

Bayesian methods for modelling non-random missing data mechanisms in longitudinal studies

Alexina Jane Mason

Imperial College London

Department of Epidemiology and Public Health

PhD Thesis

Abstract

In longitudinal studies, data are collected on a group of individuals over a period of time, and inevitably this data will contain missing values. Assuming that this missingness follows convenient ‘random-like’ patterns may not be realistic, so there is much interest in methods for analysing incomplete longitudinal data which allow the incorporation of more realistic assumptions about the missing data mechanism. We explore the use of Bayesian full probability modelling in this context, which involves the specification of a joint model including a model for the question of interest and a model for the missing data mechanism.

Using simulated data with missing outcomes generated by an informative missingness mechanism, we start by investigating the circumstances and the extent to which Bayesian methods can improve parameter estimates and model fit compared to complete-case analysis. This includes examining the impact of misspecifying different parts of the model. With real datasets, when the form of the missingness is unknown, a diagnostic that indicates the amount of information in the missing data given our model assumptions would be useful. p_D is a measure of the dimensionality of a Bayesian model, and we explore its use and limitations for this purpose.

Bayesian full probability modelling is then used in more complex settings, using real examples of longitudinal data taken from the British birth cohort studies and a clinical trial, some of which have missing covariates. We look at ways of incorporating information from additional sources into our models to help parameter estimation, including data from other studies and knowledge elicited from an expert. Additionally, we assess the sensitivity of the conclusions regarding the question of interest to varying the assumptions in different parts of the joint model, explore ways of presenting this information, and outline a strategy for Bayesian modelling of non-ignorable missing data.

Contents

Acknowledgements	14
Glossary	15
Glossary of Models	16
1 Introduction and basic concepts	20
1.1 Why is missing data an important issue?	20
1.2 Structure of thesis	21
1.3 Types of missing data in statistical models	21
1.4 Missing data mechanisms	22
1.5 Typology of adjustment methods for missing data	24
1.5.1 Selection models	24
1.5.2 Pattern Mixture models	25
1.5.3 Shared Parameter models	25
1.6 Types of Bayesian models	26
1.6.1 Non-hierarchical regression models	26
1.6.2 Random effects models	27
1.6.3 Autoregressive models	28
1.7 Scope and aims of this research	29
2 Literature review	30
2.1 Overview of ad hoc methods	30
2.2 Overview of ‘statistically principled’ methods	31
2.3 Multiple Imputation	31
2.4 Bayesian full probability modelling	34
2.4.1 Formulation of models	34
2.4.2 Implementation	36
2.4.3 Model checking	37

2.4.4	Definition of DIC	38
2.4.5	DIC constructions for missing data models	40
2.4.6	Model uncertainty and sensitivity analysis	42
3	Data	45
3.1	British birth cohort studies	45
3.1.1	NCDS educational test scores data	46
3.1.2	MCS income data	48
3.2	Antidepressant trial	51
3.3	Synthetic data	53
3.3.1	Multivariate Normal (MVN) data	54
3.3.2	Scores data	54
4	Exploring Bayesian full probability modelling using simulated missingness	56
4.1	Specification of joint model for simple case simulations	56
4.2	Model evaluation	57
4.3	Design of synthetic data for simulations	58
4.4	Linear missingness	59
4.4.1	What are the deficiencies of complete-case analysis?	59
4.4.2	What improvements do we get from adding a model of missingness?	62
4.4.3	How critical is the strength of the relationship in the model of interest?	64
4.4.4	How critical is the adequacy of the model of missingness?	64
4.4.5	How critical is the error distribution in the model of interest?	65
4.5	Non-linear missingness	68
4.6	Model choice using DIC	69
4.6.1	The effective sample size (ESS)	72
4.6.2	Choice of plug-ins	74
4.6.3	Stability of DIC_O	76
4.6.4	DIC_O calculated without the use of reweighting	77
4.6.5	Model choice with DIC_O	79
4.6.6	Comments on the use of DIC_O and DIC_F	81
4.7	Interpreting the model of missingness scaled p_D	82
4.7.1	Relationship between scaled p_D and the change in β_1 between JM and CC	83
4.7.2	Relationship between scaled p_D and the fitted slope of the MoM, θ_1	84
4.7.3	Relationship between scaled p_D and the gradient of missingness, ϕ_1	84
4.7.4	Relationship between scaled p_D and reduction in MSE from CC to JM	85

4.7.5	Use of the model of missingness part of DIC_W	87
5	Bayesian full probability modelling of real data with missing responses	90
5.1	Example 1: NCDS educational test scores	90
5.1.1	Model of interest with single covariate	91
5.1.2	Expanded model of interest	92
5.2	Example 2: Antidepressant trial	94
5.2.1	Models of interest for HAMD data	94
5.2.2	Models of missingness for HAMD data	94
5.2.3	Model comparison for HAMD data	95
5.3	Example 3: MCS income	102
5.3.1	Choosing an initial model of interest	102
5.3.2	Simplifying assumptions	107
5.3.3	Adding a response model of missingness	109
6	Modelling missing covariates	117
6.1	Issues	117
6.1.1	Correlated binary covariates	118
6.1.2	Ordered categorical covariates	120
6.1.3	Unordered categorical covariates	121
6.1.4	Mixed covariates	122
6.2	MCS income example, assuming ignorable response missingness	122
6.2.1	Imputing <i>edu</i> and <i>sing</i> as correlated binary variables	124
6.2.2	Imputing <i>edu</i> , <i>sing</i> and <i>reg</i> as correlated binary variables	128
6.2.3	Imputing <i>edu</i> as an ordered categorical variable	129
6.2.4	Imputing categorical <i>edu</i> and binary <i>sing</i> as correlated variables	131
6.2.5	Adding continuous <i>age</i> to the covariate model of missingness	133
6.3	MCS income example, with a response model of missingness	134
7	Incorporating information from additional sources	139
7.1	Data from another study	139
7.1.1	BCS70 educational level data	140
7.1.2	Use in the covariate model of missingness	141
7.2	Expert knowledge	146
7.2.1	Elicitation	148
7.2.2	Use in the response model of missingness	154

8	Designing sensitivity analyses	161
8.1	Choice of sensitivity analyses for the MCS income example	162
8.1.1	Sensitivity to model of interest assumptions	162
8.1.2	Sensitivity to response model of missingness assumptions	163
8.1.3	Sensitivity to covariate model of missingness assumptions	164
8.1.4	Sensitivity to dataset	164
8.2	Results of sensitivity analyses for the MCS income example	165
8.2.1	Assumption sensitivity	165
8.2.2	Parameter sensitivity	171
8.3	Strategy for sensitivity analyses	176
9	Bayesian modelling of missing data: conclusions and extensions	179
9.1	Strengths and limitations of Bayesian joint models	179
9.2	Substantive questions	180
9.3	Strategy for Bayesian modelling of non-ignorable missing data	181
9.3.1	Proposed model building strategy	181
9.3.2	Proposed use of DIC in the presence of missing data	182
9.4	Extensions	184
A	Datasets	186
A.1	Details of NCDS educational test scores data	186
A.2	Details of MCS income data	186
A.3	Details of BCS70 data	189
B	Simulated non-linear missingness	192
B.1	Quadratic missingness	192
B.2	Step missingness	198
C	MCS income example	201
C.1	Models of interest fitted to complete cases	201
C.2	Effective sample size	203
C.3	Additional details on the JM.C1 covariate model of missingness	204
D	Deviance Information Criteria (DIC)	205
D.1	Algorithm for DIC based on the observed data likelihood (DIC_O)	205
D.1.1	A version using the DH reweighting method	205
D.1.2	A version without reweighting (Gold Standard)	207

D.2	Algorithm for DIC based on the full data likelihood (DIC_F)	207
D.3	Implementation	208
E	WinBUGS code	210
E.1	Code for running joint model JMD	211
E.2	Code for running model of interest MoIRJ	215
	Bibliography	216

List of Figures

1.1	Graphical representation for a simple regression model	27
1.2	Graphical representation for a random effects model	28
2.1	Graphical representation for ignorable missingness in the response variable	35
2.2	Graphical representation for non-ignorable missingness in the response variable	35
2.3	Simplified graphical representation for non-ignorable missingness in the response variable	36
3.1	NCDS mathematics test scores (subset with observed values for all three scores) . . .	47
3.2	MCS hourly net pay by sweep	49
3.3	Observed mean response profiles for antidepressant trial	53
4.1	MVN: performance of CC and JM compared with the TRUE generated targets	61
4.2	Scores.linear: performance of CC and JM compared with the TRUE generated targets	62
4.3	Comparison of shapes of linear missingness and their associated logits	63
4.4	MVN, MVN(sq), MVN(sr) and MVN(log): impact of different transforms and missingness models	66
4.5	ESS for <i>Scores.LinMostPos</i> modelled using Scores.JM1	73
4.6	Posterior distributions of the first 5 <i>logitp</i> for <i>Scores.LinMostPos</i>	74
4.7	DIC _O calculations for <i>Scores.LinMostPos</i> modelled using Scores.JM1	79
4.8	DIC _O Gold Standard algorithm calculations for <i>Scores.LinMostPos</i>	80
4.9	MVN: the relationship of scaled p_D with β_1 and θ_1	83
4.10	Scores.linear: the relationship of scaled p_D to the slope in the model of interest	84
4.11	Scores.linear: the relationship of scaled p_D with the gradient of missingness	84
4.12	The relationship of scaled p_D to the % reduction in MSE from CC to JM	85
5.1	HAMD example: modelled mean response profiles - comparing the model of interest .	97
5.2	HAMD example: modelled mean response profiles - comparing the model of missingness	98
5.3	HAMD example: $\bar{D}$ and $\hat{D}$ in DIC _O calculations	101
5.4	Graphical representation for model of interest with individual and ward random effects	104
5.5	Graphical representation for model of interest with individual random effects and stratum specific intercepts	104

5.6	MCS income example: wards by sample size	106
5.7	MCS income example: mean income for selected covariates	107
5.8	Proposed strategy for the use of DIC in modelling data with missing response	116
6.1	Graphical representation for a regression model with covariate missingness	118
6.2	Graphical representation for model of interest and covariate model of missingness . . .	123
6.3	Graphical representation for joint model consisting of a model of interest, covariate model of missingness and response model of missingness	135
7.1	MCS income example: screen shots from ELICITOR of the original elicitation	149
7.2	MCS income example: prior densities generated by forward sampling using WinBUGS, based on the original elicitation for selected design points	153
7.3	MCS income example: prior and posterior distributions for parameters in the response model of missingness	157
7.4	MCS income example: prior and posterior probability of response as the change in pay varies and as the level of pay in sweep 1 varies, with the other parameters fixed	158
7.5	MCS income example: probability of response for varying <i>change</i> and <i>level</i>	159
8.1	MCS income example: posterior predictive distribution and observed value of hourly pay to show the fit of re-issued individuals using base model (JMD)	169
8.2	MCS income example: proportional increase in pay associated with selected covariates versus δ_1 , conditional on δ_2 from JMJ variants	172
8.3	MCS income example: the four quadrants of δ_1 and δ_2	173
8.4	MCS income example: posterior mean of proportional change in pay associated with selected covariates versus δ_1 and δ_2 from JMJ variants	174
8.5	MCS income example: response model of missingness mean deviance, DIC_W and scaled p_D , and the mean square error of the fit of hourly pay for the re-issued individuals versus δ_1 and δ_2 from JMJ variants	175
8.6	Proposed strategy for sensitivity analyses for models with missingness	177
9.1	Proposed strategy for Bayesian modelling of non-ignorable missing data	183
B.1	Shapes of the quadratic missingness imposed on the datasets in scenario Scores.quad and their associated logits	193
B.2	Examples of modelling quadratic shaped missingness with a quadratic logit	196
B.3	The effects of different model of missingness priors on <i>jsteep</i>	197
B.4	Examples of modelling quadratic shaped missingness with a piecewise linear logit . . .	199
B.5	Example of modelling step shaped missingness with a linear logit	200
C.1	MCS income example: residual plots for models of interest MoIA-MoID	201
C.2	MCS income example: residual plots v selected covariates for MoID	202

List of Tables

1	Models for analysing synthetic data	16
2	Models of interest for the MCS income example	17
3	Joint models with a model of interest and a response model of missingness for the MCS income example	18
4	Joint models with a model of interest and a covariate model of missingness for the MCS income example	18
5	Joint models with a model of interest, response model of missingness and covariate model of missingness for the MCS income example	19
3.1	Complete cases for models of mathematics test score at 16	48
3.2	Missingness Pattern for NCDS educational test scores	48
3.3	Description of MSC income dataset variables	50
3.4	Missingness Pattern for sweep 1 MCS income data	51
3.5	Missingness Pattern for sweep 2 MCS income data	51
3.6	Missingness Pattern for sweeps 1 and 2 of the MCS income data	52
3.7	Missingness Pattern for Antidepressant Trial Data	52
3.8	Missingness by treatment and week for Antidepressant Trial Data	53
4.1	Summary of MVN synthetic data scenarios	59
4.2	Summary of Scores synthetic data scenarios	59
4.3	Joint Models for <i>Scores.LinMostPos</i>	71
4.4	Effective sample size for <i>Scores.LinMostPos</i>	73
4.5	Skewness of posterior distribution for <i>Scores.LinMostPos</i>	75
4.6	DIC _O for <i>Scores.LinMostPos</i> using Scores.JM1 (Ksample variability)	76
4.7	DIC _O for <i>Scores.LinMostPos</i> using Scores.JM1 (Qsample variability 1)	77
4.8	DIC _O for <i>Scores.LinMostPos</i> using Scores.JM1 (Qsample variability 2)	77
4.9	DIC _O for <i>Scores.LinMostPos</i> using Scores.JM1 (Gold Standard algorithm)	78
4.10	DIC _O for <i>Scores.LinMostPos</i> (Gold Standard v Reweighting)	81
5.1	NCDS example (single covariate): parameter estimates	91

5.2	NCDS example (single covariate): MAR v MNAR model of missingness parameter estimates	92
5.3	NCDS example (multiple covariates): parameter estimates	93
5.4	HAMD example: models of missingness	95
5.5	HAMD example: parameter estimates for AR2 models	96
5.6	HAMD example: parameter estimates for RE models	96
5.7	HAMD example: model of missingness scaled p_D	98
5.8	HAMD example: skewness of posterior distributions of selected parameters	99
5.9	HAMD example: DIC_O	99
5.10	HAMD example: model of missingness DIC_W	101
5.11	MCS income example: summary of models of interest	105
5.12	MCS income example: subset of parameter estimates for models of interest MoIA-MoID	105
5.13	MCS income example: DIC for model of interest	106
5.14	MCS income example: marginal probability of inclusion of covariates in models with reversible jump	108
5.15	MCS income example: sweep 1 by sweep 2 counts of educational level for subset of 505 MCS individuals	109
5.16	MCS income example: details of joint models consisting of a model of interest and a response model of missingness	110
5.17	MCS income example: parameter estimates for joint models JM.R1-JM.R8	111
5.18	MCS income example: skewness of posterior distributions for selected parameters . . .	112
5.19	MCS income example: model of missingness DIC_W for JM.R1-JM.R8	112
5.20	MCS income example: comparison of parameter estimates for a complete case analysis (MoIE) and a joint model (JM.R8)	114
6.1	MCS income example: summary of the covariate models of missingness for two correlated binary variables (dichotomised <i>edu</i> and <i>sing</i>)	124
6.2	MCS income example: boundaries on latent variables for dichotomised <i>edu</i> and <i>sing</i> .	127
6.3	MCS income example: summary of covariate imputations for five variants of model JM.C1 (dichotomised <i>edu</i> and <i>sing</i>)	128
6.4	MCS income example: boundaries on latent variable for three category <i>edu</i>	130
6.5	MCS income example: comparison of observed and imputed educational level for sweep 2 from model JM.C3	131
6.6	MCS income example: boundaries on latent variable for four category <i>edu</i>	131
6.7	MCS income example: comparison of parameter estimates for different combinations of sub-models	136
6.8	MCS income example: comparison of covariate imputations from joint models with (JMA) and without (JM.C6) a response model of missingness and simplistic imputations	137
6.9	MCS income example: sweep 1 by sweep 2 counts of three category education level for the 505 MCS individuals	138

7.1	MCS income example: sweep 5 by sweep 6 counts of educational level for 157 selected BCS70 individuals	141
7.2	MCS income example: comparison of educational level data from MCS and BCS70 . .	141
7.3	MCS income example: comparison of parameter estimates with (JMB) and without (JMA) including external data from BCS70	143
7.4	MCS income example: comparison of covariate imputations with (JMB) and without (JMA) including external data from BCS70	144
7.5	MCS income example: comparison of interval widths of covariate model of missingness parameters for joint models run with 157 BCS70 records (JMB) and 10 replicates of the BCS70 records (JMBx10)	145
7.6	MSC income example: explanatory variables for income missingness in sweep 2	148
7.7	MSC income example: elicited values of response percentage for sweep 2 income . . .	150
7.8	MCS income example: comparison of parameter estimates using a response model of missingness with linear functional form (JMB) and piecewise linear functional form with (JMD) and without informative priors (JMC)	155
7.9	MCS income example: comparison of covariate imputations using a response model of missingness with linear functional form (JMB) and piecewise linear functional form with (JMD) and without informative priors (JMC)	156
8.1	MSC income example: summary of differences of joint models for sensitivity analysis from base model (JMD)	165
8.2	MSC income example: parameter estimates for joint models JMB-JMI (with non-negligible differences ^c from JMD highlighted in bold)	166
8.3	MCS income example: MSE of imputed hourly pay for seven re-issued individuals for joint models JMB-JMI	170
8.4	MCS income example: response model of missingness DIC_W calculated using <i>logitp</i> plug-ins for joint models JMB-JMI	170
8.5	MCS income example: estimates of proportional change in pay associated with selected covariates for JMJ variants compared with base model (JMD)	171
A.1	Details of the source of NCDS educational test scores variables	186
A.2	Criteria for inclusion of record in MCS income dataset	187
A.3	Details of the source of MCS income variables	187
A.4	Levels of National Vocational Qualification (NVQ) equivalence	189
A.5	National Statistics Socio-Economic Classification (NS-SEC)	189
A.6	Criteria for inclusion of record in BCS70 dataset	189
A.7	Details of the source of BCS70 variables	190
B.1	Scores.quad: modelling quadratic shaped missingness with a linear logit	194
B.2	Scores.quad: modelling quadratic shaped missingness with a quadratic logit	195
B.3	Scores.quad: modelling quadratic shaped missingness with a piecewise linear logit . . .	198
B.4	Scores.step: modelling step shaped missingness with a linear logit	200

C.1	MCS income example: parameter estimates for models of interest MoIA-MoIE	202
C.2	MCS income example: effective sample size for estimating posterior means	203
C.3	MCS income example: covariate model of missingness parameter estimates for five variants of JM.C1, which impute the missing values for dichotomised educational level and single/partner allowing for correlation between these two covariates	204

Acknowledgements

I would like to express my gratitude to my supervisors, Nicky Best, Sylvia Richardson and Ian Plewis, for their guidance and enthusiasm throughout my PhD. I also thank John, without whom I would neither have started nor completed this thesis, for his support and interest.

Thanks also to Mike Kenward who provided the clinical trial data analysed in this project. Lastly I would like to acknowledge the Economic and Social Research Council who provided funding.

Glossary

BCS70: British Cohort Study, the 1970 British birth cohort study

CC: A model of interest, run on Complete Cases only

CMoM: Covariate Model of Missingness

DH: Daniels and Hogan (2008)

DK: Diggle and Kenward (1994)

JM: A joint model, includes a model for the question of interest and a model for the missing data mechanism

MCS: Millennium Cohort Study, the 2000/01 British birth cohort study

MoI: Model of Interest

MoM: Model of Missingness

NCDS: National Child Development Study, the 1958 British birth cohort study

RMoM: Response Model of Missingness

TRUE: A joint model consisting of a model of interest and a model of missingness, run with a full set of values (for bench marking datasets with simulated missingness)

Glossary of Models

Synthetic data

Table 1: Models for analysing synthetic data

Model Name	Model of Interest	Model of Missingness	Response ^a	Covariate ^b
JM	✓	✓ logit - estimated	$\mathbf{y}$	$\mathbf{x}$
CC	✓	✗	$\mathbf{y}_{obs}$	$\mathbf{x}_{obs}$
TRUE	✓	✓ logit	$\mathbf{y}_{full}$	$\mathbf{x}$
JMfixed	✓	✓ logit - fixed ^c	$\mathbf{y}$	$\mathbf{x}$
JMexact	✓	✓ fixed ^d	$\mathbf{y}$	$\mathbf{x}$

^a $\mathbf{y}$ = combination of observed and missing responses; $\mathbf{y}_{obs}$ = observed responses only; $\mathbf{y}_{full}$ = completely observed responses, before missingness imposed.

^b $\mathbf{x}$ = completely observed covariates; $\mathbf{x}_{obs}$ = covariates for the individuals in $\mathbf{y}_{obs}$.

^c model of missingness parameters fixed to the posterior means of $\boldsymbol{\theta}$ estimated by TRUE.

^d parameters fixed in exact equation that was used to impose the missingness (identity link).

MCS Income Example

We use the following conventions to name the models that we use for the MCS income example:

MoI + letter: model of interest only;

JM.R + number: joint model consisting of a model of interest and a model of missingness for the response;

JM.C + number: joint model consisting of a model of interest and a model of missingness for the covariates;

JM + letter: joint model consisting of a model of interest (MoI), a covariate model of missingness (CMoM) and a response model of missingness (RMoM).

A summary of all of the models that we use is given below. The number of categories used for the categorical variables is given in brackets, with 2 indicating a binary variable. All functions are linear and all priors are non-informative or very mildly informative on a specified scale unless otherwise stated.

Table 2: Models of interest for the MCS income example

model label	response transform	ward random effects	stratum specific intercepts	covariates	error distribution
MoIA	<i>log</i>	✓	✗	<i>age, edu(6), eth(2), hsize, kids, reg(2), sing(2)</i>	Normal
MoIB	<i>log</i>	✓	✗	<i>age, edu(6), eth(2), hsize, kids, reg(2), sing(2)</i>	t_4
MoIC	<i>log</i>	✗	✓	<i>age, edu(6), eth(2), hsize, kids, reg(2), sing(2)</i>	Normal
MoID	<i>log</i>	✗	✓	<i>age, edu(6), eth(2), hsize, kids, reg(2), sing(2)</i>	t_4
MoIE	<i>log</i>	✗	✓	<i>age, edu(6), eth(2), reg(2), sing(2)</i>	t_4
MoIF	<i>log</i>	✗	✓	<i>age, edu(3), eth(2), reg(2), sing(2)</i>	t_4
MoIRJ ^a	<i>log</i>	✗	✓	<i>age, age², edu(6), eth(2), hsize, kids, reg(2), sing(2), age × edu(6)</i>	t_4
MoIRJg ^b	<i>log</i>	✗	✓	<i>age, age², edu(6), eth(2), hsize, kids, reg(2), sing(2), age × edu(6)</i>	t_4

All models include individual random effects.

^a run with a reversible jump variable selection sub-model, where the binary components of the categorical covariates can be selected separately

^b run with a reversible jump variable selection sub-model, where none or all of the binary components of the categorical covariates must be selected

Table 3: Joint models with a model of interest and a response model of missingness for the MCS income example

model label	model of interest	response model of missingness type	regressors for $\text{logit}(p_i)$
JM.R1	MoIE	MCAR	
JM.R2	MoIE	MAR	level^a
JM.R3	MoIE	MNAR	level2^b
JM.R4	MoIE	MNAR	$\text{level}^a, \text{change}^c$
JM.R5	MoIE	MAR	$\text{ctry}(4), \text{eth}(2), \text{sc}(4)$
JM.R6	MoIE	MAR	$\text{ctry}(4), \text{eth}(2), \text{sc}(4), \text{level}^a$
JM.R7	MoIE	MNAR	$\text{ctry}(4), \text{eth}(2), \text{sc}(4), \text{level2}^b$
JM.R8	MoIE	MNAR	$\text{ctry}(4), \text{eth}(2), \text{sc}(4), \text{level}^a, \text{change}^c$
JM.R9	MoIF	MNAR	$\text{ctry}(4), \text{eth}(2), \text{sc}(4), \text{level}^a, \text{change}^c$

^a level is sweep 1 untransformed hourly pay

^b level2 is sweep 2 untransformed hourly pay

^c change is $\text{level2} - \text{level}$

Table 4: Joint models with a model of interest and a covariate model of missingness for the MCS income example

model label	model of interest	covariate model of missingness					covariate types
		<i>edu</i>	<i>sing</i>	<i>reg</i>	<i>age</i>		
JM.C1	MoIE, but <i>edu</i> (2)	✓(2)	✓(2)				2 binary
JM.C2	MoIE, but <i>edu</i> (2)	✓(2)	✓(2)	✓(2)			3 binary
JM.C3	MoIF	✓(3)					1 categorical (3 levels)
JM.C4	MoIE, but <i>edu</i> (4)	✓(4)					1 categorical (4 levels)
JM.C5	MoIF	✓(3)	✓(2)				1 categorical(3); 1 binary
JM.C6	MoIF	✓(3)	✓(2)		✓		1 categorical(3); 1 binary; 1 continuous

All models allow for correlation between the ticked covariates. Missing values for the unticked variables are set off-model.

Table 5: Joint models with a model of interest, response model of missingness and covariate model of missingness for the MCS income example

model label	response transform	Model of Interest covariates ^a	error distribution	regressors ^b	Response Model of Missingness functional form ^c	includes elicitation	Covariate Model of Missingness includes BCS70 data
JMA	$\log$	as MoIF	t ₄	set1	linear	✗	as JM.C6 ✗
JMB	$\log$	as MoIF	t ₄	set1	linear	✗	as JM.C6 ✓
JMC	$\log$	as MoIF	t ₄	set2	piecewise linear	✗	as JM.C6 ✓
JMD	$\log$	as MoIF	t ₄	set2	piecewise linear	✓(final)	as JM.C6 ✓
JME	$\log$	as MoIF	t ₄	set2	piecewise linear	✓(original)	as JM.C6 ✓
JMF	$\log$	as MoIF	t ₄	set2	piecewise linear	✓(fixed ^d)	as JM.C6 ✓
JMG	$\log$	as MoIF	Normal	set2	piecewise linear	✓(final)	as JM.C6 ✓
JMH	$\log$	as MoIF + $age^2, age \times edu(3)$	t ₄	set2	piecewise linear	✓(final)	as JM.C6 ✓
JMI	$\sqrt{\cdot}$	as MoIF	t ₄	set2	piecewise linear	✓(final)	as JM.C6 ✓
JMJ	$\log$	as MoIF	t ₄	set2	piecewise linear	✓(final/fixed ^e)	as JM.C6 ✓

The model of interest of all these models has individual random effects and stratum specific intercepts.

^a the set of covariates for MoIF is $\{age, edu(3), eth(2), reg(2), sing(2)\}$

^b $set1 = \{ctry(4), eth(2), sc(4), level, change\}$, $set2 = \{ctry(2), eth(2), sc(2), level, change\}$

^c The functional form for $level$ and $change$.

^d All the response model of missingness parameters are fixed to their final elicitation median values and not estimated.

^e This model has 49 variants, with the two parameters associated with $change$ fixed to different values. All the parameters associated with the other regressors in the response model of missingness have informative priors based on the final elicitation.

Chapter 1

Introduction and basic concepts

1.1 Why is missing data an important issue?

Missing data is pervasive in many areas of scientific investigations under different forms and complicates subsequent analyses for researchers. It is generally an unavoidable nuisance, which can lead to biased and inefficient inference if ignored or handled inappropriately. Given its ubiquitous nature and potential for distorting scientific investigation, unsurprisingly an extensive literature has built up on the topic. In epidemiology and biostatistics the various approaches, covering both cross sectional and longitudinal studies, have been catalogued and reviewed in papers (Schafer and Graham, 2002; Ibrahim *et al.*, 2005) and websites (Carpenter and Kenward, 2005; van Buuren, 2005), as well as detailed in comprehensive textbooks (Schafer, 1997; Little and Rubin, 2002; Molenberghs and Kenward, 2007; Daniels and Hogan, 2008). Such is the importance of missing data, that entire issues of journals have been devoted to the subject (Everitt, 1999; van Buuren, 2003; Lynn, 2006).

In this thesis we focus on the use of Bayesian methods for modelling non-random missing data mechanisms in longitudinal studies. Longitudinal studies allow social scientists and epidemiologists to follow individuals over a period of time, and the data collected enable wide ranging analyses on diverse topics including health, education, economic and social circumstances, often with the purpose of informing policy decisions (Feinstein, 2003). Inevitably such studies lose members over time and generally suffer from missing data. Currently, there is much interest in the extent and determinants of attrition in longitudinal studies (Nicoletti and Peracchi, 2005; Behr *et al.*, 2005) and general concern regarding the impact of non-response on inference (Hawkes and Plewis, 2006; Diggle *et al.*, 2007).

1.2 Structure of thesis

The scope and aims of this thesis are discussed in this chapter, after describing some basic missing data concepts and standard Bayesian models. This is followed by a literature review in Chapter 2, and an introduction to the datasets used in Chapter 3. In Chapter 4, we explore Bayesian full probability modelling using simulated response missingness. Our investigation of response missingness continues in Chapter 5, with three increasingly complex real examples. Covariate missingness is the topic of Chapter 6, while in Chapter 7 we examine ways of incorporating additional information to inform about the missing data process and/or values, and in Chapter 8 we explore an approach to sensitivity analysis. Finally, in Chapter 9 we outline a strategy for building Bayesian models to analyse data from longitudinal studies with non-random missingness, and suggest directions for future research in this area. Chapters 2-8 conclude with a boxed summary of the key points.

Much of the material in Chapter 4 and abridged versions of Examples 1 and 2 from Chapter 5 are used in a paper on insights into the use of Bayesian models for informative missing data, which is in the final stage of preparation. Some of these insights were presented at the International Biometric Conference, held in Dublin in July 2008. The ideas and expert elicitation in Chapter 7 were used to form a presentation entitled ‘Modelling non-random missing data in longitudinal studies: how can information from additional sources help?’, that was part of the session on methodological developments for combining data at the 2008 ESRC research methods festival in Oxford.

We now introduce some basic concepts which are key in analysing missing data (Sections 1.3-1.5) and the Bayesian formulation of standard models used in this thesis (Section 1.6). Then, this chapter is concluded by a description of the scope and aims of this research.

1.3 Types of missing data in statistical models

In general, before proceeding with an analysis of data which includes missing values, it is useful to determine the type of missing data under consideration. This includes distinguishing between different patterns of missing data, such as univariate, monotone and arbitrary (Schafer and Graham, 2002; Little and Rubin, 2002). In a univariate pattern, missingness is confined to a single variable or group of variables which are either entirely observed or entirely missing for each individual. To have monotone missingness, variables or variable groups in a rectangular data set, denoted $\mathbf{y} = (y_{ij})$, for $i = 1, \dots, n$ individuals and $j = 1, \dots, k$ variables, can be ordered in such a way that if y_j is missing for a individual, then $y_{j+1}, \dots, y_k$ are missing as well. Some methods for dealing with missing data are only applicable to certain missingness patterns.

The above missing data patterns can apply to both cross sectional and longitudinal studies, but there

are also distinctions which are specific to longitudinal data. In longitudinal studies, non-response can take three forms: unit non-response (sampled individuals are absent from the outset of the study), wave non-response (where an individual does not respond in a particular wave but re-enters the study at a later stage) and attrition or drop-out (where an individual is permanently lost as the study proceeds), and these may have different characteristics (Hawkes and Plewis, 2006). Also, different kinds of non-response can often be distinguished, typically not located, not contacted and refusal. Missing data patterns in longitudinal studies may be further complicated by data missing on particular items (item non-response) or on a complete group of questions (domain non-response).

From a modelling perspective, it also makes a difference whether we are dealing with missing response, missing covariates or missingness in both the response and covariates. In the multivariate case, the problem might extend to missing responses.

1.4 Missing data mechanisms

The appropriateness of a particular approach is dependent on the mechanism that leads to the missing data, and Rubin (1976) developed a framework for inference from incomplete data that is still widely used. Following Rubin, missing data are generally classified into three types: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Informally, MCAR occurs when the missingness does not depend on observed or unobserved data, in the less restrictive MAR it depends only on the observed data, and when neither MCAR or MAR hold, the data are MNAR.

More formally (Little and Rubin, 2002), let $\mathbf{y} = (y_{ij})$ denote a rectangular data set for $i = 1, \dots, n$ individuals and $j = 1, \dots, k$ variables, and partition $\mathbf{y}$ into observed, $\mathbf{y}_{obs}$, and missing, $\mathbf{y}_{mis}$, values, i.e. $\mathbf{y} = (\mathbf{y}_{obs}, \mathbf{y}_{mis})$. Now define $\mathbf{m} = (m_{ij})$ to be a binary indicator variable such that

$$m_{ij} = \begin{cases} 0: & y_{ij} \text{ observed} \\ 1: & y_{ij} \text{ missing.} \end{cases}$$

The missing data mechanism can then be defined by the form of the conditional distribution of $\mathbf{m}$ given $\mathbf{y}$, say $f(\mathbf{m}|\mathbf{y}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the unknown parameters of the missingness function. Then, the data are MCAR if

$$f(\mathbf{m}|\mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta}) = f(\mathbf{m}|\boldsymbol{\theta}).$$

Note that this allows dependence on $\boldsymbol{\theta}$, but not on the values of any variables in $\mathbf{y}$. For MAR data, the conditional missingness distribution becomes

$$f(\mathbf{m}|\mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta}) = f(\mathbf{m}|\mathbf{y}_{obs}, \boldsymbol{\theta}). \tag{1.1}$$

For MNAR, there is no simplification of the conditional distribution of $\mathbf{m}$.

The implications of these different missing data mechanisms can be understood by considering the joint distribution of $\mathbf{y}$ and $\mathbf{m}$, i.e.

$$f(\mathbf{y}, \mathbf{m} | \boldsymbol{\beta}, \boldsymbol{\theta}) = f(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m} | \boldsymbol{\beta}, \boldsymbol{\theta}), \quad (1.2)$$

where $\boldsymbol{\beta}$ denotes the unknown parameters of the model of interest. We cannot use standard methods because the likelihood depends on the missing data. However, the marginal distribution of the observed data can be obtained by integrating out the missing data,

$$f(\mathbf{y}_{obs}, \mathbf{m} | \boldsymbol{\beta}, \boldsymbol{\theta}) = \int f(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m} | \boldsymbol{\beta}, \boldsymbol{\theta}) d\mathbf{y}_{mis}. \quad (1.3)$$

The joint distribution can be specified as the product of the distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{m}$ given $\mathbf{y}$, i.e.

$$f(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m} | \boldsymbol{\beta}, \boldsymbol{\theta}) = f(\mathbf{y}_{obs}, \mathbf{y}_{mis} | \boldsymbol{\beta}, \boldsymbol{\theta}) f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\beta}, \boldsymbol{\theta}). \quad (1.4)$$

This can be simplified to

$$f(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m} | \boldsymbol{\beta}, \boldsymbol{\theta}) = f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta}) f(\mathbf{y}_{obs}, \mathbf{y}_{mis} | \boldsymbol{\beta}) \quad (1.5: \text{Selection Model})$$

if we assume that $\mathbf{m} | \mathbf{y}, \boldsymbol{\theta}$ is conditionally independent of $\boldsymbol{\beta}$, and $\mathbf{y} | \boldsymbol{\beta}$ is conditionally independent of $\boldsymbol{\theta}$, which is usually reasonable in practice. This form of the joint distribution is known as a selection model and will be discussed further in Section 1.5. If the data are MAR (similar logic will hold for MCAR), then Equations 1.1 and 1.5 imply that Equation 1.3 can be rewritten as

$$\begin{aligned} f(\mathbf{y}_{obs}, \mathbf{m} | \boldsymbol{\beta}, \boldsymbol{\theta}) &= f(\mathbf{m} | \mathbf{y}_{obs}, \boldsymbol{\theta}) \int f(\mathbf{y}_{obs}, \mathbf{y}_{mis} | \boldsymbol{\beta}) d\mathbf{y}_{mis} \\ &= f(\mathbf{m} | \mathbf{y}_{obs}, \boldsymbol{\theta}) f(\mathbf{y}_{obs} | \boldsymbol{\beta}). \end{aligned} \quad (1.6)$$

In this case the joint distribution has factored into two terms, one involving the observed data and parameters $\boldsymbol{\beta}$, and the other involving the missingness indicator and parameters $\boldsymbol{\theta}$. The missing data mechanism is termed *ignorable* for likelihood inference about $\boldsymbol{\beta}$ if the missing data are MAR and the parameters of the data model, $\boldsymbol{\beta}$, and the missingness mechanism, $\boldsymbol{\theta}$, are distinct (Little and Rubin, 2002). For ignorability to hold in Bayesian inference, in addition the priors for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ need to be independent. Note that the term ignorable applies to the missingness model, not the missing data. If the distinctness of parameters condition does not hold, then ignoring the missingness term will lead to consistent but not fully efficient inference, as $f(\mathbf{m} | \mathbf{y}_{obs}, \boldsymbol{\theta})$ would provide information about those parameters in $\boldsymbol{\theta}$ which also belong to $\boldsymbol{\beta}$. Often the missing data mechanism is unknown and the data alone cannot determine whether we have MAR or MNAR missingness, making sensitivity analysis essential.

In the context of drop-out, the MCAR, MAR and MNAR mechanisms are sometimes referred to as completely random drop-out (CRD), random drop-out (RD) and non-random drop-out (NRD) (Diggle and Kenward, 1994; Kenward and Molenberghs, 1999).

1.5 Typology of adjustment methods for missing data

Broadly speaking, there are two types of methods for handling missing data: ad hoc methods and ‘statistically principled’ methods, which are discussed more fully in Chapter 2. Ad hoc methods may have the advantage of relative simplicity, but they usually introduce bias and do not reflect statistical uncertainty. By contrast, ‘statistically principled’ methods combine information in the observed data with assumptions about the missing value mechanism. When the missingness mechanism cannot be assumed to be ignorable, a ‘statistically principled’ approach is to build a joint model and undertake sensitivity analysis. The selection model factorisation of the joint model has already been introduced, but pattern-mixture models and shared-parameter models provide alternative frameworks. We now look at each of these factorisations, using $\mathbf{y} = (\mathbf{y}_{obs}, \mathbf{y}_{mis})$ to denote the *complete data* and $\mathbf{m}$ to denote the missing data indicators, suppressing any dependence on covariates. $(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m})$ and $(\mathbf{y}_{obs}, \mathbf{m})$ are often referred to as the *full data* and *observed data* respectively.

1.5.1 Selection models

Selection models first appeared in the econometrics literature, with Heckman’s work on sample selection bias (see Vella (1998) for review). They were introduced in Section 1.4, and factorise the joint distribution for the complete data and the missingness indicator by specifying a marginal distribution for the complete data and a conditional distribution for the missingness indicator given the complete data, as given by Equation 1.5. In the literature $f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\boldsymbol{\beta})$ is variously referred to as the full-data response model and the measurement model. One of the attractions of the selection model factorisation is that it specifies $f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\boldsymbol{\beta})$ directly, which is the distribution that the analyst is usually interested in, so we will call it the model of interest.

Selection models for informative drop-out in longitudinal data typically have a multivariate normal analysis model for the complete data, i.e. for $f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\boldsymbol{\beta})$, and a probit or logistic regression model for the drop-out process, i.e. $f(\mathbf{m}|\mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta})$, with the parameters estimated using maximum likelihood (Diggle and Kenward, 1994). Molenberghs *et al.* (1997) provide an example of their use with longitudinal ordinal data. There is an assumption that the same model structure is appropriate for individuals with and without missing data, which is untestable from the data. These models can be sensitive to the correct specification of both parts of the joint model, as shown by Kenward (1998).

1.5.2 Pattern Mixture models

By contrast, pattern-mixture models classify individuals by their missingness and allow different model structures for each pattern of missing data, using the factorisation

$$f(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m} | \boldsymbol{\beta}, \boldsymbol{\theta}) = f(\mathbf{y}_{obs}, \mathbf{y}_{mis} | \mathbf{m}, \boldsymbol{\beta}) f(\mathbf{m} | \boldsymbol{\theta}), \quad (1.7: \text{Pattern Mixture Model})$$

shown here with the assumption of distinctness of parameters. The observed data are then a mixture of these patterns. Pattern-mixture models are under-identified and require restrictions or prior information for parameter identification (Little, 1993). Molenberghs *et al.* (2003) discuss various strategies for dealing with this under-identification, suggesting model simplification as an alternative to Little's identifying restrictions.

These two approaches have been reviewed and compared in the context of longitudinal data (Hogan and Laird, 1997; Kenward and Molenberghs, 1999; Michiels *et al.*, 2002; Fitzmaurice, 2003), and both have advantages and disadvantages. Selection models lack robustness to untestable assumptions about the missing data process and often lead to flat likelihoods (Molenberghs *et al.*, 1998), while pattern-mixture models are by construction under-identified. In both types of model, assumptions must be made about the data that have not been observed, but these assumptions are more explicit for pattern-mixture models as the parameters for which the data provide information are clearly distinguished from parameters which are based on no information. As the assumptions required preclude a definitive analysis, authors generally advocate that both types of models are best used as part of a sensitivity analysis.

1.5.3 Shared Parameter models

Adding random coefficients to these models to allow for individual specific latent effects yields random-coefficient selection models and random-coefficient pattern-mixture models (Little, 1995). These models are also known as shared parameter models (Daniels and Hogan, 2008; Molenberghs and Kenward, 2007). Let $\mathbf{b}$ denote a set of random effects and $\boldsymbol{\varphi}$ associated unknown parameters. Then the joint distribution, $f(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}, \mathbf{b} | \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\varphi})$, can be factorised as a selection model,

$$f(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}, \mathbf{b} | \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\varphi}) = f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{b}, \boldsymbol{\theta}) f(\mathbf{y}_{obs}, \mathbf{y}_{mis} | \mathbf{b}, \boldsymbol{\beta}) f(\mathbf{b} | \boldsymbol{\varphi})$$

or a pattern-mixture model

$$f(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}, \mathbf{b} | \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\varphi}) = f(\mathbf{y}_{obs}, \mathbf{y}_{mis} | \mathbf{m}, \mathbf{b}, \boldsymbol{\beta}) f(\mathbf{m} | \mathbf{b}, \boldsymbol{\theta}) f(\mathbf{b} | \boldsymbol{\varphi}).$$

Demirtas and Schafer (2003) review random-coefficient pattern-mixture models and show that they can be unstable and non-robust even in simple examples. Paddock (2007) uses a conditional linear

model, which is a special case of random-coefficient pattern-mixture models, to deal with informative censoring. Pattern-set mixture models combine the ideas behind selection models and pattern-mixture models (Little, 1993; Little and Rubin, 2002).

1.6 Types of Bayesian models

We now review the standard models that are used as our model of interest in this thesis, discussing their Bayesian formulation. We start by considering the simple linear regression model that we use for non-hierarchical data, and then describe more complex models that allow us to take account of the structure of data from longitudinal studies.

1.6.1 Non-hierarchical regression models

The data for our research consists of a univariate outcome y_i and a vector of covariates $x_{1i}, \dots, x_{pi}$ for $i = 1, \dots, n$ individuals. We model this data using a linear regression model assuming normality of the errors, i.e.

$$y_i = \beta_0 + \sum_{k=1}^p \beta_k x_{ki} + \epsilon_i,$$

$$\epsilon_i \sim N(0, \sigma^2),$$

where N denotes a Normal distribution. The equivalent Bayesian formulation for this model is

$$y_i \sim N(\mu_i, \sigma^2),$$

$$\mu_i = \beta_0 + \sum_{k=1}^p \beta_k x_{ki}, \tag{1.8}$$

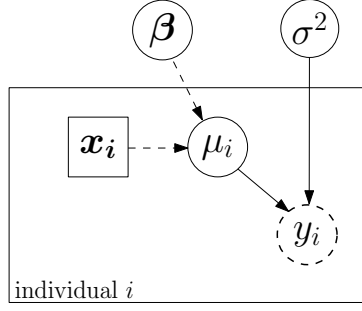
$$\beta_0, \beta_1, \dots, \beta_p, \sigma^2 \sim \text{Prior distribution.}$$

The models in the early chapters are generally specified with non-informative prior distributions, but later we look at incorporating additional information using informative priors.

It is often helpful to show the relationships in Bayesian models in graphical form, and our simple regression model is shown graphically in Figure 1.1. In this diagram, each variable is represented by a circular or square node. Circles denote random variables, and are dashed if they are only partially observed, while squares denote quantities that are regarded as constants. So here the response variable, y , has missing values, but the covariates, $\mathbf{x}$, are fully observed. Full arrows indicate stochastic links to which a probability relationship is attached, while deterministic relationships are denoted by dashed lines.

Equation 1.8 specifies our model of interest, $f(\mathbf{y}|\boldsymbol{\beta}, \sigma)$. In the case shown in Figure 1.1, where the covariates are completely observed and only the response has missing values, we do not need a miss-

Figure 1.1: Graphical representation for a simple regression model



ingness model provided we can assume that the missing data mechanism is ignorable. The missing data are treated as additional unknown parameters, and imputation of $\mathbf{y}_{mis}$ is unnecessary for valid inference about β and σ . However, if we cannot assume that the missing data mechanism is ignorable, then we need to specify a response model of missingness.

The situation is different if our data has missing covariates rather than missing responses. In this case an imputation model for the missing data is required to fully exploit all the available data, regardless of our assumptions about the missingness process.

By using a Bayesian approach, we can build our statistical models in a modular way. In subsequent chapters, we demonstrate the usefulness of this feature by adding sub-models to a model of interest, progressively allowing for different types of missingness in the response variable, missingness in the covariates and incorporating additional information. We also show how this flexibility enables us to adapt our models to carry out sensitivity analysis.

Our main examples involve data drawn from longitudinal studies which consequently have a time dimension. In these cases we have response y_{it} for $i = 1, \dots, n$ individuals measured at $t = 1, \dots, T$ time points, and an associated set of covariates $x_{1it}, \dots, x_{pit}$. To model this data we need to take account of the stochastic dependence between measurements on the same individual. We use two types of models to take account of this correlation: random effects models and autoregressive models.

