac{\beta^2\left(\frac{a_1}{a_2}\right)^2}{2\sigma_y^2} + \frac{2\beta\frac{a_1}{a_2} y}{2\sigma_y^2}. \tag{C.6}$$

By setting $\alpha/\sigma_a^2 = \beta/\sigma_y^2$ we can express (C.6) without any terms involving both (a_1, a_2) and y, indicating that for any choice of α, σ_a^2 there is a value of β, σ_y^2 for which the proposed imputation model is compatible with the analysis model.

C.2 BAYESIAN MODELS FOR AN INCOMPLETE RATIO

It is conceptually natural to model missing covariates using Bayesian methods. The problem discussed in section 5.4.3, that the imputation model and the analysis model may not correspond to any joint model, does not exist for Bayesian models, where the model for missing data and the analysis model are joint. The compatibility between the missing data model and the analysis model is thus assured.

Table C.1: Candidate fully Bayesian models for $\boldsymbol{x}_i$

Model for covariates	Label
$(\boldsymbol{z}_i, x_{pi} \mid \boldsymbol{w}_i) \sim$ MVN	B1
$(\boldsymbol{z}_i, x_{pi}, a_{1i} \mid \boldsymbol{w}_i) \sim$ MVN	B2
$(\boldsymbol{z}_i, x_{pi}, a_{2i} \mid \boldsymbol{w}_i) \sim$ MVN	B3
$(\boldsymbol{z}_i, x_{pi}, a_{1i}, a_{2i} \mid \boldsymbol{w}_i) \sim$ MVN	B4
$(\boldsymbol{z}_i, a_{1i}, a_{2i} \mid \boldsymbol{w}_i) \sim$ MVN	B5
$(\boldsymbol{z}_i, \ln(a_{1i}), \ln(a_{2i}) \mid \boldsymbol{w}_i) \sim$ MVN	B6

The practical disadvantage of fully Bayesian models for an incomplete ratio and/or its components is computation. Bayesian models are also in general more computationally demanding than MI. Further, the imputation models described above could be implemented fairly automatically using a choice of software, while the Bayesian models requires knowledge of WinBUGS and/or the ability to code the models manually in another package.

Here, we explore whether Bayesian models, by working with the full joint likelihood, will provide more coherent results than MI. In our example datasets we aim to obtain posterior means and credible intervals under various models.

C.2.1 *Models, software and priors*

A Bayesian model combines model (5.1) with a model for the incomplete covariates given the complete covariates. Candidate Bayesian models for the covariates are listed in table C.1 (again, note the *Label* column, where the number corresponds to the imputation model with equivalent motivation). Details of how the Cox model is fit are given in section 5.4.5 and appendix C.2.2. In contrast to MI, no explicit conditioning on the outcome is required for Bayesian models.

Note that, except for the lack of issues around compatibility, the critique of the imputation models given in 5.4.4 with equivalent labels applies equally to the Bayesian models given in table C.1. That is, models B1–B3 may ignore some of the observed data, while B2–B4 are likely to be misspecified to some degree.

To fit Bayesian joint models in our case studies, we used WinBUGS 1.4.3[95]. Because we are dealing with the Cox model, we used the method outlined in the WinBUGS manuals (*Leuk: survival analysis using Cox regression* in Examples Volume I) to specify the models[96].

Vague prior distributions were used for all parameters.

C.2.2 *Details on Bayesian analyses*

Below we give WinBUGS code used to demonstrate the setup of the fully Bayesian Cox model where x_p is modelled and a_1, a_2 are ignored (this is the model denoted B1 in table 5.3). Models B2–B6 differ only in that they simply specify the models for BMI, weight and height^2 differently.

The data file is made up of the covariates `age sex hb logvl sqcd4 bmi`, a vector of length N indicating death `fail`, a vector of length N of survival times for all individuals `obst`, and a