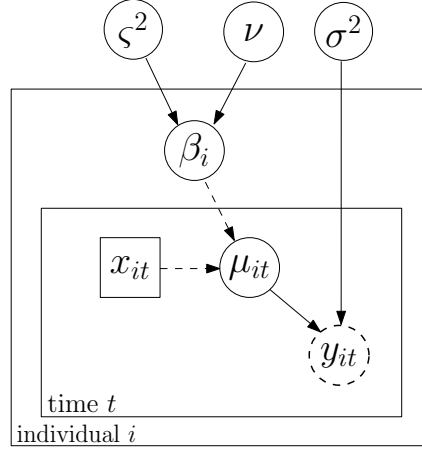
1.6.2 Random effects models

Random effects models for longitudinal data incorporate individual specific parameters, which ‘borrow strength’ from the corresponding parameters associated with the other individuals. For example, if we have a single covariate and no intercept, a random effects model is given by

$$\begin{aligned}
 y_{it} &\sim N(\mu_{it}, \sigma^2), \\
 \mu_{it} &= \beta_i x_{it}, \\
 \beta_i &\sim N(\nu, \varsigma^2), \\
 \nu, \varsigma^2, \sigma^2 &\sim \text{Prior distribution},
 \end{aligned} \tag{1.9}$$

and shown graphically in Figure 1.2, again assuming that there is missingness only in the response. By assuming that the β_i come from a common prior distribution, they are no longer modelled as independent, but as *exchangeable*. The common prior distribution has unknown parameters ν and ς , called *hyperparameters*. Such models are also known as hierarchical or multilevel models. For a fuller discussion see, for example, Gelman *et al.* (2004).

Figure 1.2: Graphical representation for a random effects model



1.6.3 Autoregressive models

Alternatively, the dependence between responses for the same individual can be modelled as a function of the time separation between the measurements. In autoregressive models, y_{it} depends on previous observations in addition to the covariates.

Consider a set of time-ordered responses for a single individual, $y_1, \dots, y_T$. Then an autoregressive Gaussian model without covariates is defined using a time lag q , and a set of autoregression coefficients $\alpha_1 \dots \alpha_q$, s.t.

$$y_t | y_{t-1}, y_{t-2}, \dots, y_{t-q} \sim N(\mu_t, \sigma^2),$$

$$\mu_t = \sum_{j=1}^q \alpha_j y_{t-j}, \quad t = q+1, \dots, T.$$

There are two possible ways of incorporating explanatory variables, depending on whether we assume that the mean responses or the error terms are serially correlated. Consider the case of an autoregressive model of order 1, with a single explanatory variable. Then we can set up the model with serial correlation about the mean response as $y_t | y_{t-1} \sim N(\beta x_t + \alpha y_{t-1}, \sigma^2)$, or with serially correlated errors as $y_t | y_{t-1} \sim N(\beta' x_t + \alpha'(y_{t-1} - \beta' x_{t-1}), \sigma'^2)$. Although both methods are reasonable, they are not equivalent (Diggle *et al.*, 2002).

In our second real example, Section 5.2, we use an autoregressive model of order 2 with explanatory

variables, implemented as

$$\begin{aligned}
y_{it} &\sim N(\mu_{it}, \sigma^2), \\
\mu_{i1} &= \beta_0 + \sum_{k=1}^p \beta_k x_{ki1}, \\
\mu_{i2} &= \beta_0 + \sum_{k=1}^p \beta_k x_{ki2} + \alpha_1 \varepsilon_{i1}, \\
\mu_{it} &= \beta_0 + \sum_{k=1}^p \beta_k x_{kit} + \alpha_1 \varepsilon_{i,t-1} + \alpha_2 \varepsilon_{i,t-2}, \quad t = 3, \dots, T \\
\varepsilon_{it} &= y_{it} - \beta_0 - \sum_{k=1}^p \beta_k x_{kit}, \\
\beta_0, \beta_1, \dots, \beta_p, \alpha_1, \alpha_2, \sigma^2 &\sim \text{Prior distribution.}
\end{aligned} \tag{1.10}$$

In this form, it is the ε_{it} that are serially correlated.

1.7 Scope and aims of this research

Given the enormous breadth of the topic of modelling missing data, we restrict our research to investigating the performance of Bayesian full probability modelling for analysing data with missing variables generated by a missingness mechanism that is thought to be informative. Our approach is to build joint models using the selection model factorisation described in Section 1.5.1. Comparisons with other factorisations are beyond the scope of this thesis.

Our particular focus is missing response. However, the majority of longitudinal datasets suffer from missing covariates as well as missing responses. Consequently, we also consider how covariate missingness can be accommodated in our joint models. One of the advantages of a Bayesian approach is that it provides an ideal framework for incorporating additional data from external sources or expert knowledge. This issue is also explored as part of this research.

The aims of this project can be summarised as:

- to carry out an in-depth investigation into the use of Bayesian full probability modelling for informative missing data, including computational issues;
- to investigate the elicitation of expert knowledge and its incorporation into Bayesian models for missing data through informative priors;
- to develop a general strategy for building joint models to analyse data from longitudinal studies with non-random missingness, including advice on sensitivity analyses and model fit diagnostics.

Chapter 2

Literature review

Before proceeding we review the literature on missing data methods, starting with brief overviews of ad hoc methods and ‘statistically principled’ methods, which include multiple imputation and Bayesian full probability modelling. We then look at multiple imputation in more detail, before reviewing Bayesian full probability modelling.

2.1 Overview of ad hoc methods

Ad hoc methods usually create a single ‘complete’ data set, which is then analysed using standard complete-data methods. A common method is complete-case analysis, in which individuals whose information is incomplete are discarded. Although this method has the advantage of simplicity, it is generally inappropriate as it leads to loss of precision and, unless the missing-data mechanism is MCAR, to bias (see Little and Rubin (2002), Chapter 3). Many statistical software packages take this approach, automatically omitting any individual that has a missing value for any variable.

There are also a variety of single imputation methods, where the missing values are simply replaced using either explicit modelling methods, for example, mean or regression imputation, or by implicit methods such as hot deck imputation. For univariate missingness, regression imputation computes a regression of the incomplete variable on the fully observed variables using the complete cases, and then substitutes the missing values with the predictions from this regression. For hot deck imputation, substitute values are drawn from “similar” observed individuals. Further detail on these methods is provided by Little and Rubin (2002), Chapter 4. With longitudinal data, an example of a simple imputation method is to carry the last observation forward. Single imputation methods can be dangerous as they fail to properly reflect statistical uncertainty and are also likely to introduce bias.

Although practitioners frequently resort to these ad hoc methods, discarding information to make the estimation problem more tractable, such methods are not recommended (Schafer and Graham, 2002).

2.2 Overview of ‘statistically principled’ methods

In contrast to ad-hoc methods, so-called ‘statistically principled’ or ‘model-based’ methods combine the available information in the observed data with explicit assumptions about the missing value mechanism, accounting for the uncertainty introduced by the missing data. These include maximum likelihood methods which are typically implemented by the EM algorithm, weighting methods, multiple imputation and Bayesian full probability modelling.

The EM algorithm is an iterative computation of maximum-likelihood estimates in which each iteration involves two steps, the expectation step in which missing values are replaced by their expectations and the maximisation step in which the parameters are estimated given the current estimates of the missing values (Dempster *et al.*, 1977; Gelman *et al.*, 2004). By including the missing-data model in the likelihood, the EM algorithm can be applied when missing data is non-ignorable.

An alternative, inverse probability weighting, uses weights based on the probability of response to reduce bias in complete-case analysis in some circumstances (Schafer and Graham, 2002; Molenberghs and Kenward, 2007). There has been a resurgence of interest in weighting in recent years, with new methods proposed for parametric and semiparametric regression (Horton and Laird, 1999; Rotnitzky and Robins, 1997). Jones *et al.* (2006) use inverse probability weighted estimators in analysing longitudinal data.

Multiple imputation has two distinct phases, the imputation of the missing data and the analysis of the model of interest. In Bayesian full probability modelling, a joint model is specified for both the relationship of interest and the missing data mechanism, and the joint posterior distribution is estimated simultaneously for both unknown parameters and missing data. We will now discuss multiple imputation, which has much in common with Bayesian full probability modelling, and Bayesian full probability modelling in greater detail.

2.3 Multiple Imputation

Multiple imputation was originally developed by Rubin, to handle the problem of missing data in public-use databases from sample surveys and censuses, where the database constructor and the data analysts are different individuals (Rubin, 1996). When the data collector has access to additional information about non-respondents, compared to the analyst, perhaps because of confidentiality constraints, imputation has obvious advantages. However, this does raise issues about the implications of differences in the imputer’s and analyst’s models, and Schafer (1999) points out that the imputation model should preserve those distributional features that will be the subject of future analyses.

There are three steps for multiple imputation: (1) create $K > 1$ complete datasets, (2) analyse each of these K datasets using complete-case methods and (3) combine the results from the K analyses using the rules developed by Rubin. The validity of multiple imputation hinges on how the imputations are generated, and the specification of the imputation model is the most complex step. This involves two modelling choices: the form of the model and the set of predictors that enter the model.

In general, while using all available information gives multiple imputations with minimal bias, including all the variables from a large dataset is infeasible (due to multicollinearity and computational problems) and unnecessary. Rubin (1996) points out that it is better to err on the side of caution and include predictors rather than leave them out. van Buuren *et al.* (1999) provide a practical strategy for choosing a useful set of imputation predictors from a large set of variables, recommending that these include the variables that appear in the complete data model, variables known to be associated with missingness, and variables that explain a considerable amount of variance in the variables to be imputed. Similar suggestions are made by King *et al.* (2001), who also warn of the dangers of ignoring non-linear or interaction terms: for example, if the square of an explanatory variable is included in the model of interest then this squared term should also be included in the imputation model. The missing data model should reflect any complexity in the analysis model (Carpenter and Goldstein, 2004).

Collins *et al.* (2001) compare strategies for the inclusion of *auxiliary variables*, i.e. those variables which are included in an analysis solely to improve the performance of the missing data method, under different MAR missing data mechanisms. Their simulation shows that an *inclusive* strategy (which makes liberal use of auxiliary variables) is preferable to a *restrictive* strategy (which makes minimal use of such variables), as there are few risks and potentially substantial gains. They also point out that the pattern of missingness affects their results.

Kenward and Carpenter (2007) provide an up-to-date overview of multiple imputation, pointing out its advantages of flexibility, generality and transparency. In contrast to earlier recommendations of 5-10 imputations, they suggest that 100-200 are sometimes required. There are two main approaches to multiple imputation: (1) methods based on a joint multivariate normal distribution and (2) methods that iterate a series of conditional univariate models. The first approach provides a theoretically sound method for generating multiple imputations assuming an underlying multivariate normal distribution, whereas the conditional models in the second approach do not necessarily correspond to a genuine joint distribution (Gelman and Raghunathan, 2001).

A variety of imputation methods have been implemented, and their appropriateness depends on the type of variables (continuous or discrete) and pattern of missingness (monotone or arbitrary). Allison (2000) provides a cautionary tale about what can go wrong, describing how estimates of regression coefficients can be badly biased if multiple imputations for missing predictor variables are produced

by an imputation model generated from an algorithm which does not include all the variables in the analysis model. A review of a number of software packages is provided by Horton and Lipsitz (2001), and updated by Horton and Kleinman (2007). Faris *et al.* (2002) compare three multiple imputation methods: Schafer’s *norm*, Harrell’s *transcan* and van Buuren’s *MICE* and find that they produce similar results for predicting clinical outcomes (1-year mortality) for cardiac care patients. Norm, which is described in detail in Schafer (1997), is an example of the first approach (1), while MICE (van Buuren and Oudshoorn, 1999) follows the second approach (2).

Multivariate Imputation by Chained Equations (MICE), i.e. approach (2), specifies separate conditional distributions for the missing data in each incomplete variable, and a variety of imputation functions are provided (van Buuren and Oudshoorn, 2005). However, because it is not always certain that an underlying multivariate distribution exists, there is an issue of incompatible conditionals (termed ‘incompatible Gibbs’ by Rubin). van Buuren *et al.* (2006) investigate the practical consequences of this, and conclude that despite the theoretical weaknesses, the performance can be quite good; but it is still an open research problem. Oudshoorn *et al.* (1999) investigate the use of MICE in a large national public use dataset and conclude that it is a suitable algorithm because it allows the incorporation of different models, providing a flexible way of implementing the complex imputation models required for handling missing data in complex surveys.

The literature contains many examples of the use of multiple imputation in different settings, e.g. Barnard and Meng (1999) in medical studies and Rassler (2003) in data fusion. Multiple imputation has been compared with other methods, for example, Schafer (2003) provides a comparison with maximum likelihood methods and Carpenter *et al.* (2006) compare it with inverse probability weighting. Rubin (2003) describes two newer ideas for multiple imputation: partially incompatible MCMC and nested multiple imputation. With partially incompatible MCMC, distributionally incompatible MCMC are only used to impute the missing values that destroy a monotone pattern. In nested multiple imputation, the missing data are split into computationally expensive and computationally inexpensive parts, and several imputations of the inexpensive part are created for each imputation of the expensive part.

Most current techniques for creating multiple imputations assume that the non-response mechanism is ignorable, however the multiple imputation paradigm does not require that non-response is ignorable. Imputations, in principle, may be created under any kind of model for the missing-data mechanism (Schafer, 1999). Schafer (2003) discusses generating imputations under selection models and pattern-mixture models. Durrant and Skinner (2006) provide an example of the use of data augmentation for non-ignorable missing data and Longford *et al.* (2006) also discuss departures from MAR. Carpenter *et al.* (2007) propose weighting imputations obtained under MAR to get an overall MNAR estimate, where the weights depend on the assumed degree of departure from MAR, and are calculated using importance sampling.

2.4 Bayesian full probability modelling

2.4.1 Formulation of models

In contrast to multiple imputation, where the imputation of the missing data and the analysis of the model of interest are performed separately, Bayesian full probability modelling performs these two stages simultaneously. To see how suitable Bayesian models are formulated, we will consider the case where our model of interest is a simple linear regression with a univariate Normal outcome y_i and a vector of covariates $x_{1i}, \dots, x_{pi}$, for $i = 1, \dots, n$ individuals, as specified by Equation 1.8 and discussed in Section 1.6.1. We start by considering the case where the covariates are fully observed and only the response has missing values, as shown in Figure 1.1.

If we assume that the missing data mechanism is ignorable, we do not need a missingness model as imputation of y_{mis} is unnecessary for valid inference about β and σ . Values for y_{mis} can be generated from the posterior predictive distribution $f(y_{mis} | y_{obs}, \beta, \sigma)$ if required, but this will not alter the posterior inference about β or σ . This is reflected in Figure 1.1, which only shows the model of interest. However, we could redraw this diagram to include a model of missingness, typically

$$\begin{aligned} m_i &\sim \text{Bernoulli}(p_i), \\ \text{link}(p_i) &= f(y_i, \theta), \\ \theta &\sim \text{Prior distribution.} \end{aligned} \tag{2.1}$$

where m_i is a binary missing value indicator for y_i , θ are the parameters of missingness and ‘link’ would typically be taken to be the logit or probit function. A representation of the joint model is shown in Figure 2.1, with y drawn as a double node, such that the observed values, y_{obs} , and missing values, y_{mis} , are shown separately. It is clear from this diagram that the probability of missingness is dependent on the observed values but not the missing values, as required for ignorable missingness. Note that a different ignorable missingness diagram could be drawn in which the probability of missingness is also dependent on the covariates, by adding an arrow from $\mathbf{x}$ to p . For ignorable missingness, the model of interest and model of missingness can be fitted separately.

A major advantage of Bayesian modelling is that these models can be adapted relatively easily, to allow for the possibility that the missing data mechanism is non-ignorable. To do this, we need to add a link from the missing values of y to the probability of missingness, as shown in Figure 2.2. The two parts of the model must now be fitted together. If we collapse the double node back to a single node representing both observed and missing values of y , we obtain Figure 2.3. This will be our representation from now on, but it should be interpreted as Figure 2.2. Carpenter *et al.* (2002) use clinical trial data to illustrate how a drop-out at random model can be adapted to become a non-random drop-out model.

Figure 2.1: Graphical representation for ignorable missingness in the response variable

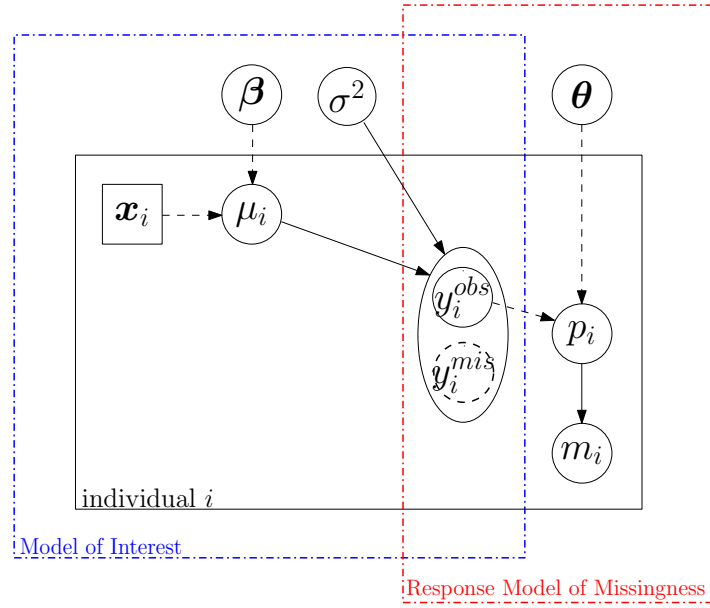
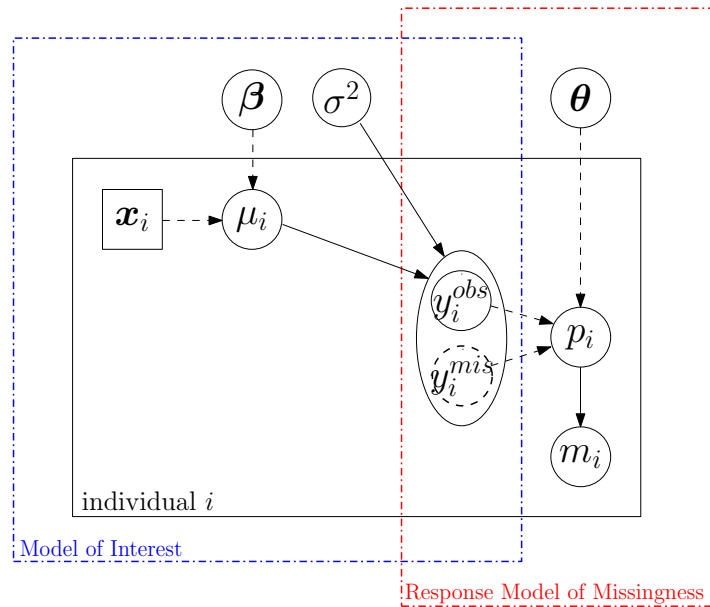


Figure 2.2: Graphical representation for non-ignorable missingness in the response variable

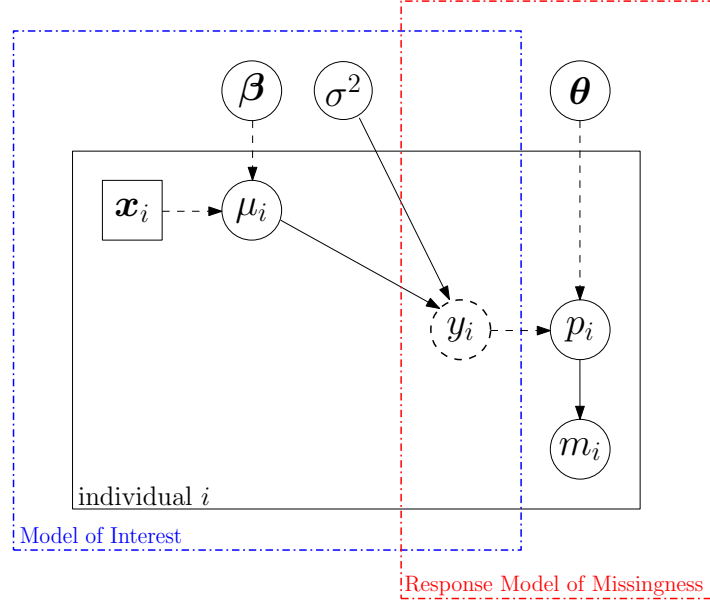


If the covariates have missing values then a covariate model of missingness is also required (this is the subject of Chapter 6). As we will see, the joint model can become quite complicated when there are missing values in both the response and the covariate, and assumptions of non-ignorable missingness are made for some or all of these.

A further advantage of Bayesian modelling is that it provides an ideal framework for incorporating data from external sources and informative prior information (Best *et al.*, 1996; White *et al.*, 2004). These issues are explored in Chapter 7.

Bayesian methods have been used to discuss missing data in different contexts, as shown by the

Figure 2.3: Simplified graphical representation for non-ignorable missingness in the response variable



following examples. Pettitt *et al.* (2006) fit Bayesian hierarchical models to categorical longitudinal data, with missing response and covariates, assuming MAR. Clayton *et al.* (1998) have evaluated Bayesian methods for using auxiliary information in the analysis of incidence data arising from a longitudinal study, where data is missing by design. Kadane (1993) provides an early example of how to conduct a Bayesian analysis of survey data with non-response, which explores sensitivity to different beliefs about the missing data. Bayesian methods have been used by Forster and Smith (1998) in modelling non-ignorable non-response for categorical survey data and by Huang *et al.* (2005) in estimating the parameters of generalised linear models with non-ignorably missing covariates. Wood *et al.* (2006) compare a Bayesian approach for using the number of failed contact attempts to adjust for non-ignorable non-response with a modified conditional likelihood method and an EM procedure.

2.4.2 Implementation

The posterior distribution is the basis of all Bayesian inference, but for complex models the required integrations are often not possible analytically. Markov chain Monte Carlo (MCMC) methods, in which sampling from the joint posterior distribution is performed, provide a widely used alternative means of implementing Bayesian models (Gilks *et al.*, 1996; Gelman *et al.*, 2004). WinBUGS software allows Bayesian analysis of complex statistical models using these techniques (Spiegelhalter *et al.*, 2003).

Sound inference based on MCMC output requires that the Markov chain being simulated has achieved a steady state or *converged*. Numerous techniques and diagnostics for assessing convergence have been developed (Cowles and Carlin, 1996; Brooks and Roberts, 1998), and the use of a variety of

diagnostic tools is generally advocated. Typically these are based on running multiple chains with different starting points. This is the approach we take, checking the Gelman-Rubin convergence statistic (Brooks and Gelman, 1998) for individual parameters and inspecting their trace plots.

Inference based on serially correlated MCMC draws is generally less precise than from the same number of independent draws (Gelman *et al.*, 2004). The suitability of the posterior sample size can be assessed by examining the effective sample size for estimating the posterior means for a range of parameters. The effective sample size, which adjusts actual MCMC sample size for autocorrelation, can be calculated using the “effectiveSize” function in the coda package written for the R software (R Development Core Team, 2007; Plummer *et al.*, 2007). In choosing the run length, we also take into account running time and the number of parameters for which we store the complete posterior samples, as opposed to summary statistics. On occasion this results in compromise, because of the large number of runs we perform for our missing data investigations. However, we aim for effective sample sizes in the high hundreds and preferably thousands, and consider this adequate given the exploratory nature of this work.

2.4.3 Model checking

A crucial part of any statistical analysis is assessing the fit of the model (Gelman *et al.*, 2004; O’Hagan, 2003). In Bayesian modelling this includes checking for any sensitivity to the prior or conflict between the prior and the data, in addition to more ‘standard’ checks regarding, for example, residuals and predictions. The plausibility of the posterior distribution of the model parameters and predictions can be checked using substantive knowledge that has not been formally included in the model. For more formal checking, ideally we would use different data for model calibration and checking, but this is not always possible.

An alternative is to use cross-validation. In its most common form, ‘leave-one-out’ cross-validation, each individual (unit) is left out of the model fitting in turn, and its predictive distribution compared with its observed value(s). However, these methods are often too computationally intensive to be practical for large datasets. An approximation to cross-validation is a technique known as posterior predictive model checking, in which the model is fitted using all the data and posterior predictive draws are compared to the data that have actually occurred. If the model fit is reasonable, then the predictions generated by the model should look similar to the observed data.

The basic technique is to draw simulated values from the posterior predictive distribution, and compare these to the observed data, either graphically or using *discrepancy measures*. The posterior predictive distribution is given by

$$f(\mathbf{y}^{pred}|\mathbf{y}) = \int f(\mathbf{y}^{pred}|\boldsymbol{\beta})f(\boldsymbol{\beta}|\mathbf{y})\mathrm{d}\boldsymbol{\beta}. \quad (2.2)$$

For graphical model checking the observed data and simulated data are displayed side by side. Various types of graphical display are possible including a display of all the data, data summaries and residuals (Gelman, 2004; Gelman *et al.*, 2005; Kerman *et al.*, 2007; Abayomi *et al.*, 2008).

A *discrepancy measure* or *test statistic*, $T(\mathbf{y}, \boldsymbol{\theta})$, is a scalar summary of parameters ($\boldsymbol{\theta}$) and data ($\mathbf{y}$) that is used as a standard when comparing data to predictive simulations. Lack of fit can be measured using a Bayesian p -value, i.e. the probability that the predicted data or discrepancy measure could be more extreme than the observed data or discrepancy measure, as described by Gelman *et al.* (2004), Chapter 6.

In standard posterior predictive checks since the data has influenced the parameter estimation, the data is used twice, for fitting and checking, resulting in conservatism. Alternative mixed predictive checks which are less conservative have been proposed by Marshall and Spiegelhalter (2003) for hierarchical models.

2.4.4 Definition of DIC

Likelihood

The *likelihood*, $L(\boldsymbol{\phi}|\mathbf{y})$, summarises the information in the data $\mathbf{y}$ about parameters $\boldsymbol{\phi}$, and is defined as any function proportional in $\boldsymbol{\phi}$ to the joint density evaluated at the observed sample, i.e.

$$L(\boldsymbol{\phi}|\mathbf{y}) \propto f(y_1, \dots, y_n|\boldsymbol{\phi}) = \prod_{i=1}^N f(y_i|\boldsymbol{\phi}). \quad (2.3)$$

For complex models it is not always obvious what should be considered as forming the likelihood and what as forming the prior. For example, for Bayesian hierarchical models, we can define the prior and likelihood in different ways depending on the quantities of interest. The chosen separation of the joint density into prior and likelihood determine what Spiegelhalter *et al.* (2002) refer to as the *focus* of the model.

For example, if we are modelling repeated observations for an individual, y_{it} , we may be interested in subject specific means or in the population mean, and this makes a difference to the focus. Suppose the y_{it} are modelled according to a modified version of Equation 1.9 with no covariates but a random intercept, i.e.

$$\begin{aligned} y_{it} &\sim N(\mu_i, 1), \\ \mu_i &\sim N(\nu, 1), \\ \nu &\sim f(\nu). \end{aligned}$$

Then for estimation of the mean for individual i , μ_i , we may consider the prior to be $f(\mu_i) = \int f(\mu_i|\nu)f(\nu) d\nu$ and the likelihood to be $f(\mathbf{y}_i|\mu_i)$. Alternatively, for estimation of the population

mean, ν , a natural formulation is to consider the prior as $f(\nu)$ and the likelihood as $f(\mathbf{y}|\nu) = \int f(\mathbf{y}|\boldsymbol{\mu})f(\boldsymbol{\mu}|\nu) d\boldsymbol{\mu}$.

With missing data, there is also the question of what is to be included, just the observed data or the missing data as well.

Deviance

Deviance is a measure of overall fit of a model, and is defined here as -2 times the log-likelihood,

$$D(\boldsymbol{\phi}) = -2\log L(\boldsymbol{\phi}|\mathbf{y}). \quad (2.4)$$

Larger values of the deviance indicate poorer fit. In Bayesian statistics deviance can be summarised in different ways, with the posterior mean of the deviance, $\overline{D(\boldsymbol{\phi})} = E\{D(\boldsymbol{\phi})|\mathbf{y}\}$, suggested as a sensible Bayesian measure of fit (Dempster, 1973) (reprinted as Dempster (1997)), though this is not penalised for model complexity. Alternatively, the deviance can be calculated using a point estimate such as the posterior means for $\boldsymbol{\phi}$, $D(\bar{\boldsymbol{\phi}}) = D\{E(\boldsymbol{\phi}|\mathbf{y})\}$.

In general we use the notation $E(h(\boldsymbol{\phi})|\mathbf{y})$ to denote the expectation of $h(\boldsymbol{\phi})$ with respect to the posterior distribution of $\boldsymbol{\phi}|\mathbf{y}$. However, in more complex formula, we will occasionally use the alternative notation, $E_{\boldsymbol{\phi}|\mathbf{y}}(h(\boldsymbol{\phi}))$.

p_D

Spiegelhalter *et al.* (2002) propose that the difference between these two measures, $p_D = \overline{D(\boldsymbol{\phi})} - D(\bar{\boldsymbol{\phi}})$, is an estimate of the ‘effective number of parameters’ in the model, i.e. p_D is given by

$$p_D = E\{D(\boldsymbol{\phi})|\mathbf{y}\} - D\{E(\boldsymbol{\phi}|\mathbf{y})\}. \quad (2.5)$$

Spiegelhalter *et al.* (2002) point out that p_D can be thought of as a measure of the ratio of the information in the likelihood about the parameters to the information in the posterior (likelihood plus prior). For example, for a normal linear model $y_i \sim N(\beta_0 + \beta_1 x_i, 1)$, where β_0 and β_1 have uninformative prior distributions all the information will come from the data and $p_D \approx 2$, i.e. the true number of parameters in the model.

DIC

Further, Spiegelhalter *et al.* (2002) also propose that adding p_D to the posterior mean deviance gives a measure of fit that is penalised for complexity,

$$\text{DIC} = \overline{D(\boldsymbol{\phi})} + p_D. \quad (2.6)$$

This Bayesian model comparison criterion which is called a *Deviance Information Criterion*, DIC, can be used to compare models in a similar way to the Akaike Information Criterion (for non-hierarchical

models with vague priors on all parameters, $DIC \approx AIC$), with the model taking the smallest value of DIC being preferred. DIC can also be written as a function of the log likelihood, i.e.

$$DIC = 2\log L\{E(\phi|\mathbf{y})|\mathbf{y}\} - 4E_{\phi|\mathbf{y}}\{\log L(\phi|\mathbf{y})\}. \quad (2.7)$$

DIC can often be calculated automatically by WinBUGS, taking $\overline{D(\phi)}$ to be the posterior mean of $-2\log L(\phi|\mathbf{y})$, and evaluating $D(\bar{\phi})$ as -2 times the log-likelihood at the posterior mean of the stochastic nodes.

2.4.5 DIC constructions for missing data models

We are interested in DIC constructions for missing data models, and in particular for joint models using a selection model factorisation (Section 1.5.1). For the examples considered in this thesis, such models typically consist of a linear regression model of interest with Normal errors and a linear logit model of missingness, i.e.

$$MoI : f(\mathbf{y}|\beta, \sigma) = N(\mathbf{X}\beta, \sigma^2)$$

$$MoM : f(\mathbf{m}|\mathbf{y}, \theta) = Bernoulli(p) \text{ where } p = h(y, \theta).$$

Sometimes our model of interest will have t rather than Normal errors.

For complete data, DIC can be constructed using a likelihood based on the full data model of interest, i.e. $f(\mathbf{y}|\beta, \sigma)$. With models for non-ignorable missing data, not only do we have the complication of not observing the full data, but we must also take account of the missing data mechanism, $f(\mathbf{m}|\mathbf{y}, \theta)$. Celeux *et al.* (2006) discuss different DIC constructions for missing data models, in the context of mixtures of distributions and random effects models. More pertinently, Daniels and Hogan (2008) (DH), Chapter 8, discuss two different constructions for selection models, which we now review.

DH point out that for missing data models, DIC is based on the fit of the joint model $f(\mathbf{y}, \mathbf{m}|\beta, \sigma, \theta)$ to the observed data $(\mathbf{y}_{obs}, \mathbf{m})$. DIC will indicate how consistent the modelling assumptions are with the observed data, but provide no information about the fit to the missing data.

DIC based on the observed data likelihood (DIC_O)

One construction is based on the observed data likelihood, $L(\beta, \sigma, \theta|\mathbf{y}_{obs}, \mathbf{m})$,

$$DIC_O = 2\log L\{E(\beta, \sigma, \theta|\mathbf{y}_{obs}, \mathbf{m})|\mathbf{y}_{obs}, \mathbf{m}\} - 4E_{\beta, \sigma, \theta|\mathbf{y}_{obs}, \mathbf{m}}\{\log L(\beta, \sigma, \theta|\mathbf{y}_{obs}, \mathbf{m})\}$$

where

$$L(\beta, \sigma, \theta|\mathbf{y}_{obs}, \mathbf{m}) \propto \int f(\mathbf{y}, \mathbf{m}|\beta, \sigma, \theta) d\mathbf{y}_{mis}.$$

For a selection model, recalling Equation 1.5:

$$\begin{aligned}
L(\beta, \sigma, \theta | \mathbf{y}_{obs}, \mathbf{m}) &\propto \int f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \theta) f(\mathbf{y}_{obs}, \mathbf{y}_{mis} | \beta, \sigma) d\mathbf{y}_{mis} \\
&= \int f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \theta) f(\mathbf{y}_{obs} | \beta, \sigma) f(\mathbf{y}_{mis} | \mathbf{y}_{obs}, \beta, \sigma) d\mathbf{y}_{mis} \\
&= f(\mathbf{y}_{obs} | \beta, \sigma) \int f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \theta) f(\mathbf{y}_{mis} | \mathbf{y}_{obs}, \beta, \sigma) d\mathbf{y}_{mis} \\
&= f(\mathbf{y}_{obs} | \beta, \sigma) E_{\mathbf{y}_{mis} | \mathbf{y}_{obs}, \beta, \sigma} \{f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \theta)\}.
\end{aligned}$$

DIC based on the full data likelihood (DIC_F)

An alternative construction is based on the posterior predictive expectation of the full data likelihood, $L(\beta, \sigma, \theta | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m})$. Assuming complete response data, a DIC based on the full data likelihood can be constructed as

$$\begin{aligned}
\text{DIC}_{full}(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}) &= 2\log L\{E(\beta, \sigma, \theta | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}) | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}\} \\
&\quad - 4E_{\beta, \sigma, \theta | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}} \{\log L(\beta, \sigma, \theta | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m})\}.
\end{aligned}$$

DIC_F is obtained by taking the expectation of DIC_{full}($\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}$) with respect to the posterior predictive distribution of $f(\mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{m})$, i.e.

$$\begin{aligned}
\text{DIC}_F &= E_{\mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{m}} \{\text{DIC}_{full}(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m})\} \\
&= 2E_{\mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{m}} \{\log L\{E(\beta, \sigma, \theta | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}) | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}\}\} \\
&\quad - 4E_{\mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{m}} \{E_{\beta, \sigma, \theta | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}} \{\log L(\beta, \sigma, \theta | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m})\}\}.
\end{aligned}$$

DH provide a broad outline of algorithms for calculating DIC_O and DIC_F, which we tailor for the applications considered in this thesis. Details of these can be found in Appendix D. In general, some of the expectations in DIC_O and DIC_F cannot be evaluated directly from the output of a standard MCMC run. For these, either “nested” MCMC is required, or some other simulation method. DH propose the use of reweighting to avoid repeatedly generating samples during the evaluation of these expectations.

Reweighting

Suppose we wish to make inference about a distribution $f^*(\cdot)$ using Monte Carlo integration. However, suppose instead that we actually have available a sample, $z^{(1)}, \dots, z^{(Q)}$, from a different distribution $f(\cdot)$. This sample can be reweighted to make some inference based on $f^*(\cdot)$, using weights of the form

$$w_q = \frac{f^*(z^{(q)})}{f(z^{(q)})}.$$

This is a form of importance sampling, and its success is known to be critically dependent on the variability of the sampling weights (Peruggia, 1997), with greater variability leading to poorer estimates.

For the method to be successful, we require that the two distributions $f(\cdot)$ and $f^*(\cdot)$ are reasonably close, and in particular that $f(\cdot)$ has a heavier tail than $f^*(\cdot)$ (Liu, 2001; Gelman *et al.*, 2004).

Liu (2001), Chapter 2, describes a rule of thumb that uses effective sample size (ESS) as a measure of the efficiency of importance sampling, and partially justifies this using the delta method (Liu, 1996). Effective sample size has also been used to measure the efficiency of sequential imputation (Kong *et al.*, 1994). We will consider ESS in further detail in our applications using DIC_O and DIC_F in Section 4.6.1.

2.4.6 Model uncertainty and sensitivity analysis

Greenland (1996) describes sensitivity analysis as a “quantitative extension of the qualitative speculations”, and the importance of assessing all major sources of uncertainty is emphasised by Greenland (2005) and discussants. Draper (1995) points out that models generally have parametric and structural uncertainty, and that the latter is less routinely acknowledged. When data are incomplete, taking account of structural uncertainty is crucial as the results from the types of models required display great sensitivity to the assumptions made. The uncertainty extends to an inability to distinguish between a MNAR and MAR missingness from the data (Molenberghs and Kenward, 2007). Indeed, the fit of a MNAR model to the observed data can always be reproduced exactly by a MAR counterpoint, but the two models will produce different predictions for the unobserved data and different inferences (Molenberghs *et al.*, 2008).

A model’s fit to the observed data can be assessed, but its fit to the unobserved data given the observed data cannot be assessed. Consequently, there is a need for sensitivity analysis, in which, for example, several statistical models are considered. The degree of stability of inferences across these models provides an indication of how much confidence can be placed in them. Molenberghs and Kenward (2007), Chapter 20, suggest that a sensitivity analysis for a selection model might involve varying the distributional assumptions, or supplementing the analysis with several pattern mixture models which allow the missing data part to be explicitly modified.

DH advocate what they call the *extrapolation factorisation*, where the full data model is factored into identified and non-identified components, i.e.

$$f(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}|\boldsymbol{\omega}) = f(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \mathbf{m}, \boldsymbol{\omega}_E)f(\mathbf{y}_{obs}, \mathbf{m}|\boldsymbol{\omega}_O).$$

Here $\boldsymbol{\omega}_E$ and $\boldsymbol{\omega}_O$ denote the parameters of the extrapolation model, $f(\mathbf{y}_{mis}|\mathbf{y}_{obs}, \mathbf{m}, \boldsymbol{\omega}_E)$, and observed data model, $f(\mathbf{y}_{obs}, \mathbf{m}|\boldsymbol{\omega}_O)$, respectively. They argue that sensitivity analysis is facilitated if models for non-ignorable drop-out have one or more parameters with transparent interpretations which are completely non-identified by the observed data. To formalise this idea, DH define *sensitivity parameters*, and fixing these to particular values identifies the full data model but does not affect the fit of the

model to the observed data. Sensitivity analysis regarding the missing data mechanism can then be carried out by examining inferences across a range of values for these sensitivity parameters. However, fully specified parametric selection models cannot be factored in this way, as all the parameters can be identified using the parametric assumptions in both parts of the model. Using a simple example, DH illustrate how the model of missingness parameters are identified by the observed data and distributional assumptions about the model of interest, and therefore are not sensitivity parameters according to their criteria. By contrast, pattern mixture models are amenable to the extrapolation factorisation.

There are also a range of proposals for sensitivity analysis in which a MAR model is taken as a starting point, and the impact of potential deviations towards a MNAR mechanism are explored. One tool that uses a perturbation scheme around a MAR mechanism is local influence (Verbeke *et al.*, 2001; Jansen *et al.*, 2006; Molenberghs and Kenward, 2007). Although the original idea of local influence was to detect observations with a large impact on conclusions due to their aberrant missingness mechanism, it actually detects observations that have a large impact on the model of interest or model of missingness parameters for a wide variety reasons not necessarily related to the missingness mechanism, for example an unusual mean profile or autocorrelation structure. Local influence, which changes the missingness process for one observation from MAR to MNAR, is different to case deletion, where one observation is removed entirely, and can yield different results (Molenberghs *et al.*, 2008). Troxel *et al.* (2004) propose a screening tool that measures the potential impact of non-ignorability on an analysis, called an *index of sensitivity to non-ignorability* (ISNI), which studies the sensitivity in the neighbourhood of a MAR model and avoids estimating a full non-ignorable model. Ma *et al.* (2005) extend the ISNI to longitudinal studies.

White *et al.* (2008a) quantify the degree of departure from the MAR assumption for binary outcomes by the informative missingness odds ratio (IMOR), which they define as the ratio of the odds of success among individuals with missing outcomes to the odds of success among individuals with observed outcomes. They use the IMOR to set up sensitivity analyses using pattern-mixture models in a meta-analysis context. In a companion paper, White *et al.* (2008b) show that the IMOR can also be used in a selection model parameterisation, as they develop a hierarchical model for meta-analysis which is implemented using WinBUGS software.

Summary: literature review

‘Statistically principled’ methods are generally recommended for modelling missing data, and Bayesian full probability modelling falls in this category. If the missingness mechanism is thought to be non-ignorable, one approach entails building a joint model based on the selection model factorisation, which consists of a model of interest and a model of missingness. Such Bayesian models can be implemented using MCMC methods, and as with any statistical model their fit should be checked. However, only their fit to the observed data and not to the missing data can be assessed, so sensitivity analysis is regarded as crucial for assessing model uncertainty.

DIC is widely used in Bayesian statistics for comparing models. For non-ignorable missing data models, its construction is complicated by the unobserved data and the need to take account of the missing data mechanism. In this case, constructions based on the observed data likelihood and the full data likelihood have been proposed.

Chapter 3

Data

In this chapter we introduce the main datasets that are used throughout this research. These include three real data examples, two from the British birth cohort studies and one from a clinical trial, and some synthetic datasets that we use in our initial methodological exploration.

3.1 British birth cohort studies

Our main sources of data for developing and testing strategies for Bayesian full probability modelling are the three most recent British birth cohort studies, each of which has their own ‘attrition’ characteristics. The earliest, the National Child Development Study (NCDS), originated as the Perinatal Mortality Survey, which was designed to examine the social and obstetric factors associated with still-birth and early infant mortality. It subsequently evolved into a longitudinal study, following all those living in Great Britain who were born in one week in March 1958, and currently has eight sweeps of data available including the original birth survey. The next, the British Cohort Study (BCS70), was designed along similar lines to the NCDS to survey all those living in Great Britain who were born in a particular week in April 1970, and currently has available seven sweeps of data. The BCS70 was originally named the British Births Survey and had a similar initial focus to the NCDS. Both these studies surveyed over 17,000 babies, and their scope has broadened over time to encompass their subjects’ health, education, social development, and various aspects of adult life. The response pattern for BCS70 is rather different from that of NCDS, with generally lower response rates for BCS70 (Plewis *et al.*, 2004). For both these studies, cohort members move in and out of the study over time, allowing the possibility of using data from subsequent as well as previous sweeps in predicting missing data, in contrast to approaches based on the assumption that attrition is an absorbing state.

The Millennium Cohort Study (MCS) was set up to provide information about children living and growing up in each of the four countries of the UK, including sufficient data for various sub-groups, such as those living in disadvantaged circumstances and of ethnic minorities, to be investigated. Data

about the children’s families is provided in addition to data about the children themselves, through interviews and self-completion forms undertaken by a main respondent (usually the cohort member’s mother) and a partner respondent (usually the father). In order to meet these aims, the design of MCS is very different from the other two studies, with over 18,000 cohort members being selected at random following stratification and clustering from children born in the UK between specified dates at the start of the Millennium (Plewis, 2007a). This has led to substantial non-response in the initial sweep (Plewis (2007a) categorise 15% of the issued sample as non-respondents), in contrast to NCDS and BCS70. Two further sweeps of this cohort are now also available. Non-response in the initial sweep and first follow-up (sweep 2) is discussed by Plewis (2007b), Ketende (2008) concentrates on response in sweeps 2 and 3, and Calderwood *et al.* (2008) discuss participation in the first three sweeps.

All three studies are multi-disciplinary, with domains of interest including health, family background and education. The response patterns for the different domains vary (Plewis *et al.*, 2004), with unusually low response for education in sweep 3 for NCDS, when the cohort members were age 16, which was affected by a change of school leaving age. For sweep 3, 87% of the target sample responded, of which 82% answered questions on education. In view of this sizeable missingness, it seems appropriate to use a subset of the educational data from NCDS to motivate our initial research, using models proposed by Goldstein (1979) for investigating the effects of social class on educational attainment as a starting point. Our other major example from the British birth cohort studies uses data taken from MCS to predict income, for which there is evidence of non-ignorable non-response (Plewis *et al.*, 2008). We will now discuss each of these datasets in turn.

3.1.1 NCDS educational test scores data

The relationship between changes in educational attainment and social class was first investigated using NCDS data by Fogelman and Goldstein (1976) for cohort members at ages 7 and 11, and re-examined with the addition of data at age 16 by Fogelman *et al.* (1978). They used scores from educational attainment tests for reading and mathematics given by NCDS to cohort members. Some of these tests were developed specifically for NCDS, while others were originally intended for other purposes. Technical and interpretational problems arising from the original analysis were examined by Goldstein (1979). Goldstein initially fitted a pair of linear equations which regressed 11 year test scores on 7 year test scores, and 16 year test scores on both 7 and 11 year test scores, modelling reading and mathematics separately. These models were then adjusted to allow for measurement error, a major issue in his paper. Further variants incorporated fixed social class as an additive variable, social class changes, and mathematics scores as explanatory variables for the reading scores and vice versa.

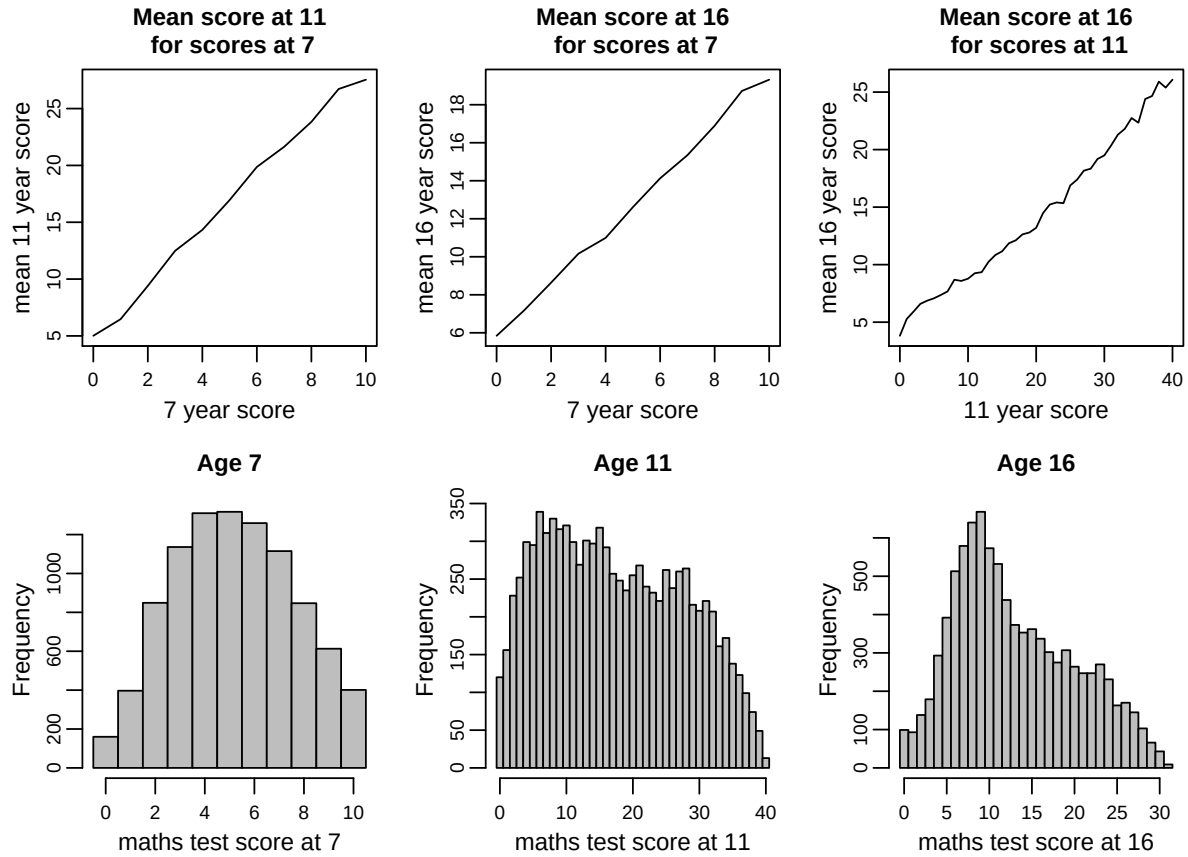
We restrict our attention to the mathematics test scores, using the scores at age 16 as our response variable. It is thought that the missingness in the age 16 test score may be informative, with individuals

less likely to have taken the test if they were likely to perform poorly. The top row of plots in Figure 3.1 shows that the relationships between the mathematics test scores at each age are approximately linear, while the histograms show that while the mathematics test score at age 7 is approximately normally distributed, the distributions of the test scores at 11 and 16 are asymmetric, with lower scores more prevalent than higher scores. We use centred and standardised versions of these test score variables, which is recommended good practice to improve mixing in MCMC estimation. Social class in NCDS is based on the occupation of the male head of the household and has slightly different splits at each sweep. We follow Goldstein in aggregating to three consistent groupings as follows:

- non-manual workers (social classes I, II and III non-manual);
- skilled and semi-skilled manual workers (social classes III manual and IV);
- unskilled manual workers (social class V).

If there is no male head of the household, then social class is set to missing.

Figure 3.1: NCDS mathematics test scores (subset with observed values for all three scores)



In his analysis, Goldstein used complete cases only, resulting in sample sizes decreasing rapidly with increasing model complexity, with just over 9,000 cohort members for the simpler models and approximately 5,200 for the most complex models. Given that the cross sectional target sample at age 16 was 16,901 (Plewis *et al.*, 2004), this implies that a very sizeable proportion of data is missing for some variables, offering plenty of scope for research into the impact of missing data. Table 3.1 shows the number of records available for complete case analysis, for a selection of the models for the

mathematics test scores analysed by Goldstein. The full complexity of the missing data patterns for modelling the test score at age 16 using the three covariates test score at age 7, test score at age 11 and social class at age 11 (2^4 patterns) is revealed by Table 3.2. The source of these four variables is given in Table A.1.

Table 3.1: Complete cases for models of mathematics test score at 16

Covariates	Complete Cases	Percentage*
mathematics test score at 7 and 11	9404	55.6%
+ social class at 11	8195	48.5%
+ social class at 7 and 16	6185	36.6%

* percentage of cross sectional target sample at age 16

Table 3.2: Missingness Pattern for NCDS educational test scores

	number of records ^b	maths7	maths11	sc11 ^d	maths16	number of missing variables ^c
Pattern 1	8195	1	1	1	1	0
Pattern 2	706	0	1	1	1	1
Pattern 3	508	1	0	1	1	1
Pattern 4	2749	1	1	1	0	1
Pattern 5	1209	1	1	0	1	1
Pattern 6	75	0	0	1	1	2
Pattern 7	334	0	1	1	0	2
Pattern 8	256	1	0	1	0	2
Pattern 9	202	0	1	0	1	2
Pattern 10	574	1	0	0	1	2
Pattern 11	576	1	1	0	0	2
Pattern 12	68	0	0	1	0	3
Pattern 13	452	0	0	0	1	3
Pattern 14	158	0	1	0	0	3
Pattern 15	831	1	0	0	0	3
Pattern 16	1665	0	0	0	0	4
MissObs ^a		3660	4429	5667	6637	

^a total missing observations for the variable

^b number of records for the missingness pattern

^c number of missing variables in the missingness pattern

^d social class at age 11

3.1.2 MCS income data

Income data from surveys is widely analysed, and income non-response is of considerable concern. We take income data from the MCS to examine the impact of using different Bayesian models on the conclusions to some typical research questions. Analysis of data from the MCS needs to take account of its design, as it is clustered geographically, and disproportionately stratified (Plewis, 2007a). The

population is stratified by UK country (England, Wales, Scotland and Northern Ireland), with England further stratified into three strata (ethnic minority, disadvantaged and advantaged) and the other three countries into two strata (disadvantaged and advantaged). For each stratum individuals are clustered by electoral ward.

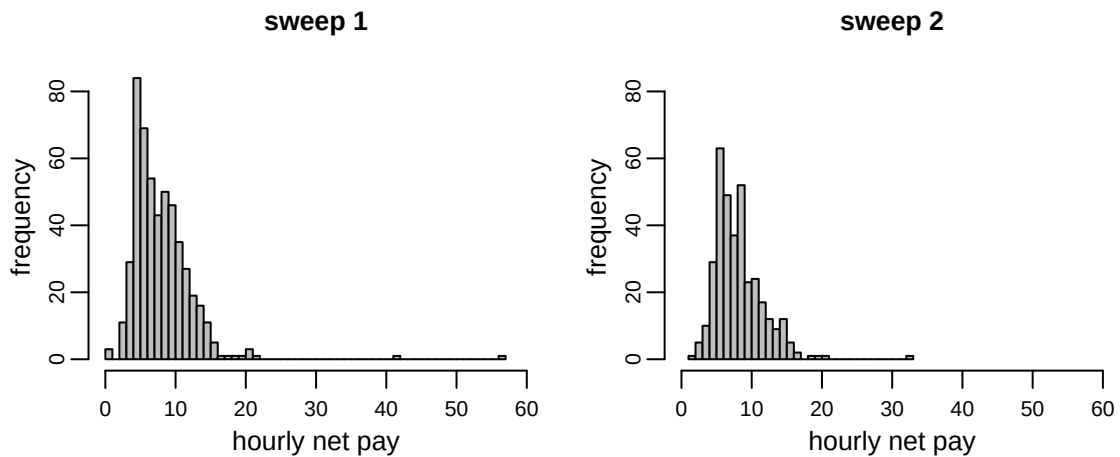
Using data from sweeps 1 and 2, we predict income for the subset of main respondents who are single in sweep 1, in paid work and not self-employed. We will consider a number of motivating questions which include:

- How much extra does an individual earn if they have a degree?
- Does change in partnership status affect income?
- What effect does ethnicity have on an individual's rate of pay?

Our dataset is formed from individuals who are single (3,194 individuals), in paid work (741) and not self-employed in sweep 1 (removes a further 27). We also exclude those who are known to be self-employed or not working in sweep 2, and four records with extreme pay values which look suspicious, leaving 559 individuals. Further details of the data selection process can be found in Table A.2.

By definition we are looking at a set of individuals who are, with two exceptions (see Appendix A), the mothers of very young children, so it is hardly surprising that many are working part-time. To simplify our models, we choose hourly net pay as our response variable. Hourly net pay, *hpay*, is calculated by dividing annual pay by number of hours worked in a year, and further details of this calculation can be found in Appendix A. The distribution of the observed *hpay* can be seen from Figure 3.2.

Figure 3.2: MCS hourly net pay by sweep



Drawing on the existing literature, we select potential covariates with our motivating questions and the structure of the survey in mind. Our dataset also includes variables which may help to explain the missingness (Hawkes and Plewis, 2008). These variables are detailed briefly in Table 3.3, with more extensive descriptions and source given in Table A.3.

Table 3.3: Description of MSC income dataset variables

name	description	details
<i>age</i>	age at interview	continuous ^a - median = 26, range = (15, 48)
<i>hsize</i>	household size	continuous ^a - median = 3, range = (2, 11)
<i>kids</i>	number of children	continuous ^a - median = 1, range = (1, 8)
<i>edu</i>	educational level	6 levels ^b
<i>sc</i>	social class	4 levels (NS-SEC 5 classes with 3 omitted) ^c
<i>eth</i>	ethnic group	2 levels (1 = white; 2 = non-white)
<i>sing</i> ^d	single/partner	2 levels (1 = single; 2 = partner)
<i>reg</i>	region of country	2 levels (1 = London; 2 = other)
<i>ctry</i>	country	4 levels ^e
<i>stratum</i>	country by ward type ^f	9 levels
<i>ward</i> ^g	group or single electoral ward	

^a all the continuous covariates are centred and standardised; the median and ranges are for sweep 1 on the original scale.

^b the level of National Vocational Qualification (NVQ) equivalence of the individual's highest academic or vocational educational qualification

^c social class 3 is small employers and own account workers, and these individuals are excluded by definition

^d always 1 for sweep 1 by dataset definition

^e 1 = England; 2 = Wales; 3 = Scotland; 4 = Northern Ireland

^f three strata for England (advantaged, disadvantaged and ethnic minority); two strata for Wales, Scotland and Northern Ireland (advantaged and disadvantaged)

^g the sample is clustered by ward

Our educational level variable, *edu*, is the level of National Vocational Qualification (NVQ) equivalence of the main respondent's highest academic or vocational educational qualification, and details of these levels can be found in Table A.4. We regard individuals with only other or overseas qualifications as missing. The main respondent's social class uses the National Statistics Socio-Economic Classification (NS-SEC) grouped into 5 categories, but since we have excluded the self-employed from our dataset, there are no individuals in category 3 and *sc* has 4 levels (see Table A.5). Note that *sing* is always 1 in sweep 1 from the definition of our dataset, but is used to indicate whether the individual has acquired a partner by sweep 2.

Ctry, *stratum* and *ward* are fully observed by survey design, and the pattern of missingness for the remaining variables in the dataset is shown for sweeps 1 and 2 in Tables 3.4 and 3.5 respectively. In sweep 1, 8% of individuals have missing *hpay*, a very small number have missing *edu*, *sc* or *eth*, and the remaining variables are completely observed. In sweep 2 missingness is substantially higher, with 32% of individuals having no sweep 2 data due to wave missingness (Pattern 6 in Table 3.5), and a small amount of item missingness, predominantly for *hpay*. The pattern of missingness across both sweeps is shown in Table 3.6. We restrict our analysis of this dataset to modelling the missingness in sweep 2.

Some sweep 2 data was collected from individuals who were originally non-contacts or refusals in sweep 2, after they were re-issued by the field work agency. In our dataset, seven individuals have a

Table 3.4: Missingness Pattern for sweep 1 MCS income data

	number of records ^b	<i>age</i> ₁	<i>hsize</i> ₁	<i>kids</i> ₁	<i>reg</i> ₁	<i>eth</i> ₁	<i>sc</i> ₁	<i>edu</i> ₁	<i>hpay</i> ₁	number of missing variables ^c
Pattern 1	505	1	1	1	1	1	1	1	1	0
Pattern 2	43	1	1	1	1	1	1	1	0	1
Pattern 3	3	1	1	1	1	1	1	0	1	1
Pattern 4	2	1	1	1	1	1	0	1	1	1
Pattern 5	1	1	1	1	1	0	1	1	1	1
Pattern 6	4	1	1	1	1	1	1	0	0	2
Pattern 7	1	1	1	1	1	1	0	0	1	2
MissObs ^a		0	0	0	0	1	3	8	47	

^a total missing observations for the variable^b number of records for the missingness pattern^c number of missing variables in the missingness pattern

Table 3.5: Missingness Pattern for sweep 2 MCS income data

	number of records ^b	<i>age</i> ₂	<i>hsize</i> ₂	<i>kids</i> ₂	<i>eth</i> ₂	<i>reg</i> ₂	<i>sing</i> ₂	<i>edu</i> ₂	<i>sc</i> ₂	<i>hpay</i> ₂	number of missing variables ^c
Pattern 1	343	1	1	1	1	1	1	1	1	1	0
Pattern 2	22	1	1	1	1	1	1	1	1	0	1
Pattern 3	2	1	1	1	1	1	1	0	1	1	1
Pattern 4	9	1	1	1	1	1	1	1	0	1	1
Pattern 5	2	1	1	1	1	1	1	1	0	0	2
Pattern 6	181	0	0	0	0	0	0	0	0	0	9
MissObs ^a		181	181	181	181	181	181	183	192	205	

^a total missing observations for the variable^b number of records for the missingness pattern^c number of missing variables in the missingness pattern

complete set of sweep 2 variables as a result of these re-issues. We will set these data to missing for the purpose of fitting our models, so they can be used for model checking.

3.2 Antidepressant trial

To complement the examples taken from the British birth cohort studies, we also analyse data from a six centre clinical trial comparing three treatments of depression. These data were previously analysed by Diggle and Kenward (1994) (DK) and Yun *et al.* (2007). DK found evidence of informative missingness given their modelling assumptions, and we will examine the evidence provided by Bayesian models.

In this clinical trial, 367 subjects were randomised to one of three treatments and rated on the Hamilton depression score (HAMD) on five weekly visits, the first before treatment, week 0, and the remaining

Table 3.6: Missingness Pattern for sweeps 1 and 2 of the MCS income data

	number of records ^b	oth_1^d	eth_1	sc_1	edu_1	$hpay_1$	oth_2^e	edu_2	sc_2	$hpay_2$	number of missing variables ^c
Pattern 1	319	1	1	1	1	1	1	1	1	1	0
Pattern 2	21	1	1	1	1	0	1	1	1	1	1
Pattern 3	1	1	1	1	0	1	1	1	1	1	1
Pattern 4	2	1	1	0	1	1	1	1	1	1	1
Pattern 5	20	1	1	1	1	1	1	1	1	0	1
Pattern 6	8	1	1	1	1	1	1	1	0	1	1
Pattern 7	2	1	1	1	1	0	1	1	1	0	2
Pattern 8	2	1	1	1	0	1	1	0	1	1	2
Pattern 9	1	1	1	1	1	0	1	1	0	1	2
Pattern 10	1	1	1	1	1	1	1	1	0	0	2
Pattern 11	1	1	1	1	1	0	1	1	0	0	3
Pattern 12	157	1	1	1	1	1	0	0	0	0	4
Pattern 13	18	1	1	1	1	0	0	0	0	0	5
Pattern 14	1	1	0	1	1	1	0	0	0	0	5
Pattern 15	4	1	1	1	0	0	0	0	0	0	6
Pattern 16	1	1	1	0	0	1	0	0	0	0	6
MissObs ^a		0	1	3	8	47	181	183	192	205	

^a total missing observations for the variable(s)^b number of records for the missingness pattern^c number of missing variables and groups of variables in the missingness pattern^d other sweep1 variables - age_1 , $hsize_1$, $kids_1$ and reg_1 ^e other sweep2 variables - age_2 , $hsize_2$, eth_2 , $kids_2$, reg_2 and $sing_2$

four during treatment, weeks 1-4. The HAMD score is the sum of 16 test items and takes values between 0 and 50, where the higher the score the more severe the depression. Our question of interest is whether there are differences in the time trend of the HAMD score between treatments. We see from Table 3.7 that some subjects dropped out of the trial from week 2 onwards, with approximately one third lost by the end of the study.

Table 3.7: Missingness Pattern for Antidepressant Trial Data

	number of records ^b	treatment	centre	wk0	wk1	wk2	wk3	wk4	number of missing variables ^c
Pattern 1	246	1	1	1	1	1	1	1	0
Pattern 2	42	1	1	1	1	1	1	0	1
Pattern 3	27	1	1	1	1	1	0	0	2
Pattern 4	52	1	1	1	1	0	0	0	3
MissObs ^a		0	0	0	0	52	79	121	

^a total missing observations for the variable^b number of records for the missingness pattern^c number of missing variables in the missingness pattern

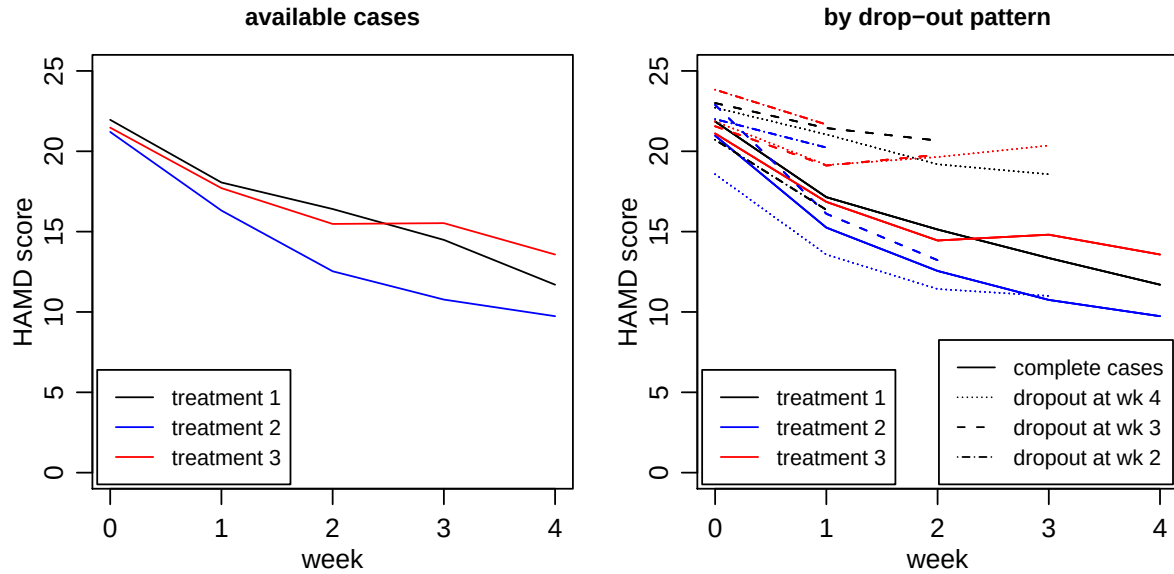
Similar numbers of subjects received each treatment (120, 118 and 129 for treatments 1, 2 and 3 respectively), but the levels and timing of drop-out differ as shown in Table 3.8. In particular, fewer subjects drop out of treatment 3, and although the missingness percentage is similar for treatments 1 and 2 by week 4, more of the drop-out occurs earlier for treatment 2.

Table 3.8: Missingness by treatment and week for Antidepressant Trial Data

	treatment 1	treatment 2	treatment 3	all treatments
% missingness at week 2	11.7	22.0	9.3	14.2
% missingness at week 3	19.2	29.7	16.3	21.5
% missingness at week 4	36.7	35.6	27.1	33.0

Mean response profiles, calculated as a simple average of the observed data, are shown in Figure 3.3. The left plot is based on all the available data and indicates an overall decrease in the HAMD score over time. From the right plot, which distinguishes between the different drop-out patterns, we see that individuals allocated treatments 1 and 3 generally have higher profiles if they dropped out rather than remained in the study, but that this difference is not apparent for treatment 2.

Figure 3.3: Observed mean response profiles for antidepressant trial



3.3 Synthetic data

Datasets where the actual values of the missing data are known are extremely useful for evaluating methods for modelling missing data, so we create some synthetic data scenarios by imposing missingness on complete sets of both simulated and real data.

3.3.1 Multivariate Normal (MVN) data

In our initial investigations we create a number of scenarios using simulated multivariate Normal data. For each scenario, 50 datasets each with 1000 records comprising a response, y , and a single covariate, x , are simulated s.t.

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix} \right). \quad (3.1)$$

For these datasets our model of interest is

$$y_i \sim N(\mu_i, \sigma^2) \quad (3.2)$$

$$\mu_i = \beta_0 + \beta_1 x_i$$

and the true values of the parameters are $\beta_0 = 1$ and $\beta_1 = \alpha$. We then delete some of the responses, y , according to different models of missingness which will be described later, so that the missing values are also known. This simple setup is useful to highlight key characteristics of the performance of our joint models.

3.3.2 Scores data

We also use some synthetic data scenarios based on real data but with simulated missingness. These are constructed from the 10,312 NCDS individuals with observed mathematics test scores at ages 11 and 16. Again our datasets have a response and a single covariate, with the model of interest regressing the test score at age 16 on the test score at age 11. However the size of each dataset varies slightly, as they are created by sampling individuals completely at random with a 0.1 probability of inclusion (each dataset is drawn from the described set of 10,312 cohort members with replacement). Some of the responses are then deleted according to different missingness criteria.

Although these synthetic data are based on educational test scores, such scores are typical of data arising in many settings, including medical and epidemiological studies as well as social science applications. To emphasis this generality and avoid confusion with the real NCDS educational test score example introduced in Section 3.1.1, we will simply refer to these data as *Scores* henceforth.

Summary: data

A combination of synthetic data and real data examples are used in the research described in this thesis. Two types of synthetic data scenarios are used: (1) **MVN data scenarios** which are formed by simulating missingness on simulated full datasets and (2) **Scores data scenarios** which have simulated missingness on fully observed real data taken from NCDS.

Three real examples are analysed, which we refer to as NCDS educational test scores, MCS income and Antidepressant trial. The motivating questions for each of these can be summarised as follows:

NCDS educational test scores: does social class affect educational attainment?

MCS income: do educational attainment, partnership status or ethnicity affect income?

Antidepressant trial: does treatment affect the change in depression score over time?

Chapter 4

Exploring Bayesian full probability modelling using simulated missingness

Having introduced our data, we now investigate how well Bayesian joint models perform when data have missing responses generated by an informative missingness mechanism. To understand the problems caused by this type of missing response, we look at what happens to the parameter estimates and fit of the model of interest when we ignore the missingness and perform a complete-case analysis. We then compare these results with those obtained when a model of missingness is added to the model of interest, and discuss the improvements. Next we consider some of the issues that influence the performance of the joint model, namely the strength of the relationship in the model of interest, the adequacy of the model of missingness and skewness in the data. The chapter is concluded with a discussion of potential diagnostics.

All the models in this chapter have been fitted in WinBUGS (Spiegelhalter *et al.*, 2003), which uses MCMC methods (see Section 2.4.2), and run for 15,000 iterations including 10,000 burn-in, with three chains initialised using diffuse starting values. For each run we have looked at the Gelman-Rubin convergence statistic (Brooks and Gelman, 1998) for the individual parameters, and have assumed convergence if all of these are less than 1.05 and a visual inspection of the trace plots is satisfactory. On this basis, all the runs discussed in this chapter converged unless stated otherwise.

4.1 Specification of joint model for simple case simulations

In this chapter, we restrict ourselves to the case where only the response has missing values, and assume that the missingness mechanism is non-ignorable. We consider the case where the model of interest is a linear regression with a univariate outcome y_i and a vector of covariates $x_{1i}, \dots, x_{pi}$, for

$i=1, \dots, n$ individuals, as specified by Equation 1.8, and the model of missingness has the form

$$\begin{aligned} m_i &\sim \text{Bernoulli}(p_i), \\ \text{logit}(p_i) &= \theta_0 + \theta_1 y_i \end{aligned} \tag{4.1}$$

where m_i is a binary missing value indicator for y_i , unless otherwise stated.

We shall refer to a joint model of this form, consisting of a model of interest and a missingness model, run with missing response values, as JM. For simulated datasets where the missing response values are known, we also run a joint model of the same form but with a full set of response values, which we shall call TRUE. This is used for bench marking simulated datasets, as it gives targets for the fit of both the model of interest and the missingness model. In addition, we run the model of interest (Equation 1.8) on complete cases only, referred to as CC, providing a comparison of the model of interest fit with a commonly used method.

Vague priors are specified for the unknown parameters of the model of interest: the β parameters are assigned $N(0, 10000)$ priors and the precision, $\frac{1}{\sigma^2}$, a $\text{Gamma}(0.001, 0.001)$ prior (WinBUGS parameterises the Normal distribution in terms of the precision, rather than the variance). There is an ongoing debate regarding prior distributions for variance parameters (Gelman, 2005), but this is most pertinent for random effects variances. Following Wakefield (2004) and Jackson *et al.* (2006), we specify a $\text{logistic}(0, 1)$ prior for θ_0 and a weakly informative $N(0, 0.68)$ prior for θ_1 , which corresponds to an approximately flat prior on the scale of p_i .

4.2 Model evaluation

As part of the assessment of our models, we look at the bias and efficiency of the parameter estimates. We define the percentage bias of a parameter estimate as

$$\% \text{ bias} = \frac{(\hat{\beta} - \beta)}{\beta} \times 100 \tag{4.2}$$

where β is the parameter estimate based on the full dataset (modelled by TRUE) and $\hat{\beta}$ is the parameter estimate for some other model, i.e. JM or CC. In each case, the parameter estimates are taken to be the posterior mean of β . Note that this is slightly different to the usual definition of bias as the expectation of the difference between a parameter estimate and its theoretical true value. The efficiency of a parameter estimate is defined to be the width of the 95% interval given by TRUE divided by the 95% interval from some other model (JM or CC). The 95% intervals are calculated from the 2.5 and 97.5 percentiles of the posterior distribution of the parameter.

Additionally, we use mean square error (MSE), i.e.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - E(y_i | \beta))^2, \tag{4.3}$$

as a measure of overall fit for comparing the performance of CC and JM. We calculate this quantity using the posterior means of the β parameters as plug-ins. MSE does not penalise for model complexity.

4.3 Design of synthetic data for simulations

We use the MVN and Scores synthetic data, which were introduced in Section 3.3, to explore the performance of these models. For both we create a number of scenarios by deleting some of the responses according to different missingness criteria. Each scenario comprises a number of replicate datasets.

Three shapes of missingness have been imposed to approximate a range of patterns that may occur in real datasets, namely linear, quadratic and step. The linear rule allows the probability of being missing to increase or decrease at a constant rate, i.e.

$$p_i = \phi_0 + \phi_1 y_i. \quad (4.4)$$

Of course, the values of ϕ_0 and ϕ_1 are chosen to ensure that $0 \leq p_i \leq 1$ for all y_i . Quadratic missingness is formed according to the equation

$$p_i = \phi_0 + \phi_1 y_i + \phi_2 y_i^2 \quad (4.5)$$

and has three variants: u-shaped (the probability of missingness is highest in both tails), n-shaped (the probability of missingness is highest for mid values) and j-shaped (an asymmetric shape, with the probability of missingness highest in one of the tails). Again ϕ_0 , ϕ_1 and ϕ_2 are chosen so that $0 \leq p_i \leq 1$. The step missingness is created using the equation

$$p_i = \begin{cases} 0 & \text{if } y_i < \text{step} \\ 1 & \text{otherwise} \end{cases} \quad (4.6)$$

where *step* is some relevant threshold value. It is simplistic and provides an ‘extreme’ pattern, with all values before the step observed and all values after missing. The potential for bias in the parameters of the model of interest is dependent on the location as well as the percentage of the missingness. In general for a linear model of interest, there is greater sensitivity to data missing from the tails rather than the middle of the distribution (Collins *et al.*, 2001).

Our scenarios differ in the shapes of missingness imposed, and in a number of other aspects that are thought to have an influence on the performance of a joint model. Each scenario is described in detail as we analyse them, and summarised for the MVN and Scores data in Tables 4.1 and 4.2 respectively.

All the MVN synthetic data scenarios have linear shaped missingness, and vary in respect of

- the correlation between the response and covariate, α (Equation 3.1);

- the transform of the response, y ;
- the number of records;
- the number of datasets.

The basic setup, the scenario labelled MVN, is $\alpha = 0.5$, no transform of the response, 1000 records and 50 datasets. From the literature (Daniels and Hogan, 2008), we know that selection models are sensitive to the assumption about the distribution of the errors in the model of interest, and we will investigate this using different transforms of the response. We use the Scores synthetic data scenarios to investigate the different missingness shapes. However, we start with an in depth investigation of linear missingness.

Table 4.1: Summary of MVN synthetic data scenarios

Synthetic Data Scenario	Correlation ^a	Transform of Response	Number of Records	Number of Datasets
MVN	0.5	none	1000	50
MVN(0.1)	0.1	none	1000	50
MVN(0.25)	0.25	none	1000	50
MVN(0.75)	0.75	none	1000	50
MVN(0.9)	0.9	none	1000	50
MVN(sq)	0.5	square	1000	50
MVN(sr)	0.5	square root	1000	50
MVN(log)	0.5	log	1000	50
MVN(n100)	0.5	none	100	50
MVN(n10000)	0.5	none	10000	10

^a the correlation between the response and covariate

Table 4.2: Summary of Scores synthetic data scenarios

Synthetic Data Scenario	Missingness Shape	Transform of Response	Number of Datasets
Scores.linear	linear	none	119
Scores.quad	quadratic	none	18
Scores.step	step	none	6

the number of records in each simulation dataset varies

4.4 Linear missingness

4.4.1 What are the deficiencies of complete-case analysis?

The deficiencies of complete-case analysis are well known, see for example Little and Rubin (2002), Chapter 3. Nevertheless, we begin our investigation by reviewing these using the MVN and Scores.linear scenarios.

MVN scenario

For each dataset in the MVN scenario, we impose three forms of linear missingness on y , using Equation 4.4 with varying values of ϕ , where p_i is the probability of being missing. We shall refer to these three forms of missingness as MCAR, posMNAR and negMNAR, defined as follows.

MCAR: the probability of being missing is set to 0.5 for all values of y , i.e. $\phi_0 = 0.5$, $\phi_1 = 0$.

posMNAR: linear non-ignorable missingness with a steep positive gradient, such that the probability of being missing for the lowest value of y is 0 and the probability of being missing for the highest value of y is 1, i.e. ϕ_1 is positive.

negMNAR: linear non-ignorable missingness with a steep negative gradient, such that the probability of being missing for the lowest and highest values of y are 1 and 0 respectively, i.e. ϕ_1 is negative.

The design of this scenario means that while the gradient of missingness, ϕ_1 , is always zero for MCAR, it varies slightly across the 50 replicates for posMNAR (0.13 to 0.18) and negMNAR (-0.18 to -0.13) as it is dependent on the range of the generated responses. The percentage of missingness has a mean of 50% for all forms of missingness, with a range of 43-58%. A complete-case analysis was performed on each form of missingness for each dataset by running CC using WinBUGS as described in Section 4.1. To get a target fit, TRUE (see Section 4.1) was also run.

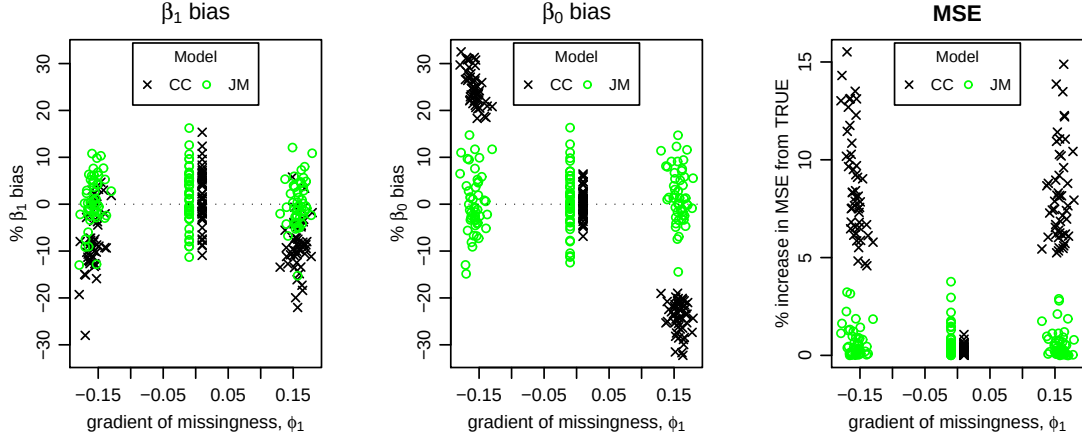
We know that complete-case analysis assumes that the missingness mechanism is MCAR, and therefore expect bias in our parameter estimates for MNAR missingness. The extent and pattern of this bias is shown by the CC points (black crosses) in the β_0 and β_1 bias plots of Figure 4.1. (We will discuss the JM points (open green circles) in Section 4.4.2.) The MCAR estimates for β_0 and β_1 look unbiased. By contrast, on average, CC under-estimates the slope of the model of interest, β_1 , for posMNAR and negMNAR by similar amounts. When high responses are more likely to be lost (posMNAR) β_0 is always substantially under-estimated, while when low responses are more likely to be lost (negMNAR) CC always substantially over-estimates β_0 . The % bias for β_0 is about three times the % bias for β_1 .

As expected, the deterioration in the overall fit compared to TRUE, as measured by MSE, is slight for MCAR, but has a mean of 8-9% for the two forms of MNAR missingness (Figure 4.1).

Scores.linear scenario

Having seen how MNAR missingness biases the intercept of the model of interest either up or down, depending on whether low or high responses are more likely to be lost, but always biases the slope downwards, we carried out a further simulation using the Scores data (Scores.linear scenario). We simulate the missingness, with varying gradients and proportions of missingness by specifying 11 ‘start’ probabilities of missingness, from 0 to 1 at 0.1 intervals, and 11 ‘end’ probabilities of missingness,

Figure 4.1: MVN: performance of CC and JM compared with the TRUE generated targets



The points grouped on the left of each graph correspond to the 50 negMNAR runs ($-0.18 \leq \phi_1 \leq -0.13$), the points in the middle correspond to the 50 MCAR runs ($\phi_1=0$, but CC and JM offset for clarity) and the points grouped on the right correspond to the 50 posMNAR runs ($0.13 \leq \phi_1 \leq 0.18$).

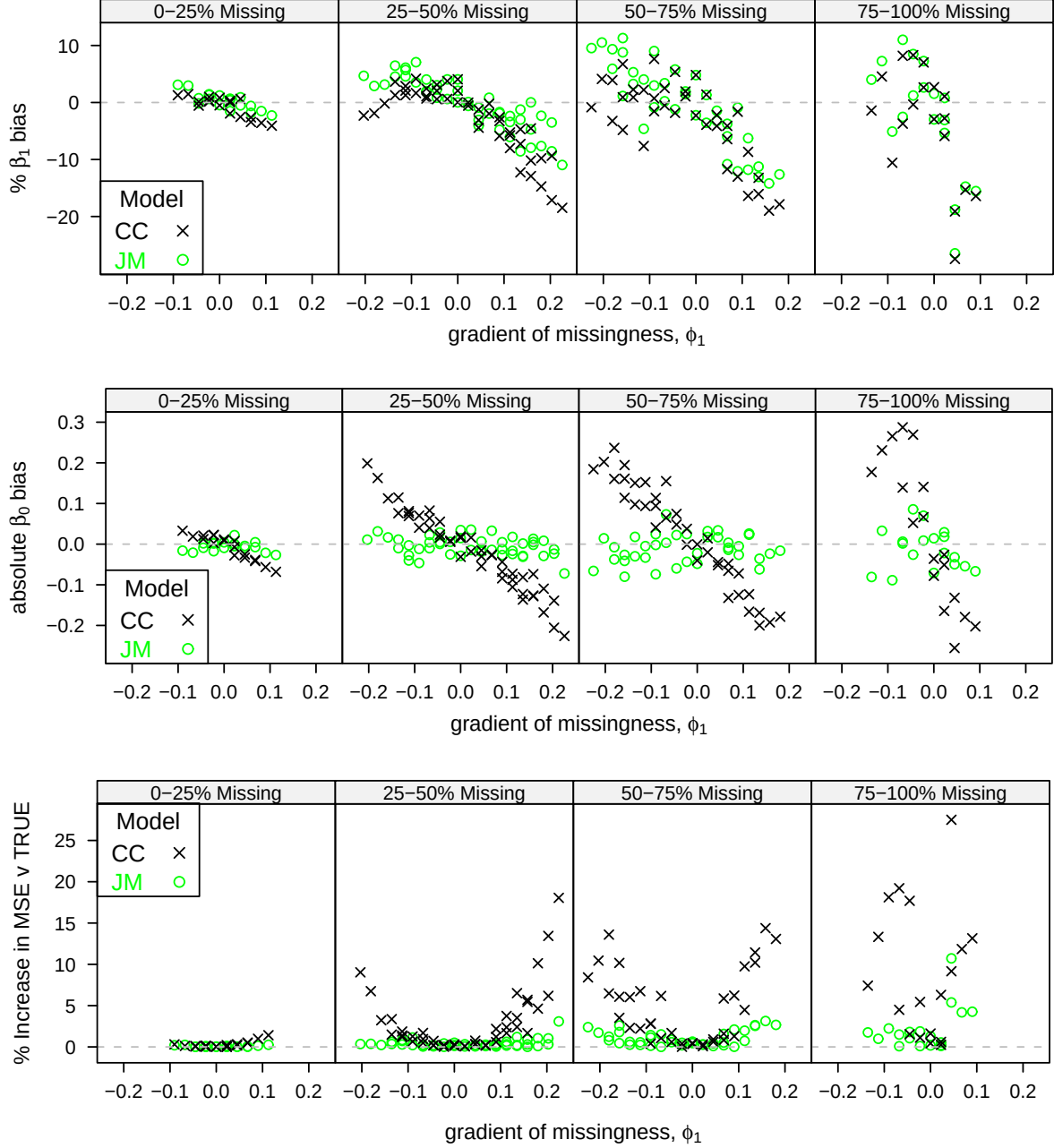
from 0 to 1 at 0.1 intervals. Every start probability is crossed with every end probability, with two exceptions, giving a comprehensive coverage of all possibilities and 119 subsets. The two exceptions were combining a start probability of 0 with an end probability of 0 (no missing data), and a start probability of 1 and an end probability of 1 (no observed data). The gradients varied from -0.23 to 0.23 including 0 (which is equivalent to a MCAR mechanism), and the percentage of missingness from 4.9% to 96.5%. As before the models TRUE and CC are run for each dataset.

The % differences between the CC and TRUE estimates of β_1 are plotted against the gradient of missingness, ϕ_1 , as black crosses for four levels of missingness in the top panel in Figure 4.2. (The JM points (open green circles) which are also shown will be discussed in Section 4.4.2.) When ϕ_1 is positive (individuals are more likely to be missing if they have high scores at age 16) CC always under-estimates β_1 , apart from a few datasets with gradient close to 0. The magnitude of this under-estimation increases with ϕ_1 and the percentage of missing values. From our multivariate Normal simulations we would expect to see a similar degree of under-estimation when ϕ_1 is negative, but in fact CC sometimes under-estimates and sometimes over-estimates β_1 , with over-estimation more likely, and the bias is less for negative ϕ_1 compared to positive ϕ_1 . The Scores.linear datasets differ from our MVN datasets in that both the response and covariate distributions are skewed, and we will explore the implications of this asymmetry later.

Our findings for the β_0 bias are as expected, as shown in the middle panel in Figure 4.2. Note that absolute, rather than percentage, bias is shown for β_0 , because TRUE β_0 is very close to 0 and leads to very large percentages. CC increasingly over-estimates β_0 as ϕ_1 becomes increasingly negative and increasingly under-estimates β_0 as ϕ_1 becomes increasingly positive at all levels of missingness.

As regards MSE (bottom panel in Figure 4.2), the overall fit of CC compared to TRUE deteriorates

Figure 4.2: Scores.linear: performance of CC and JM compared with the TRUE generated targets



as ϕ_1 becomes steeper or the percentage of missing values increases.

4.4.2 What improvements do we get from adding a model of missingness?

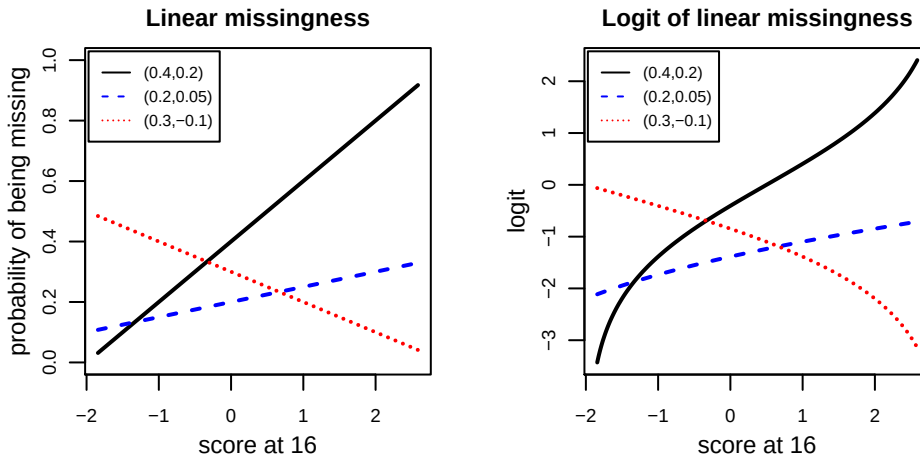
We have seen that with missing responses, complete-case analysis results in biased parameter estimates unless the missingness is MCAR. Further, we suspect that the direction of this bias is affected by the shape of the distribution of the original data (observed and missing) in addition to the shape of the missingness pattern. We now investigate the extent to which these biases are removed by adding a model of missingness to our model of interest. Again we start by looking at the impact on the

simulated MVN data before examining the more realistic Scores data.

MVN scenario

A third model, JM, was run as described in Section 4.1. Note that although the true model of missingness is the linear equation $p_i = \phi_0 + \phi_1 y_i$, we are modelling this by the linear logistic equation $\text{logit}(p_i) = \theta_0 + \theta_1 y_i$ to ensure that the probabilities lie in the range $[0,1]$. A logistic transformation of a linear line gives a sigmoid curve which is essentially linear for probabilities between 0.2 and 0.8, but non-linear outside this range (Collett, 2003). Figure 4.3 shows examples of linear missingness and their associated logits. This means that our imposed pattern of missingness can at best only be approximately rather than exactly modelled.

Figure 4.3: Comparison of shapes of linear missingness and their associated logits



The figures in brackets in the graph legends are the values of the ϕ parameters in Equation 4.4.

Looking at the JM points in Figure 4.1 (i.e. the open green circles), we see that the bias in the β estimates is almost eliminated and the overall fit of our model of interest is close to the TRUE model. Interestingly, for MNAR missingness, the estimate of β_1 is always higher for JM than CC, resulting in a reduced β_1 difference from TRUE for most but not all repetitions. For β_0 and the MSE the improvement for all repetitions is very clear.

So far we have concentrated on the bias of the parameter estimates, but the efficiency of these estimates is also of interest. The loss of records for CC results in efficiency of approximately 70% for the estimates of both parameters. Similar efficiency is achieved for the JM estimate of β_1 , with the additional information from the partially observed cases being offset by greater uncertainty. However, for the JM estimate of β_0 , efficiency is further reduced to just over 40%.

Scores.linear scenario

JM was run for Scores.linear. The dataset with the highest percentage of missingness is excluded from the results as it failed to converge. Looking again at Figure 4.2, we see that, consistent with our

findings for MVN, JM always pulls β_1 upwards from the CC estimate (apart from MCAR or almost MCAR), not always giving an improved estimate. However JM consistently produces an estimate for β_0 which is much closer to the target from TRUE, correcting both under-estimation and over-estimation. The addition of a model of missingness leads to an improvement in the overall fit (MSE) for all but the shallowest gradients.

It is known that selection models can be sensitive to the correct specification of both parts of the joint model (Kenward, 1998), and we now explore the sensitivity of our findings to a related issue, the strength of the relationship in the model of interest.

4.4.3 How critical is the strength of the relationship in the model of interest?

Because MVN is simulated data, we know that our assumption of a linear relationship between the covariate and response is correct. Our findings so far are based on a true correlation between the response and covariate of 0.5. To investigate how the strength of this relationship impacts our results, we repeat the simulation using correlations of 0.1, 0.25, 0.75 and 0.9 (scenarios MVN(0.1), MVN(0.25), MVN(0.75) and MVN(0.9) respectively). We find that as the relationship between the variables gets stronger, the CC bias is reduced for both β_0 and β_1 , there is less variation between replicates and the efficiency of the estimation of the JM β_0 increases towards the CC level. Reassuringly, JM seems to correct the bias in the parameter estimates regardless of the strength of the relationship. However, as the correlation gets weaker we start to encounter MCMC convergence problems, with 10 of the 150 JM runs failing to converge with the 0.25 correlation and 110 of the 150 JM runs not converging when the correlation is 0.1. This suggests that JM identifiability is driven by a strong relationship in the model of interest.