vector of length T of distinct failure times `t`. Note that the data must be sorted in ascending order of `obst` before being passed to WinBUGS. All covariates are centred at their mean.

```
model
{
# Set up data
for(i in 1:N) {
    for(j in 1:T) {
        # risk set = 1 if obst >= t
        Y[i,j] <- step(obst[i] - t[j] + eps)
        # counting process jump = 1 if obst in [ t[j], t[j+1] )
        # i.e. if t[j] <= obst < t[j+1]
        dN[i, j] <- Y[i, j] * step(t[j + 1] - obst[i] - eps) * fail[i]
    }
}
# Analysis model
for(j in 1:T) {
    for(i in 1:N) {
        dN[i, j]   ~ dpois(Idt[i, j])              # Likelihood
        Idt[i, j] <- Y[i, j] * exp(eta[i]) * dL0[j] # Intensity
    }
    dL0[j] ~ dgamma(mu[j], c)
    mu[j] <- dL0.star[j] * c    # prior mean hazard
}
c <- 0.1
r <- 0.1
for (j in 1 : T) { dL0.star[j] <- r * (t[j + 1] - t[j]) }
for(i in 1:N) {
    eta[i] <- (beta1*age[i]) + (beta2*sex[i]) + (beta3*hb[i]) + (beta4*logvl[i])
        + (beta5*sqcd4[i]) + (beta6*(bmi[i]))
}
# Model for covariates.
# The specified univariate distributions imply marginal multivariate normality
for(i in 1:N) {
    # model for augmenting bmi
    bmi[i] ~ dnorm(mubmi[i],0.01)
    mubmi[i] <- dabmi0 + (dabmi1*age[i]) + (dabmi2*sex[i]) + (dabmi3*hb[i])
        + (dabmi4*logvl[i]) + (dabmi5*sqcd4[i])
    # model for augmenting cd4 count
    sqcd4[i] ~ dnorm(mucd4[i],0.01)
    mucd4[i] <- dacd40 + (dacd41*age[i]) + (dacd42*sex[i]) + (dacd43*hb[i])
        + (dacd44*logvl[i])
    # model for augmenting hb
    logvl[i] ~ dnorm(muvl[i],0.01)
    muvl[i] <- davl0 + (davl1*age[i]) + (davl2*sex[i]) + (davl3*hb[i])
    # model for augmenting hb
```

```
    hb[i] ~ dnorm(muhb[i],0.01)
    muhb[i] <- dahb0 + (dahb1*age[i]) + (dahb2*sex[i])
}
beta1 ~ dnorm(0,0.01) # priors
beta2 ~ dnorm(0,0.01)
... [these priors are used for all parameters]
}
```

The priors for regression coefficients are $\sim \mathrm{N}(0, 100)$. The prior for dL_0, the baseline intensity, requires slightly more explanation. This is modelled as $dL_0 \sim \Gamma(cr\{t_{(j+1)} - t_{(j)}\}, c)$, that is, a gamma distribution with mean $r\{t_{(j+1)} - t_{(j)}\}$ and variance $r\{t_{(j+1)} - t_{(j)}\}/c$. The expression $\{t_{(j+1)} - t_{(j)}\}$ is the time increment between the jth and $j+1$th failure times; in the Aurum data, the mean time increment was 8 days. Note that r is not invariant to the scale of t, although c is. We used $c = 0.1$ and $r = 0.1$. A change of time scale would require r to be altered to obtain an equivalent prior distribution.

C.2.3 Results

Fitting the Bayesian models in WinBUGS was troublesome.

For the Aurum data, all MCMC chains ran slowly and some stalled persistently. The simplest models (for example B1) took 5–10 hours to produce 5,000 iterations of the MCMC sampler. Model B5 took 10 days to produce 1,000 iterations and would only update under a very specific set of initial values. WinBUGS stalled repeatedly and the need to set the model updating again inflated the run time. We present results for model B5 but do not claim the MCMC sampling converged to the true posterior distribution. Results for model B6 are absent because WinBUGS was unable to sample at all; the reason for this was unclear. WinBUGS ran a lot faster when fitting models that imputed missing values of x_p actively, that is B1–B4.

Results for the Aurum data are given in figure C.1 (contrasting with the results obtained via MI in figure 5.1). Posterior distributions obtained from different fully Bayesian analyses give diverse results. For hæmoglobin, posterior means for all models except B5 are slightly closer to 0 than any of the MI models, and the 95% credible intervals tend to be slightly shorter than the MI confidence intervals. This may in part be the effect of the prior for the hazard, as seen in the comparison of Bayesian and frequentist analysis of complete cases. Under model B5 the posterior distribution for the log hazard ratio had mean much closer to zero with smaller posterior variance than under other models.

For BMI, posterior means from B1–B5 are very variable. B1 and B2 largely agree with the MI and (Bayesian) CC estimates, although the intervals are longer than those obtained after MI. Posterior means from B3 and B4 are closer to 0 and have shorter credible intervals than MI models or the other Bayesian models. For B4 this perhaps reflects the incorrect assumption made about the joint distribution of x_p, a_1, a_2 (this is surprising because the issue does not appear to affect model M4). Model B5 shows an effect in the *opposite* direction to all other estimates. This was the model that was very difficult to run in WinBUGS. As noted above we do not claim B5 ever converged to the true posterior density.