4.4.4 How critical is the adequacy of the model of missingness?

We now turn our attention to the other part of the joint model, and consider the adequacy of the model of missingness. Our results are potentially affected by two sources of error in our model of missingness: the use of a linear logit model to approximate a linear relationship and failure to fit the ‘best possible linear logit’. The ‘best possible linear logit’ is assumed to be the linear logit fitted by TRUE.

In order to gain insight into the relative importance of the two sources of error, we now compare the performance of JM with two other versions of the joint model, JMexact and JMfixed. To do this, we run JMexact and JMfixed on MVN. Whereas JM is set up with the parameters of the missingness model unknown so they are estimated, for JMexact and JMfixed the missingness model parameters are assumed known and therefore fixed. For JMexact we replace the linear logit with the exact equation

which was originally used to select the missing responses, i.e. $p_i = \phi_0 + \phi_1 y_i$. The design of the simulation ensures that the constraint $0 \leq p_i \leq 1$ is satisfied for all i . For JMfixed we retain the linear logit for the model of missingness, but fix its θ parameters to the posterior means of θ_0 and θ_1 that were estimated by TRUE.

The results for both JMexact and JMfixed are similar to the results for JM for the bias and efficiency of β_1 , and the bias of β_0 . However, the efficiency of the β_0 estimates are now similar to CC rather than JM (66% for JMexact and 67% for JMfixed), and the MSE is lower. This suggests that in Bayesian joint modelling the use of a linear logit adequately models linear missingness, but there are some benefits to improving the fit of the model of missingness if possible. This might be achieved by the use of informative priors to incorporate additional knowledge on the shape of the missingness model.

Having established that a linear logit is a good choice for modelling linear missingness, we now look at what happens when we use an incorrect model of missingness. We again repeat the JM analysis of the MVN datasets, this time using a restricted linear missingness model in which θ_1 is restricted to positive values for negMNAR and negative values for posMNAR. The missingness model slope parameter, θ_1 , is now estimated to be close to 0, and our joint model no longer removes the bias in the model of interest parameter estimates, producing similar estimates to CC. This can be seen from comparing the linear (blue) and restricted linear (red) points labelled “no” (indicating no transformation) in Figure 4.4, and seeing that the red dots and crosses are almost on top of each other, whereas the blue dots and crosses are definitely different with the blue dots showing significant improvement. (The remaining points in this graph will be discussed below and in Section 4.4.5.)

We also run the MVN simulations using the quadratic logistic equation

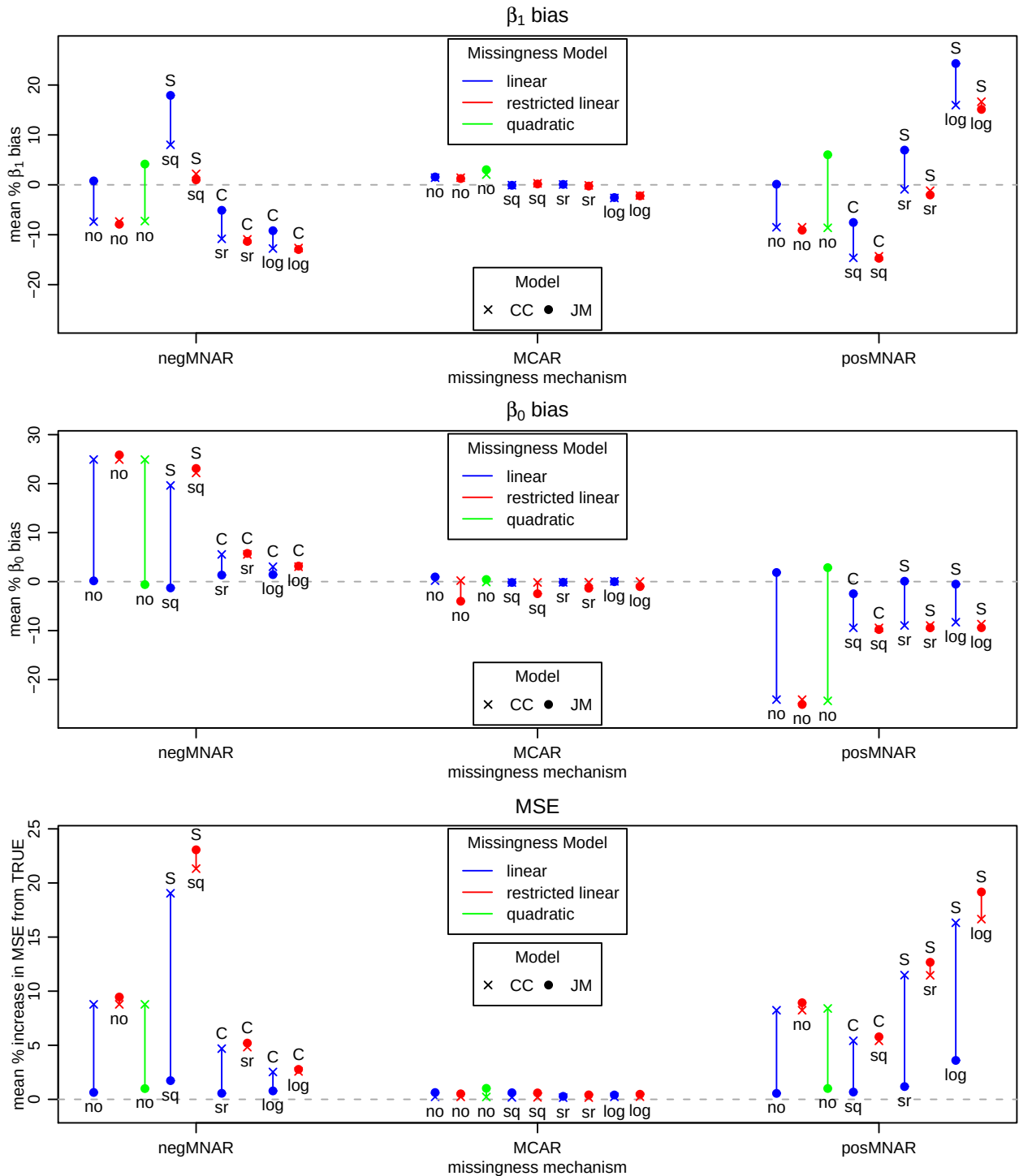
$$\text{logit}(p_i) = \theta_0 + \theta_1 y_i + \theta_2 y_i^2 \quad (4.7)$$

as the model of missingness, with a $\text{logistic}(0,1)$ prior for θ_0 and weakly informative $N(0,0.68)$ priors for θ_1 and θ_2 . About 20% of the repetitions failed to converge. Those which did converge failed to correct the bias in the β_1 estimates, as shown by the green dots in the top panel of Figure 4.4 which are not close to 0. The green dots are closer to 0 for β_0 and the MSE, but the quadratic logistic equation does not do as well as the linear logistic equation.

4.4.5 How critical is the error distribution in the model of interest?

In Section 4.4.2 we found that JM was much better at correcting bias in the β_1 estimate for symmetric MVN than skewed Scores.linear, and we now consider the reasons for this. In setting up our model of interest, we have assumed that the errors follow a Normal distribution. This assumption is used by JM when it fills in the missing responses, in a way that will attempt to correct any skewness in the observed responses given their covariates (Little and Rubin, 2002: Chapter 15; Daniels and

Figure 4.4: MVN, MVN(sq), MVN(sr) and MVN(log): impact of different transforms and missingness models



(1) The set of repetitions varies, as only converged runs are included. (2) In the Missingness Model legend, restricted linear indicates θ_1 is restricted to positive values for negMNAR and negative values for posMNAR. (3) The transformation used is indicated beneath each pair of points (no=none, sq=square, sr=square root and log=log). (4) A letter above a pair of points indicates whether the skewness from the transformation and the missingness are in the same direction (S) or in conflict (C). (5) The length of the line joining a dot and cross indicates the size of the change in the mean % β_i bias (top two plots) or the change in the increase in MSE from TRUE (bottom plot). If the dot is closer to the zero line than the cross, then JM performs better than CC for the plotted measure. Our target is for the dot to lie on the zero line.

Hogan, 2008: Section 8.3). For MVN this is fine, because all the skewness in the observed responses stems from the missingness mechanism and so JM does well. By contrast, the Scores response is already skewed and JM now tries to compensate for the combined effects of the original skewness and the added skewness from the imposed missingness. It has no way of distinguishing between the two sources of skewness, so cannot limit its correction to the skewness from the missingness mechanism as required.

To better understand what is happening, we transform the response in our original MVN data and then impose MCAR, posMNAR and negMNAR linear missingness as defined in Section 4.4.1 on this transformed data. We use three transformations, namely square (sq), square root (sr) and log (generating scenarios MVN(sq), MVN(sr) and MVN(log) respectively), and run CC, TRUE and JM. JM is run twice, once with a linear model of missingness and once with the restricted model of missingness described in Section 4.4.4. JM failed to converge for a few repetitions, mainly for the square transformation. Using the converged runs, the performance of these models in terms of the mean % bias of the parameters of the model of interest and the MSE is compared for the transformed and untransformed data in Figure 4.4.

The skewness from the transformation and the missingness are in the same direction when we have a square transformation and a negMNAR missingness mechanism, or a log or square root transformation and a posMNAR missingness mechanism (indicated by “S” label). When we have a square transformation and a posMNAR missingness mechanism, or a log or square root transformation and a negMNAR missingness mechanism, the two sources of skewness are in conflict (indicated by “C” label).

We start by considering the JM with the linear model of missingness (blue symbols in Figure 4.4). For the transformed data our model of interest has an incorrect error distribution, and the addition of a good model of missingness reduces the MSE when the missingness mechanism is MNAR, but the bias in the individual model of interest parameter estimates may not be removed or even reduced. For β_0 , the bias is removed if the skewness from the transformation and the missingness are in the same direction (see, for example, the blue dots labelled “sr” and “log” plotted against posMNAR in the β_0 panel of Figure 4.4), but only reduced if the two sources of skewness are in conflict (e.g. see the blue dot labelled “sq” for posMNAR). As regards β_1 , if the two sources of skewness are in the same direction, the complete case analysis over-estimation is increased (e.g. compare the blue cross and dot labelled “sr” for posMNAR in the β_1 panel of Figure 4.4), while for conflicting skewness the bias is reduced but not removed (e.g. compare the blue cross and dot labelled “sq” for posMNAR).

If we choose an incorrect model of missingness for JM, then there are only slight changes in the parameter estimates from CC, but they may result in a small deterioration in the fit of the model as measured by MSE. This is demonstrated by the JM with the restricted linear model of missingness,

shown in Figure 4.4 by the red dots and crosses.

This provides an explanation of the performance of JM with Scores.linear, which has a positively skewed response. When the gradient of the imposed missingness, ϕ_1 , is positive, we add negative skewness which is in conflict with the original skewness. From our findings from MVN we expect JM to reduce the bias in β_1 , which is confirmed by Figure 4.2. For negative ϕ_1 the two sources of skewness are in the same direction and as expected the β_1 bias is generally increased. This suggests that we should transform the Scores response to make it more Normal, for example by using a Box Cox power transformation (Box and Cox, 1964), i.e.

$$y = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1} & : \lambda_1 \neq 0 \\ \log(y + \lambda_2) & : \lambda_1 = 0 \end{cases} \quad (4.8)$$

where λ_2 is set to 2, to ensure that $y + \lambda_2$ is always positive. However, choosing λ_1 is difficult. If we use only the observed data to determine λ_1 , we are making assumptions about the missing data that are not justified from the data at hand. A possible approach is to carry out a sensitivity analysis to explore the impact of using different λ_1 values, and we demonstrate this in Section 5.1 with real examples.

Transformation is one of many ways of dealing with skewed data. Alternatively, we could use non-Gaussian distributions, such as the skew-Normal, to allow for the asymmetry, and estimate additional parameters. However, we restrict ourselves to transformations.

4.5 Non-linear missingness

With real data, the shape of the missingness mechanism is unknown, so it is useful to gain insight into how a joint model performs when the shape of the missingness deviates from linearity. To this end, we investigate non-linear missingness, using scenarios formed from the Scores data by imposing quadratic missingness according to Equation 4.5 (Scores.quad) and step missingness using Equation 4.6 (Scores.step).

We fit joint models with linear, quadratic and piecewise linear logit models of missingness to the datasets in Scores.quad, and details are given in Appendix B.1. Not surprisingly, we find that we get little benefit from using a linear logit to model quadratic shaped missingness, but reassuringly neither does it do much harm. While we expected to do better with a quadratic logit, we find that these models are fragile and need constrained priors. Even when such a model has been successfully fitted, it seldom results in much bias reduction. An alternative, piecewise linear logit models are even more difficult to fit, but if successfully estimated usually reduce the β_1 bias and improve the fit of the model of interest.

For step missingness, where all the responses above or below a certain threshold are missing, the β_1 bias from complete-case analysis is potentially much more serious than for linear missingness. We find that step models can give substantial bias reductions if the form of the missingness really is a step function, but their use is clearly limited and a linear logit model seems to work almost as well for step shaped missingness (see Appendix B.2 for details).

We expect the results of modelling the non-linear missingness datasets Scores.quad and Scores.step to be affected by the original skewness in the Scores data, given this skewness affects the performance of modelling Scores.linear with a linear logit as discussed in Section 4.4.5. Whereas we found that modelling the MVN datasets with a quadratic logit missingness model gives improvements over a complete case analysis, though not as large as a linear logit missingness model (Figure 4.4), modelling the Scores.linear datasets with a quadratic logit missingness model results in convergence problems and can be damaging in terms of the fit of the model of interest. This shows the dangers of misspecifying both the model of interest and the model of missingness.

In conclusion, it seems that if we have no prior knowledge about the shape of the missingness, then the safest strategy is to use a linear logit for the model of missingness part of a joint model. If the missingness is linear or close to linear there are potential benefits, and if not, its effect will be reasonably benign. If there is external evidence that the missingness is non-linear, then informative priors are needed to help with the estimation of the model of missingness parameters. We demonstrate this in Chapter 7, where we use a piecewise linear logit with fixed knots and informative priors on the remaining parameters based on expert knowledge. However, it is impossible to know how wrong our assumptions about the shape of the missingness might be, and in the next section we investigate how DIC can help in assessing the relative fit of models based on different assumptions.

4.6 Model choice using DIC

We now use our Scores data to investigate how DIC, which was defined in Sections 2.4.4 and 2.4.5, can be used to help choose between different joint models. DIC can be constructed in different ways, and we consider three formulations:

DIC_W: the DIC automatically produced by WinBUGS. This is a conditional DIC which treats the missing data as additional parameters (Celeux *et al.*, 2006). WinBUGS produces DIC values for the model of interest and model of missingness separately.

DIC_O: an observed DIC, based on the observed data likelihood (introduced in Section 2.4.5).

DIC_F: a complete DIC, based on the full data likelihood (also introduced in Section 2.4.5).

To evaluate DIC, we require the posterior mean of the deviance, $\bar{D}$, and a *plug-in deviance*, $\hat{D}$. $\hat{D}$ is calculated using point estimates of the parameters, which we will refer to as *plug-ins*. In order to see how the three formulations differ, we consider the likelihoods on which they are based.

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}, \mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{m}) &\propto f(\mathbf{y}_{obs}, \mathbf{m} | \boldsymbol{\beta}, \sigma, \boldsymbol{\theta}, \mathbf{y}_{mis}) \\ &= f(\mathbf{y}_{obs} | \boldsymbol{\beta}, \sigma) f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta}) \end{aligned} \quad (4.9: \text{DIC}_W)$$

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} | \mathbf{y}_{obs}, \mathbf{m}) &\propto \int f(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m} | \boldsymbol{\beta}, \sigma, \boldsymbol{\theta}) d\mathbf{y}_{mis} \\ &= f(\mathbf{y}_{obs} | \boldsymbol{\beta}, \sigma) E_{\mathbf{y}_{mis} | \mathbf{y}_{obs}, \boldsymbol{\beta}, \sigma} \{ f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta}) \} \end{aligned} \quad (4.10: \text{DIC}_O)$$

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}) &\propto f(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m} | \boldsymbol{\beta}, \sigma, \boldsymbol{\theta}) \\ &= f(\mathbf{y}_{obs}, \mathbf{y}_{mis} | \boldsymbol{\beta}, \sigma) f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta}) \end{aligned} \quad (4.11: \text{DIC}_F)$$

In DIC_W , the part of the likelihood calculated for the model of interest is based on $\mathbf{y}_{obs}$ only and the part for the model of missingness is conditional on $\mathbf{y}_{mis}$. The plug-ins for DIC_W are evaluated as $E(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}, \mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{m})$. Conditional DICs have asymptotic and coherency difficulties (Celeux *et al.*, 2006; Little and Rubin, 1983). However, the model of missingness part of DIC_W may be useful in helping to choose between joint models with the same model of interest but different models of missingness.

DIC_O differs from DIC_W in the model of missingness part, which is evaluated by integrating over $\mathbf{y}_{mis}$. It is the calculation of the expectation in Equation 4.10 which creates complexity in the DIC_O computation. Further detail on the derivation of Equation 4.10 can be found in Section 2.4.5. The plug-ins for DIC_O are $E(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} | \mathbf{y}_{obs}, \mathbf{m})$.

For DIC_F the posterior mean of the deviance is straightforward to calculate using the full data likelihood, $L(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m})$. However, in contrast to DIC_W and DIC_O , the expectation for the plug-ins is also conditional on $\mathbf{y}_{mis}$, i.e. $E(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m})$, giving a distribution of plug-ins given $\mathbf{y}_{mis}$, which we must then integrate over the posterior predictive distribution of $\mathbf{y}_{mis}$. So the complications arise in calculating the plug-in deviance, which requires integrating the distribution of $\hat{D}$ with respect to $\mathbf{y}_{mis}$, where the plug-in parameters are conditional on $\mathbf{y}_{mis}$.

We hope that each of these DICs will provide a different insight into the comparative fit of various models. DIC_W can be calculated automatically by WinBUGS, but we need a way of calculating DIC_O and DIC_F . So we now examine how best to implement and interpret DIC_O and DIC_F , using a dataset from the Scores.linear scenario. This dataset, which we shall call *Scores.LinMostPos*, has the steepest positive linear gradient (i.e. it was formed with a ‘start’ missingness probability of 0 and

an ‘end’ missingness probability of 1). Our investigation is based on the four joint models specified in Table 4.3. After exploring methods of implementation, we will look at how they can be used to help choose between models.

Table 4.3: Joint Models for *Scores.LinMostPos*

Model Name	Model of Interest	Model of Missingness
Scores.JM1	$y_i \sim N(\mu_i, \sigma^2); \quad \mu_i = \beta_0 + \beta_1 x_i$	$\text{logit}(p_i) = \theta_0 + \theta_1 y_i$
Scores.JM2	$y_i \sim N(\mu_i, \sigma^2); \quad \mu_i = \beta_0$	$\text{logit}(p_i) = \theta_0 + \theta_1 y_i$
Scores.JM3	$y_i \sim N(\mu_i, \sigma^2); \quad \mu_i = \beta_0 + \beta_1 x_i$	$\text{logit}(p_i) = \theta_0 + \theta_1 y_i + \theta_2 y_i^2$
Scores.JM4	$y_i \sim t_4(\mu_i, \sigma^2); \quad \mu_i = \beta_0 + \beta_1 x_i$	$\text{logit}(p_i) = \theta_0 + \theta_1 y_i$

The likelihood for the model of interest is calculated using

$$f(\mathbf{y}|\boldsymbol{\beta}, \sigma) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu_i)^2\right) \quad \text{for Scores.JM1-3}$$

$$\text{and } f(\mathbf{y}|\boldsymbol{\beta}, \sigma) = \prod_{i=1}^n \frac{\Gamma(\frac{5}{2})}{2\sigma\sqrt{\pi}} \left[1 + \left(\frac{y_i - \mu_i}{2\sigma}\right)^2\right]^{-\frac{5}{2}} \quad \text{for Scores.JM4.}$$

In the t_4 distribution σ is a scale parameter, s.t. $\sigma = \sqrt{\frac{\text{var}}{2}}$, where var is the variance of the distribution.

For all four models, the likelihood for the model of missingness is calculated using

$$f(\mathbf{m}|\mathbf{y}, \boldsymbol{\theta}) = \prod_{i=1}^n p_i^{m_i} (1 - p_i)^{1-m_i}$$

$$\text{where } p_i = \frac{e^{\theta_0 + \theta_1 y_i}}{1 + e^{\theta_0 + \theta_1 y_i}} \quad \text{for Scores.JM1,2\&4}$$

$$\text{and } p_i = \frac{e^{\theta_0 + \theta_1 y_i + \theta_2 y_i^2}}{1 + e^{\theta_0 + \theta_1 y_i + \theta_2 y_i^2}} \quad \text{for Scores.JM3.}$$

As an initial attempt to calculate DIC_O and DIC_F for these four models, we implement the algorithms suggested by Daniels and Hogan (2008), using the R software (R Development Core Team, 2007) and making use of the R2WinBUGS package (Sturtz *et al.*, 2005) to call WinBUGS to carry out the required MCMC runs. The algorithm for DIC_O is sketched below, and given in fuller detail in Appendix D.

DIC_O Algorithm using reweighting

- 1 Call WinBUGS to carry out a standard MCMC run on the selection model, and save samples of length K of the model of interest and model of missingness parameters, denoted $\boldsymbol{\beta}^{(k)}$, $\sigma^{(k)}$ and $\boldsymbol{\theta}^{(k)}$, $k = 1, \dots, K$, ($K\text{sample}$).
- 2 Evaluate the $K\text{sample}$ posterior means of the model parameters, $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}$ and $\hat{\boldsymbol{\theta}}$.
- 3 Generate a sample of length Q of missing responses from the appropriate likelihood evaluated at the posterior means of the model of interest parameters using $f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}|\hat{\boldsymbol{\beta}}, \hat{\sigma})$, ($Q\text{sample}$).

- 4 For each member, k , of the Ksample, and each missing response in the Qsample, calculate a weight and evaluate the expectation term from Equation 4.10, $E_{\mathbf{y}_{mis}|\mathbf{y}_{obs},\boldsymbol{\beta}^{(k)},\sigma^{(k)}}\{f(\mathbf{m}|\mathbf{y}_{obs},\mathbf{y}_{mis},\boldsymbol{\theta}^{(k)})\}$, by averaging over its weighted Qsample. (See step 4 in Appendix D for the required equations.)
- 5 Using these expectations, calculate the posterior mean of the deviance, $\bar{D}$, by averaging over the Ksample. (See step 5 in Appendix D for the required equations.)
- 6 Evaluate the expectation term of the plug-in deviance by averaging over the unweighted Qsample, and calculate the plug-in deviance, $\hat{D}$, using the posterior means from the Ksample. (See step 6 in Appendix D for the required equations.)
- 7 Finally, calculate $DIC_O = 2\bar{D} - \hat{D}$.

The algorithm for DIC_F is detailed in Appendix D, along with some notes on implementation. We shall refer to the samples generated at steps 1 and 3/4 of the algorithms as our ‘Ksample’ and ‘Qsample’ respectively, with K and Q denoting the size of these samples. The Ksample is produced by a WinBUGS MCMC run, and the Qsample is the importance sampling sample (Section 2.4.5, paragraph headed “Reweighting”). Initially both sample sizes are set to 2,000, i.e. $K = Q = 2,000$. The samples produced by the WinBUGS runs are based on 2 chains of 15,000 iterations, with 10,000 burn-in and the thinning parameter set to 5. These run lengths are consistent with those used throughout our research with the synthetic data, although previously we did not thin our samples. Based on the Gelman-Rubin convergence statistic and a visual inspection of the chains, the WinBUGS runs required for calculating DIC_O and DIC_F for the four models for *Scores.LinMostPos* all converged. We start checking the usability of these DIC_O and DIC_F by looking at the adequacy of their sample sizes.

4.6.1 The effective sample size (ESS)

Following Liu (2001), Chapter 2, we define ESS as

$$ESS = \frac{Q}{1 + var(w)} \quad (4.12)$$

where Q is the number of samples generated from a distribution, $f(\cdot)$, and $var(w)$ is the variance of normalised importance weights. A small variance is good, for ESS approaches the actual sample size, Q , as $var(w)$ gets smaller. This variance can be estimated by

$$\frac{\sum_{q=1}^Q (w_q - \bar{w})^2}{(Q - 1)\bar{w}^2} \quad (4.13)$$

where $\bar{w}$ is the mean of the sample of w_q s.

In the calculation of both DIC_O and DIC_F , a set of sampling weights, $w_q^{(k)}$, is produced for each member of the Ksample. So for our examples, we calculate 2,000 values of ESS corresponding to

the 2,000 Ksample members. The distribution of these for DIC_O and DIC_F is shown in Figure 4.5 for *Scores.LinMostPos* modelled using Scores.JM1, with the median and 95% interval (the latter calculated from the 2.5 and 97.5 percentiles) given below each plot. The inadequacy of the sampling weights for the DIC_F calculation is obvious, with all the ESS values between 1 and 6, even though the chains are combined to give 2,000 samples. The ESS for DIC_O is more varied, ranging from 2 to 1,963, but includes a sizeable proportion which are sufficiently low to be of concern. From the ESS summary statistics in Table 4.4 for *Scores.LinMostPos*, we see that the ESS values for the other three models are similar to Scores.JM1.

Figure 4.5: ESS for *Scores.LinMostPos* modelled using Scores.JM1

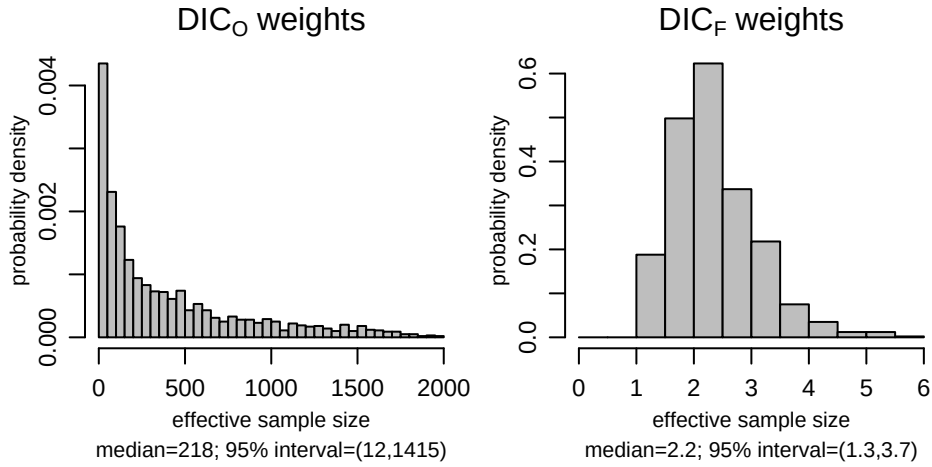


Table 4.4: Effective sample size for *Scores.LinMostPos*

	DIC _O		DIC _F	
	median	95% interval	median	95% interval
Scores.JM1	218.1	(11.8,1414.7)	2.2	(1.3,3.7)
Scores.JM2	323.6	(13.4,1773.1)	1.0	(1.0,1.0)
Scores.JM3	174.8	(7.4,1371.2)	1.0	(1.0,1.6)
Scores.JM4	226.3	(5.0,1347.4)	2.2	(1.2,3.2)

Given the evident problems with the weights used in the DIC_F calculations for these examples, we do not pursue this formulation of the DIC any further and concentrate on the DIC_O . We note that Daniels and Hogan also experienced difficulty with calculating DIC_F in practice (Mike Daniels, personal correspondence).

As regards the DIC_O calculations, restricting the Ksample to large ESS is not valid because it distorts the posterior distribution. Two versions of DIC_O are produced, based on different choices of plug-in parameters for the calculation of $\hat{D}$, which we call the θ and *logitp* plug-ins. For the θ plug-ins, the posterior means $\hat{\theta}_0 = E(\theta_0)$, $\hat{\theta}_1 = E(\theta_1)$, and $\hat{\theta}_2 = E(\theta_2)$ if required, are used to calculate the $\text{logit}(\hat{p}_i)$, while for the *logitp* plug-ins the posterior means of the $\text{logit}(p_i)$ are the plug-ins. We now compare these two sets of plug-ins.

4.6.2 Choice of plug-ins

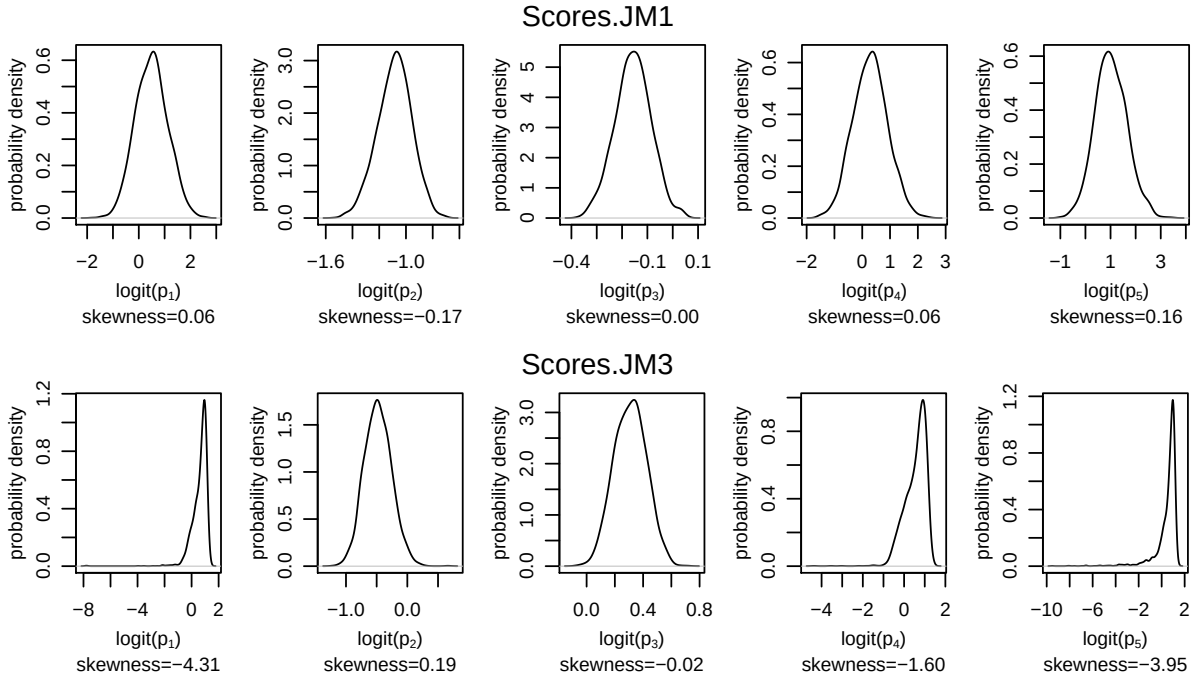
Spiegelhalter *et al.* (2002) discuss how p_D is dependent on the choice of plug-in estimator and Celeux *et al.* (2006) comment that the choice of plug-in parameters is delicate. Spiegelhalter *et al.* (2002) also point out that using the posterior mean leads to a lack of invariance to transformations of the parameters, so we expect different results from working with *logitp* plug-ins (the linear predictor) compared to the θ plug-ins (the stochastic parameters used to calculate the linear predictor).

The reasonableness of the choice of the posterior mean depends on the approximate normality of the parameter's posterior distribution. Figure 4.6 compares the posterior distributions of the *logitp* of the first 5 individuals in *Scores.LinMostPos* for models *Scores.JM1* and *Scores.JM3*. The coefficient of skewness which is a measure of the asymmetry of a distribution is shown beneath each plot, and has been calculated as follows:

$$\text{coefficient of skewness} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\text{sd}(x)^3}. \quad (4.14)$$

This coefficient of skewness should be close to 0 for a symmetric distribution. As a guide to interpreting the values for our Ksample of size 2,000, 95% of 10,000 simulated Normal datasets with 2,000 members had skewness in the interval (-0.1,0.1). All five distributions are approximately Normal for *Scores.JM1*, but individuals 1, 4 and 5 have a very skewed distribution for *Scores.JM3*. These three individuals have missing response.

Figure 4.6: Posterior distributions of the first 5 *logitp* for *Scores.LinMostPos*



Individuals 1, 4 and 5 have missing response.

Table 4.5 shows the coefficient of skewness for a selection of parameters that are used as plug-ins for *Scores.LinMostPos*. Mean and 95% interval values are given for y_{mis} , and the $\text{logit}(p_i)$ for all

individuals, observed individuals and missing individuals. The posterior distributions for the $\text{logit}(p_i)$ for the missing individuals for Scores.JM2, Scores.JM3 and Scores.JM4 are clearly skewed, as are the posterior distributions for the Scores.JM2 and Scores.JM4 $\mathbf{y}_{mis}$. σ , $\log(\sigma)$ and τ are all included in the table, and the difference in their skewness demonstrates sensitivity to the choice of the form of the scale parameter plug-in. This provides evidence that using a log transformation for σ is appropriate as argued by Spiegelhalter *et al.* (2002). (All our DIC calculations work with plug-in values for σ calculated on the log scale.)

Table 4.5: Skewness of posterior distribution for *Scores.LinMostPos*

	Scores.JM1	Scores.JM2	Scores.JM3	Scores.JM4
β_0	0.129	0.018	-0.009	0.000
β_1	0.043		0.062	0.006
σ	0.238	0.110	0.236	0.098
$\log(\sigma)$	0.146	-0.027	0.137	-0.014
τ	0.036	0.299	0.058	0.237
θ_0	-0.007	-0.200	-0.014	-0.103
θ_1	0.170	-0.158	0.131	0.102
θ_2			-0.179	
$\mathbf{y}_{mis}^a$	0.004	-0.454	0.098	1.331
	(-0.084,0.086)	(-0.563,-0.350)	(-0.040,0.218)	(0.556,2.600)
$\text{logitp}(\text{all})^a$	-0.021	0.145	-0.751	0.551
	(-0.182,0.164)	(-0.267,0.658)	(-4.091,0.491)	(-0.198,2.159)
$\text{logitp}(\text{obs})^a$	-0.085	-0.207	0.067	-0.133
	(-0.185,0.107)	(-0.279,-0.024)	(-0.316,0.274)	(-0.198,0.033)
$\text{logitp}(\text{mis})^a$	0.057	0.573	-1.748	1.385
	(-0.092,0.198)	(0.458,0.694)	(-4.526,0.787)	(0.637,2.760)

^a mean value is shown, with 95% interval below in brackets

The validity of DIC_O is dependent on a good choice of the plug-in parameters, and these should come from approximately symmetric unimodal distributions. As a future investigation of the DIC_O , the posterior mean could be compared to the posterior median, which was investigated at some length by Spiegelhalter *et al.* (2002), and the posterior mode, which was considered as an alternative by Celeux *et al.* (2006).

For Scores.JM1 the two sets of plug-ins produce very different $\hat{D}$ (2411 for the $\boldsymbol{\theta}$ plug-ins versus 2310 for the logitp plug-ins) and hence different p_D and DIC_O . This difference is not explained by skewness, as neither set of parameters is particularly skewed. To understand what is happening, we look at the formulae used for evaluating the expectation term in the plug-in log likelihood. For the $\boldsymbol{\theta}$ plug-ins

$$E_{\mathbf{y}_{mis}|\mathbf{y}_{obs},\hat{\boldsymbol{\beta}},\hat{\sigma}}\{f(\mathbf{m}|\mathbf{y}_{obs},\mathbf{y}_{mis},\hat{\boldsymbol{\theta}})\} \approx \frac{1}{Q} \sum_{q=1}^Q f(\mathbf{m}|\mathbf{y}_{obs},\mathbf{y}_{mis}^{(q)},\hat{\boldsymbol{\theta}})$$

so $\hat{D}$ is evaluated by integrating over $\mathbf{y}_{mis}$, as in the calculation of $\bar{D}$. However, for the logitp plug-ins

this formula collapses to

$$E_{\mathbf{y}_{mis}|\mathbf{y}_{obs},\hat{\beta},\hat{\sigma}}\{f(\mathbf{m}|\mathbf{y}_{obs},\mathbf{y}_{mis},\hat{\theta})\} \approx f(\mathbf{m}|\hat{logitp})$$

and we no longer integrate over $\mathbf{y}_{mis}$. We now have a conditional $\hat{D}$, the same $\hat{D}$ calculated for DIC_W . DIC_O based on *logitp* plug-ins does not have a sensible interpretation because there is an inconsistency between $\bar{D}$, which integrates over $\mathbf{y}_{mis}$, and $\hat{D}$ which does not. Henceforth, all DIC_O calculations will use the θ plug-ins. By contrast, both sets of plug-ins allow consistent calculation of $\bar{D}$ and $\hat{D}$ for DIC_W (and DIC_F , which we have discarded for other reasons).

4.6.3 Stability of DIC_O

Next we investigate the stability of DIC_O for our initial sample sizes of 2,000. Model Scores.JM1 is re-run on the *Scores.LinMostPos* data four times using different random number seeds for the step 1 MCMC run which produces the Ksample. The Qsample is generated from the same random number seed, but will also differ between runs due to the Ksample differences. Table 4.6 shows small differences between the DIC_O based on the θ plug-ins for the original run (Repetition1) and the four repeated runs (Repetition2 - K2000 to Repetition5 - K2000). Increasing the Ksample size will slightly decrease this variation, but in this example we consider a Ksample of 2,000 to be adequate.

Table 4.6: DIC_O for *Scores.LinMostPos* using Scores.JM1 (Ksample variability)

	$\bar{D}$	θ plug-ins		
		$\hat{D}$	p_D	DIC
Repetition1	2418.5	2411.0	7.4	2425.9
Repetition2 - K2000	2418.1	2410.8	7.3	2425.4
Repetition3 - K2000	2418.2	2411.0	7.2	2425.4
Repetition4 - K2000	2418.1	2410.9	7.2	2425.3
Repetition5 - K2000	2418.3	2411.0	7.4	2425.7

Each repetition uses a different random number seed to generate the Ksample, but the same random number seed to generate the Qsample. $K=Q=2000$.

Next, we turn our attention to the Qsample, again re-running Scores.JM1 on *Scores.LinMostPos*, but this time using different random number seeds for the sample generated at step 3. In these repetitions the Ksample is generated from the same random number seed, so any differences are attributable to variation in the Qsample. From Table 4.7 we see that sampling variation in the Qsample results in differences that are sufficiently large to be a concern. $\bar{D}$ varies between 2,418 and 2,425, and $\hat{D}$ exhibits similar levels of variation. The effect is a difference of 6.2 between the highest and lowest DIC_O . Rules of thumb suggest that differences of 1-2 in DIC should not be regarded as important (Spiegelhalter *et al.*, 2002), but the effects of Qsample variability are considerably larger than this, which undermines the usability of this formulation of DIC.

Table 4.7: DIC_O for *Scores.LinMostPos* using Scores.JM1 (Qsample variability 1)

	$\bar{D}$	θ plug-ins		
		$\hat{D}$	p_D	DIC
Repetition1	2418.3	2411.1	7.2	2425.5
Repetition2 - Q2000	2419.1	2410.4	8.8	2427.9
Repetition3 - Q2000	2420.0	2413.5	6.4	2426.4
Repetition4 - Q2000	2423.0	2416.0	7.0	2430.0
Repetition5 - Q2000	2424.6	2417.6	7.0	2431.7
mean	2421.0	2413.7	7.3	2428.3

Each repetition uses the same random number seed to generate the Ksample, but a different random number seed to generate the Qsample. $K=Q=2000$.

To see whether increasing the size of the Qsample helps, we run Scores.JM1 on *Scores.LinMostPos* five more times, setting Q to 10,000. From Table 4.8 we see that the levels of variation are only slightly lower than those based on $Q=2,000$, resulting in a DIC difference of 5.8. We also note that on average the larger sample size produced lower $\bar{D}$, $\hat{D}$ and DIC figures.

Table 4.8: DIC_O for *Scores.LinMostPos* using Scores.JM1 (Qsample variability 2)

	$\bar{D}$	θ plug-ins		
		$\hat{D}$	p_D	DIC
Repetition1 - Q10000	2418.5	2410.6	7.9	2426.4
Repetition2 - Q10000	2415.0	2407.0	8.1	2423.1
Repetition3 - Q10000	2420.8	2413.5	7.3	2428.0
Repetition4 - Q10000	2416.5	2408.2	8.2	2424.7
Repetition5 - Q10000	2420.3	2413.1	7.2	2427.5
mean	2418.2	2410.5	7.7	2425.9

Each repetition uses the same random number seed to generate the Ksample, but a different random number seed to generate the Qsample. $K=2000$; $Q=10000$.

Given the level of instability in the figures and the low ESS associated with some sets of sampling weights, we decide that our implementation of the reweighting approach to calculate DIC_O is not providing DIC that are of practical use for model comparison.

4.6.4 DIC_O calculated without the use of reweighting

To remove the problems associated with reweighting, we implement a second algorithm that generates a Qsample for every member of the Ksample, the Gold Standard DIC_O algorithm described in Appendix D.1.2. One of the original reasons for using reweighting was to speed up the computation of the DIC, since the gold standard method involves generating $K + (K \times Q) + Q$ samples, whereas the importance

sampling method just generates $K + Q$ samples, and then reweights the single Q sample for every replicate in the K sample. However, for equivalent sample sizes we find that our implementation of the Gold Standard DIC_O algorithm runs in about the same time as the original algorithm that uses reweighting for this example (both algorithms take approximately 2 hours on a desktop computer with a dual core 2.4GHz processor and 3.5GB of RAM for $K = 2,000$ and $Q = 10,000$), so there appears to be no computational advantage to using reweighting in practice. This is because the computational time saved in the reweighting algorithm by not generating the extra samples, is offset by calculating $K \times Q$ model of interest likelihoods to evaluate the weights.

We perform five repetitions fitting Scores.JM1 to *Scores.LinMostPos* using this algorithm, with $K = 2,000$ and $Q = 10,000$, and different random number seeds to generate the Qsample but not the Ksample. From the first five rows of Table 4.9 (GS Repetition1 - Q10000 to GS Repetition5 - Q10000) we see that $\bar{D}$ is now almost stable, but $\hat{D}$ is still variable. The value for GS10000rep1 demonstrates just how sizeable this variation can be.

Table 4.9: DIC_O for *Scores.LinMostPos* using Scores.JM1 (Gold Standard algorithm)

	$\bar{D}$	θ plug-ins		
		$\hat{D}$	p_D	DIC
GS Repetition1 - Q10000	2415.1	2394.9	20.2	2435.2
GS Repetition2 - Q10000	2414.9	2407.7	7.2	2422.1
GS Repetition3 - Q10000	2415.0	2404.6	10.3	2425.3
GS Repetition4 - Q10000	2414.8	2408.5	6.3	2421.1
GS Repetition5 - Q10000	2415.1	2407.9	7.2	2422.2
GS Scores.JM1 - Q40000 ^a	2413.0	2403.8	9.2	2422.1

Each repetition uses the same random number seed to generate the Ksample, but a different random number seed to generate the Qsample. $K=2000$; $Q=10000$ unless otherwise stated.

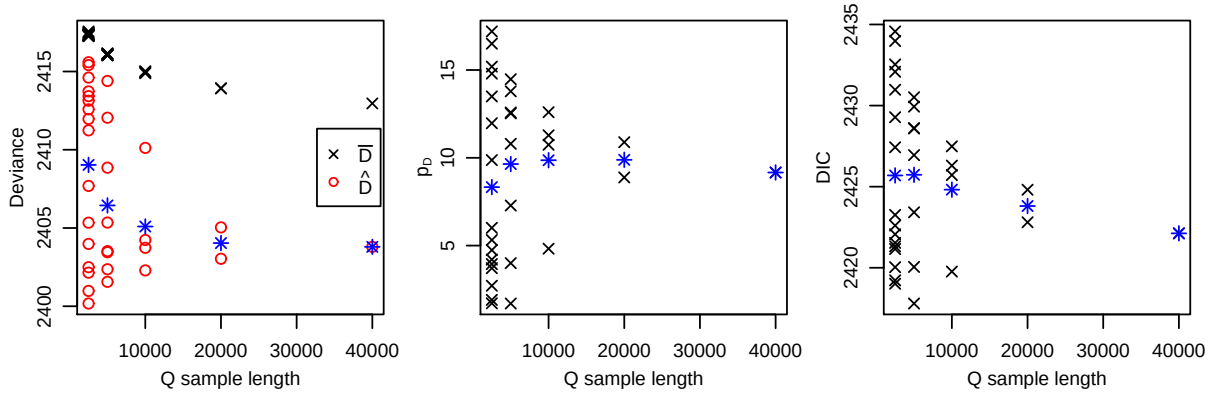
^a $Q=40000$.

As a further test of the Gold Standard DIC_O algorithm, we carry out a single run with Q set to 40,000. This run takes approximately 8 hours on a desktop computer with a dual core 2.4GHz processor and 3.5GB of RAM. The result is shown in the final line of Table 4.9 (GS Scores.JM1 - Q40000). As with the weighted algorithm, we find that the increase in sample size leads to a lower $\bar{D}$. Comparisons for the $\hat{D}$, p_D and DIC are affected by the GS Repetition1 - Q10000 results.

To investigate the stability of our results and the effects of sample size, we also calculate DIC_O using subsets of the GS Scores.JM1 - Q40000 Qsample. These subsets are created by splitting the complete 40,000 sample in half, and then successively splitting the resulting subsets in half until we reach a sample size of 2,500. This gives 2 non-overlapping Qsamples of length 20,000, 4 non-overlapping Qsamples of length 10,000, 8 non-overlapping Qsamples of length 5,000 and 16 non-overlapping Qsamples of length 2,500. These extra calculations can be carried out with negligible

additional cost in running time. $\bar{D}$, $\hat{D}$, p_D and DIC are plotted against the size of the Qsample in Figure 4.7. Mean values at each value of Q are shown as blue stars for $\hat{D}$, p_D and DIC. Looking at the left plot, we observe little variation in $\bar{D}$ at each Q (for $Q=2,500$ the range is 2,417.2 to 2,417.6), but a clear downwards trend as Q increases, which seems to be converging towards a limit. However, $\hat{D}$ exhibits considerable instability, with a range of 2,400.2 to 2,415.6 at $Q=2,500$, that decreases as Q increases. A similar downwards trend to $\bar{D}$ is indicated by the mean values of $\hat{D}$. The centre and right plots show the resulting instability in p_D and DIC respectively. There is no obvious trend for p_D , but evidence of decreasing DIC. In the future, we will look only at the Deviance plot, as p_D and DIC are functions of deviance and provide no additional information.

Figure 4.7: DIC_O calculations for *Scores.LinMostPos* modelled using Scores.JM1



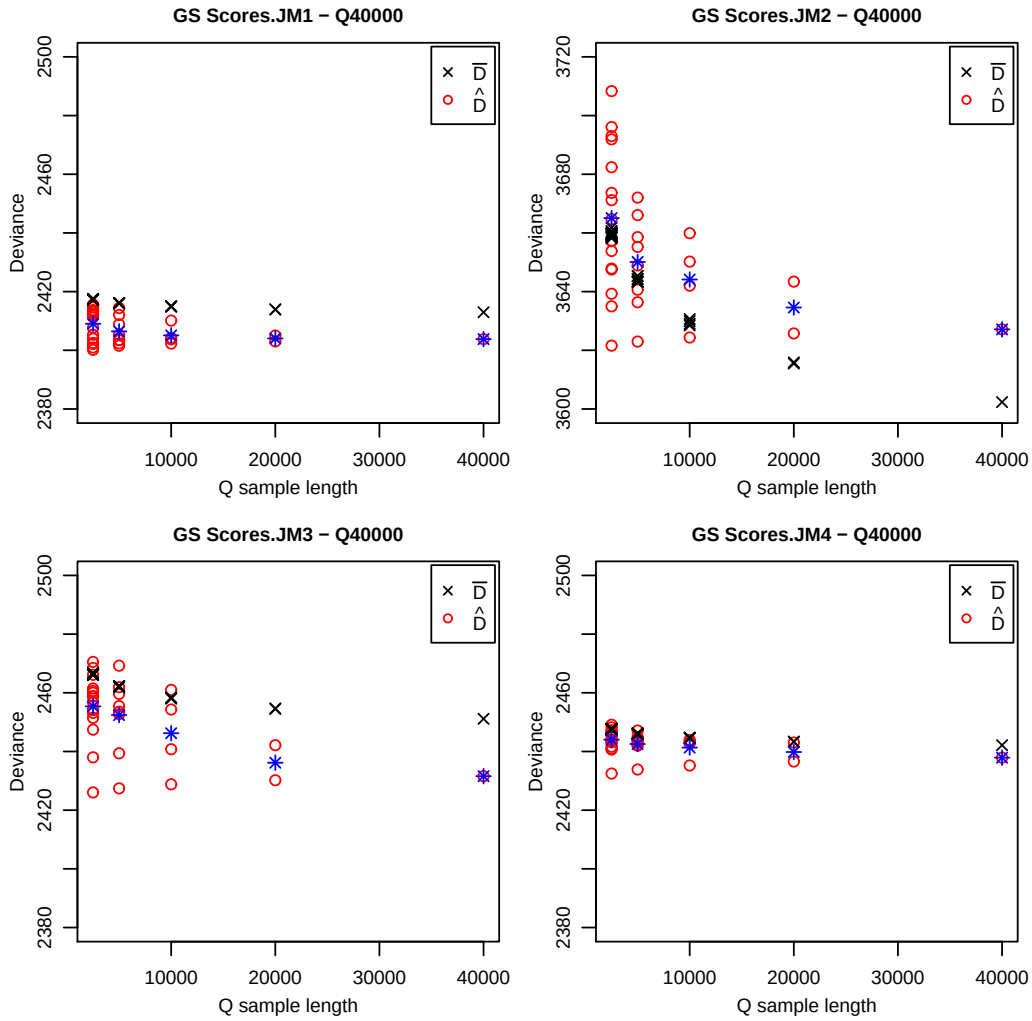
Calculations use θ plug-ins. The mean $\hat{D}$, p_D and DIC at each value of Q are indicated by blue stars.

We now recalculate DIC_O using the Gold Standard algorithm for *Scores.LinMostPos* modelled using the other three joint models, with $K=2,000$ and $Q=40,000$ (GS Scores.JM2 - Q40000, GS Scores.JM3 - Q40000 and GS Scores.JM4 - Q40000). Figure 4.8 allows us to assess the adequacy of our Qsample for the four different models by splitting it into subsets and plotting $\bar{D}$ and $\hat{D}$ against the sample lengths. The scale ranges of the four plots are consistent, but the badly fitting GS Scores.JM2 - Q40000 has a different magnitude to the other three. Increased instability in the $\hat{D}$ estimates compared to GS Scores.JM1 - Q40000, is evident for GS Scores.JM2 - Q40000 and GS Scores.JM3 - Q40000. Downward trends in the deviance estimates, converging towards a limit, are shown for all models. However, for GS Scores.JM2 - Q40000 and to some extent GS Scores.JM3 - Q40000, this limit has not yet been reached indicating that a larger Q is required for these models. Taken together, these plots suggest that higher variability and slower convergence to a limit are associated with poorer fitting models.

4.6.5 Model choice with DIC_O

Recall that since the missingness is simulated, we know that Scores.JM3 is an incorrect model. From exploratory plots (see Figure 3.1) Scores.JM2 is also incorrect, but we do not know whether Scores.JM1 or Scores.JM4 are true models. Table 4.10 shows the DIC_O calculated by the Gold Standard algorithm

Figure 4.8: DIC_O Gold Standard algorithm calculations for *Scores.LinMostPos*



Calculations use θ plug-ins. The mean $\hat{D}$ at each value of Q are indicated by blue stars.

for the four models for *Scores.LinMostPos*. Two versions of the MSE for the model of interest, as defined by Equation 4.3, are also displayed in this table. One version (obs) is based only on the observed data, and the other (all) uses both the observed and missing data which is possible with the Scores data because we have simulated missingness and know the true values of the missing y_i . Based on the observed data, the MSE slightly favours Scores.JM1, but there is little to choose between the fit of any of the models except Scores.JM2, which is a very poor choice. If the missing data are also brought into consideration, then Scores.JM1 and Scores.JM4 fit better than Scores.JM3. In contrast to the MSE, DIC_O assesses the joint fit of both parts of the model, penalised for complexity, although, as with MSE obs, the fit of the model of interest is only considered with respect to the observed responses. DIC_O clearly favours Scores.JM1.

Table 4.10 also shows the DIC_O obtained using the algorithm with reweighting. As expected, given previous observations about the effects of increasing Q , the Gold Standard algorithm produces lower $\bar{D}$, $\hat{D}$ and DIC figures for all the models. However, their ordering is unchanged.

For the Gold Standard algorithm, we note that p_D for Scores.JM2 is negative. Negative p_D can

Table 4.10: DIC_O for *Scores.LinMostPos* (Gold Standard v Reweighting)

	Gold Standard algorithm ^a				Reweighting algorithm ^b				MSE	
	$\bar{D}$	$\hat{D}$	p_D	DIC	$\bar{D}$	$\hat{D}$	p_D	DIC	all ^c	obs ^d
Scores.JM1	2413.0	2403.8	9.2	2422.1	2418.3	2411.1	7.2	2425.5	0.406	0.340
Scores.JM2	3602.3	3627.1	-24.8	3577.5	3680.6	3671.7	8.9	3689.5	2.243	1.153
Scores.JM3	2451.1	2431.6	19.6	2470.7	2461.0	2446.5	14.5	2475.5	0.434	0.341
Scores.JM4	2442.2	2437.9	4.3	2446.4	2450.3	2442.1	8.2	2458.5	0.402	0.342

^a Gold Standard DIC_O algorithm (θ plug-ins): $K=2000$; $Q=40000$.

^b Reweighting DIC_O algorithm (θ plug-ins): $K=2000$; $Q=2000$.

^c MSE based on all data (observed and missing)

^d MSE based on observed data only

indicate a poor model, which is certainly the case with Scores.JM2. Of the remaining models, DIC_O identifies Scores.JM1 and Scores.JM4 as better models than Scores.JM3. However, unlike the MSE, it suggests that the model assuming Normal errors is preferable to the one with t_4 errors.

4.6.6 Comments on the use of DIC_O and DIC_F

The calculation of DIC is known to be sensitive to the choice of the plug-in estimates used in calculating the point estimate of the deviance, and this sensitivity is heightened by the presence of missing data. In addition, if we use reweighting in the algorithms for calculating DIC_O or DIC_F , we must also contend with issues surrounding the variability of the sampling weights, and checks that these are satisfactory are crucial. Based on our exploration with a simulated example, our implementations of DIC_O and DIC_F using reweighting are not usable.

However, DIC_O can be calculated without using reweighting, and this approach is more promising, although we must choose the size of the Q_{sample} to be sufficiently large to avoid overestimating DIC and problems with instability in the plug-in deviance. Based on limited exploration of synthetic data, we tentatively propose working with a Q_{sample} of at least 40,000. The adequacy of this can easily be assessed by splitting this sample as described in Section 4.6.4 and plotting the associated $\bar{D}$ and $\hat{D}$ against sample size. We will look further at this in the context of a real example in Section 5.2, using our antidepressant clinical trial data. Again based on this synthetic example, we tentatively suggest that even with a well chosen Q_{sample} size, a DIC difference of at least 5 is required to provide some evidence of a genuine difference in the fit of two models, as opposed to reflecting sampling variability.

In our Scores example, we know that Scores.JM2 has a totally inadequate model of interest, and this is reflected in the high DIC_O and picked up by both versions of the MSE. Since we simulated the missingness for this example, we know that the model of missingness in Scores.JM3 is too complex. DIC_O correctly suggests that Scores.JM3 is a worse model than Scores.JM1 and Scores.JM4, which is not something that the MSE based only on the observed response data picks up. As the Scores

data is taken from a real dataset, rather than generated from some distribution, we cannot be sure whether Normal or t_4 errors are more appropriate. The MSE on the observed data suggests that the fit is essentially the same. By contrast, DIC_O distinctly favours the Normal distribution.

In using DIC_O we must remember that it will only tell us about the fit of our model to the observed data and nothing about the fit to the missing data. However, it does seem reasonable to use it to compare joint models with different models of interest but the same models of missingness, and we will explore this further with a real example. It can provide useful information about a set of models, but should not be used in isolation to choose a single model. Sensitivity analysis to check that conclusions are robust to a range of assumptions about the missing data remains crucial.

4.7 Interpreting the model of missingness scaled p_D

We now consider DIC_W , the DIC calculated automatically in WinBUGS, and in particular its associated p_D . As stated in Section 4.6, WinBUGS produces separate values for the model of interest and model of missingness, with the model of interest values based on the records with observed responses only. For the model of interest, p_D can be interpreted as approximately the true number of parameters in the model if vague priors have been used. However, we are interested in possible interpretations of p_D for the model of missingness, which is less straightforward. The missing observations are treated as additional parameters, so in a saturated model we would expect each to contribute 1 to p_D , while if they provide no information they will contribute 0. If *scaled* p_D is defined as

$$\text{scaled } p_D = \frac{p_D - \text{number of parameters}}{\text{number of missing observations}}, \quad (4.15)$$

it allows us to assess the information content being derived per missing observation, without being influenced by the number of missing observations. In this formula, the number of parameters is the number of regression coefficients in the linear predictor of the model of missingness part of the selection model. Strictly speaking, this only holds if we have uninformative priors on these coefficients. If the priors are informative, then each parameter will not contribute 1 to p_D .

In Section 4.6.2, we discussed the importance of the choice of plug-ins for calculating p_D , and it is possible to get negative p_D values when the posterior distribution for a parameter is skewed or there is substantial conflict between the prior and the data.

For a Bernoulli model the deviance is given by

$$\text{Deviance} = -2 \sum_{i=1}^n (m_i \log(p_i) + (1 - m_i) \log(1 - p_i)), \quad (4.16)$$

where m_i and p_i are as defined by Equation 4.1 and the ‘link’ is taken to be a logit function. The version calculated by WinBUGS uses plug-ins defined by the stochastic parameters in the likelihood,

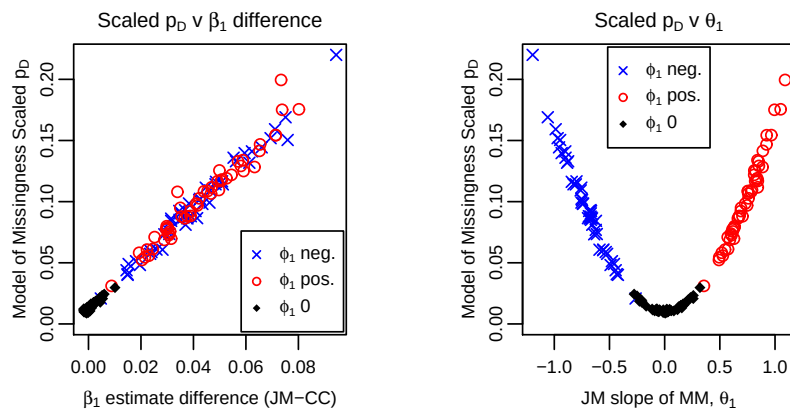
i.e. it calculates $\text{logit}(\hat{p}_i) = \hat{\theta}_0 + \hat{\theta}_1 y_i$ using the posterior means $\hat{\theta}_0 = E(\theta_0)$ and $\hat{\theta}_1 = E(\theta_1)$ as the plug-ins, assuming stochastic prior distributions were specified on θ_0 and θ_1 . We find with our synthetic data scenarios that this sometimes leads to negative p_D s for the missingness model. To avoid this, we calculate the posterior means of the $\text{logit}(p_i)$ and use these as our plug-ins, which is the canonical parameterisation (Spiegelhalter *et al.*, 2002).

4.7.1 Relationship between scaled p_D and the change in β_1 between JM and CC

MVN scenario

For MVN, the left graph in Figure 4.9 plots the model of missingness scaled p_D against the difference in the β_1 estimates between JM and CC. This provides evidence that scaled p_D is indicative of the size of the change of the slope parameter estimate in our model of interest, i.e. evidence of the ‘influence’ of the joint model. (The plot on the right in Figure 4.9 is discussed separately, Section 4.7.2.)

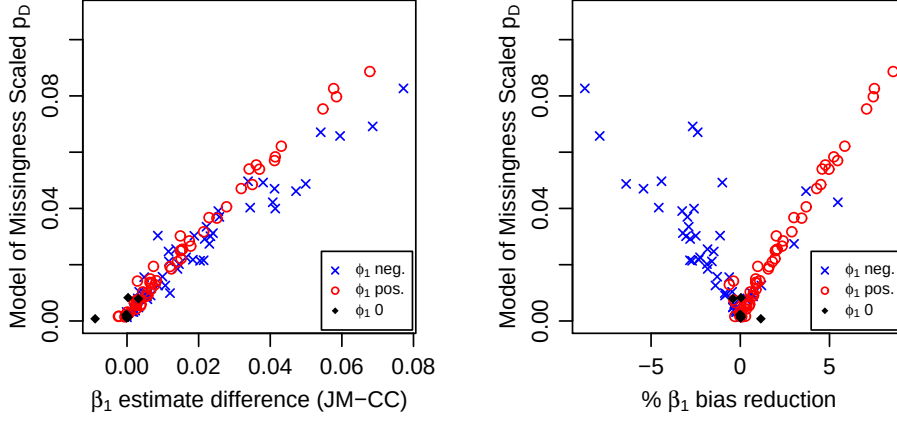
Figure 4.9: MVN: the relationship of scaled p_D with β_1 and θ_1



Scores.linear scenario

The left plot in Figure 4.10 shows the relationship between scaled p_D and the difference in the β_1 estimates from JM to CC, this time using Scores.linear. There is a similar relationship to that in Figure 4.9 for MVN. The right plot suggests that scaled p_D is also indicative of the size of the change in the % bias of β_1 . However, it tells us nothing about the direction of this change, JM β_1 could be closer or further away from the TRUE β_1 than the CC β_1 . So p_D is not helpful in determining whether the estimation of the model of interest slope parameter has improved. We should not assume that higher values of p_D necessarily indicate ‘better’ β_1 estimates, but *interpret it as an indicator of the magnitude of effect of adding a missingness model on their estimation.*

Figure 4.10: Scores.linear: the relationship of scaled p_D to the slope in the model of interest



% bias reduction is the absolute percentage reduction in bias from CC to JM, and is calculated as $\% \text{ bias reduction} = \text{abs}(\text{CC } \% \text{ bias}) - \text{abs}(\text{JM } \% \text{ bias})$. Positive numbers indicate that JM is doing better in terms of bias than CC, while negative numbers indicate that JM is doing worse.

4.7.2 Relationship between scaled p_D and the fitted slope of the MoM, θ_1

MVN scenario

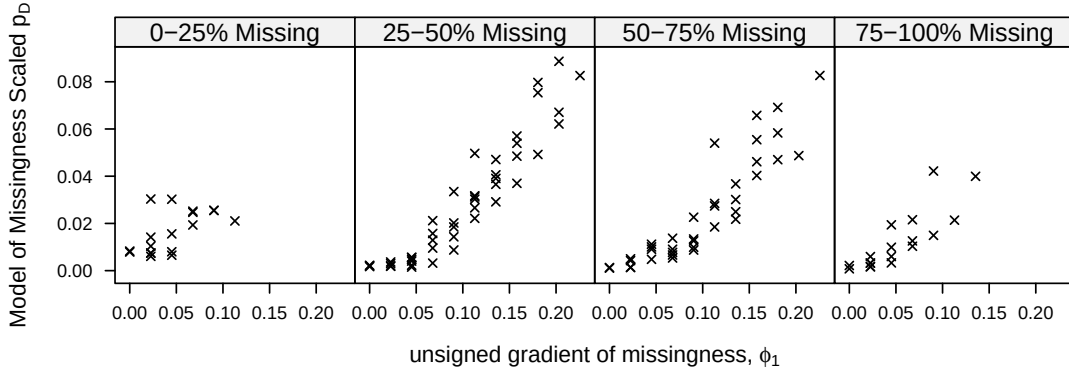
From the right plot in Figure 4.9 we see that there is also a relationship between scaled p_D and the fitted slope, θ_1 , of the missingness model, indicating that scaled p_D increases as the strength of the association between the probability of missingness and the true value of the response increases.

4.7.3 Relationship between scaled p_D and the gradient of missingness, ϕ_1

Scores.linear scenario

Figure 4.11 shows how scaled p_D increases as the magnitude of the gradient of missingness increases, and that for similar gradients of missingness, scaled p_D tends to decrease as the percentage of missingness increases. If we replace the gradient of missingness by the fitted slope of the model of missingness, θ_1 , which is on the logit scale, we find a sharper version of the same relationship. This is consistent with our MVN simulation findings.

Figure 4.11: Scores.linear: the relationship of scaled p_D with the gradient of missingness

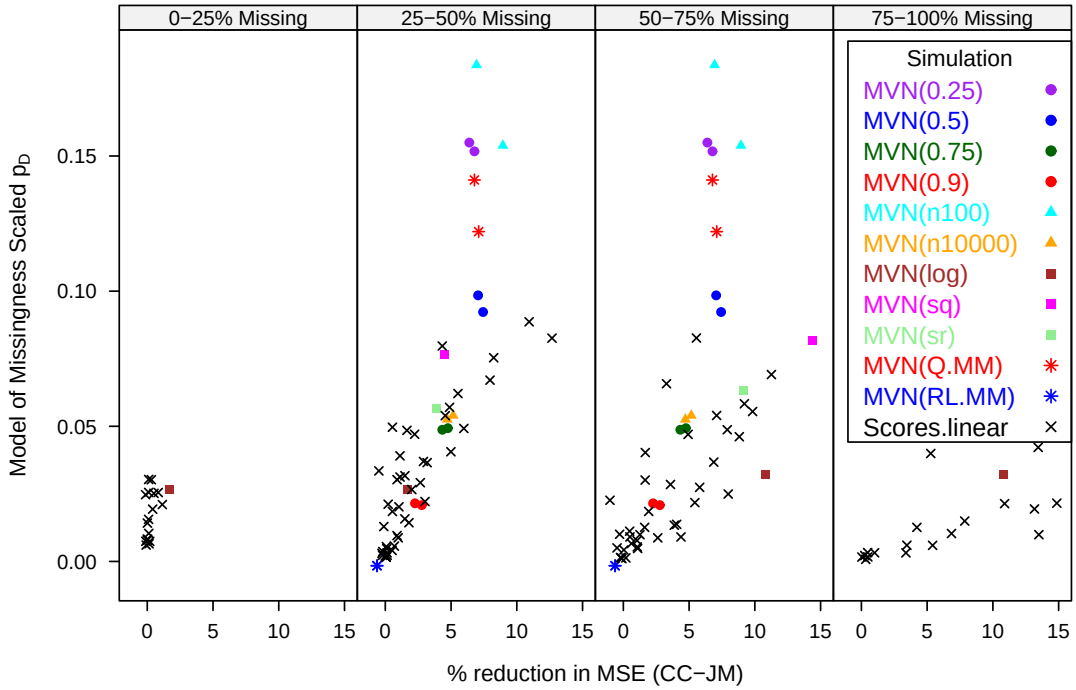


4.7.4 Relationship between scaled p_D and reduction in MSE from CC to JM

Scores.linear scenario

From the black crosses in Figure 4.12 we see that the percentage reduction in MSE from CC to JM increases as scaled p_D increases, so scaled p_D is correlated with the improvement in overall fit, as measured by MSE, from CC to JM. Since we have also seen that scaled p_D is correlated with the gradient of missingness, one interpretation is that this reflects the amount of information in the missingness model that can be used to improve the fit of the model of interest. The higher scaled p_D , the further our missing data is from MAR, and the greater the potential for extracting information from the joint model.

Figure 4.12: The relationship of scaled p_D to the % reduction in MSE from CC to JM



The MVN points are means across multiple repetitions and are shown in two panels if they are close to a boundary (all simulations except MVN(sq) and MVN(sr)). MVN(0.25), MVN(0.5), MVN(0.75) and MVN(0.9) are all simulations with 1000 records and the correlation shown in brackets; MVN(n100) and MVN(n10000) are both simulations with correlation 0.5, but have 100 and 10,000 records respectively; MVN(log), MVN(sq) and MVN(sr) are simulations with correlation 0.5, but the response was transformed using the log, square and square root transform respectively; MVN(Q.MM) and MVN(RL.MM) are both simulations with correlation 0.5, but have a quadratic and restricted linear missingness model respectively. Two points are shown for each MVN simulation, one for posMNAR and one for negMNAR. All unconverged runs are excluded from the calculations.

MVN - does the correlation between response and covariate affect the relationship?

The purple, blue, green and red circles in Figure 4.12 show the mean model of missingness scaled p_D against the mean % reduction in MSE from CC to JM taken over the 50 replicates from the MVN simulations with 0.25, 0.5, 0.75 and 0.9 correlation between the response and covariate respectively. There are two points for each simulation, one for posMNAR and one for negMNAR, which are always close together. The mean percentage of missingness for these simulations is 50%, and so they have

been placed in both the 25-50% missing and 50-75% missing panels, where they generally reinforce the pattern seen with Scores.linear. The exception is the simulations with 0.25 correlation, which have higher scaled p_D for the level of MSE reduction than seen with the Scores.linear simulations. So there is some evidence that the relationship between scaled p_D and reduction in MSE from CC to JM is affected by the strength of the correlation between response and covariate.

MVN - does the number of records affect the relationship?

To see what impact the number of records in a dataset has on the magnitude of scaled p_D , two more simulations are carried out on MVN, both with 0.5 correlation. The first, MVN(n100), has 50 replicates with 100 records, and all but one run converge. The second, MVN(n10000), has 10 replicates with 10,000 records, the reduced number of replicates being necessitated by computational speed. The posMNAR and negMNAR mean points are shown on Figure 4.12 by the orange and cyan triangles. The small number of records leads to higher scaled p_D for the level of MSE reduction seen previously. The larger number of records is positioned with the other points, but below the MVN(0.5) blue circles which are based on scenarios with 1,000 records. This suggests that the number of records does affect the relationship, despite having attempted to adjust for sample size by scaling.

MVN - does transforming the response affect the relationship?

The means for the MVN simulations using log, square and square root transforms of the response (shown as brown, pink and light green squares) are positioned within or close to the black crosses, which suggests that the relationship is not affected by transforming the response.

MVN - does the choice of model of missingness affect the relationship?

When the restricted linear missingness model is used (blue stars), both scaled p_D and % reduction in MSE are close to 0, but the position of the quadratic missingness model points (red stars) are slightly higher than the main body of points. So the choice of model of missingness may affect the relationship.

Scores.linear - does changing the model of interest affect the relationship?

As a final experiment, the Scores.linear simulation is rerun with the responses artificially dichotomised and a logistic regression model of interest fitted. An equivalent plot to Figure 4.12 shows similar shape and variable range. This provides some evidence that the relationship is robust to the choice of model of interest.

To summarise, this research suggests that Figure 4.12 is not data or model specific and provides some idea of the magnitude of scaled p_D in certain circumstances, although the number of data points, percentage of missing data and strength of the relationship of the model of interest all have some effect. When scaled p_D is about 0.02 we expect that a joint model will not change the fit of our model of interest very much, but if it is about 0.1 then we expect the joint model to make a substantial

difference.

In conclusion, scaled p_D provides an indication of how far the missing responses are from MAR, given the model assumptions. As the model of interest influences the estimation of the model of missingness parameters through $\mathbf{y}_{mis}$ and vice versa, we can never consider one part of the model in isolation and our use of scaled p_D must acknowledge this. Scaled p_D appears useful as a measure of overall improvement in fit, but is not so helpful in determining whether the estimation of individual parameters has improved.

4.7.5 Use of the model of missingness part of DIC_W

As discussed in Section 4.6, the model of missingness part of DIC_W may be useful in comparing joint models with the same model of interest but different models of missingness. To gain some insight into its usefulness, we calculate the mean DIC for the model of missingness for the 43 MVNsim repetitions with posMNAR missingness which converged with linear, restricted linear and quadratic missingness models. This gives values of 1311, 1385 and 1324 respectively, so the correct missingness model has the lowest DIC. Although from our interpretation of the scaled p_D of the missingness model we might be tempted to prefer the quadratic missingness model because of its higher value, the DIC points us back towards the linear missingness model. This suggests that both the scaled p_D and the DIC for the model of missingness DIC_W should be used to help evaluate our models.

Our investigations with simulated missingness suggest that the DIC_O , and the model of missingness part of DIC_W and its associated scaled p_D can provide different insights into proposed models. For the DIC_O , we are restricted in our choice of plug-ins to ensure consistency in the calculation of the posterior mean deviance and the plug-in deviance, but have more options for the DIC_W . However, we cannot have confidence in any of these measures unless we are sure of the adequacy of these plug-ins in terms of skewness and multi-modality. We consider their practical usage further in a real data example in Section 5.2, and our proposed strategy for their use is shown at the end of Chapter 5 and discussed in Chapter 9.

Summary:**exploring Bayesian full probability modelling using simulated missingness**

For non-ignorable missingness, complete case analysis results in bias in the model of interest parameter estimates and adversely affects the overall fit of the model of interest. Bayesian joint models reduce this bias and improve the fit, provided the correct model of interest and model of missingness are specified. Unfortunately, we can never be sure that either model is correctly specified. The two most critical assumptions concern the error distribution of the model of interest and the functional form of the model of missingness.

If we incorrectly specify a linear model of missingness, then little further harm is done to the fit of the model of interest (compared to complete case analysis), but the potential benefits from using a joint model are reduced or lost. However, incorrectly specifying more complex models of missingness, for example with quadratic or piecewise linear forms, can add bias and lead to a further deterioration in fit. Even if correctly specified, the extra parameters in these more complex models are very difficult to estimate with non-informative priors. From our investigations, piecewise linear models appear a more promising option than quadratic models, but should only be implemented if we have expert knowledge to justify and inform their use.

By contrast, the effects of misspecifying the error distribution of the model of interest are much less predictable. In this case, there are two sources of skewness in our model, and as Skinner comments in discussion of Diggle and Kenward (1994), disentangling informative non-response and distributional skewness is difficult. The bias in the β_0 parameter is still mostly removed and the MSE reduced with a joint model. However, the estimation of β_1 is not so robust and the behaviour of the joint model depends on whether or not the two sources of skewness are in conflict. If they are in conflict, we get a reduction rather than removal of the bias, but if they are in the same direction the bias is greater than in a complete case analysis. This suggests that joint models need careful interpretation if our primary concern is the relationship between the response and a particular covariate rather than response and several covariates as in the case of prediction.

Summary continued:**exploring Bayesian full probability modelling using simulated missingness**

The conditional DIC automatically generated by WinBUGS, DIC_W , has asymptotic and coherency difficulties for selection models. However its model of missingness part can be used as part of a comparison of joint models with the same model of interest but different models of missingness. The associated scaled p_D can be used to get some idea of how far the missing data departs from MAR given that the other assumptions are correct, but is also affected by the size of the dataset, the proportion of missing data and the strength of the relationship in the model of interest. However, higher values should not necessarily be taken as an indication of ‘better’ model of interest parameter estimates, but of the magnitude of effect of the missingness model on their estimation.

Proposed DIC based on the observed data likelihood, DIC_O , and the full data likelihood, DIC_F , need to be calculated from WinBUGS output using other software. DIC_O , which we have successfully implemented, appears best suited to comparing joint models with the same model of missingness but different models of interest. Implementation of DIC_F proved more difficult, and our attempts were unsuccessful.

Chapter 5

Bayesian full probability modelling of real data with missing responses

We now apply our Bayesian joint models to the three real datasets introduced in Chapter 3, continuing to focus on missing response and assuming that the missingness mechanism is non-ignorable. Where these datasets also have missing covariates, we either exclude the records with covariate missingness or impute the missing covariates using simplistic assumptions. Chapter 6 considers more realistic models to handle both missing response and missing covariates. As before we will compare CC and JM, but we can no longer run model TRUE. In our third example which models MCS income, we form an initial model to serve as a foundation for the increasingly complex models that will be built in subsequent chapters.

5.1 Example 1: NCDS educational test scores

Our first real example builds models using the NCDS educational test score data, which were introduced in Section 3.1.1. As this dataset underpins the scores data used extensively in the previous chapter, it allows us to start by using the models explored with the synthetic missingness. Hence our joint model continues to consist of a model of interest of the form given by Equation 1.8, i.e.

$$y_i \sim N(\mu_i, \sigma^2),$$
$$\mu_i = \beta_0 + \sum_{k=1}^p \beta_k x_{ki},$$

and a linear logit response model of missingness as specified by Equation 4.1, i.e.

$$m_i \sim \text{Bernoulli}(p_i),$$
$$\text{logit}(p_i) = \theta_0 + \theta_1 y_i.$$

Vague priors are specified for the model of interest parameters and mildly informative priors specified for the model of missingness parameters, as for the simulation studies in Chapter 4. Since we are

still focussing on the impact of missing response, each analysis uses only the records from the NCDS dataset in which the relevant covariates are fully observed.

In this example, all the models described are run with 3 chains for 15,000 iterations of which 10,000 are burn-in and a thinning rate of 5. According to our usual criteria, as discussed at the beginning of Chapter 4, all the runs converged.

5.1.1 Model of interest with single covariate

We begin by using a model of interest which has a single covariate as used throughout Chapter 4, and regress the mathematics test score at age 16 against the mathematics test score at age 11. In this analysis, we discard the 4,429 records which have missing age 11 mathematics test scores (24% of the full data), leaving 14,129 records for JM. 27% of the records used for JM are discarded for fitting CC, leaving 10,312 records for which both test scores are observed. Each model is run five times, with the response transformed according to the Box Cox transformation (Equation 4.8), with λ_2 always set to 2 to ensure that the response is always positive, and λ_1 taking values of 0 to 1 at 0.25 intervals. The observed data suggests that the response is normalised when λ_1 is a half. The regression parameter estimates change monotonically as λ_1 changes, and so the results for only three of the runs (no transform, square root transform and log transform) are shown.

From the parameter estimates in Table 5.1 we see that the addition of a missingness model results in a small decrease in the constant parameter and minimal increase to the slope parameter, regardless of the transformation of the response. The θ_1 parameter from the model of missingness provides evidence that lower test scores are more likely to be missing. Intuitively, this seems reasonable.

Table 5.1: NCDS example (single covariate): parameter estimates

Parameter	Box-Cox λ_1	CC		JM		% diff ^a
β_0	0	0.54	(0.53,0.55)	0.50	(0.49,0.50)	-8.7
MoI	0.5	0.73	(0.72,0.74)	0.68	(0.67,0.69)	-7.6
intercept	1	1.00	(0.99,1.01)	0.92	(0.91,0.94)	-7.6
β_1	0	0.42	(0.41,0.43)	0.43	(0.42,0.44)	1.3
MoI slope	0.5	0.55	(0.54,0.56)	0.56	(0.55,0.57)	1.1
(for age 11 score)	1	0.77	(0.75,0.78)	0.78	(0.76,0.79)	1.0
θ_0	0			-0.78	(-0.83,-0.73)	
MoM	0.5			-0.75	(-0.81,-0.70)	
intercept	1			-0.75	(-0.80,-0.69)	
θ_1	0			-0.46	(-0.55,-0.38)	
MoM slope	0.5			-0.38	(-0.45,-0.31)	
(for response)	1			-0.29	(-0.34,-0.24)	

Table shows the posterior mean, with the 95% interval in brackets.

^a % difference in parameter estimate from CC to JM.

Scaled p_D for the model of missingness, calculated using the *logitp* plug-ins, is between 0.007 and 0.008 for the joint models run with and without transforming the response. These values are very slightly lower if θ plug-ins are used. In neither case does this provide much evidence of informative missingness in this application.

There seems to be a contradiction between the messages from the θ_1 estimates and the scaled p_D . The θ_1 estimates are significantly negative which we might interpret as evidence of informative missingness, while the scaled p_D is tiny suggesting minimal informative missingness (consistent with the very small changes in the model of interest parameters when a model of missingness is added). The explanation lies in the high correlation (0.77) between the covariate and response in this dataset.

In this longitudinal example, x is a good proxy for y , so we could fit a MAR model of missingness, using the age 11 mathematics test score as the regressor, i.e. $\text{logit}(p_i) = \theta_0 + \theta_2 x_i$, instead of $\text{logit}(p_i) = \theta_0 + \theta_1 y_i$. Diagrammatically, this means replacing the arrow from y to p in Figure 2.3 with an arrow from x to p . The two models can now be estimated separately. In this case, the model of missingness p_D will just be the number of θ parameters and hence scaled p_D should be 0 by definition. From Table 5.2 we see that the 95% intervals for the slope parameters for the MAR model (θ_2) and MNAR model (θ_1) just overlap. So in this simple case, there is not much added value in using y over x .

Table 5.2: NCDS example (single covariate): MAR v MNAR model of missingness parameter estimates

	MAR		MNAR ^a	
θ_0 (MoM intercept)	-1.00	(-1.04,-0.97)	-0.75	(-0.80,-0.69)
θ_1 (MoM slope: age 16 score)			-0.29	(-0.34,-0.24)
θ_2 (MoM slope: age 11 score)	-0.22	(-0.25,-0.18)		

Table shows the posterior mean, with the 95% interval in brackets

^a MNAR model is JM with Box-Cox $\lambda_1 = 1$

This demonstrates that a clearly non-zero coefficient associated with a partially observed variable in a model of missingness is not in itself indicative of informative missingness. There may be other fully observed variables not included in the model of missingness which explain much of the missingness. However, if the scaled p_D also differs significantly from 0, then we do have evidence of informative missingness given the model assumptions.

5.1.2 Expanded model of interest

Next we rerun the models for the five transforms using an expanded model of interest with multiple covariates, where mathematics test score at 7 and social class at age 11 are added to the mathematics test score at 11. This is one of the models used by Goldstein (1979). Following Goldstein, we aggregate social class into three groupings: non-manual workers (social classes I, II and III non-manual), skilled

and semi-skilled manual workers (social classes III manual and IV) and unskilled manual workers (social class V). After discarding all records in which one or more of the covariates are missing, we are left with 10,944 records for JM, of which 25% cannot be used for CC. Our model of missingness is unchanged from Equation 4.1, having no additional covariates.

Again the results are monotonic across the λ_1 values, and so only a subset are given in Table 5.3. As with the simpler model, there is evidence that lower test scores are more likely to be missing, but apart from the constant all the parameter estimates are very similar for CC and JM. The missingness model gradient is slightly shallower for the enlarged model of interest.

Table 5.3: NCDS example (multiple covariates): parameter estimates

Parameter	Box-Cox λ_1	CC		JM		% diff ^a
β_0	0	0.60	(0.58,0.62)	0.56	(0.54,0.58)	-6.6
MoI	0.5	0.82	(0.80,0.84)	0.77	(0.75,0.79)	-5.7
intercept	1	1.12	(1.10,1.15)	1.06	(1.04,1.09)	-5.6
β_1	0	0.38	(0.37,0.40)	0.39	(0.38,0.40)	1.4
MoI slope	0.5	0.50	(0.49,0.51)	0.51	(0.49,0.52)	1.2
(for age 11 score)	1	0.69	(0.68,0.71)	0.70	(0.68,0.72)	1.2
β_2	0	0.05	(0.04,0.06)	0.05	(0.04,0.06)	-0.1
MoI slope	0.5	0.06	(0.05,0.07)	0.06	(0.05,0.07)	-0.4
(for age 7 score)	1	0.07	(0.06,0.09)	0.07	(0.06,0.09)	-0.6
β_3	0	-0.09	(-0.11,-0.07)	-0.09	(-0.11,-0.07)	-0.1
MoI slope	0.5	-0.13	(-0.15,-0.10)	-0.13	(-0.15,-0.10)	-0.6
(social class skilled & semi-skilled)	1	-0.18	(-0.22,-0.15)	-0.18	(-0.22,-0.15)	-0.3
β_4	0	-0.15	(-0.19,-0.11)	-0.16	(-0.20,-0.11)	3.5
MoI slope	0.5	-0.18	(-0.23,-0.14)	-0.19	(-0.24,-0.14)	2.8
(for social class unskilled)	1	-0.24	(-0.30,-0.17)	-0.24	(-0.31,-0.18)	3.2
θ_0	0			-0.90	(-0.96,-0.85)	
MoM	0.5			-0.87	(-0.93,-0.81)	
intercept	1			-0.86	(-0.93,-0.80)	
θ_1	0			-0.41	(-0.50,-0.32)	
MoM slope	0.5			-0.35	(-0.43,-0.27)	
(for response)	1			-0.26	(-0.32,-0.20)	

Table shows the posterior mean, with the 95% interval in brackets.

^a % difference in parameter estimate from CC to JM.

Scaled p_D for the model of missingness is slightly lower using the expanded model of interest compared to the single covariate version (0.006 using the *logitp* plug-ins). This suggests that less information is extracted from each missing observation to improve the fit of the model of interest as additional covariates are added, as we would expect. The clear effect of social class on educational attainment remains robust to the sensitivities carried out.

As we have found little evidence of informative missingness in this example, we will not consider it further but move on to our next example.

5.2 Example 2: Antidepressant trial

For our second example, we analyse the antidepressant clinical trial data that were introduced in Section 3.2.

5.2.1 Models of interest for HAMD data

Exploratory plots (Figure 3.3) indicate a downwards trend in the HAMD score over time, so for our model of interest we follow Diggle and Kenward (1994) and regress HAMD against time, allowing a quadratic relationship and a different intercept for each centre, s.t.

$$\begin{aligned} y_{ictw} &\sim N(\mu_{ictw}, \sigma^2) \\ \mu_{ictw} &= \beta_c + \eta_t w + \xi_t w^2 \end{aligned} \tag{5.1}$$

where i =individual, t =treatment ($1, \dots, 3$), c =centre ($1, \dots, 6$) and w =week ($0, \dots, 4$). We use two variants of this model: an autoregressive model (Section 1.6.3) and a random effects model (Section 1.6.2). In the first (AR2) we assume an autoregressive covariance structure of order 2 on the residuals, as specified by Equation 1.10. In the second (RE) we allow individual random effects on the intercept, replacing β_c with β_i . Each β_i is drawn from a Normal distribution, with mean and variance dependent on the centre, $c(i)$, where individual i was treated. For both variants we assign vague priors to the unknown parameters: giving the regression coefficients $N(0, 10000)$ priors and the precision ($\frac{1}{\sigma^2}$) a $\text{Gamma}(0.001, 0.001)$ prior. In the RE version, $\beta_i \sim N(\gamma_{c(i)}, \rho_{c(i)}^2)$ with each $\gamma_{c(i)}$ assigned $N(0, 10000)$ priors and the hierarchical standard deviations $\rho_{c(i)}$ assigned noninformative uniform priors as suggested by Gelman (2005).

In this example, we are interested in any differences between the effects of the three treatments on the change in depression score (HAMD) over time. Hence our primary interest is the time trend of the HAMD score for each treatment, which is determined by the η_t and ξ_t parameters.

5.2.2 Models of missingness for HAMD data

We specify three models of missingness as detailed in Table 5.4, and assign a logistic prior to θ_0 and weakly informative Normal priors to all the other θ parameters as previously discussed (Section 4.1). The simplest form of informative drop-out is given by HAMD.MoM1 where missingness depends on the current value of the HAMD score, while the form of HAMD.MoM2 allows us to interpret θ_3 as the effect of the change in the HAMD score on the response missingness. HAMD.MoM3 has the same

form as HAMD.MoM2, but includes separate θ for each treatment, which allows treatment to directly affect the missingness process. Joint models combining these models of missingness with the RE and AR2 models of interest will be referred to as HAMD.JM*(RE) and HAMD.JM*(AR2) respectively.

Table 5.4: HAMD example: models of missingness

Model Name	Model of Missingness equation
HAMD.MoM1	$\text{logit}(p_{iw}) = \theta_0 + \theta_2 y_{iw}$
HAMD.MoM2	$\text{logit}(p_{iw}) = \theta_0 + \theta_1 y_{i(w-1)} + \theta_3 (y_{iw} - y_{i(w-1)})$
HAMD.MoM3	$\text{logit}(p_{itw}) = \theta_{0t} + \theta_{1t} y_{i(w-1)} + \theta_{3t} (y_{iw} - y_{i(w-1)})$

5.2.3 Model comparison for HAMD data

These six joint models and the models of interest on complete cases only, CC(RE) and CC(AR2), have been run with 2 chains for 110,000 iterations with 100,000 burn-in and thinning set to 10. All the models converged according to our usual criteria. The parameter estimates are shown in Tables 5.5 and 5.6. Note that the γ in Table 5.6 are the hierarchical means of the β_i across centres in the random effects models, and comparable with the β in Table 5.5. Adding a missingness model makes little difference to the β or γ estimates, but there are substantial changes in some of the η and ξ parameters associated with the effect of treatment over time. The positive θ_2 estimates for the HAMD.JM1 models suggest that drop-out is associated with high HAMD scores, while the negative θ_3 in the HAMD.JM2 models indicate that change in the HAMD score is informative, with individuals more likely to drop-out if their HAMD score goes down. These two complementary messages are that the more severely depressed subjects, and those for whom the treatment appears most successful are more likely to drop-out. The HAMD.JM3 models provide some evidence that the missingness process is affected by the treatment.

How much difference does the choice of model of interest make?

To see whether using AR2 or RE as our model of interest makes a difference, we compare the mean response profiles for each pair of models (Figure 5.1). For a complete case analysis the solid (AR2) and dashed (RE) lines for each treatment are almost identical, but there are small differences for HAMD.JM1 and HAMD.JM2, and these increase for the more complex HAMD.JM3. This is consistent with the discussion in Daniels and Hogan (2008) concerning the increased importance of correctly specifying the dependence structure in the model of interest when dealing with missing data.

How much difference does the choice of model of missingness make?