Figure C.1: Results from analyses of Aurum data under different Bayesian models for BMI.

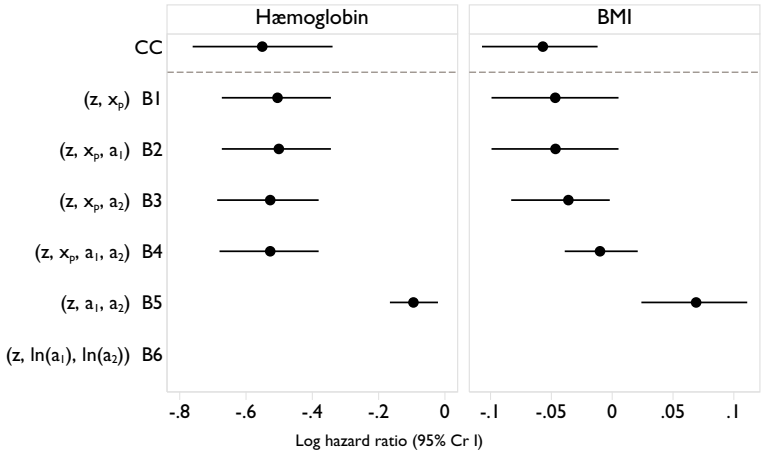
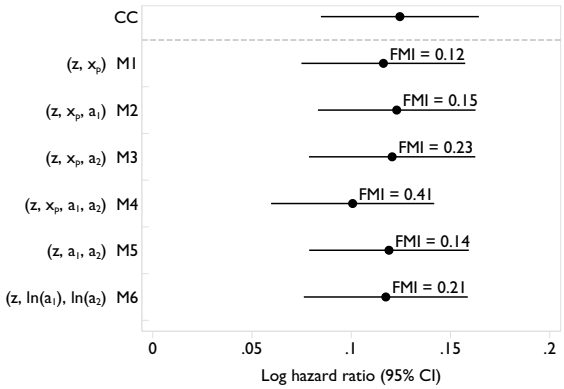


Figure C.2: Results from analyses of Epic-Norfolk data under different models for cholesterol ratio using predictive mean matching. The estimated fraction of missing information (FMI) is given next to MI analyses.



For the Epic-Norfolk data it was not possible to compile any of the fully Bayesian models in WinBUGS, even for complete cases. We tried compiling the complete cases model for subsets of the data of gradually increasing size (starting with $n = 1000$); model compilation failed beyond $n > 4000$. The Epic-Norfolk dataset is too large for WinBUGS and so attempts to fit the fully Bayesian models were abandoned. This is a setting where a fully Bayesian analysis is impractical to any but the most dedicated.

C.3 PREDICTIVE MEAN MATCHING TO IMPUTE CHOLESTEROL IN EPIC-NORFOLK

As described in section 5.5.2.1, we re-ran the imputation models for Epic-Norfolk using *predictive mean matching*. Figure C.2 gives the full results analogous to those given in figure 5.2. Note that, with the exception of model M5, there is less consistency between models than between the models that did not use PMM. Note also that the fraction of missing information is uniformly greater for the models that use PMM.

D NOTES ON SMC FCS

D.1 SMC FCS IN CHAPTERS 4 AND 5

The work presented in chapters 4 and 5 involved scenarios where there was no standard approach to imputation that would be considered satisfactory. In particular, this work threw up concerns about the use of 'passive' imputation. At the time, the idea of SMC FCS and its implementation as a Stata command were in the early stages of development and so its inclusion was not viable. However, the method lends itself well to the set up of this work involved in both chapters.

In chapter 4 the simulation study was set up such that the marginal distribution of covariates was easy to specify, but conditioning on outcome meant all imputation models considered were incorrectly specified. This would have been the ideal setting for SMC FCS: the proposal distribution would use $\boldsymbol{f}(\boldsymbol{x}) \sim$ MVN and specify the analysis model as a logistic regression of $y \mid \boldsymbol{g}(\boldsymbol{x})$ for rejection sampling.

In chapter 5 the marginal distribution of covariates was simulated as $(\log(a_1), \log(a_2)) \sim$ BVN. The analysis model then used the ratio a_1/a_2. Again, it is obvious that the SMC FCS approach could be used in this setting by using the correct marginal distribution of the covariates as the proposal distribution and the correct analysis model for rejection sampling.