The slight upwards impact of the addition of the HAMD.MoM1 model of missingness to the AR2 model of interest can be seen by comparing the CC (solid lines) and JM1 (dot-dash lines) in Figure 5.2; the impact is hardly discernable when the RE model of interest is used. By contrast, there

Table 5.5: HAMD example: parameter estimates for AR2 models

	CC (AR2)		HAMD.JM1(AR2)		% ^a	HAMD.JM2(AR2)		% ^a	HAMD.JM3(AR2)		% ^a
β_1	21.73	(20.64,22.90)	21.69	(20.56,22.81)	-0.2	21.76	(20.61,22.88)	0.1	21.77	(20.65,22.91)	0.2
β_2	22.46	(21.45,23.48)	22.48	(21.34,23.58)	0.1	22.42	(21.35,23.52)	-0.2	22.46	(21.41,23.59)	0.0
β_3	19.36	(18.36,20.38)	19.34	(18.30,20.43)	-0.1	19.42	(18.39,20.45)	0.3	19.41	(18.39,20.42)	0.3
β_4	23.94	(22.90,25.01)	23.94	(22.83,25.00)	0.0	24.02	(22.88,25.18)	0.3	23.98	(22.95,25.02)	0.2
β_5	20.70	(19.67,21.75)	20.66	(19.59,21.72)	-0.2	20.78	(19.72,21.88)	0.4	20.75	(19.73,21.82)	0.2
β_6	20.81	(19.77,21.89)	20.87	(19.83,21.97)	0.3	20.71	(19.69,21.81)	-0.5	20.77	(19.71,21.85)	-0.2
η_1	-3.50	(-4.31,-2.64)	-3.53	(-4.35,-2.72)	0.9	-3.45	(-4.27,-2.62)	-1.5	-3.49	(-4.27,-2.71)	-0.3
η_2	-5.31	(-6.18,-4.51)	-5.15	(-6.04,-4.25)	-3.0	-5.56	(-6.45,-4.67)	4.6	-5.74	(-6.61,-4.88)	8.0
η_3	-3.71	(-4.53,-2.91)	-3.68	(-4.47,-2.89)	-0.9	-3.71	(-4.51,-2.92)	0.1	-3.72	(-4.51,-2.94)	0.3
ξ_1	0.33	(0.12,0.53)	0.37	(0.17,0.58)	13.1	0.26	(0.03,0.47)	-22.4	0.33	(0.11,0.56)	-0.9
ξ_2	0.65	(0.44,0.85)	0.64	(0.42,0.86)	-1.0	0.65	(0.45,0.86)	0.8	0.66	(0.44,0.87)	1.5
ξ_3	0.52	(0.32,0.72)	0.54	(0.34,0.74)	3.5	0.48	(0.28,0.68)	-7.8	0.50	(0.30,0.70)	-3.8
θ_0			-3.12	(-3.72,-2.53)		-3.19	(-3.80,-2.62)				
$\theta_0(t_1)$									-2.65	(-3.91,-1.58)	
$\theta_0(t_2)$									-3.75	(-5.20,-2.56)	
$\theta_0(t_3)$									-3.89	(-5.10,-2.81)	
θ_1						0.04	(0.00,0.09)				
$\theta_1(t_1)$									0.04	(-0.04,0.12)	
$\theta_1(t_2)$									0.01	(-0.10,0.10)	
$\theta_1(t_3)$									0.08	(0.00,0.15)	
θ_2			0.08	(0.04,0.11)							
θ_3						-0.14	(-0.27,-0.02)				
$\theta_3(t_1)$									0.00	(-0.21,0.27)	
$\theta_3(t_2)$									-0.34	(-0.59,-0.10)	
$\theta_3(t_3)$									-0.08	(-0.28,0.08)	

Table shows the posterior mean, with the 95% interval in brackets

^a % difference in parameter estimate from CC to JM

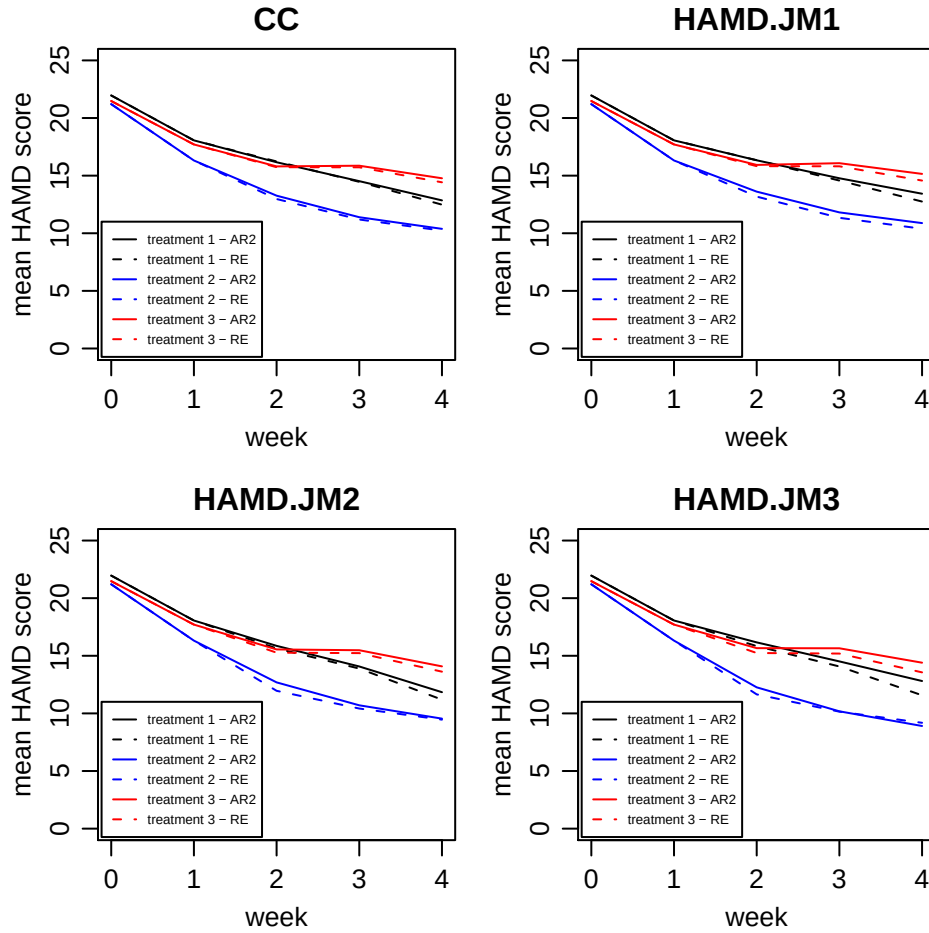
Table 5.6: HAMD example: parameter estimates for RE models

	CC (RE)		HAMD.JM1(RE)		% ^a	HAMD.JM2(RE)		% ^a	HAMD.JM3(RE)		% ^a
γ_1	22.45	(21.28,23.59)	22.46	(21.30,23.64)	0.0	22.45	(21.30,23.61)	0.0	22.47	(21.37,23.55)	0.1
γ_2	21.88	(20.77,23.03)	21.87	(20.77,22.99)	0.0	21.91	(20.80,23.03)	0.1	21.97	(20.91,23.05)	0.4
γ_3	19.19	(18.19,20.22)	19.16	(18.11,20.20)	-0.2	19.37	(18.33,20.47)	0.9	19.38	(18.32,20.49)	1.0
γ_4	24.08	(22.65,25.36)	24.08	(22.77,25.46)	0.0	24.18	(22.75,25.52)	0.4	24.14	(22.75,25.54)	0.2
γ_5	20.71	(19.52,21.92)	20.71	(19.49,21.96)	0.0	20.90	(19.67,22.18)	0.9	20.88	(19.66,22.07)	0.8
γ_6	20.09	(18.76,21.48)	20.12	(18.80,21.47)	0.1	19.94	(18.69,21.20)	-0.7	19.96	(18.73,21.23)	-0.6
η_1	-3.22	(-4.00,-2.44)	-3.21	(-4.00,-2.41)	-0.2	-3.29	(-4.10,-2.43)	2.3	-3.22	(-3.99,-2.39)	0.0
η_2	-5.47	(-6.29,-4.68)	-5.38	(-6.18,-4.55)	-1.6	-6.02	(-6.83,-5.15)	10.2	-6.20	(-7.08,-5.35)	13.5
η_3	-3.67	(-4.40,-2.93)	-3.63	(-4.36,-2.93)	-0.9	-3.86	(-4.61,-3.07)	5.2	-3.88	(-4.64,-3.12)	5.7
ξ_1	0.25	(0.05,0.46)	0.27	(0.06,0.48)	5.8	0.19	(-0.04,0.40)	-24.3	0.20	(-0.02,0.39)	-22.2
ξ_2	0.68	(0.48,0.90)	0.67	(0.46,0.88)	-1.6	0.77	(0.55,0.98)	12.1	0.79	(0.57,1.02)	15.8
ξ_3	0.51	(0.32,0.69)	0.51	(0.33,0.70)	0.4	0.50	(0.30,0.70)	-1.0	0.50	(0.31,0.70)	-1.0
θ_0			-2.61	(-3.22,-2.03)		-3.10	(-3.75,-2.50)				
$\theta_0(t_1)$									-2.22	(-3.22,-1.31)	
$\theta_0(t_2)$									-3.79	(-5.38,-2.41)	
$\theta_0(t_3)$									-3.57	(-4.87,-2.38)	
θ_1						-0.01	(-0.07,0.04)				
$\theta_1(t_1)$									-0.02	(-0.10,0.05)	
$\theta_1(t_2)$									-0.10	(-0.27,0.03)	
$\theta_1(t_3)$									-0.01	(-0.12,0.09)	
θ_2			0.04	(0.01,0.08)							
θ_3						-0.28	(-0.39,-0.18)				
$\theta_3(t_1)$									-0.17	(-0.32,-0.04)	
$\theta_3(t_2)$									-0.54	(-0.87,-0.30)	
$\theta_3(t_3)$									-0.32	(-0.54,-0.13)	

Table shows the posterior mean, with the 95% interval in brackets

^a % difference in parameter estimate from CC to JM

Figure 5.1: HAMD example: modelled mean response profiles - comparing the model of interest



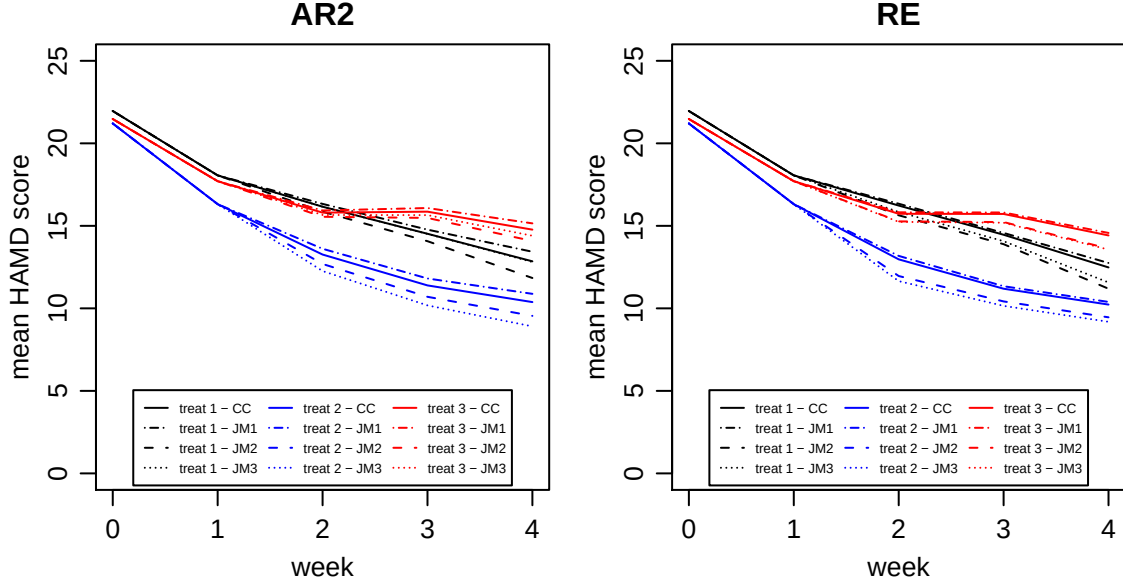
is a consistent downwards impact when HAMD.MoM2 is used for both models of interest (dashed lines). However, with HAMD.MoM3 (dotted lines) the profiles shift by different amounts resulting in increased differences between treatments, particularly for the AR2 model of interest.

What does the model of missingness scaled p_D tell us?

We start by choosing *logitp* plug-ins to calculate the model of missingness scaled p_D , and the results are shown in Table 5.7 (we will discuss the version of model of missingness scaled p_D based on θ plug-ins later). From our previous investigations with simulated data, we interpret the model of missingness scaled p_D as providing evidence of informative missingness for HAMD.JM2 and HAMD.JM3, but very little for HAMD.JM1. For HAMD.JM2 and HAMD.JM3 this evidence is stronger for the RE version of the joint model than the AR2. These findings are consistent with the amount of shift from CC in the mean response profiles shown in Figure 5.2.

We check the reasonableness of the *logitp* plug-ins by looking at their coefficients of skewness (Equation 4.14), as shown at the bottom of Table 5.8. They suggest that the *logitp* plug-ins are not a particularly good choice for these models.

Figure 5.2: HAMD example: modelled mean response profiles - comparing the model of missingness



In the AR2 plot, the CC and JM3 lines for treatment 1 are almost coincident.
In the RE plot, the CC and JM1 lines are almost coincident for all treatments,
and the JM2 and JM3 lines for treatment 3 are almost coincident.

Table 5.7: HAMD example: model of missingness scaled p_D

	θ plug-ins		<i>logitp</i> plug-ins	
	AR2	RE	AR2	RE
HAMD.JM1	0.009	-0.001	0.019	0.011
HAMD.JM2	0.016	0.230	0.114	0.275
HAMD.JM3	0.018	0.324	0.280	0.397

As an alternative, we choose θ plug-ins, and the resulting model of missingness scaled p_D s are shown in Table 5.7. There is still virtually no evidence of informative missingness for both versions of HAMD.JM1. We continue to suspect informative missingness for HAMD.JM2 and HAMD.JM3 for the RE version, though the evidence is a little weaker. However, a different picture emerges for the AR2 versions of HAMD.JM2 and HAMD.JM3, with the evidence of informative missingness disappearing for the θ plug-ins, which does not happen for the *logitp* plug-ins. However, there are also some question marks about some of the θ plug-ins in terms of skewness (Table 5.8), notably for the two HAMD.JM3 models.

In this example, both sets of plug-ins are suspect, and we are not particularly comfortable with either. It suggests that the model of missingness scaled p_D is not robust if the plug-ins used in its calculation are not sound. An extension of this work would be to investigate whether substantial prior belief giving rise to symmetric, informative priors helps reduce the skewness in the posterior distributions on which these plug-ins are based, and using alternative plug-ins such as posterior modes.

Table 5.8: HAMD example: skewness of posterior distributions of selected parameters

	H.JM1(AR2)	H.JM2(AR2)	H.JM3(AR2)	H.JM1(RE)	H.JM2(RE)	H.JM3(RE)
θ_0	-0.064	-0.079		-0.121	-0.159	
$\theta_0(t_1)$			-0.412			-0.212
$\theta_0(t_2)$			-0.686			-0.348
$\theta_0(t_3)$			-0.223			-0.260
θ_1		-0.198			-0.213	
$\theta_1(t_1)$			-0.038			-0.196
$\theta_1(t_2)$			-0.385			-0.902
$\theta_1(t_3)$			-0.092			-0.511
θ_2	-0.009			-0.001		
θ_3		-0.047			-0.245	
$\theta_3(t_1)$			0.400			-0.134
$\theta_3(t_2)$			-0.044			-0.756
$\theta_3(t_3)$			-0.282			-0.433
y_{mis}^a	0.001	0.041	0.097	-0.002	0.086	0.088
	(-0.093,0.090)	(-0.055,0.142)	(-0.035,0.290)	(-0.092,0.093)	(-0.044,0.205)	(-0.090,0.261)
$logitp^a$	-0.055	-0.325	-0.571	-0.105	-0.316	-0.518
	(-0.116,0.200)	(-0.722,0.459)	(-1.332,0.671)	(-0.208,0.297)	(-0.549,0.129)	(-1.089,0.419)

^a mean value is shown, with 95% interval below in brackets

Can DIC help us choose between models?

DIC_O is calculated for the six HAMD models using the Gold Standard DIC_O algorithm given in Appendix D. The runs using HAMD.MoM1 and HAMD.MoM2 take approximately 5 hours on a desktop computer with a dual core 2.4GHz processor and 3.5GB of RAM, while the more complex models with HAMD.MoM3 run in about 24 hours. As before we shall refer to the samples generated at steps 1 and 3 as our ‘Ksample’ and ‘Qsample’ respectively, with K and Q denoting the lengths of these samples. The Ksample is set to 2000, formed from 2 chains of 110,000 iterations, with 100,000 burn-in and the thinning parameter set to 10, and Q is set to 40,000. Table 5.9 shows DIC_O for our six models for the HAMD data. Recall that DIC_O is based on θ plug-ins, as *logitp* plug-ins do not produce a DIC with a sensible interpretation. The likelihood for the model of missingness is calculated for the weeks with drop-out, and for each of these weeks excludes individuals who have already dropped out.

Table 5.9: HAMD example: DIC_O

	$\bar{D}$	θ plug-ins		
		$\hat{D}$	p_D	DIC
HAMD.JM1(AR2)	9995.8	9978.6	17.2	10013.0
HAMD.JM1(RE)	9663.2	9359.6	303.7	9966.9
HAMD.JM2(AR2)	9991.0	9965.5	25.5	10016.5
HAMD.JM2(RE)	9680.6	9372.6	308.0	9988.5
HAMD.JM3(AR2)	9995.1	9965.0	30.1	10025.2
HAMD.JM3(RE)	9698.1	9392.1	306.0	10004.2

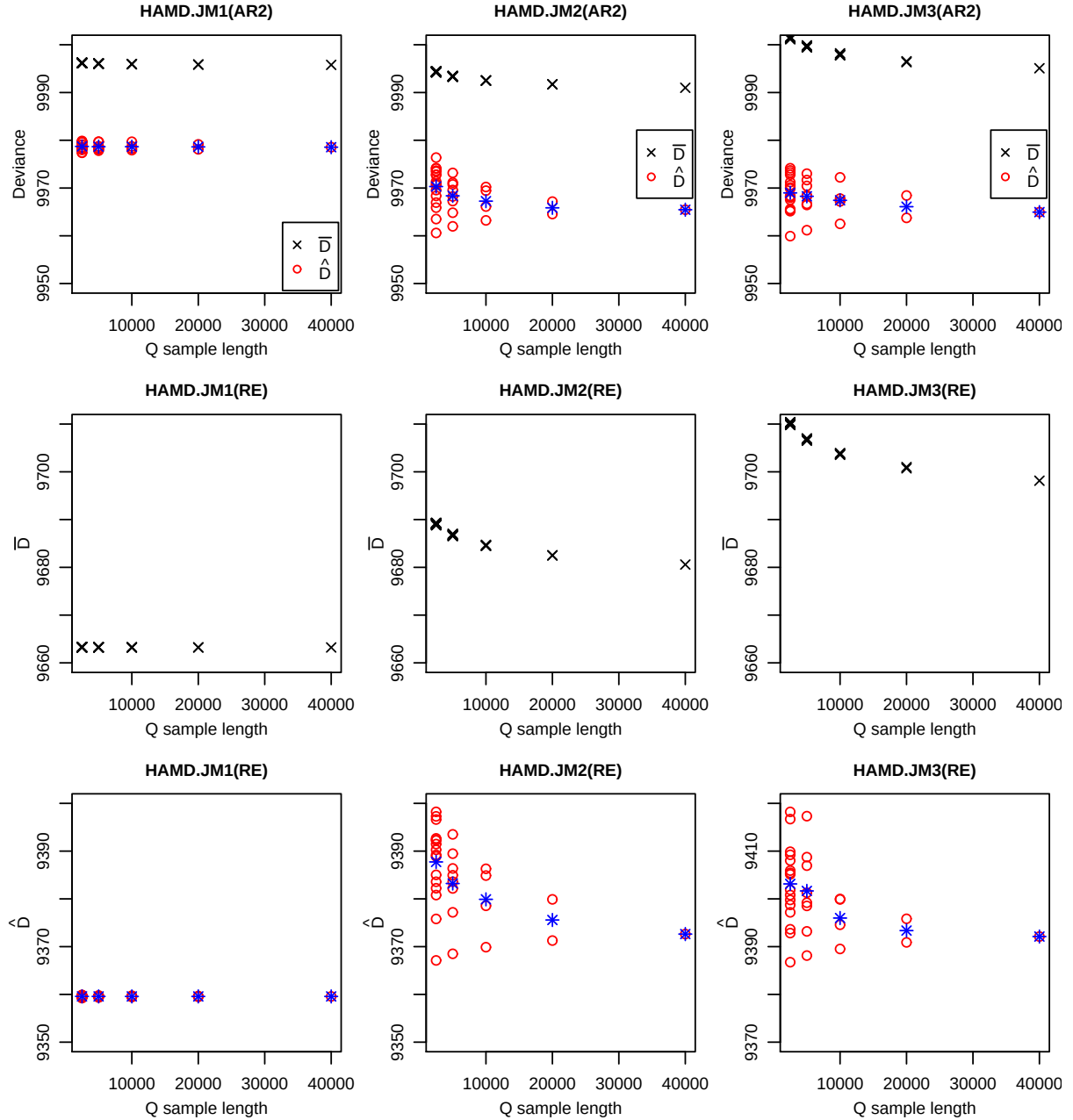
Before discussing these results, we check that they are soundly based. We have already commented on our concerns regarding the skewness in the posterior distributions of the plug-ins in the context of evaluating the scaled p_D . Here we examine the adequacy of our Qsample, by splitting it into subsets and plotting $\bar{D}$ and $\hat{D}$ against the sample lengths as described in Section 4.6.4. These plots for the six models are displayed in Figure 5.3. The mean and plug-in deviances are shown on the same plot for the AR2 models, but separate plots are used for the RE models, where the difference between the two deviances is much larger, to maintain consistent scales. We see that both $\bar{D}$ and $\hat{D}$ are stable and show little variation even for small Q for both HAMD.JM1 models. For the other models, trends similar to those exhibited by our synthetic data (see Section 4.6.4) are evident, but again there is convergence to a limit suggesting the adequacy of $Q=40,000$. As before, we also see that the instability associated with small Q decreases with increased sample size. The trends and variation are more pronounced for the RE models than the AR2 models.

Our investigation with simulated data suggests that DIC_O can give useful information about which model of interest to select. For the HAMD example DIC_O provides evidence that the random effects model of interest is preferable to the autoregressive model of interest when combined with each model of missingness. However, as discussed with the simulated examples, it should not be used to select a model of missingness because it only tells us about the fit to the observed data. If the model adjusts for informative missingness, then the fit of the observed data necessarily deteriorates, because in a selection model the same model of interest is applied to both the observed and the missing data. DIC_O is based on the joint model, with the model of interest and model of missingness contributing equally to differences in DIC_O between two joint models. The model of interest only takes account of the observed data, and the fit of this is worse if there is informative missingness. This can lead to a higher DIC_O if not offset by improvements in the fit of the model of missingness part, as happens with the HAMD models. Although DIC_O suggests that HAMD.JM1 fits the observed data best, we actually want to know which joint model fits all the data best. Since this is unknowable with missing data, a range of results reflecting different plausible models of missingness should be presented. DIC_F may be able to provide additional insight, but as discussed in Section 4.6 we are not able to calculate this due to technical issues.

Finally we look at the model of missingness part of DIC_W . Table 5.10 shows two versions, one automatically generated by WinBUGS using the θ plugins and the other calculated using *logitp* plugins. We have already discussed the model of missingness scaled p_D , which is not very robust to the choice of plug-ins in this example. In contrast, the message from the model of missingness DIC is consistent for both sets of plug-ins. HAMD.MoM3, used in HAMD.JM3(AR2) and HAMD.JM3(RE), provides a better fit to this part of the model than the two alternatives that have been used.

To conclude, DIC_O suggests that the RE model of interest is better than the AR2. For RE models,

Figure 5.3: HAMD example: $\bar{D}$ and $\hat{D}$ in DIC_O calculations



Calculations use θ plug-ins. The mean $\hat{D}$ at each value of Q are indicated by blue stars.

Table 5.10: HAMD example: model of missingness DIC_W

	$\bar{D}$	θ plug-ins			<i>logitp</i> plug-ins		
		$\hat{D}$	p_D	DIC	$\hat{D}$	p_D	DIC
HAMD.JM1(AR2)	698.6	695.5	3.1	701.8	694.3	4.3	703.0
HAMD.JM2(AR2)	653.4	649.5	3.9	657.3	636.5	16.9	670.2
HAMD.JM3(AR2)	626.0	621.8	4.2	630.2	583.1	42.9	668.9
HAMD.JM1(RE)	719.6	717.8	1.9	721.5	716.3	3.3	723.0
HAMD.JM2(RE)	547.5	517.7	29.8	577.3	511.2	36.3	583.8
HAMD.JM3(RE)	521.6	480.4	41.2	562.8	464.6	57.0	578.6

the model of missingness scaled p_D provides evidence of informative missingness that depends on the change in HAMD. This is also supported by substantial improvements in the model of missingness DIC_W for HAMD.JM2 and HAMD.JM3 over HAMD.JM1, i.e. HAMD.JM2 and HAMD.JM3 better explain the missingness pattern than HAMD.JM1. Overall, of the joint models explored, those with a RE model of interest and a model of missingness that depends on the change in HAMD (either treatment specific or not) seem most appropriate for the HAMD data.

If we based our analysis of this clinical trial data on a complete case analysis, we would conclude that treatment 2 lowers the HAMD score more than treatments 1 and 3 throughout the trial, and treatment 1 is more successful than treatment 3 in lowering HAMD in the later weeks. Using our preferred joint models, although all the treatments appear a little more effective in lowering HAMD, their ordering is unaltered compared to the CC analysis.

5.3 Example 3: MCS income

Our third example uses data from the first two sweeps of the MCS to predict income, and was introduced in Section 3.1.2. We do not use DIC_O for model comparison in this example, as the joint models become too complex for this to be done easily, since each model change requires adaption of the functions in the R software that we developed for calculating DIC_O .

5.3.1 Choosing an initial model of interest

The model of interest needs to take account of the design of the survey and the hierarchical structure of the data. We also need to choose a transform for our response variable and suitable error structure, and select a set of covariates. Each of these issues is considered in turn.

Transform of response and error structure

From our simulation studies, we know that the selection of an appropriate transformation for the response variable is very important when we are modelling missing data using Bayesian joint models. It is clear from Figure 3.2 that *hpay* should be transformed, and at this stage we base our choice on the observed values for both sweeps. We choose a transform based on maximising the log likelihood of a constant model fitted over a range of Box Cox power transformations (Equation 4.8). The maximum log likelihood is given by $\lambda_1 = 0.15, 0.08$ and 0.15 for the 512 observed sweep 1 values, 354 observed sweep 2 values and combined observed values for sweeps 1 and 2 respectively. As these values are close to 0, we take a log transformation, which is often used with income measures. We know that there are drawbacks in selecting a transform based purely on observed data, and this assumption will be revisited in Chapter 8. As regards the error structure, we start by considering Normal and t_4 errors.

Random effects

We choose between two models which account for the survey design and data structure. The first with individual and ward random effects, has the general form with non-informative priors given by Equation 5.2 and is shown diagrammatically in Figure 5.4.

$$\begin{aligned}
y_{it} &\sim N(\mu_{it}, \sigma^2) \text{ or } y_{it} \sim t_4(\mu_{it}, \sigma^2) \\
\mu_{it} &= \alpha_i + \vartheta_{w(i)} + \sum_{k=1}^p \beta_k x_{kit} + \sum_{k=p+1}^q \beta_k z_{ki} \\
\alpha_i &\sim N(0, \varsigma^2) \text{ individual random effects} \\
\varsigma &\sim N(0, 10000^2)I(0,) \\
\vartheta_{w(i)} &\sim N(\gamma_{s(i)}, \rho_{s(i)}^2) \text{ ward random effects} \\
\gamma_{s(i)} &\sim N(0, 10000^2) \\
\rho_{s(i)} &\sim N(0, 10000^2)I(0,) \\
\beta_k &\sim N(0, 10000^2) \\
\frac{1}{\sigma^2} &\sim \text{Gamma}(0.001, 0.001)
\end{aligned} \tag{5.2}$$

for $t = 1, 2$ sweeps, $i = 1, \dots, n$ individuals, $w = 1, \dots, 193$ wards and $s = 1, \dots, 9$ strata. $w(i)$ denotes the ward of individual i , and $s(i)$ the stratum of individual i .

$N(\text{mean}, \text{variance})I(0,)$ denotes a half Normal distribution restricted to positive values.

The second is a simplified model with individual random effects only and stratum specific intercepts, as given by Equation 5.3 and Figure 5.5.

$$\begin{aligned}
y_{it} &\sim N(\mu_{it}, \sigma^2) \text{ or } y_{it} \sim t_4(\mu_{it}, \sigma^2) \\
\mu_{it} &= \alpha_i + \gamma_{s(i)} + \sum_{k=1}^p \beta_k x_{kit} + \sum_{k=p+1}^q \beta_k z_{ki} \\
\alpha_i &\sim N(0, \varsigma^2) \text{ individual random effects} \\
\varsigma &\sim N(0, 10000^2)I(0,) \\
\gamma_{s(i)} &\sim N(0, 10000^2) \text{ stratum specific intercepts} \\
\beta_k &\sim N(0, 10000^2) \\
\frac{1}{\sigma^2} &\sim \text{Gamma}(0.001, 0.001)
\end{aligned} \tag{5.3}$$

for $t = 1, 2$ sweeps, $i = 1, \dots, n$ individuals and $s = 1, \dots, 9$ strata.

To select an initial model of interest, we run four models, detailed as MoIA to MoID in Table 5.11, on individuals who have a completely observed set of responses and covariates for both sweeps, excluding the seven who resulted from sweep 2 re-issues (322 complete cases). In all these models we use $\{age$ (main respondent's age), $hsize$ (household size), $kids$ (number of children), edu (educational level), reg (London/other), $sing$ (single/partner) $\}$ as our set of $\mathbf{x}$ covariates, and eth (ethnic group) as our

Figure 5.4: Graphical representation for model of interest with individual and ward random effects

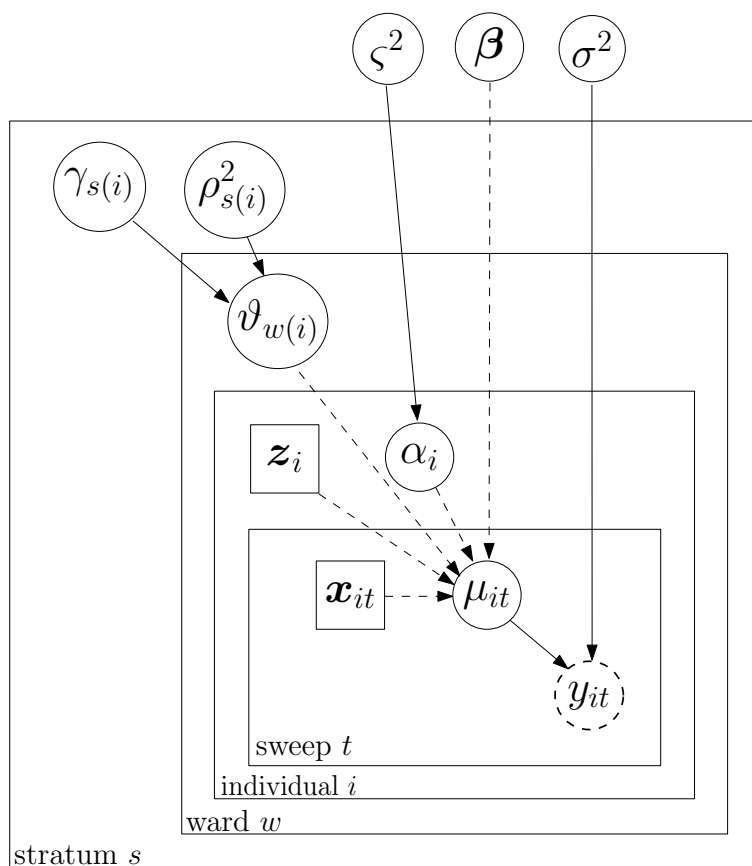
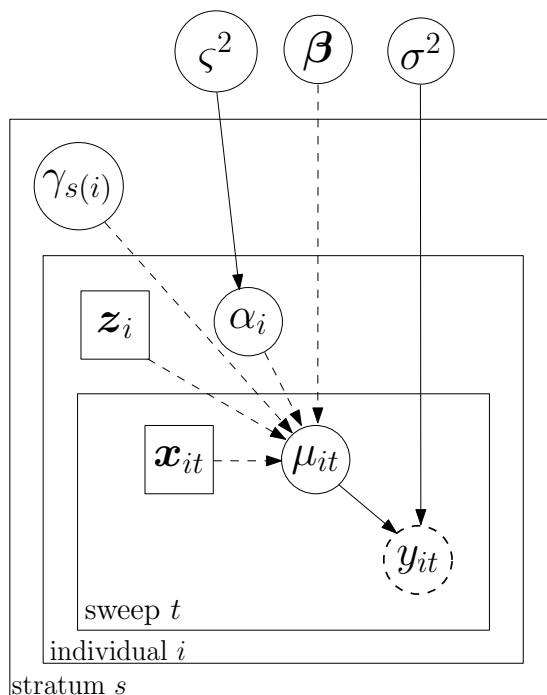


Figure 5.5: Graphical representation for model of interest with individual random effects and stratum specific intercepts



z covariate, assuming that a linear relationship holds for each continuous covariate (we return to this assumption later). The Gelman-Rubin convergence statistics for individual parameters suggest

convergence of all the runs, and a visual inspection of the trace plots is satisfactory.

Table 5.11: MCS income example: summary of models of interest

label	random effects	errors
MoIA	individual and ward random effects	Normal
MoIB	individual and ward random effects	t_4
MoIC	individual random effects and stratum specific intercepts	Normal
MoID	individual random effects and stratum specific intercepts	t_4

All models run with 2 chains for 500,000 iterations of which 400,000 are burn-in and a thinning rate of 10.

The parameter estimates are similar for the four models (a subset is shown in Table 5.12 and the full set can be found in Table C.1). All these models suggest that higher levels of hourly pay are associated with increasing age and higher levels of education, and lower levels of hourly pay are associated with being non-white, living outside London and gaining a partner between sweeps (recall that we are modelling mothers of young children, so intuitively, gaining a partner reduces their dependence on their own income). There is negligible evidence of an association with either the household size or the number of children. We examine the fit of each of these models using a series of residual plots, and consider their fit satisfactory for an initial model (examples are shown in Figures C.1 and C.2).

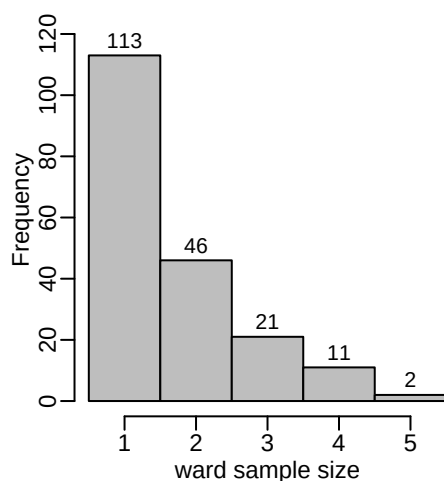
Table 5.12: MCS income example: subset of parameter estimates for models of interest MoIA-MoID

	MoIA		MoIB		MoIC		MoID	
β_{age}	0.11	(0.07,0.15)	0.12	(0.08,0.16)	0.11	(0.07,0.15)	0.12	(0.08,0.16)
β_{hsize}	0.00	(-0.04,0.04)	0.00	(-0.04,0.03)	0.00	(-0.03,0.04)	0.00	(-0.04,0.03)
β_{kids}	0.00	(-0.04,0.05)	0.00	(-0.04,0.05)	0.01	(-0.04,0.05)	0.01	(-0.04,0.05)
$\beta_{edu[1]}$	-0.17	(-0.33,0.00)	-0.14	(-0.31,0.03)	-0.17	(-0.33,0.00)	-0.15	(-0.31,0.02)
$\beta_{edu[2]}$	0.04	(-0.09,0.17)	0.07	(-0.07,0.21)	0.03	(-0.10,0.16)	0.06	(-0.07,0.20)
$\beta_{edu[3]}$	0.08	(-0.06,0.22)	0.10	(-0.05,0.25)	0.08	(-0.06,0.22)	0.09	(-0.05,0.24)
$\beta_{edu[4]}$	0.23	(0.09,0.38)	0.26	(0.11,0.40)	0.24	(0.10,0.38)	0.25	(0.10,0.40)
$\beta_{edu[5]}$	0.42	(0.15,0.69)	0.47	(0.20,0.74)	0.42	(0.15,0.69)	0.45	(0.19,0.72)
β_{eth}	-0.09	(-0.25,0.07)	-0.09	(-0.24,0.06)	-0.05	(-0.19,0.11)	-0.07	(-0.21,0.08)
β_{reg}	-0.21	(-0.37,-0.05)	-0.18	(-0.32,-0.03)	-0.20	(-0.34,-0.06)	-0.18	(-0.31,-0.05)
β_{sing}	-0.07	(-0.15,0.00)	-0.08	(-0.15,0.00)	-0.07	(-0.15,0.00)	-0.07	(-0.15,0.00)

Table shows the posterior mean, with the 95% interval in brackets

At this stage we wish to work with a single model of interest. As the inclusion of the ward random effects has little noticeable impact, we choose the simpler model with individual random effects and stratum specific intercepts. We expect that the extra layer of complexity given by the ward effects would be more necessary if the entire MCS cohort is being modelled. The effect of selecting a subset (Section 3.1.2) has resulted in many wards containing a single individual, and only 36% of individuals coming from wards with three or more individuals (see Figure 5.6). Further, we choose t_4 errors over Normal errors, for robustness to outliers. These decisions are supported by the DIC values calculated by WinBUGS, which are shown in Table 5.13 (MoIE will be discussed shortly, at the end of the paragraphs headed “Variable selection”).

Figure 5.6: MCS income example: wards by sample size



Based on 322 complete case individuals

Table 5.13: MCS income example: DIC for model of interest

	$\bar{D}$	$\hat{D}$	pD	DIC
MoIA	359.95	205.97	153.98	513.93
MoIB	283.95	94.38	189.57	473.52
MoIC	358.58	209.56	149.02	507.60
MoID	285.98	104.38	181.60	467.58
MoIE	283.28	102.90	180.38	463.67

values automatically generated by WinBUGS

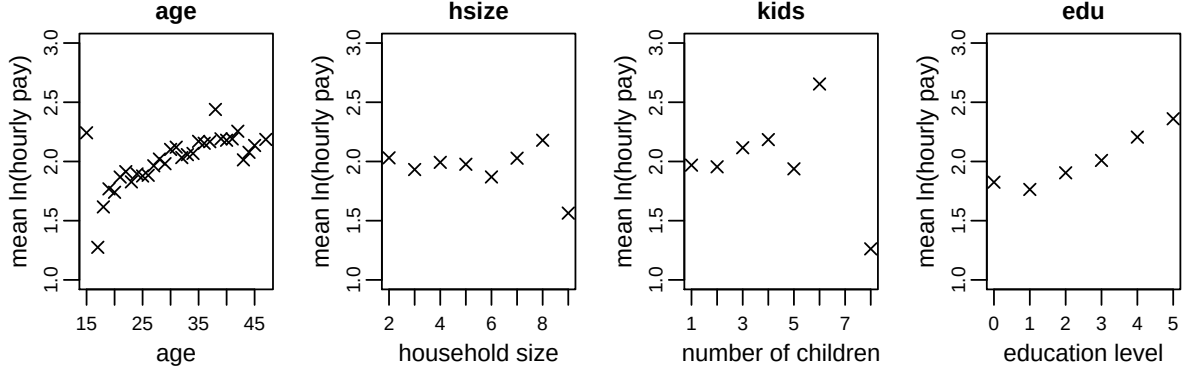
Variable selection

Next we consider which covariates should be included in our model of interest and the form they should take in a little more detail. Figure 5.7 suggests a quadratic relationship between age and income, and provides evidence for excluding both household size and number of children (the erratic points for large household size and high numbers of children are based on one or a few individuals). As regards educational level, mean pay for individuals with no qualifications is slightly higher than those with level 1 qualifications, but otherwise mean pay increases with educational level.

To investigate further, we run a model with a *reversible jump variable selection* sub-model (Lunn *et al.*, 2005; Lunn *et al.*, 2006). This is based on MoID in that it assumes t_4 errors and includes individual random effects and stratum specific intercepts, but has an expanded set of potential covariates within a variable selection sub-model. An age^2 term and $age \times edu$ interaction terms are the additional inclusions. All the other terms are linear as before. In a variable selection sub-model the number of selected covariates and the covariates themselves are unknown parameters, and the MCMC sampler ‘jumps’ between different settings.

We run this model in two forms: in one the binary components of the categorical covariates can be

Figure 5.7: MCS income example: mean income for selected covariates



selected separately (MoIRJ), and in the other none or all of its components must be selected (MoIRJg). The WinBUGS code for MoIRJ is given in Appendix E.2. WinBUGS output includes the posterior probability of different model structures and the marginal probability of inclusion for each variable. With so many possible combinations for MoIRJ, the highest ranked model structure had a posterior probability of 2.5%.

The marginal probabilities of inclusion for the covariates are detailed in Table 5.14. We choose $\{age, edu, reg, sing, eth\}$ as the covariate set for our initial model, i.e. all the covariates with a marginal probability of inclusion > 0.5 plus *eth*. We retain *eth* because it is required to answer one of our motivating questions (Section 3.1.2). Despite evidence for their inclusion, we exclude the $age \times edu$ interaction terms for simplicity, but will consider them further along with age^2 in Chapter 8.

Before considering other parts of the joint model, we run our selected initial model of interest, which we shall call MoIE, on the 322 complete cases. To recap, this model regresses the log of hourly pay (*hpay*) against covariates $\{age, edu, reg, sing, eth\}$, with individual random effects and stratum specific intercepts assuming t_4 errors. The parameter estimates are almost identical to MoID, which contained *kids* and *hsize*, (see Table C.1) and the DIC is slightly lower compared to MoID (Table 5.13). MoIE is run with 2 chains for 100,000 iterations of which 50,000 are burn-in and a thinning rate of 5, and the usual checks suggest convergence. The resulting effective samples sizes for estimating the posterior means are in the thousands for all parameters, apart from the stratum specific intercepts which are high hundreds (see Table C.2), so we consider this shorter run to be of adequate length. Ideally we would run our models for more iterations, but we have to be a bit pragmatic about the run length because of the number of models investigated.

5.3.2 Simplifying assumptions

As discussed in Section 3.1.2, our MCS income data has a complicated missingness pattern, with substantial levels of missingness in the covariates as well as the response, caused by both item and

Table 5.14: MCS income example: marginal probability of inclusion of covariates in models with reversible jump

	MoIRJ	MoIRJg ^a
<i>age</i>	0.982	0.969
<i>age</i> ²	0.475	0.451
<i>hsize</i>	0.101	0.094
<i>kids</i>	0.123	0.110
<i>edu</i> _[NVQ1]	0.980	1.000
<i>edu</i> _[NVQ2]	0.238	
<i>edu</i> _[NVQ3]	0.273	
<i>edu</i> _[NVQ4]	0.986	
<i>edu</i> _[NVQ5]	0.934	
<i>eth</i>	0.313	0.321
<i>sing</i>	0.675	0.695
<i>reg</i>	0.906	0.903
<i>age</i> × <i>edu</i> _[NVQ1] <i>interaction</i>	0.258	0.106
<i>age</i> × <i>edu</i> _[NVQ2] <i>interaction</i>	0.235	
<i>age</i> × <i>edu</i> _[NVQ3] <i>interaction</i>	0.249	
<i>age</i> × <i>edu</i> _[NVQ4] <i>interaction</i>	0.685	
<i>age</i> × <i>edu</i> _[NVQ5] <i>interaction</i>	0.356	

^a The *edu*_[NVQ1] and *age* × *edu*_[NVQ1] *interaction* probabilities apply to all *edu* and *age* × *edu* *interaction* levels respectively.

sweep missingness (Tables 3.4, 3.5 and 3.6). To avoid over-complexity at this stage, we now make some simplifying assumptions.

Limiting missingness to sweep 2

All individuals who are not completely observed in sweep 1 are excluded, so we are only concerned with modelling the missingness in sweep 2. Ways of extending our models to accommodate the sweep 1 missingness are discussed in Section 9.4. This assumption reduces our dataset to 505 individuals.

Imputing missing covariates in sweep 2

Our model will not run with missing covariates unless we specify a model for their imputation. However, we wish to avoid specifying a covariate model of missingness at this stage. Instead we impute values for the missing covariates using simplistic assumptions, before running a joint model consisting of a model of interest and model of missingness for the response. Of the six explanatory variables which we propose to use in our initial model of interest, two (*eth* and *stratum*) do not change between sweeps. Missing sweep 2 *age* is set to the individual's sweep 1 age, plus the mean difference in the ages between sweeps 1 and 2 for individuals with observed age at both sweeps. All the remaining missing covariates in sweep 2 are assumed to be unchanged from their sweep 1 values, and set accordingly. We now consider the implications for each in turn.

edu: from Table 5.15 we find that most of the individuals with observed *edu* in sweep 2 have the same educational level as they did in sweep 1. However, 5% do increase their educational level between sweeps by one or more levels, so we would prefer a more realistic imputation for this variable.

reg: of the 348 individuals with observed *reg* in sweep 2, only three moved from London to another region and just one moved into London. Therefore, setting the missing *reg* in sweep 2 to their sweep 1 values seems reasonable.

sing: a third of the individuals with observed *sing* in sweep 2 have a partner, and setting all the missing sweep 2 values to be single will reduce this proportion to 23%. So, this assumption needs to be revisited.

We return to the problem of generating more realistic imputations of missing covariates in Chapter 6.

Table 5.15: MCS income example: sweep 1 by sweep 2 counts of educational level for subset of 505 MCS individuals

	<i>edu</i> ₂ =0	<i>edu</i> ₂ =1	<i>edu</i> ₂ =2	<i>edu</i> ₂ =3	<i>edu</i> ₂ =4	<i>edu</i> ₂ =5	<i>edu</i> ₂ =missing
<i>edu</i> ₁ =0	25	0	2	0	0	0	23
<i>edu</i> ₁ =1	0	26	3	0	0	1	16
<i>edu</i> ₁ =2	0	0	118	6	1	0	63
<i>edu</i> ₁ =3	0	0	0	73	3	0	38
<i>edu</i> ₁ =4	0	0	0	0	75	1	23
<i>edu</i> ₁ =5	0	0	0	0	0	5	3

5.3.3 Adding a response model of missingness

We form eight joint models, JM.R1 to JM.R8, by adding a response model of missingness to MoIE. The response model of missingness takes the form $m_i \sim \text{Bernoulli}(p_i)$, with $\text{logit}(p_i)$ defined by one of the equations in Table 5.16, where m_i is a binary missing value indicator for *hpay*₂, set to 1 when hourly pay in sweep 2 for individual *i* is observed and 0 otherwise. (This is the reverse of the definition used for the missing value indicator in our previous examples, and the change is required for consistency with the expert elicitation which will be described in Chapter 7.) The inclusion of variables *sc* (social class), *eth* (ethnic group) and *ctry* (country) is suggested by Hawkes and Plewis (2008). Note that in the response model of missingness we use an untransformed version of *hpay*.

The priors for the θ and δ parameters are specified as $\theta_0 \sim \text{Logistic}(0, 1)$, $\theta_k \sim N(0, 10000^2)$ and $\delta \sim N(0, 0.68)$. All these models are run with 2 chains for 100,000 iterations of which 50,000 are burn-in and a thinning rate of 5, and the usual checks suggest convergence. The effective sample sizes for estimating the posterior means of the parameters of JM.R8 are all above 480 (see Table C.2), so

Table 5.16: MCS income example: details of joint models consisting of a model of interest and a response model of missingness

label	model of interest	equation/details for response model of missingness
JM.R1	MoIE	$\text{logit}(p_i) = \theta_0$
JM.R2	MoIE	$\text{logit}(p_i) = \theta_0 + \theta_{level}hpay_{i1}$
JM.R3	MoIE	$\text{logit}(p_i) = \theta_0 + \delta_{level}hpay_{i2}$
JM.R4	MoIE	$\text{logit}(p_i) = \theta_0 + \theta_{level}hpay_{i1} + \delta_{change}(hpay_{i2} - hpay_{i1})$
JM.R5	MoIE	JM.R1 + sc_1^a , <i>eth</i> and <i>ctry</i>
JM.R6	MoIE	JM.R2 + sc_1^a , <i>eth</i> and <i>ctry</i>
JM.R7	MoIE	JM.R3 + sc_1^a , <i>eth</i> and <i>ctry</i>
JM.R8	MoIE	JM.R4 + sc_1^a , <i>eth</i> and <i>ctry</i>

the second index on *hpay* indicates the sweep

^a sweep 1 social class is used

we consider the run length to be adequate for our investigative purposes.

Comparison of different models of missingness

The model of interest parameter estimates for these eight joint models are very similar, and can be found in Table 5.17. We will compare them to MoIE after discussing the model of missingness part of the joint models.

We start by considering the first four models, which do not include the variables *sc*, *eth* and *ctry*. In the MCAR model (JM.R1), θ_0 results in a constant probability of missingness of 0.37, reflecting the percentage of missingness. In the MAR model (JM.R2), we find very little evidence that missing sweep 2 pay is associated with the level of pay in sweep 1. However, the MNAR model (JM.R3) suggests that it is associated with low sweep 2 pay (the posterior mean of δ_{level} is 0.11, and a positive coefficient in the model of missingness indicates that sweep 2 pay is more likely to be missing for lower values of its associated covariate.). The second form of MNAR model (JM.R4) suggests a relationship with both the level of pay in sweep 1 and the change in pay between the two sweeps, with sweep 2 pay more likely to be missing for individuals with low pay in sweep 1 (posterior mean of $\theta_{level} = 0.13$) and a decrease in pay between sweeps (posterior mean of $\delta_{change} = 0.16$).

When we add other variables which are thought to be associated with the missingness into the response model of missingness, these findings remain although the 95% intervals of θ_{level} , δ_{level} and δ_{change} are wider. JM.R5-JM.R8 suggest that pay is more likely to be missing at sweep 2 if the individual has social class at level 2, 3 or 4 (although the evidence is weaker in the MNAR models) and if they are non-white. The evidence for country differences is weak.

Table 5.17: MCS income example: parameter estimates for joint models JM.R1-JM.R8

	JM.R1	JM.R2	JM.R3	JM.R4	JM.R5	JM.R6	JM.R7	JM.R8
σ	0.35 (0.32,0.39)	0.35 (0.32,0.39)	0.35 (0.32,0.39)	0.35 (0.32,0.39)	0.35 (0.32,0.39)	0.35 (0.32,0.39)	0.35 (0.32,0.39)	0.35 (0.32,0.39)
ς	0.23 (0.19,0.27)	0.23 (0.20,0.27)	0.23 (0.19,0.27)	0.23 (0.19,0.27)	0.23 (0.19,0.27)	0.23 (0.19,0.27)	0.23 (0.19,0.27)	0.23 (0.19,0.27)
γ_1	2.01 (1.83,2.19)	2.02 (1.84,2.19)	2.01 (1.84,2.18)	2.00 (1.82,2.17)	2.02 (1.84,2.19)	2.02 (1.84,2.20)	2.01 (1.84,2.19)	2.00 (1.83,2.17)
γ_2	1.98 (1.81,2.14)	1.98 (1.82,2.14)	1.96 (1.81,2.12)	1.95 (1.79,2.11)	1.98 (1.82,2.14)	1.98 (1.82,2.15)	1.97 (1.81,2.13)	1.96 (1.80,2.12)
γ_3	1.99 (1.79,2.19)	1.99 (1.79,2.19)	1.97 (1.77,2.16)	1.95 (1.75,2.15)	2.00 (1.80,2.19)	1.99 (1.79,2.20)	1.97 (1.78,2.17)	1.96 (1.76,2.16)
γ_4	2.04 (1.81,2.28)	2.05 (1.81,2.28)	2.02 (1.79,2.25)	2.01 (1.78,2.24)	2.05 (1.81,2.28)	2.03 (1.81,2.29)	2.03 (1.80,2.26)	2.01 (1.78,2.25)
γ_5	1.95 (1.76,2.13)	1.95 (1.77,2.13)	1.93 (1.76,2.11)	1.92 (1.74,2.10)	1.95 (1.77,2.13)	1.95 (1.76,2.14)	1.94 (1.76,2.12)	1.93 (1.75,2.11)
γ_6	1.95 (1.76,2.15)	1.96 (1.76,2.15)	1.95 (1.76,2.14)	1.94 (1.75,2.13)	1.96 (1.76,2.15)	1.96 (1.76,2.16)	1.95 (1.76,2.15)	1.94 (1.75,2.14)
γ_7	1.88 (1.70,2.06)	1.89 (1.70,2.06)	1.87 (1.70,2.05)	1.86 (1.68,2.04)	1.89 (1.70,2.06)	1.89 (1.70,2.07)	1.87 (1.70,2.05)	1.86 (1.68,2.04)
γ_8	1.94 (1.70,2.18)	1.94 (1.70,2.18)	1.93 (1.70,2.17)	1.92 (1.69,2.16)	1.94 (1.70,2.18)	1.94 (1.70,2.19)	1.93 (1.70,2.17)	1.93 (1.69,2.16)
γ_9	1.99 (1.81,2.17)	1.99 (1.81,2.17)	1.98 (1.80,2.15)	1.96 (1.79,2.14)	1.99 (1.81,2.17)	1.99 (1.81,2.18)	1.98 (1.81,2.16)	1.97 (1.79,2.15)
β_{age}	0.12 (0.08,0.15)	0.12 (0.08,0.15)	0.11 (0.08,0.14)	0.11 (0.08,0.14)	0.12 (0.08,0.15)	0.12 (0.08,0.15)	0.11 (0.08,0.15)	0.11 (0.08,0.14)
$\beta_{edu}[NVQ1]$	-0.03 (-0.17,0.12)	-0.03 (-0.18,0.11)	-0.03 (-0.18,0.11)	-0.03 (-0.17,0.11)	-0.03 (-0.17,0.11)	-0.03 (-0.18,0.12)	-0.03 (-0.17,0.11)	-0.03 (-0.17,0.11)
$\beta_{edu}[NVQ2]$	0.12 (0.00,0.23)	0.11 (0.00,0.23)	0.12 (0.01,0.23)	0.12 (0.01,0.23)	0.11 (0.00,0.23)	0.12 (0.00,0.23)	0.12 (0.01,0.23)	0.12 (0.01,0.23)
$\beta_{edu}[NVQ3]$	0.21 (0.08,0.33)	0.20 (0.08,0.32)	0.20 (0.08,0.32)	0.20 (0.08,0.32)	0.20 (0.09,0.33)	0.21 (0.08,0.33)	0.20 (0.08,0.32)	0.20 (0.08,0.32)
$\beta_{edu}[NVQ4]$	0.27 (0.15,0.39)	0.27 (0.15,0.39)	0.28 (0.16,0.40)	0.28 (0.16,0.41)	0.27 (0.15,0.40)	0.27 (0.15,0.40)	0.28 (0.16,0.40)	0.28 (0.16,0.41)
$\beta_{edu}[NVQ5]$	0.49 (0.25,0.72)	0.48 (0.25,0.72)	0.49 (0.26,0.72)	0.48 (0.26,0.72)	0.48 (0.25,0.72)	0.49 (0.25,0.72)	0.49 (0.26,0.72)	0.49 (0.26,0.72)
β_{eth}	-0.09 (-0.22,0.04)	-0.09 (-0.22,0.04)	-0.09 (-0.22,0.04)	-0.09 (-0.22,0.04)	-0.09 (-0.22,0.04)	-0.09 (-0.22,0.04)	-0.09 (-0.22,0.04)	-0.09 (-0.22,0.04)
β_{reg}	-0.14 (-0.26,-0.01)	-0.14 (-0.26,-0.02)	-0.14 (-0.26,-0.02)	-0.14 (-0.26,-0.01)	-0.14 (-0.26,-0.02)	-0.14 (-0.27,-0.02)	-0.14 (-0.27,-0.02)	-0.14 (-0.26,-0.02)
β_{sing}	-0.07 (-0.14,0.01)	-0.07 (-0.14,0.01)	-0.06 (-0.13,0.01)	-0.06 (-0.13,0.01)	-0.07 (-0.14,0.01)	-0.07 (-0.14,0.01)	-0.06 (-0.14,0.01)	-0.06 (-0.13,0.01)
θ_0	0.55 (0.37,0.73)	0.48 (0.09,0.87)	-0.31 (-1.19,0.56)	-0.38 (-1.25,0.47)	1.26 (0.78,1.76)	1.28 (0.62,1.95)	0.39 (-1.20,1.91)	0.11 (-1.49,1.73)
θ_{level}	0.01 (-0.03,0.05)			0.13 (0.01,0.26)		0.00 (-0.05,0.04)		0.13 (-0.04,0.31)
δ_{level}			0.11 (0.00,0.24)				0.09 (-0.05,0.26)	
δ_{change}				0.16 (0.02,0.31)				0.16 (-0.04,0.37)
$\theta_{ctry}[Wales]$					-0.14 (-0.70,0.41)	-0.14 (-0.71,0.43)	-0.15 (-0.72,0.42)	-0.15 (-0.73,0.44)
$\theta_{ctry}[Scotland]$				0.23 (-0.31,0.78)	0.22 (-0.31,0.78)	0.22 (-0.31,0.78)	0.24 (-0.31,0.80)	0.24 (-0.31,0.82)
$\theta_{ctry}[NI]$				0.20 (-0.37,0.77)	0.19 (-0.37,0.77)	0.19 (-0.37,0.77)	0.20 (-0.38,0.79)	0.21 (-0.37,0.81)
θ_{eth}				-0.80 (-1.34,-0.26)	-0.81 (-1.35,-0.27)	-0.81 (-1.35,-0.27)	-0.91 (-1.51,-0.33)	-0.96 (-1.58,-0.37)
$\theta_{sc[NS-SEC2]}$				-0.67 (-1.25,-0.11)	-0.68 (-1.25,-0.11)	-0.68 (-1.25,-0.11)	-0.48 (-1.15,0.17)	-0.40 (-1.07,0.26)
$\theta_{sc[NS-SEC4]}$				-0.81 (-1.63,-0.01)	-0.82 (-1.64,0.02)	-0.82 (-1.64,0.02)	-0.53 (-1.49,0.42)	-0.44 (-1.40,0.52)
$\theta_{sc[NS-SEC5]}$				-0.82 (-1.35,-0.32)	-0.83 (-1.37,-0.30)	-0.83 (-1.37,-0.30)	-0.55 (-1.25,0.16)	-0.44 (-1.17,0.28)

Table shows the posterior mean, with the 95% interval in brackets.

Initially we choose *logitp* plug-ins for calculating the DIC_W for the model of missingness for the eight joint models. The skewness in the *logitp* for the MNAR models is shown in Table 5.18, which includes the median in addition to the mean and 95% intervals, because the mean is inflated by the extreme skewness in the *logitp* posterior distributions for a few individuals in some models. In particular, for JM.R7 and JM.R8, the *logitp* mean lies outside the 95% interval, and examining their *logitp* trace plots we find that a few have very large spikes. As the *logitp* plug-ins are affected by skewness, we also consider using θ plug-ins, but find even more skewness in the missing *hpay* used in their calculation (see Table 5.18).

Table 5.18: MCS income example: skewness of posterior distributions for selected parameters

		JM.R3	JM.R4	JM.R7	JM.R8
logitp	mean	0.232	0.256	-1.402	-0.699
	median	0.338	0.378	0.092	0.154
	95% interval	(0.016,1.091)	(0.031,0.829)	(-1.061,0.483)	(-0.638,0.530)
hpay _{mis}	mean	-0.802	-0.903	-0.456	-0.645
	median	-0.620	-0.775	-0.396	-0.587
	95% interval	(-1.470,-0.340)	(-1.520,-0.440)	(-1.060,-0.076)	(-1.276,-0.267)

Table 5.19 shows the model of missingness DIC_W for all the joint models, calculated using both plug-ins. The choice of plug-in makes a difference to the DIC for the four MNAR models (JM.R3, JM.R4, JM.R7 and JM.R8), and although JM.R8 has the lowest DIC in both cases, the rankings of the other models are affected. As the *logitp* plug-ins are affected by lower levels of skewness than the θ plug-ins, we will use the *logitp* plug-ins from now on, although we do not totally trust either.

Table 5.19: MCS income example: model of missingness DIC_W for JM.R1-JM.R8

	$\bar{D}$	θ plug-ins				<i>logitp</i> plug-ins			
		$\hat{D}$	p_D	DIC	scaled p_D	$\hat{D}$	p_D	DIC	scaled p_D
JM.R1	664.5	663.5	1.0	665.5		663.5	1.0	665.5	
JM.R2	665.5	663.5	2.0	667.4		663.5	2.0	667.4	
JM.R3	649.7	634.2	15.5	665.2	0.073	641.7	8.0	657.7	0.032
JM.R4	641.4	619.1	22.3	663.7	0.104	628.9	12.5	653.9	0.051
JM.R5	649.5	641.5	8.0	657.6		641.5	8.0	657.6	
JM.R6	650.6	641.5	9.1	659.7		641.5	9.1	659.6	
JM.R7	639.7	622.9	16.8	656.5	0.042	623.6	16.1	655.8	0.038
JM.R8	630.0	604.8	25.2	655.3	0.082	608.1	21.9	651.9	0.064

Without the additional covariates, *sc*₁, *eth* and *ctry*, the models allowing informative missingness (JM.R3 and JM.R4) explain the missingness pattern better than either the MCAR model (JM.R1) or the MAR model (JM.R2) according to the DIC_W . When *sc*₁, *eth* and *ctry* are included, there is some evidence from the model of missingness DIC_W that JM.R8 is preferable. As we start to make

more realistic assumptions about the covariate missingness, we may find more evidence for informative missingness. We explore this possibility using JM.R8, which encompasses the simpler models.

The scaled p_D for the model of missingness for the MNAR models, which is also shown in Table 5.19, provides some evidence of informative missingness that depends either on the level of pay in sweep 2 or on the change in pay between sweeps, given the model assumptions. However, there are only very slight differences in the estimates of the model of interest parameters for the eight joint models (Table 5.17). We may find that this changes as we improve the covariate imputation.

Comparison with complete case analysis

We now compare joint model JM.R8 with a complete case analysis (MoIE), as shown in Table 5.20, and find that including partially observed individuals makes a difference. The largest impact is on the education covariate, with each level of qualification now having a stronger association with higher pay, particularly for levels 2 and 3. The anomaly of individuals with level 1 qualifications having lower pay than those without qualifications disappears.

JM.R8 provides a basic joint model consisting of two sub-models, a model of interest and a response model of missingness. Our next step is to add a third sub-model, a covariate model of missingness, so that imputing the covariates becomes part of the joint model. This is the subject of the next chapter.

Table 5.20: MCS income example: comparison of parameter estimates for a complete case analysis (MoIE) and a joint model (JM.R8)

	MoIE		JM.R8		% diff ^a
σ	0.33	(0.30,0.37)	0.35	(0.32,0.39)	5.8
ς	0.21	(0.17,0.25)	0.23	(0.19,0.27)	9.5
γ_1	2.14	(1.95,2.33)	2.00	(1.83,2.17)	-6.5
γ_2	2.10	(1.93,2.28)	1.96	(1.80,2.12)	-7.0
γ_3	2.09	(1.86,2.31)	1.96	(1.76,2.16)	-6.0
γ_4	2.18	(1.91,2.45)	2.01	(1.78,2.25)	-7.4
γ_5	2.07	(1.87,2.27)	1.93	(1.75,2.11)	-6.9
γ_6	2.06	(1.85,2.26)	1.94	(1.75,2.14)	-5.5
γ_7	2.08	(1.88,2.29)	1.86	(1.68,2.04)	-10.5
γ_8	2.06	(1.81,2.32)	1.93	(1.69,2.16)	-6.7
γ_9	2.10	(1.90,2.30)	1.97	(1.79,2.15)	-6.2
β_{age}	0.12	(0.09,0.16)	0.11	(0.08,0.14)	-10.2
$\beta_{edu[NVQ1]}$	-0.16	(-0.32,0.00)	-0.03	(-0.17,0.11)	-79.1
$\beta_{edu[NVQ2]}$	0.05	(-0.08,0.18)	0.12	(0.01,0.23)	116.6
$\beta_{edu[NVQ3]}$	0.09	(-0.06,0.23)	0.20	(0.08,0.32)	131.0
$\beta_{edu[NVQ4]}$	0.24	(0.10,0.38)	0.28	(0.16,0.41)	16.5
$\beta_{edu[NVQ5]}$	0.44	(0.18,0.70)	0.49	(0.26,0.72)	10.9
β_{eth}	-0.07	(-0.21,0.08)	-0.09	(-0.22,0.04)	26.1
β_{reg}	-0.18	(-0.31,-0.05)	-0.14	(-0.26,-0.02)	-22.7
β_{sing}	-0.07	(-0.14,0.00)	-0.06	(-0.13,0.01)	-20.2
θ_0			0.11	(-1.49,1.73)	
θ_{level}			0.13	(-0.04,0.31)	
δ_{change}			0.16	(-0.04,0.37)	
$\theta_{ctry[Wales]}$			-0.15	(-0.73,0.44)	
$\theta_{ctry[Scotland]}$			0.24	(-0.31,0.82)	
$\theta_{ctry[Northern Ireland]}$			0.21	(-0.37,0.81)	
θ_{eth}			-0.96	(-1.58,-0.37)	
$\theta_{sc[intermediate]}$			-0.40	(-1.07,0.26)	
$\theta_{sc[lower supervisory \& technical]}$			-0.44	(-1.40,0.52)	
$\theta_{sc[semi-routine \& routine]}$			-0.44	(-1.17,0.28)	

Table shows the posterior mean, with the 95% interval in brackets.

^a % difference in parameter estimates from MoIE to JM.R8

Summary:**Bayesian full probability modelling of real data with missing responses**

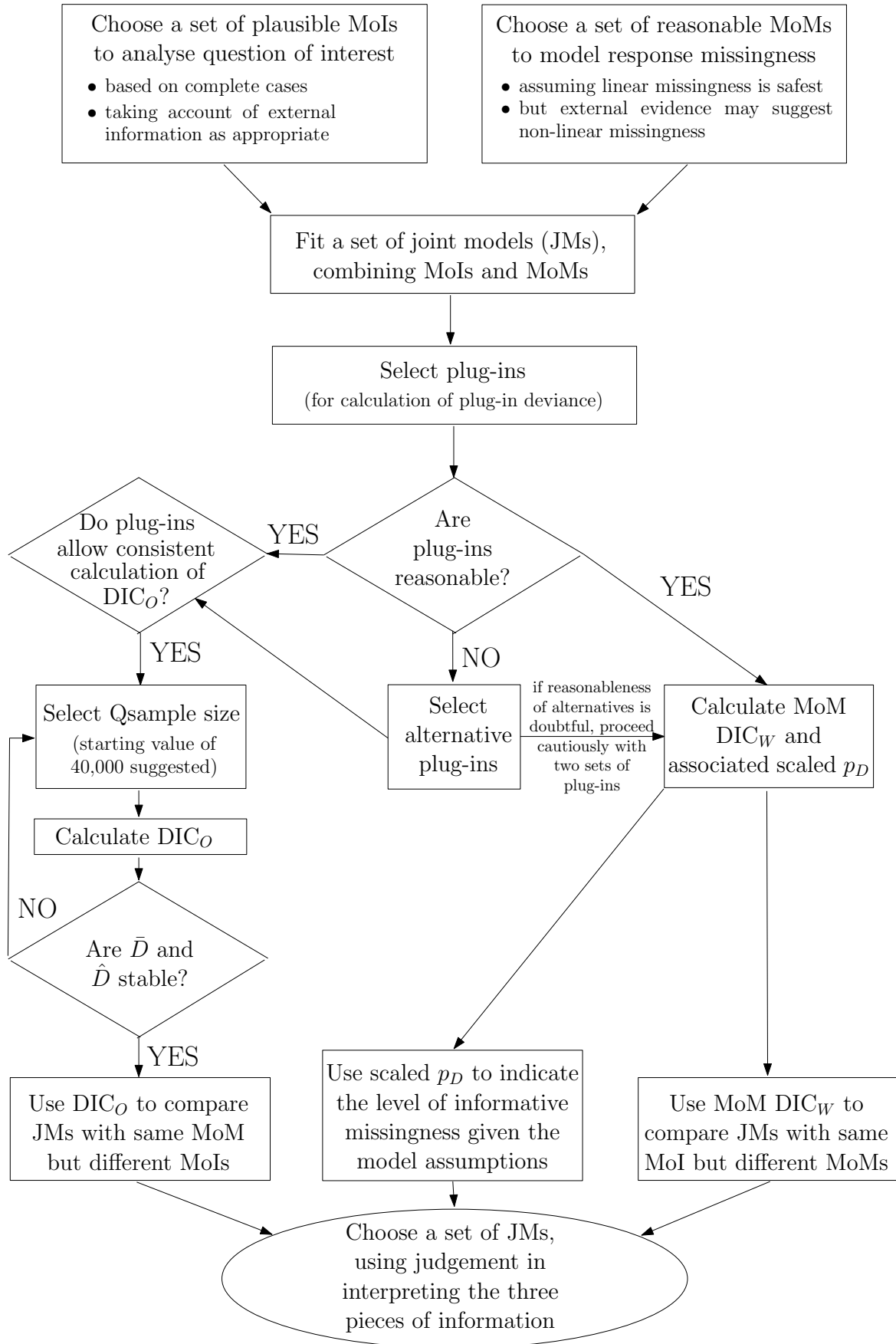
In Example 1 on NCDS educational test scores, we vary the transformation of the response. Regardless of the transformation, we find that social class clearly affects educational attainment and little evidence of informative missingness.

We model the antidepressant trial data in Example 2 using joint models which combine two different models of interest with three different models of missingness. Using DIC_O , and model of missingness DIC_W and its associated scaled p_D , we find that the joint models with a random effects model of interest and an informative model of missingness that depends on the change in HAMD are most appropriate for the HAMD data. There is strong evidence that treatment affects the change in HAMD score over time. Compared with a complete case analysis, our preferred joint models suggest that all the treatments are a little more effective in lowering HAMD (which is consistent with the findings of DK), but their ordering is unaltered.

For our third example on MCS income, we have built a basic joint model, JM.R8, for use in the remainder of this thesis. Its model of interest regresses the log of hourly pay (*hpay*) against covariates *age* (main respondent's age), *edu* (educational level), *reg* (London/other), *sing* (single/partner) and *eth* (ethnic group), with individual random effects and stratum specific intercepts assuming t_4 errors. Our chosen response model of missingness comprises a logit equation with regressors *sc* (social class), *eth* (ethnic group) *ctry* (country), *level* (level of pay in sweep 1) and *change* (change in pay between sweeps 1 and 2). We find strong evidence that higher levels of education are associated with higher pay and gaining a partner between sweeps is associated with lower pay, but little evidence of a relationship between pay and ethnicity. We also find evidence of informative missingness.

These three real examples have allowed the strategy that we developed in Chapter 4, for the use of DIC in modelling data with missing response, to be tested and refined. Our resulting proposed strategy is summarised diagrammatically in Figure 5.8.

Figure 5.8: Proposed strategy for the use of DIC in modelling data with missing response



Chapter 6

Modelling missing covariates

So far, we have concentrated on modelling missing responses, making simplistic assumptions about any missing covariates where necessary. In this Chapter, we consider the issues involved in adding a covariate model of missingness to our joint model, so the missing covariates are imputed simultaneously with the analysis of the model of interest and missing response imputation. We investigate this by revisiting our MCS income example, and creating a new joint model which incorporates more realistic assumptions about the covariate missingness.

6.1 Issues

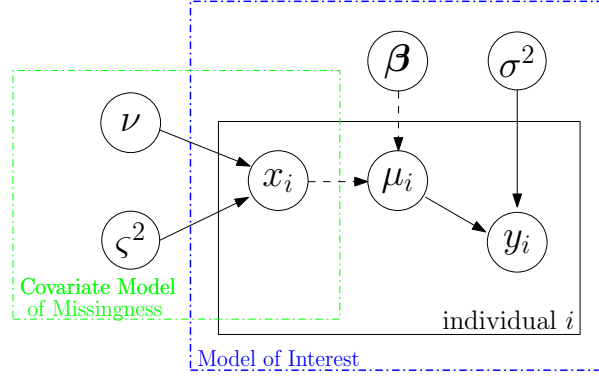
Consider a regression model with a single covariate, as specified by Equation 1.8, but let us now assume that the response variable, y , is fully observed and the covariate, x , has missing values. We need to build a model for x , and there are two obvious ways of doing this:

- 1 specify a distribution, e.g. if x is a continuous covariate, then specify $x_i \sim N(\nu, \varsigma^2)$ and assume vague priors for ν and ς^2 ;
- 2 build a regression model relating x_i to other observed covariates.

A graphical representation of the first option is shown in Figure 6.1. This set-up allows the missing x_i to be MAR, as their estimation will be influenced by feedback from the response variable. (The missingness mechanism would be MCAR if feedback from the model of interest to the covariate model of missingness is prevented, which can be achieved in WinBUGS by using the *cut* function.) Values of the missing covariates will now be automatically simulated at each iteration. In the second option, the missing x_i are also assumed to be MAR.

Unfortunately, difficulties arise when we have more than one covariate with missing values, as is usually the case with real data, and these covariates are correlated. This problem is further complicated when some or all of these covariates are not continuous. For instance, in our MCS income example, of the

Figure 6.1: Graphical representation for a regression model with covariate missingness



four covariates with missingness, one is continuous, two are binary and one is ordered categorical. We begin by considering the case of correlated binary covariates with missingness, before looking at more complicated combinations.

6.1.1 Correlated binary covariates

One strategy for modelling correlated binary covariates uses the multivariate probit model (Chib and Greenberg, 1998; Chib, 2000). In this approach each binary variable is linked to an underlying latent variable with a Normal distribution, and thresholds are specified to determine whether the original variable's value is 0 or 1.

Let $\mathbf{x}$ be the set of binary covariates with missing values, $x_1, \dots, x_p$, that we wish to model, and denote the corresponding latent variables as $\mathbf{x}^* = \{x_1^*, \dots, x_p^*\}$. Then for individual i ,

$$\mathbf{x}_i^* \sim MVN(\boldsymbol{\nu}_i, \Sigma)$$

where MVN denotes a Multivariate Normal distribution, and for covariate k

$$x_{ik} = \begin{cases} 0: & x_{ik}^* \leq 0 \\ 1: & x_{ik}^* > 0. \end{cases} \quad (6.1)$$

To complete the specification of a model for imputing the missing values of binary covariates, equations are defined for calculating the means, $\boldsymbol{\nu}_i$, of the MVN distribution. These are used to predict the latent variables, $\mathbf{x}_i^*$, from a linear combination of some fully observed explanatory variables (see Equation 6.3 for an example), and hence to predict the $\mathbf{x}_i$. In our case, the $\mathbf{x}_i$ being predicted are part observed and part missing. Although the model is fully defined, this leads to complications in implementation in WinBUGS. These can be overcome by implementing Equation 6.1 directly only for the missing covariates, and using constraints on the MVN distribution when the $\mathbf{x}_i$ are observed to force their underlying latent variables to lie in the correct interval (i.e. if x_{ik} is observed to be 0, then x_{ik}^* must lie in the interval $(-\infty, 0]$, and if x_{ik} is 1, then x_{ik}^* must be constrained to the interval $(0, \infty)$). This

is not a typical use of the probit, which is usually used as an alternative link function to the logit in a generalised linear model for binary data.

The means, $\boldsymbol{\nu}_i$, and covariance, Σ , are not uniquely likelihood identified. To see why, consider modelling a single binary covariate using a probit model, such that $x_i^* \sim N(\nu_i, \sigma)$, x_i is defined as in Equation 6.1, $\nu_i = \eta_0 + \sum_{j=1}^q \eta_j z_{ij}$ and $z_1, \dots, z_q$ are fully observed variables. If all the $\boldsymbol{\eta}$ parameters and the standard deviation, σ , are multiplied by the same constant, then the probability of the latent variable exceeding 0 will be unchanged. Similarly, in the case of multiple binary covariates, the probability of each latent variable exceeding 0 will not change if the parameters in the equation for ν_i are multiplied by a constant, c_i , and the covariance matrix is set to $C\Sigma C$ where C is a diagonal matrix with diagonal elements $\{c_1, c_2, \dots, c_p\}$ (Chib and Greenberg, 1998). If Σ is constrained to be in correlation form, the parameters in the ν_i equations and remaining elements of Σ are identifiable.

The standard conjugate prior for a covariance matrix is the Wishart distribution, but this does not allow structure to be imposed. To set the diagonal elements of a matrix to 1, requires separate priors for each element of that matrix. The difficulty is then ensuring that the matrix generated is positive definite. For two binary variables this is easy, but it becomes non-trivial for more than two variables. The unusual shapes of the sets of correlation coefficients for three and four variables are discussed by Rousseeuw and Molenberghs (1994), who give some intuition about the constraints that exist between the correlations. Molitor *et al.* (2009) use this approach for imputing the missing values for two binary covariates, in an application to low birth weight and water disinfection by-products.

An alternative approach is to set the model up as a sequence of conditional univariate models, which allows separate priors to be placed on the individual elements of the covariance matrix and guarantees that the covariance matrix of the underlying joint model is always positive definite. This can be done using the following properties of a MVN distribution (see, for example, Congdon (2001), Section 2.8). If a set of variables, $\mathbf{x}$, can be modelled as $\mathbf{x} \sim MVN(\boldsymbol{\nu}, \Sigma)$, then $\mathbf{x}$ can be partitioned into two subsets of variables, $\mathbf{x}_1$ and $\mathbf{x}_2$, such that

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim MVN \left(\begin{pmatrix} \boldsymbol{\nu}_1 \\ \boldsymbol{\nu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

Further, if $\mathbf{x}_1$ has a marginal MVN distribution, $\mathbf{x}_1 \sim MVN(\boldsymbol{\nu}_1, \Sigma_{11})$, then the conditional distribution of $\mathbf{x}_2$ when $\mathbf{x}_1$ is known is

$$\mathbf{x}_2 | \mathbf{x}_1 \sim MVN \left(\boldsymbol{\nu}_2 + \Sigma_{21} \Sigma_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\nu}_1), \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \right). \quad (6.2)$$

The covariance matrix can be constrained to correlation form by setting Σ_{11} and Σ_{22} to 1. Similarly, using an obvious extension of notation, it can be shown that the conditional distribution of $x_j | x_{j-1}, \dots, x_1$ is univariate Normal with mean and variance given by particular functions of $x_{j-1}, \dots, x_1, \nu_j, \dots, \nu_1$ and $\Sigma_{11}, \dots, \Sigma_{1j}, \dots, \Sigma_{jj}$. Examples of the implementation of this approach

are provided in Section 6.2. Unfortunately the required equations rapidly become very complicated as the number of variables increases.

A third option is not to restrict the covariance matrix to correlation form, so that a Wishart distribution can be used as its prior. Although the means and covariances in such a model are not likelihood identified, they will be posterior identified if proper priors are specified (Gelfand and Sahu, 1999). This approach is taken by Lunn *et al.* (2006), who argue that using an unrestricted covariance matrix is not a major issue provided they have no interest in making inferences about the means or covariances (personal correspondence). It has been pointed out that even with improper priors, using MCMC it is still possible to obtain convergence of a lower-dimensional subset of parameters embedded in a non-identifiable model (Gelfand and Sahu, 1999). While the unidentified parameters will not converge, the identified parameters may be very well behaved (Kass *et al.*, 1998).

In Section 6.2.1 we compare these different strategies, as a preliminary to setting up a covariate model of missingness for our MCS income example.

6.1.2 Ordered categorical covariates

The method of modelling binary variables using the multivariate probit model can be generalised to variables with more than two categories (Albert and Chib, 1993). Suppose $\mathbf{x}$ is a set of covariates, $x_1, \dots, x_p$, such that x_k takes one of $C > 2$ ordered categories. Then we introduce a set of corresponding continuous latent variables, $\mathbf{x}^* = \{x_1^*, \dots, x_p^*\}$, as in Section 6.1.1. Now, however, for individual i and covariate k ,

$$x_{ik} = c, \text{ if } \xi_{k,c-1} < x_{ik}^* \leq \xi_{k,c},$$

where $-\infty = \xi_{k,0} \leq \xi_{k,1} \leq \xi_{k,2} \leq \dots \leq \xi_{k,C-1} \leq \xi_{k,C} = \infty$ are “cut-points” which divide the real line into C intervals and $\xi_{k,1}, \dots, \xi_{k,C-1}$ are unknown. A different set of cut-points can be used for each variable. To ensure identifiability, as in the binary case, we need to impose one restriction on each set of cut-points, which is usually taken to be $\xi_{k,1} = 0$, and restrict the variances to 1.

Generating the unknown cut-points is considered very difficult, and Chen and Dey (2000) consider a re-parameterisation to ease the problem, in which each cut-point, x_{ik}^* and regression coefficient in the ν_{ik} equation is divided by $\xi_{k,C-1}$. This results in only $C - 3$ unknown cut-points for each covariate, lying between 0 and 1, and an unrestricted covariance matrix, Σ . We explore this suggestion in setting up models for imputing missing *edu* values for the MCS income example in Section 6.2.3.

6.1.3 Unordered categorical covariates

Now suppose that we have one unordered categorical covariate, x , with > 2 categories, $\{0, 1, 2, \dots, C\}$. Then a Multinomial Probit model can be set up with $C+1$ latent variables, $\mathbf{x}^* = \{x_0^*, x_1^*, \dots, x_C^*\}$, one corresponding to each category. However, there is a location problem because x only provides information about the differences in these underlying latent variables, and the scale cannot be determined. This leads to restrictions on the number of estimable parameters (Dansie, 1985; Bunch, 1991).

The standard solution to the location problem is to collapse the system by differencing with respect to one category, leaving C latent variables, $\mathbf{x}^* = \{x_1^*, \dots, x_C^*\}$, which are modelled with a MVN structure and unknown covariance matrix. An unobserved x_i is deemed to fall in the first category if all the latent variables are less than zero, or otherwise in the category whose latent variable has the highest value (Albert and Chib, 1993). More formally, for individual i

$$x_i = \begin{cases} 0: & \max(\mathbf{x}_i^*) \leq 0 \\ c: & \max(\mathbf{x}_i^*) = x_{ic}^* > 0. \end{cases}$$

However, the parameters in this Multinomial Probit model, consisting of a vector of coefficients for calculating the means of the MVN distribution and a covariance matrix, are still not fully identified because of the scale indeterminacy (multiplying the coefficients and covariance matrix by a positive constant does not change the x_i). This model differs from those discussed in Sections 6.1.1 and 6.1.2, in that the covariance matrix is now across categories of a single covariate rather than across multiple covariates. If there are just two categories, then the model reduces to that specified in Section 6.1.1 for a single covariate.

McCulloch and Rossi (1994) (MR) discuss the scale identification problem and possible solutions. One solution is to set the (1,1) element of the covariance matrix to 1, but in Bayesian analysis this approach leads to difficulties in defining a prior for the covariance matrix, Σ , similar to the issues raised in Section 6.1.1. An alternative method of identification, suggested by MR, is to fix one of the coefficients; but this also has drawbacks, including the requirement of having a coefficient whose sign can be determined a priori. MR provide an algorithm with non-identified parameters so that a Wishart prior can be used for the covariance matrix, and report the marginal posterior of the identified parameters. In a follow-up paper McCulloch *et al.* (2000) present an algorithm with fully identified parameters, that specifies a prior on Σ that fixes its (1,1) element to 1.

We do not use any unordered categorical covariates with missing values in our MCS income example, but the above suggests that there are ways of incorporating such variables in a covariate model of missingness, although they may not be easy. This is a possible future extension of this work.

6.1.4 Mixed covariates

Often, as in our MCS income example, the covariates will be a mixture of types, possibly including continuous as well as binary and categorical variables. In this case the approaches outlined in Sections 6.1.1-6.1.3 can be combined. Dunson (2000) developed Bayesian latent variable models for joint binary, categorical and continuous outcomes which accommodate multilevel data structures and different link functions, and implemented an example for modelling a binary and a count outcome simultaneously using the BUGS software.

More recently, Goldstein *et al.* (2008) have formulated a class of models for multivariate mixtures of categorical and continuous responses for multilevel data, and have developed freely available software that fits these using MCMC algorithms. Their framework creates an underlying set of latent variables and has a number of applications, including creating multiple imputations of missing data by treating all the variables with missing data as a set of multivariate responses. However, rather than using full Bayesian joint modelling, they follow the usual multiple imputation two stage process and fit their model of interest separately.

6.2 MCS income example, assuming ignorable response missingness

We now explore the various strategies for modelling correlated covariates using our MCS income example. Initially, to keep things simple, we form a joint model without a response model of missingness, i.e. we assume that the missingness in *hpay* is non-informative. Our model of interest is based on MoIE (slight modifications are required to accommodate a reduced number of categories for *edu*, the version with 3 categories being labelled MoIF), which was discussed in Section 5.3.1, and we combine this with different covariate models of missingness. The combinations of different types of covariates that we consider, and the MCS variables that are used for this purpose are listed below, with the model labels in brackets:

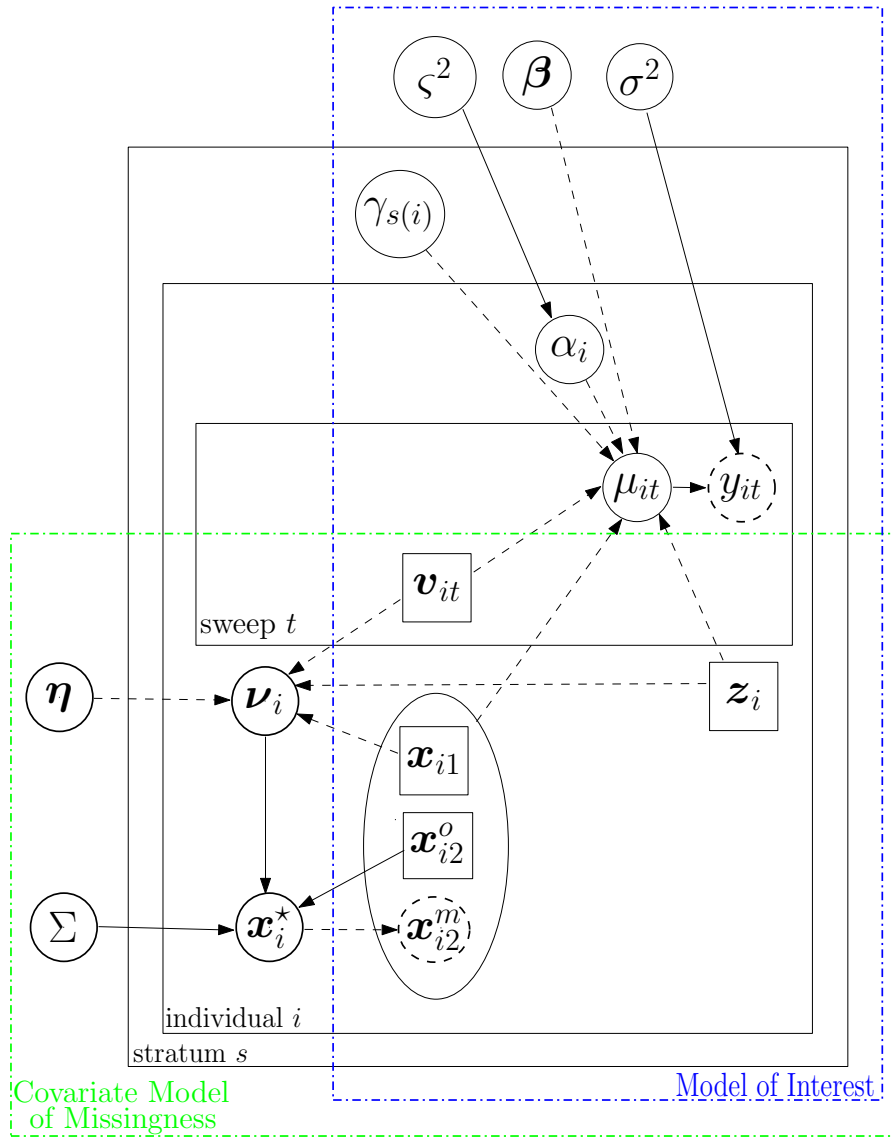
- two binary variables: dichotomised *edu* and *sing* (JM.C1);
- three binary variables: dichotomised *edu*, *sing* and *reg* (JM.C2);
- one ordered categorical variable: *edu* (JM.C3 and JM.C4);
- one ordered categorical variable and one binary variable: *edu* and *sing* (JM.C5);
- one ordered categorical variable, one binary variable and one continuous variable: *edu*, *sing* and *age* (JM.C6).

All the models in this section are run with 2 chains for 100,000 iterations of which 50,000 are burn-in and a thinning rate of 5. Based on our usual checks, these are assumed to have converged unless otherwise stated. Our convergence checks do not include any parameters that cannot be fully identified.

In each case, sweep 2 missing values of any covariates that are not included in the covariate model of missingness continue to be imputed off model using the simplistic assumptions described in Section 5.3.2.

A graphical representation of a general combined model of interest and covariate model of missingness for our MCS income example is shown in Figure 6.2. The covariates which vary between sweeps, denoted by $\mathbf{x}_{it}$ in Figure 5.5, are now subdivided into $\mathbf{x}_{it}$ which have missing sweep 2 values and $\mathbf{v}_{it}$ which are fully observed. The $\mathbf{x}_{it}$ node is displayed as a triple node, subdivided as sweep 1 values ($\mathbf{x}_{i1}$), observed sweep 2 values ($\mathbf{x}_{i2}^o$) and missing sweep 2 values ($\mathbf{x}_{i2}^m$), to more clearly show the role of each part. For example, our first covariate model of missingness will be built with $\mathbf{x} = \{edu, sing\}$ and $\mathbf{v} = \{age, reg\}$.

Figure 6.2: Graphical representation for model of interest and covariate model of missingness



6.2.1 Imputing *edu* and *sing* as correlated binary variables

We start by considering the two covariates that most concern us, namely *edu* and *sing*. The drawbacks of assuming that these variables are unchanged between sweeps when their sweep 2 value is unobserved, were discussed in Section 5.3.2. To investigate different ways of modelling two correlated binary variables, we dichotomise *edu* by combining categories 0 to 3 and categories 4 to 5, so that the *edu* levels become “no degree” and “degree”.

The covariate model of missingness for these two variables, can be set up in four different ways. We call the joint models combining these four variants with a model of interest JM.C1a-JM.C1d. A summary of the covariate model of missingness part is shown in Table 6.1, and we consider each in turn.

Table 6.1: MCS income example: summary of the covariate models of missingness for two correlated binary variables (dichotomised *edu* and *sing*)

	<i>edu</i>	<i>sing</i>	identifiability constraints
JM.C1a	joint MVN distribution		✓
JM.C1b	marginal Normal	conditional Normal	✓
JM.C1c	conditional Normal	marginal Normal	✓
JM.C1d	joint MVN distribution		✗

Using a multivariate model with Σ restricted to correlation form

In JM.C1a *edu* and *sing* are modelled for sweep 2 only, using equations

$$\begin{aligned}
 x_{i,1:2}^* &\sim MVN(\nu_{i,1:2}, \Sigma)I(L_{i,1:2}, R_{i,1:2}) \\
 \nu_{i,1} &= edu.\eta_0 + (edu.\eta_{age} \times age_{i1}) + (edu.\eta_{edu} \times edu_{i1}) \\
 \nu_{i,2} &= sing.\eta_0 + (sing.\eta_{age} \times age_{i1})
 \end{aligned} \tag{6.3}$$

for $i=1, \dots, n$ individuals, where $MVN(mean, variance)I(L, R)$ denotes a truncated MVN distribution restricted coordinatewise to values in the vector interval $[L, R]$. The second index on the *age* and *edu* variables indicates that the sweep 1 values of these variables are used. The means of the latent variables for *edu* and *sing* are $\nu_{i,1}$ and $\nu_{i,2}$ respectively. They are regressed against selected sweep 1 covariates, and we choose *age* and *edu* for the *edu* latent variable, and just *age* for the *sing* latent variable ($edu.\eta_0$ is the intercept in the $\nu_{i,1}$ equation, $edu.\eta_{age}$ is the coefficient associated with *age* in the $\nu_{i,1}$ equation, etc.). Other variables can be incorporated into these equations, provided they are fully observed, but we prefer to restrict the number of explanatory variables at this stage, as these models are for illustration purposes.

Note that we use a truncated MVN distribution to model the latent variables, because we are using the same model for both observed and missing covariates. There is no need of the truncation for the missing covariates, but it is required when the covariates are observed. This enables information to

be extracted from the observed covariates to estimate the parameters of the covariate model of missingness. Implementing the truncated distribution involves setting boundaries for the latent variables when their associated covariate is observed. The details for this model are discussed a little later in this subsection under the heading “setting the boundaries on the truncated MVN distribution”.