Any future work on these topics will include SMC FCS.

D.2 A VISUAL EXPLORATION OF ISSUES WITH SMC FCS FOR FRACTIONAL POLYNOMIALS

The simulation work in chapter 6 shunned SMC FCS, despite its apparent promise. In preliminary runs to calibrate the simulation studies the method performed extremely badly. This appendix provides a sketch of when and how failures arose. Future work will consider how best to use SMC FCS in practice.

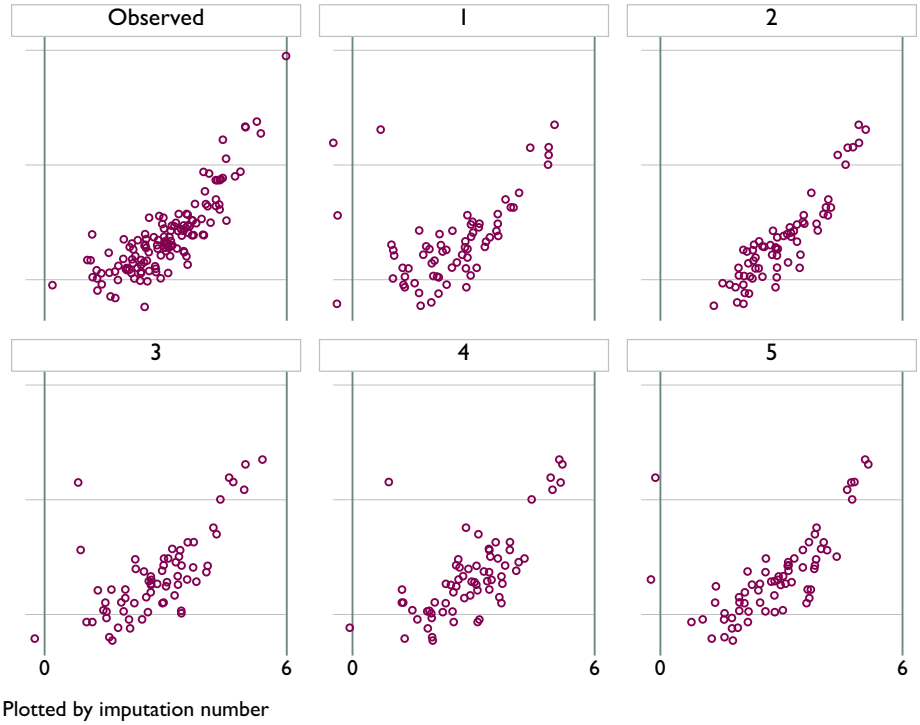
It is important to test SMC FCS in real datasets to understand when and how problems arise; ongoing work is beginning to demonstrate how and when it succeeds and fails.

Below, I demonstrate the difficulties that can arise using simples plots of y vs. x for observed and imputed data. In all cases the complete data are simulated and missing values are introduced completely at random in 40% of x observations.

Figure D.2 shows a dataset simulated as $x \sim \mathrm{N}(2.9, 1)$ and $y \sim \mathrm{N}(0.1x^{-1} + 0.02x^3, 0.3^2)$. The horizontal lines at 0 and 6 are placed just outside the range of x_{fi}. SMC FCS is used for imputation using $x \sim$ N for the proposal distribution, which is correct. The rejection probabilities come from fitting a linear regression of y on $x^{-2}, x^{-1}, x^{-0.5}, \ln(x), x^{0.5}, x, x^2, x^3$. For all imputed datasets except $m = 2$ some values of x are negative.

To deal with the problem of negative imputed values in figure D.2, an alternative would be to use $\log(x) \sim$ N for the proposal distribution. This leads to problems of rare large positive

Figure D.1: Simulated data where $x \sim \text{N}$ and $y = 0.1x^{-1} + 0.02x^3 + \sim \text{N}(0, 0.3)$ along with five SMC FCS imputations $x \sim \text{N}$ for the proposal distribution



values, which may introduce some bias to the estimation of exponents and coefficients.

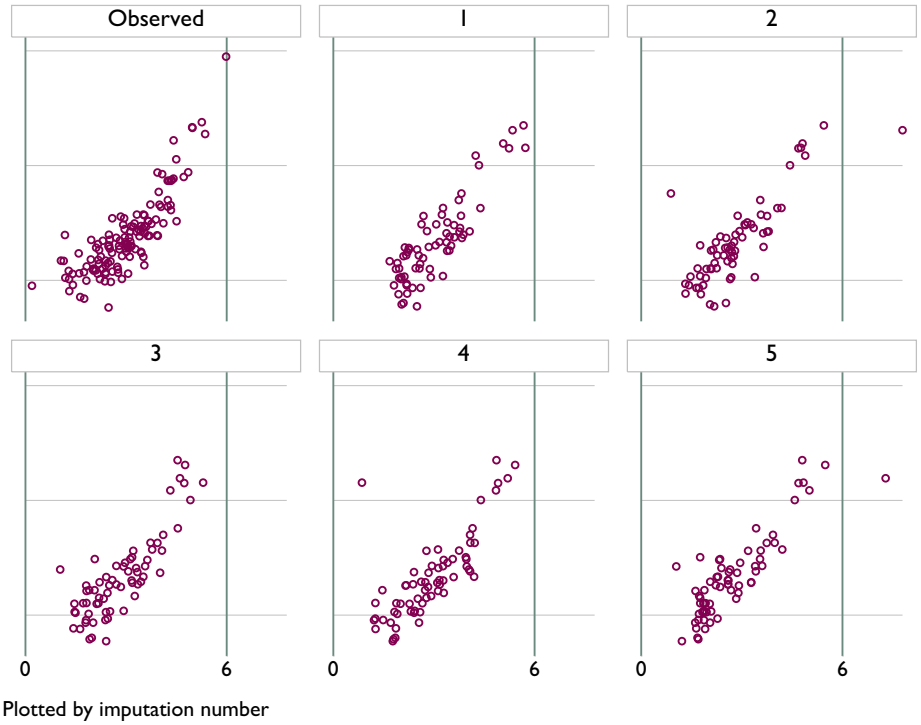
Figure D.2 plots simulated data mimicking the set up used in section 6.5, again with 40% of x values MCAR. The top row displays two imputed datasets using $x^{-0.5} \sim \text{N}$ for the proposal distribution; the middle row displays two imputed datasets using $\log(x) \sim \text{N}$ for the proposal distribution; and the bottom row displays two imputed datasets using $x^{-1} \sim \text{N}$ for the proposal distribution. While the top two rows appear to produce quite reasonable imputed values, specifying $x^{-1} \sim \text{N}$ for the proposal distribution clearly causes problems, with extreme negative outliers. Aside from these few values, the plot is very similar to the top two rows.

The above problems meant that the SMC FCS method was not used for any of the results shown in chapter 6.

The various Stata commands for MFP models involve automatic scaling of x^1 before applying FP transformations by using educated guesses based on the observed values, such that $\text{CV}(x)$ is high, but all x_i are positive. The above examples show that when $\text{CV}(x)$ is high, negative imputed values are more likely. This causes problems for fractional polynomials because it is a requirement that values of x^1 are positive; with negative values of x, it is impossible to calculated the required transformations.

Rescaling x *after* imputation is an unsatisfactory solution, because this alters the fit of the various models (as noted in chapter 6), meaning the rejection probabilities used by SMC FCS will be invalid on this new scale.

Figure D.2: Simulated data where $x \sim \text{N}$ and $y = 0.1x^{-1} + 0.02x^3 + \sim \text{N}(0, 0.3)$ along with five SMC FCS imputations $\log(x) \sim \text{N}$ for the proposal distribution



Two practical solutions would be:

1. Perform scaling of x *prior* to imputation. This will be more conservative than the default with complete data. For example, in figure D.2 a preliminary rescaling might be $(x_i + 1)$.

2. Use PMM to draw from the proposal distribution. PMM tends to be good at preserving marginal distributions, which is the requirement of the proposal distribution. It also forces imputed values to lie within the range of observed data, meaning that the scaling used prior to imputation is the same as the scaling that would be chosen post-imputation.

Use of PMM seems to be the better solution, because it does not require alteration of the analysis model to deal with missing data (this remark is not a scientific statement but an opinion I hold on multiple imputation). This is an area of current development for the `smcfcs` command. However, before recommending its use, simulation studies similar to those performed by Bartlett et al.[26] are required as a formal evaluation of its performance.

Figure D.3: Simulated data using the setup of section6.5 along with two imputations for three proposal distributions. Left column uses identical (observed) data for all three rows. Top row correctly uses $x^{-0.5} \sim \mathrm{N}$ for the proposal distribution; middle row uses $\log(x) \sim \mathrm{N}$; bottom row uses $x^{-1} \sim \mathrm{N}$