Uninformative priors are used for all the parameters in the $\nu_{i,j}$ equations, i.e. $\eta_k \sim N(0, 10000^2)$. We wish to restrict Σ to correlation form for identifiability. This is easily achieved for a 2×2 matrix,

$$\Sigma = \begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix},$$

by placing a Uniform(-1,1) prior on κ . In WinBUGS we work with the precision rather than covariance, so we define the elements of $\Omega = \Sigma^{-1}$ as follows:

$$\Omega_{11} = \Omega_{22} = \frac{1}{1 - \kappa^2}; \quad \Omega_{12} = \Omega_{21} = \frac{-\kappa}{1 - \kappa^2}.$$

Using a marginal and conditional model

Using Equation 6.2 for two binary variables, the joint model

$$\begin{pmatrix} x_{i,1}^* \\ x_{i,2}^* \end{pmatrix} \sim MVN \left(\begin{pmatrix} \nu_{i,1} \\ \nu_{i,2} \end{pmatrix}, \begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix} \right)$$

can be written as a marginal distribution,

$$x_{i,1}^* \sim N(\nu_{i,1}, 1), \tag{6.4}$$

and a conditional distribution

$$x_{i,2}^* | x_{i,1}^* \sim N(\nu_{i,2} + \kappa(x_{i,1}^* - \nu_{i,1}), 1 - \kappa^2). \tag{6.5}$$

For JM.C1b, x_1^* is the latent variable for *edu* and x_2^* is the latent variable for *sing*. The JM.C1a equations for $\nu_{i,j}$ and prior distributions continue to be used. JM.C1c takes the same form as JM.C1b, but we now use a marginal distribution for the *sing* latent variable, and a conditional distribution for the *edu* latent variable. In both cases, the marginal and conditional Normal distributions are truncated as for Equation 6.3.

Using a multivariate model with Σ unrestricted

For our final variant, JM.C1d, latent variables for *edu* and *sing* are modelled using the truncated MVN distribution and $\nu_{i,j}$ defined by Equation 6.3. The only difference from JM.C1a is the specification of the priors on Σ . We no longer restrict Σ to correlation form, and instead use a Wishart distribution as a prior for the precision matrix, $\Omega = \Sigma^{-1}$, such that $\Omega \sim Wishart(k, T)$.

The Wishart distribution is the multivariate generalisation of the Gamma distribution and has two parameters: a real scalar k and a symmetric positive definite matrix, T . It is a common default choice

for the prior of a covariance matrix (DH, p.48). However, the choice of scale matrix T is difficult, and it is often chosen to be either a diagonal matrix with marginal variances consistent with the variation in the data, or is based on the maximum likelihood estimate of Σ . In the MCMC estimation part of MLwiN, the default is to set k to the number of rows in the covariance matrix Σ , and T to kS where S is the maximum likelihood estimate of the true value of the unknown Σ (Browne, 2003). As k , which represents the size of the sample on which the prior belief is based, is set to its smallest possible value, this results in a weakly informative prior. Browne and Draper (2000) examine the performance of different diffuse settings of the Wishart parameters for a 2×2 variance matrix, using simulation studies in the context of fitting multilevel models. These include (1) setting k to 2 and T to the 2×2 identity matrix and (2) setting k to 4 and T to the maximum likelihood estimate of the variance matrix, which they describe as “gently data-determined”. They found that both sets of priors performed reasonably well as regards bias unless the number of level two units was small. Kass and Natarajan (2006) suggest an alternative strategy in which T is based on the data variability at the first level of the hierarchical model.

We set $k = 2$ to give the least informative prior possible, and set T to the identity matrix. In Section 6.1.1, we discussed how models with an unrestricted covariance matrix as used in the covariate model of missingness part of JM.C1d are likelihood unidentified. However, because we have given the parameters of Σ proper priors the corresponding posteriors will also be proper, but the use of vague priors means that the estimates of the covariate model of missingness parameters will not be meaningful. We are only interested in the imputed values of the missing covariates, which are fully identified, so this version of the model will be suitable for our purposes.

Setting the boundaries on the truncated MVN distribution

We set left and right boundaries for each latent variable modelled by a truncated MVN or Normal distribution. These boundaries only apply when the associated covariate is observed, and the distribution is unbounded when it is missing. This allows information to be extracted from the observed values to estimate the parameters of the $\nu_{i,j}$ equations. For example, if x_{ik} is observed to be in category 0, then x_{ik}^* must be less than or equal to zero and needs to be constrained to be in the interval $(-\infty, 0]$ (Equation 6.1). Additionally, we set boundaries for the missing values of the latent education variable, to prevent educational level from going down between sweeps 1 and 2. These boundaries are summarised in Table 6.2, distinguishing between observed and missing data. Note that the conditions for the observed data are based on sweep 2 data, while those for missing educational level data are based on sweep 1 data. In our WinBUGS implementation we use a limit of 10 rather than infinity.

Creating initial values for the latent variables

Because of all the constraints, limited data in some areas and in the case of JM.C1d lack of identifiability, we require good starting values for $\mathbf{x}^*$ to get the model running. So, initial values for $\mathbf{x}^*$ are

Table 6.2: MCS income example: boundaries on latent variables for dichotomised *edu* and *sing*

equation	observed data			missing data		
	condition	left	right	condition	left	right
$x_{i,1}^*$	$edu_2 = 0$	$-\infty$	0	$edu_1 = 0^a$	$-\infty$	∞
	$edu_2 = 1$	0	∞	$edu_1 = 1^a$	0	∞
$x_{i,2}^*$	$sing_2 = 0$	$-\infty$	0	none	$-\infty$	∞
	$sing_2 = 1$	0	∞	none	$-\infty$	∞

The subscript on *edu* and *sing* indicates sweep.

^a Educational level is not allowed to decrease between sweeps.

generated by running an adapted version of JM.C1a, with no model of interest and all the ν_i equations set to 0. In this run, the MCMC learns from the data using the limits imposed on the truncated MVN distribution. Two sets of initial values for $\mathbf{x}^*$ are created using iteration 2000 from a run with two chains. We use this technique to create $\mathbf{x}^*$ initial values for subsequent models as required.

The imputations of missing *edu* and *sing*

The parameter estimates for the different variants of JM.C1 are discussed in Appendix C.3. However, our real interest lies in the imputations of missing *edu* and *sing*, and we consider these based on JM.C1a. Among the individuals without a degree at sweep 1 and who have observed sweep 2 *edu*, five gained a degree between sweeps, which is equivalent to 1.9%. The posterior mean of the number of individuals with missing sweep 2 *edu* who gained a degree is 3.6, or 2.6%. For *sing*, 34% of those with observed *sing* at sweep 2 gained a partner, compared to an estimated 35% of those with missing sweep 2 *sing*, based on the posterior means of the imputations.

Table 6.3 summarises the imputations for models JM.C1a-d and JM.C1e, which we will introduce and discuss shortly. In this table, edu_{ij} is the number of individuals with missing sweep 2 *edu*, who have educational level i at sweep 1 and are imputed to have educational level j at sweep 2. The zero values for edu_{21} demonstrate that the boundary restrictions have successfully prevented educational level from falling between sweeps. In the same way, $sing_i$ is the number of individuals with missing sweep 2 *sing* and imputed value i at sweep 2 (recall that *sing* in sweep 1 is always 1, so there is no need for the double index notation here). The $edusing_{ij}$ variables are a cross tabulation of sweep 2 *edu* versus sweep 2 *sing*, and are calculated across all individuals, rather than just those with missing covariates. Here index i indicates the value of sweep 2 *edu* and the index j the value of sweep 2 *sing*.

The parameterisations of JM.C1a-c are equivalent, so we expect their imputations to be the same, allowing for Monte Carlo error, provided we have used uninformative priors as intended. Table 6.3 shows this is the case. We are particularly interested in whether restricting the covariance matrix, Σ , to correlation form makes a difference. In this example, the imputations using the unrestricted covariance matrix (JM.C1d) are virtually identical to those based on the correlation form (JM.C1a-c).

Table 6.3: MCS income example: summary of covariate imputations for five variants of model JM.C1 (dichotomised *edu* and *sing*)

	JM.C1a		JM.C1b		JM.C1c		JM.C1d		JM.C1e	
edu_{11}^a	136.4	(131,140)	136.4	(131,140)	136.4	(131,140)	136.4	(131,140)	136.6	(131,140)
edu_{12}^a	3.6	(0,9)	3.6	(0,9)	3.6	(0,9)	3.6	(0,9)	3.4	(0,9)
edu_{21}^a	0.0	(0,0)	0.0	(0,0)	0.0	(0,0)	0.0	(0,0)	0.0	(0,0)
edu_{22}^a	26.0	(26,26)	26.0	(26,26)	26.0	(26,26)	26.0	(26,26)	26.0	(26,26)
$sing_1^b$	107.4	(92,122)	107.4	(92,122)	107.3	(92,122)	107.3	(92,122)	105.0	(89,120)
$sing_2^b$	58.6	(44,74)	58.6	(44,74)	58.7	(44,74)	58.7	(44,74)	61.0	(46,77)
$edusing_{11}^c$	250.4	(236,264)	250.4	(237,264)	250.3	(236,264)	250.3	(236,264)	249.2	(235,263)
$edusing_{12}^c$	139.0	(126,153)	138.9	(126,153)	139.0	(126,153)	139.0	(126,153)	140.4	(127,154)
$edusing_{21}^c$	82.0	(76,88)	82.0	(76,88)	82.0	(76,88)	82.0	(76,88)	80.8	(75,87)
$edusing_{22}^c$	33.6	(28,40)	33.6	(28,40)	33.7	(28,40)	33.7	(28,40)	34.6	(29,41)

Table shows the posterior mean, with the 95% interval in brackets

^a The first index indicates the value of sweep 1 *edu*, and the second index the value of sweep 2 *edu*.

^b The index indicates the value of sweep 2 *sing*.

^c The first index indicates the value of sweep 2 *edu*, and the second index the value of sweep 2 *sing*; count includes individuals with observed *edu* and *sing* in sweep 2.

This suggests that there is no need to restrict Σ to correlation form provided we are not interested in the parameter values, which is potentially useful for models with higher dimension Σ , where imposing conditions to limit it to correlation form requires complex equations.

As a sensitivity, we also run a version of this model in which the means of the latent variables, $\nu_{i,j}$, are regressed against *eth* in addition to the variables used for JM.C1a-d. This is model JM.C1e, and its resulting covariate imputations are shown in Table 6.3. The differences are small, and do not have a discernable impact on the model of interest parameter estimates, so we will revert to the simpler $\nu_{i,j}$ equations.

6.2.2 Imputing *edu*, *sing* and *reg* as correlated binary variables

Next we look at models for three correlated binary variables, modelling *reg* in addition to the dichotomised *edu* and *sing* used in Section 6.2.1. We set up a covariate model of missingness for these three variables in two ways. Joint model JM.C2a has a covariate model of missingness set up as a sequence of conditional univariate models, and JM.C2b models the three variables jointly using a MVN distribution with an unrestricted covariance matrix.

For three binary variables, the joint model

$$\begin{pmatrix} x_{i,1}^* \\ x_{i,2}^* \\ x_{i,3}^* \end{pmatrix} \sim MVN \left(\begin{pmatrix} \nu_{i,1} \\ \nu_{i,2} \\ \nu_{i,3} \end{pmatrix}, \begin{pmatrix} 1 & \kappa_{12} & \kappa_{13} \\ \kappa_{12} & 1 & \kappa_{23} \\ \kappa_{13} & \kappa_{23} & 1 \end{pmatrix} \right)$$

can be written as a marginal distribution, $x_{i,1}^*$, and two conditional distributions $x_{i,2}^*|x_{i,1}^*$ and $x_{i,3}^*|x_{i,1}^*, x_{i,2}^*$. The distributions for $x_{i,1}^*$ and $x_{i,2}^*|x_{i,1}^*$, are given by Equations 6.4 and 6.5 respectively, with κ replaced

by κ_{12} . Using Equation 6.2, $x_{i,3}^*|x_{i,1}^*, x_{i,2}^*$ has a Normal distribution with mean

$$\nu_{i,3} + (1 - \kappa_{12}^2)^{-1} \{(\kappa_{13} - \kappa_{12}\kappa_{23})(x_{i,1}^* - \nu_{i,1}) + (\kappa_{23} - \kappa_{12}\kappa_{13})(x_{i,2}^* - \nu_{i,2})\},$$

and variance

$$1 - (1 - \kappa_{12}^2)^{-1}(\kappa_{13}^2 - 2\kappa_{12}\kappa_{13}\kappa_{23} + \kappa_{23}^2).$$

The $\nu_{i,j}$ equations corresponding to *edu* and *sing* are as specified for JM.C1a, and the *reg* $\nu_{i,j}$ equation is

$$\nu_{i,3} = \text{reg}.\eta_0 + (\text{reg}.\eta_{\text{age}} \times \text{age}_{i1}) + (\text{reg}.\eta_{\text{reg}} \times \text{reg}_{i1})$$

with $\nu_{i,3}$ regressed against sweep 1 *reg* and *age*. Uninformative priors, initial values and boundaries are set up as described in Section 6.2.1. No special restrictions are required for the *reg* imputations, so the boundaries for $x_{i,3}^*$ are identical to those for $x_{i,2}^*$.

Our usual burn-in length is insufficient to achieve satisfactory convergence for JM.C2a, so we run it with 500,000 iterations of which 400,000 are burn-in. This longer run converges according to our usual criteria.

The two models provide almost identical patterns of missing covariate imputations and estimates of the model of interest parameters, in line with our finding in Section 6.2.1 that whether or not we restrict the covariance matrix to correlation form makes no difference to our quantities of interest. The *edu* and *sing* imputations are very similar to those produced by the two binary covariate models, JM.C1. For *reg*, 8% of those with observed *reg*₂ who lived in London in sweep 1, moved to the regions, compared to an imputed 10% for individuals with missing *reg*₂. For those who lived in the regions in sweep 1, the percentage who moved into London in sweep 2 was less than 1% for both the observed and imputed data.

6.2.3 Imputing *edu* as an ordered categorical variable

Three category *edu*: no unknown cut-points

We now turn to the problem of imputing the missing values for an ordered categorical variable, using *edu* as our example. We start by grouping categories 0-1, 2-3 and 4-5 to create a three category version of *edu*. This has the advantage of allowing us to start by fitting a model in which all the cut-points are known, as described in Section 6.1.2, with the two required cut-points fixed at 0 and 1. Our covariate model of missingness for JM.C3 takes the form

$$x_i^* \sim N(\nu_i, \sigma^2)I(L_i, R_i)$$

$$\nu_i = \text{edu}.\eta_0 + (\text{edu}.\eta_{\text{age}} \times \text{age}_{i1}) + \sum_{k=2}^3 (\text{edu}.\eta_{\text{edu}[k]} \times \text{edu}_{[k]i1}) \quad (6.6)$$

for $i=1, \dots, n$ individuals, with ν_i regressed against sweep 1 *age* and *edu*, where k is educational level category and $edu_{[k]}$ is a binary indicator for *edu* category k (1 if edu_i is category k , 0 otherwise). Uninformative priors are used for all the parameters, and the boundaries are shown in Table 6.4.

Table 6.4: MCS income example: boundaries on latent variable for three category *edu*

observed data			missing data		
condition	left	right	condition	left	right
$edu_2 = 1$	$-\infty$	0	$edu_1 = 1^a$	$-\infty$	∞
$edu_2 = 2$	0	1	$edu_1 = 2^a$	0	∞
$edu_2 = 3$	1	∞	$edu_1 = 3^a$	1	∞

The subscript on *edu* indicates sweep.

^a Educational level is not allowed to decrease between sweeps.

The Gelman-Rubin convergence statistics and a visual inspection of a trace of the chains for a selection of parameters suggest convergence, although $edu.\eta_{edu[3]}$ mixes slowly. We also note that $edu.\eta_{edu[3]}$ has a wide 95% interval, (1.9,7.5). In other models that impute a three category version of *edu*, we find similar behaviour and often this is also manifested by a high Gelman-Rubin convergence statistic for $edu.\eta_{edu[3]}$. This is more to do with high autocorrelation than lack of convergence. In JM.C3 we find an effective sample size of just 11 for estimating the posterior mean of $edu.\eta_{edu[3]}$. The likelihood is quite flat above the threshold that satisfies the constraint that an individual with a degree at sweep 1 will also have a degree at sweep 2. While using an informative prior might help, we continue using a vague prior as the issue is not having much of an impact on the rest of the model. Although this behaviour is present in the models discussed subsequently, we will not comment on it further.

Table 6.5 enables a comparison of the imputations of missing edu_2 with the observed data, by showing the percentage of individuals with each educational level at sweep 2 for a particular value at sweep 1, split by observed (left table) and missing (right table) edu_2 . The pattern of changes in educational level between sweeps is different for individuals with actual and imputed sweep 2 values. Using the posterior mean, the percentage of individuals increasing *edu* by one level for those with missing edu_2 is double that of those with observed edu_2 , though there is a lot of uncertainty about this percentage given that increasing *edu* from level 1 to level 2 has a 95% interval (3%,31%) and increasing from level 2 to level 3 has a 95% interval (0%,10%). Further, while the observed data shows that some individuals increased *edu* by two levels, no imputations are shown with a two level increase. This may reflect differences in the age profiles of those who provided educational data at sweep 2 and those who did not, or it may indicate some shortcomings in this model. To ascertain the influence of *age* on the imputations, we rerun JM.C3, removing age_1 as an explanatory variable of ν_i . We find that the percentages of individuals with missing edu_2 increasing their educational level are only slightly lower than for the original run, so we have reservations about JM.C3.

Table 6.5: MCS income example: comparison of observed and imputed educational level for sweep 2 from model JM.C3

observed edu_2 (% of row total)					imputed edu_2 (% of row total)				
	$edu_2=1$	$edu_2=2$	$edu_2=3$	total		$edu_2=1$	$edu_2=2$	$edu_2=3$	total
$edu_1=1$	89.5	8.8	1.8	100.0	$edu_1=1$	84.3	15.7	0.0	100.0
$edu_1=2$	0.0	98.0	2.0	100.0	$edu_1=2$	0.0	96.1	3.9	100.0
$edu_1=3$	0.0	0.0	100.0	100.0	$edu_1=3$	0.0	0.0	100.0	100.0

Posterior means are used to generate the “imputed” table, which includes individuals that have observed sweep 1 edu , but missing sweep 2 edu . The subscript on edu indicates sweep.

Four category edu : unknown cut-point

Next, we work with a four category version of edu in which categories 0-1 and 4-5 are combined. We now have one unknown cut-point to estimate as part of the modelling process. Our new model, JM.C4 (four category edu), is an adapted version of JM.C3 in which the left and right boundaries for the truncated Normal distribution are as shown in Table 6.6, and the unknown cut-point, ξ , is given an uninformative prior on the interval $(0, 1)$. We find that WinBUGS is “unable to choose updater” if we use a Uniform(0,1) prior, but will run for a few iterations if we specify a Normal distribution truncated to interval $(0, 1)$, $N(0, 10000^2)I(0, 1)$. However, the run then fails with the message “cannot sample from slice mode xstar[.]”. We retry with increasingly simplified versions of the equation for ν_i , but the run continues to fail with the same error message. Given the known difficulty in generating cut-points, as discussed in Section 6.1.2, we conclude that these types of models are probably beyond the current capability of the WinBUGS software and require more specialised MCMC algorithms. For this thesis, we decide not to pursue methods for modelling ordinal categorical covariates with more than three categories any further.

Table 6.6: MCS income example: boundaries on latent variable for four category edu

observed data			missing data		
condition	left	right	condition	left	right
$edu_2 = 1$	$-\infty$	0	$edu_1 = 1^a$	$-\infty$	∞
$edu_2 = 2$	0	ξ	$edu_1 = 2^a$	0	∞
$edu_2 = 3$	ξ	1	$edu_1 = 3^a$	ξ	∞
$edu_2 = 4$	1	∞	$edu_1 = 4^a$	1	∞

The subscript on edu indicates sweep.

^a Educational level is not allowed to decrease between sweeps.

6.2.4 Imputing categorical edu and binary $sing$ as correlated variables

Our next step is to set up a model which jointly imputes the three category version of edu and $sing$, to demonstrate combining a latent variable for an ordered categorical variable and a latent variable for a binary variable (JM.C5). We use Equations 6.3, but the (1,1) element is no longer restricted

to 1, because it is the variance of a three category variable and we constrain an additional cut-point instead, i.e. we require

$$\Sigma = \begin{pmatrix} \varrho & \kappa \\ \kappa & 1 \end{pmatrix}.$$

Using Equation 6.2 we implement the joint model as a marginal distribution for the *edu* latent variable,

$$x_{i,1}^* \sim N(\nu_{i,1}, \varrho) I(L_{i,1}, R_{i,1}),$$

and a conditional distribution for the *sing* latent variable

$$x_{i,2}^* | x_{i,1}^* \sim N\left(\nu_{i,2} + \frac{\kappa}{\varrho}(x_{i,1}^* - \nu_{i,1}), 1 - \frac{\kappa^2}{\varrho}\right) I(L_{i,2}, R_{i,2}).$$

We place a Gamma(0.001, 0.001) prior on ϱ^{-1} , and as before a Uniform(-1, 1) prior on κ and uninformative priors on all the parameters in the $\nu_{i,j}$ equations. Consistent with our previous models, we impose cut-points at 0 and 1 for *edu* and at 0 for *sing* for identifiability. The boundaries for *edu* and *sing* are as given in Tables 6.4 and 6.2 respectively. We find that the percentages of individuals with missing *edu*₂ increasing their educational level are very similar to that shown in Table 6.5, and the imputations for *sing*₂ are very similar to those shown in Table 6.3. It is not surprising that the results are similar given that there is not much correlation between *edu*^{*} and *sing*^{*} (the posterior mean of κ is 0.02).

We also successfully run a form of this model with an unrestricted covariance matrix, similar to JM.C1d described in Section 6.2.1. In order to get it running, care is needed in setting the T parameter of the Wishart distribution. Instead of using an identity matrix, we set the variance components to estimates of the variance for *edu*^{*} and *sing*^{*}, but continue setting the covariance values to 0. These variance estimates are derived by running a covariate model of missingness for *edu* and *sing* separately. However, we find that this version produces different imputations for educational level, in particular for those with category 2 level qualifications in sweep 1 (17% rather than 4% are imputed to gain a degree, which is less plausible). We suspect that the difference is attributable to the priors, which may be more informative than anticipated. We test this theory by using WinBUGS to forward sample from both sets of priors for the precision matrix. In each case the mean (95% interval) of *sing* is 1.5 (1, 2), but for *edu* is 1.75 (1, 3) for the marginal and conditional distribution set-up and 1.9 (1, 3) for the unrestricted covariance matrix. The direction of this difference is consistent with the direction of the changes in the posterior. This suggests that unintended information can be incorporated through these types of covariance priors and have a sizeable impact if information from the data is limited. We leave further exploration of this issue as a future research question, but it would be useful to understand when and how these priors have different impacts on the model results. An unrestricted covariance matrix is not used in any subsequent models.

6.2.5 Adding continuous *age* to the covariate model of missingness

We would also like to be able to impute the missing values of a continuous covariate, allowing for correlation with binary and categorical covariates with missingness. We explore this by adding *age* to the covariate model of missingness part of JM.C5, forming model JM.C6. Latent variables are not required for continuous variables, instead they enter the joint covariate model of missingness directly. Hence our required joint model is

$$\begin{pmatrix} edu_i^* \\ sing_i^* \\ age_i \end{pmatrix} \sim MVN \left(\begin{pmatrix} \nu_{i,1} \\ \nu_{i,2} \\ \nu_{i,3} \end{pmatrix}, \begin{pmatrix} \varrho & \kappa_{12} & \kappa_{13} \\ \kappa_{12} & 1 & \kappa_{23} \\ \kappa_{13} & \kappa_{23} & \varphi \end{pmatrix} \right).$$

Using Equation 6.2 we implement this as a sequence of conditional univariate models, s.t.

$$\begin{aligned} edu_i^* &\sim N(\nu_{i,1}, \varrho)I(L_{i,1}, R_{i,1}) \\ sing_i^* | edu_i^* &\sim N(\nu_{i,2} + \varrho^{-1}\kappa_{12}(edu_i^* - \nu_{i,1}), 1 - \varrho^{-1}\kappa_{12}^2)I(L_{i,2}, R_{i,2}) \\ age_i | edu_i^*, sing_i^* &\sim N(A_i, B_i) \\ A_i &= \nu_{i,3} + (\varrho - \kappa_{12}^2)^{-1} \{ (\kappa_{13} - \kappa_{12}\kappa_{23})(edu_i^* - \nu_{i,1}) \\ &\quad + (\varrho\kappa_{23} - \kappa_{12}\kappa_{13})(sing_i^* - \nu_{i,2}) \} \\ B_i &= \varphi - (\varrho - \kappa_{12}^2)^{-1} (\kappa_{13}^2 - 2\kappa_{12}\kappa_{13}\kappa_{23} + \varrho\kappa_{23}^2) \\ \nu_{i,1} &= edu.\eta_0 + \sum_{k=2}^3 (edu.\eta_{edu[k]} \times edu_{[k]i1}) \\ \nu_{i,2} &= sing.\eta_0 \\ \nu_{i,3} &= age.\eta_0 + (age.\eta_{age} \times age_{i1}). \end{aligned} \tag{6.7}$$

Since our covariate model of missingness now allows for correlation between *edu* and *age*, and *sing* and *age*, we no longer regress $\nu_{i,1}$ and $\nu_{i,2}$ against sweep 1 *age*. So $\nu_{i,1}$ is regressed against edu_1 only, the second term disappears for $\nu_{i,2}$, and $\nu_{i,3}$ is regressed against age_1 . This model is successfully run.

The posterior mean for parameter $age.\eta_0$ is 0.35, which is the mean difference in the ages on the transformed scale between sweeps 1 and 2 for individuals with observed age at both sweeps. For $age.\eta_{age}$, the posterior mean is 1.00, as we would expect. The posterior mean of the correlation between edu^* and $sing^*$ is 0.02 as for JM.C5, and the posterior means for the other two correlations are almost 0. There are only slight differences in the imputations and model of interest parameter estimates from JM.C5, and we defer a more detailed look at these until the next section.

The covariate model of missingness part of JM.C6 will be our proposed covariate model of missingness for the main analysis of the MCS income data. In common with the other covariate models of missingness considered to date, it does not allow for the possibility of informative missingness in any of the covariates. This issue will be discussed in Section 8.1.3.

6.3 MCS income example, with a response model of missingness

In Section 5.3.3, we built a joint model consisting of a model of interest and a response model of missingness, and so far in this chapter we have formed a joint model consisting of a model of interest and a covariate model of missingness. We now combine three sub-models, so that simultaneously with fitting a model of interest, values for any missing covariates are imputed and possible informative missingness in our response variable is taken into account. The sub-models which we combine are:

model of interest: a version of MoIE with covariate *edu* restricted to three categories (categories 0-1, 2-3 and 4-5 grouped), which we will call MoIF;

covariate model of missingness: as described for JM.C6 in Section 6.2.5;

response model of missingness: as described for JM.R8 in Section 5.3.3.

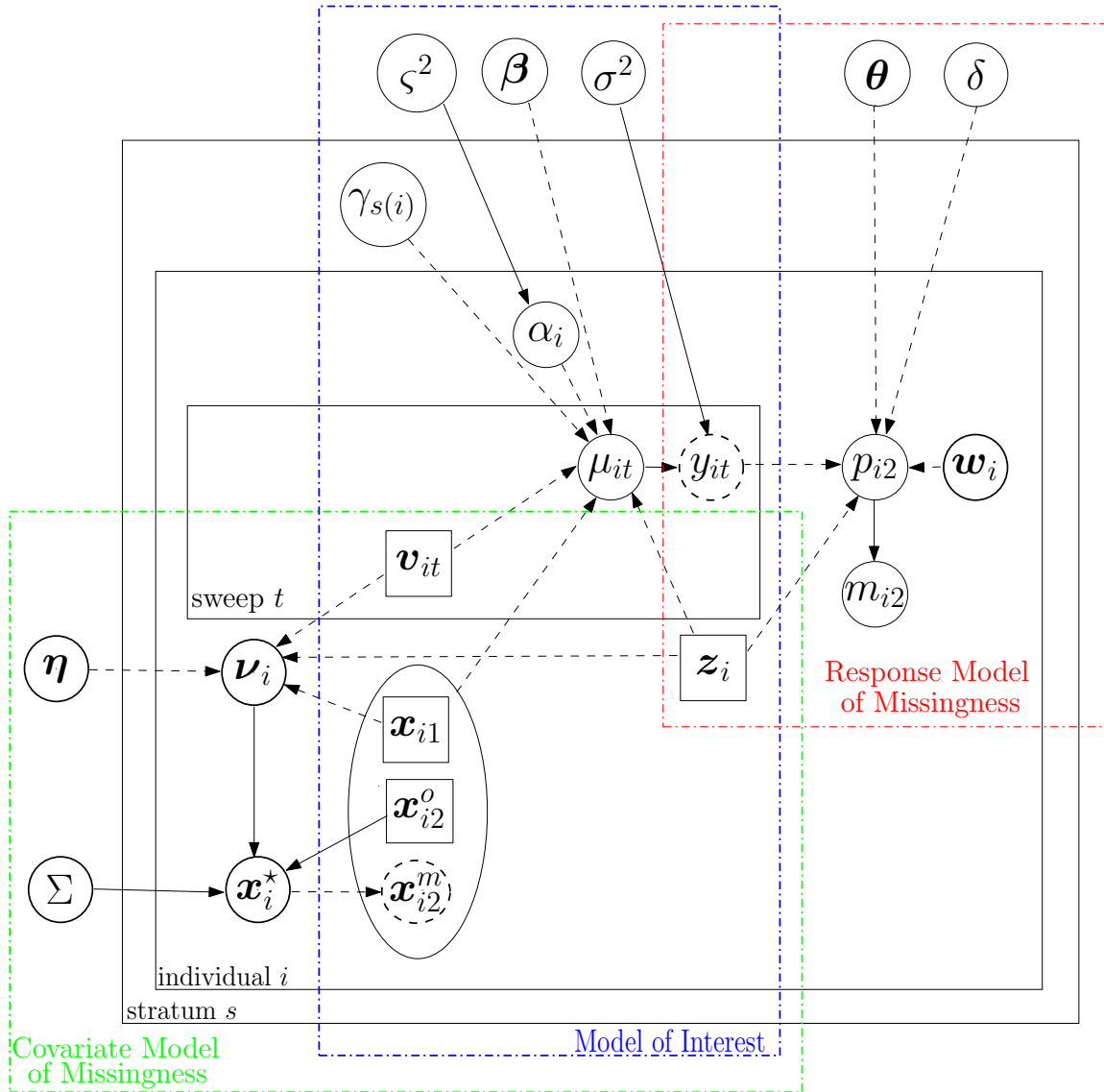
We will refer to this first three part joint model as JMA. Of our four covariates with missing values, three are dealt with by the covariate model of missingness. The exception is *reg*, whose value changes between sweeps for just four observed individuals, and we continue to set missing sweep 2 *reg_i* to their sweep 1 values, as described in Section 5.3.2, prior to running our joint model. Figure 6.3 provides a graphical representation of this type of joint model. For notational convenience in the covariate model of missingness, we assume that an *age** is set to equal *age* (*age* is continuous and does not require a latent variable). The $\mathbf{w}_i$ node in the response model of missingness denotes variables which are used in this sub-model, but not in the model of interest. In our MCS income example, $\mathbf{w}_i = \{ctry_i, sc_{i1}\}$, the country and social class variables which may help explain the response missingness.

JMA is successfully run and its parameter estimates are shown in Table 6.7. The estimates from five other models are also given for comparison. Of these, MoIE and JM.R8 have already been discussed in Sections 5.3.1 and 5.3.3 respectively, and JM.C6 was described in Section 6.2.5. In addition, we run MoIF in isolation using complete cases, and run a joint model, JM.R9, consisting of model of interest MoIF and the response model of missingness used in JM.R8, on the dataset used for running JM.R8.

Comparing model of interest parameters

The differences between MoIE and JM.R8 were discussed at the end of Section 5.3.3. Comparing MoIE and MoIF, we find that reducing the number of categories in *edu* lowers the stratum intercepts a little and reduces the association of higher pay with being white, but has negligible impact on the other variables. The strong evidence that higher pay is associated with higher levels of education remains. Whereas adding a MNAR response model of missingness to MoIE (JM.R8) had a substantial impact on some of the β_{edu} estimates, this impact is reduced when the same response model of missingness is added to MoIF (JM.R9). Adding a covariate model of missingness to MoIF (JM.C6) has a similar impact on the model of interest parameter estimates to the addition of a response model of missingness.

Figure 6.3: Graphical representation for joint model consisting of a model of interest, covariate model of missingness and response model of missingness



In both JM.R9 and JM.C6 we are no longer restricted to complete cases. Looking at JMA compared to JM.R9 and JM.C6, we find that the greatest difference is in the estimate of β_{sing} , whose magnitude more than doubles, providing much greater evidence that lower pay is associated with gaining a partner between sweeps.

Comparing response model of missingness parameters

A comparison of the response model of missingness parameter estimates for JM.R8 and JM.R9 reveals only small differences. However, there are substantial changes in some of these parameters for JMA. In particular, the strength of the relationships between sweep 2 pay being missing and the level of sweep 1 pay and change in pay between sweeps is increased.

Table 6.7: MCS income example: comparison of parameter estimates for different combinations of sub-models

	MoIE	JM.R8	MoIF	JM.R9	JM.C6	JMA
σ	0.33 (0.30,0.37)	0.35 (0.32,0.39)	0.33 (0.30,0.37)	0.35 (0.32,0.39)	0.35 (0.32,0.39)	0.36 (0.33,0.40)
ς	0.21 (0.17,0.25)	0.23 (0.19,0.27)	0.21 (0.17,0.25)	0.23 (0.19,0.27)	0.24 (0.20,0.27)	0.23 (0.19,0.26)
γ_1	2.14 (1.95,2.33)	2.00 (1.83,2.17)	2.04 (1.88,2.20)	1.98 (1.83,2.13)	1.99 (1.84,2.15)	1.96 (1.81,2.11)
γ_2	2.10 (1.93,2.28)	1.96 (1.80,2.12)	2.01 (1.86,2.16)	1.94 (1.80,2.08)	1.96 (1.82,2.11)	1.92 (1.78,2.06)
γ_3	2.09 (1.86,2.31)	1.96 (1.76,2.16)	1.96 (1.76,2.17)	1.92 (1.74,2.10)	1.95 (1.77,2.13)	1.89 (1.71,2.07)
γ_4	2.18 (1.91,2.45)	2.01 (1.78,2.25)	2.07 (1.82,2.33)	1.98 (1.77,2.21)	2.02 (1.79,2.24)	1.95 (1.73,2.16)
γ_5	2.07 (1.87,2.27)	1.93 (1.75,2.11)	1.98 (1.80,2.15)	1.90 (1.74,2.07)	1.93 (1.76,2.09)	1.88 (1.72,2.04)
γ_6	2.06 (1.85,2.26)	1.94 (1.75,2.14)	1.97 (1.78,2.16)	1.93 (1.76,2.11)	1.94 (1.76,2.12)	1.92 (1.75,2.10)
γ_7	2.08 (1.88,2.29)	1.86 (1.68,2.04)	1.99 (1.80,2.17)	1.84 (1.68,2.01)	1.86 (1.69,2.03)	1.82 (1.66,1.98)
γ_8	2.06 (1.81,2.32)	1.93 (1.69,2.16)	1.97 (1.74,2.21)	1.89 (1.67,2.12)	1.90 (1.67,2.13)	1.88 (1.66,2.10)
γ_9	2.10 (1.90,2.30)	1.97 (1.79,2.15)	2.00 (1.83,2.17)	1.95 (1.79,2.11)	1.97 (1.81,2.13)	1.93 (1.77,2.08)
β_{age}	0.12 (0.09,0.16)	0.11 (0.08,0.14)	0.12 (0.09,0.16)	0.11 (0.08,0.14)	0.12 (0.08,0.15)	0.11 (0.08,0.14)
$\beta_{edu}[\text{NVQ1}]$	-0.16 (-0.32,0.00)	-0.03 (-0.17,0.11)				
$\beta_{edu}[\text{NVQ2}/\text{NVQ2\&3}]^a$	0.05 (-0.08,0.18)	0.12 (0.01,0.23)	0.15 (0.06,0.25)	0.16 (0.08,0.24)	0.16 (0.08,0.25)	0.16 (0.08,0.24)
$\beta_{edu}[\text{NVQ3}]$	0.09 (-0.06,0.23)	0.20 (0.08,0.32)				
$\beta_{edu}[\text{NVQ4}/\text{NVQ4\&5}]^a$	0.24 (0.10,0.38)	0.28 (0.16,0.41)	0.35 (0.24,0.45)	0.31 (0.22,0.41)	0.31 (0.21,0.40)	0.32 (0.22,0.41)
$\beta_{edu}[\text{NVQ5}]$	0.44 (0.18,0.70)	0.49 (0.26,0.72)				
β_{eth}	-0.07 (-0.21,0.08)	-0.09 (-0.22,0.04)	-0.04 (-0.18,0.10)	-0.06 (-0.19,0.07)	-0.06 (-0.19,0.07)	-0.05 (-0.18,0.07)
β_{reg}	-0.18 (-0.31,-0.05)	-0.14 (-0.26,-0.02)	-0.18 (-0.31,-0.05)	-0.14 (-0.26,-0.02)	-0.14 (-0.26,-0.01)	-0.12 (-0.24,-0.01)
β_{sing}	-0.07 (-0.14,0.00)	-0.06 (-0.13,0.01)	-0.07 (-0.14,0.00)	-0.06 (-0.13,0.01)	-0.07 (-0.14,0.01)	-0.16 (-0.25,-0.07)
θ_0		0.11 (-1.49,1.73)	-0.02 (-1.65,1.62)			-1.45 (-3.16,0.13)
θ_{level}		0.13 (-0.04,0.31)	0.14 (-0.03,0.33)			0.33 (0.13,0.56)
δ_{change}		0.16 (-0.04,0.37)	0.17 (-0.03,0.39)			0.40 (0.17,0.67)
$\theta_{ctry}[\text{Wales}]$		-0.15 (-0.73,0.44)	-0.15 (-0.74,0.44)			-0.14 (-0.77,0.49)
$\theta_{ctry}[\text{Scotland}]$		0.24 (-0.31,0.82)	0.24 (-0.31,0.82)			0.25 (-0.36,0.87)
$\theta_{ctry}[\text{Northern Ireland}]$		0.21 (-0.37,0.81)	0.20 (-0.38,0.81)			0.20 (-0.41,0.84)
θ_{eth}		-0.96 (-1.58,-0.37)	-0.98 (-1.61,-0.39)			-1.17 (-1.86,-0.52)
$\theta_{sc}[\text{intermediate}]$		-0.40 (-1.07,0.26)	-0.38 (-1.05,0.31)			-0.07 (-0.77,0.66)
$\theta_{sc}[\text{lower supervisory \& technical}]$		-0.44 (-1.40,0.52)	-0.41 (-1.36,0.57)			0.04 (-0.93,1.07)
$\theta_{sc}[\text{semi-routine \& routine}]$		-0.44 (-1.17,0.28)	-0.41 (-1.14,0.32)			0.04 (-0.68,0.80)
$edu.\eta_0$					-0.26 (-0.37,-0.16)	-0.27 (-0.38,-0.17)
$edu.\eta_{edu}[\text{NVQ2\&3}]$					0.83 (0.71,0.95)	0.83 (0.72,0.96)
$edu.\eta_{edu}[\text{NVQ4\&5}]$					5.01 (1.96,9.20)	7.56 (2.92,9.63)
$sing.\eta_0$					-0.42 (-0.56,-0.29)	-0.36 (-0.51,-0.21)
$age.\eta_0$					0.35 (0.34,0.36)	0.35 (0.34,0.36)
$age.\eta_{age}$					1.00 (0.99,1.00)	1.00 (0.99,1.00)

Table shows the posterior mean, with the 95% interval in brackets

^a a first definition (NVQ2 and NVQ4) applies to models MoIE and JM.R8; the second definition (NVQ2&3 and NVQ4&5) applies to MoIF, JM.R9, JM.C6 and JMA.

Comparing covariate model of missingness parameters and the covariate imputations

The covariate model of missingness parameter estimates change little between the models with (JMA) and without (JM.C6) a response model of missingness. The covariate imputations for JM.C6 and JMA are summarised in Table 6.8, which takes the same form as Table 6.3. Mostly the differences between the two models are small, although JMA imputes 59% of individuals with missing sweep 2 *sing* to remain single, compared to the JM.C6 posterior mean of 66%. However, there are substantial changes from our original simplistic assumptions (Section 5.3.2), shown under the heading “Simplistic”, particularly for *sing*.

Table 6.8: MCS income example: comparison of covariate imputations from joint models with (JMA) and without (JM.C6) a response model of missingness and simplistic imputations

	Simplistic ^d	JM.C6		JMA	
<i>edu</i> ₁₁ ^a	39	33.4	(27,38)	34.6	(29,39)
<i>edu</i> ₁₂ ^a	0	5.6	(1,12)	4.4	(0,10)
<i>edu</i> ₁₃ ^a	0	0.0	(0,0)	0.0	(0,0)
<i>edu</i> ₂₁ ^a	0	0.0	(0,0)	0.0	(0,0)
<i>edu</i> ₂₂ ^a	101	97.3	(92,101)	98.2	(94,101)
<i>edu</i> ₂₃ ^a	0	3.7	(0,9)	2.8	(0,7)
<i>edu</i> ₃₁ ^a	0	0.0	(0,0)	0.0	(0,0)
<i>edu</i> ₃₂ ^a	0	0.0	(0,0)	0.0	(0,0)
<i>edu</i> ₃₃ ^a	26	26.0	(26,26)	26.0	(26,26)
<i>sing</i> ₁ ^b	166	109.9	(95,124)	98.5	(79,116)
<i>sing</i> ₂ ^b	0	56.1	(42,71)	67.5	(50,87)
<i>edusing</i> ₁₁ ^c	73	56.5	(50,63)	55.1	(48,62)
<i>edusing</i> ₁₂ ^c	17	27.9	(22,34)	30.5	(24,37)
<i>edusing</i> ₂₁ ^c	230	197.1	(185,209)	189.7	(176,203)
<i>edusing</i> ₂₂ ^c	73	107.9	(97,119)	114.8	(102,129)
<i>edusing</i> ₃₁ ^c	88	81.4	(76,88)	78.6	(73,85)
<i>edusing</i> ₃₂ ^c	24	34.3	(29,40)	36.2	(30,43)

Table shows the posterior mean of the number of subjects imputed in each category, with the 95% interval in brackets

^a The first index indicates the value of sweep 1 *edu*, and the second index the value of sweep 2 *edu*.

^b The index indicates the value of sweep 2 *sing*.

^c The first index indicates the value of sweep 2 *edu*, and the second index the value of sweep 2 *sing*; count includes individuals with observed *edu* and *sing* in sweep 2.

^d The simplistic imputations described in Section 5.3.2.

A concern with JMA is the estimation of the parameters associated with imputing *edu*, stemming from the lack of individuals gaining higher level educational qualifications between sweeps (see Table 6.9). We would like some additional data to help with the estimation of this part of the model, and in the next Chapter we explore ways of doing this.

Table 6.9: MCS income example: sweep 1 by sweep 2 counts of three category education level for the 505 MCS individuals

	$edu_2=0$	$edu_2=1$	$edu_2=2$	$edu_2=missing$
$edu_1=0$	51	5	1	39
$edu_1=1$	0	197	4	101
$edu_1=2$	0	0	81	26

Summary: modelling missing covariates

Imputing missing values for multiple covariates is difficult if they are not all Gaussian or easily transformed to Gaussian, and we wish to allow for correlation between them. Using our MCS example, we investigate an approach that uses a multivariate probit model and links each covariate with an underlying latent variable. We explore implementation issues, looking at different parameterisations and in particular whether we need to restrict the covariance matrix of the underlying MVN distribution to correlation form for our purposes. Our experience suggests the restriction is unnecessary. For our example, we did not attempt to model more than three covariates simultaneously, but the most promising avenue for modelling larger numbers of covariates is by using an unrestricted covariance matrix, taking care to ensure that unintended information is not incorporated through the covariance priors. This can be checked by forward sampling from the prior. Unresolved difficulties in implementing a model that includes an ordered categorical variable with more than three categories are encountered.

In multiple imputation, one approach involves specifying a joint multivariate normal distribution for variables with missing values (Section 2.3), and this effectively requires solving the problems discussed in this chapter. The difficulties of this help explain the popularity of the alternative approach that specifies separate conditional distributions for each variable, which although not theoretically rigorous, is easier to implement.

Our preferred covariate model of missingness for the MCS income example simultaneously imputes the missing values for edu , $sing$ and age allowing for correlation. This is combined with JM.R8 (our preferred two part joint model comprising a model of interest and response model of missingness) to form a three part joint model, JMA, which is used as the starting point for the next phase of our research. To implement this model we have to reduce the number of categories of edu from six to three.

Chapter 7

Incorporating information from additional sources

We have seen that estimating the parameters in the covariate model of missingness and response model of missingness can be difficult when we have limited information. By incorporating data from other sources we can increase the amount of available information, and in this chapter we explore ways of doing this. For our covariate model of missingness, where problems with imputing education are caused by so few individuals gaining higher level qualifications between sweeps, we add data from another study on individuals with similar characteristics to those being modelled. By contrast, we incorporate expert knowledge to help with the response model of missingness parameter estimation.

All the models discussed in this chapter take JMA (Joint Model A), which was discussed at the end of Section 6.3, as a starting point. For consistency, all these models are run with 2 chains for 100,000 iterations of which 50,000 are burn-in and a thinning rate of 5.

7.1 Data from another study

Combining information from multiple surveys can be beneficial in a number of ways. For example, Schenker and Raghunathan (2007) describe four projects where information from multiple surveys is combined to adjust for various non-sampling errors, including measurement error and coverage error. They also discuss potential sources of incomparability, which include differences in the types of respondents, modes of interviewing, survey contexts, sample designs and survey questions. Raghunathan and Siscovick (1998) develop a hybrid strategy for combining information on an explanatory variable from several sources, and show that in their application the estimates of interest obtained using the multiple sources were better in terms of bias, mean-square error and confidence interval coverage than those based on a single source. Their hybrid approach is a two stage strategy, in which a Bayesian model is used to estimate the explanatory variable by combining different sources of information, and then these estimates are used in a standard analysis allowing for the additional uncertainty from the first stage.

Combining data from different studies using models estimated in a Bayesian framework, has been the subject of a number of recent papers. Jackson *et al.* (2006) combine related aggregate-level and individual-level data from different sources, by carrying out simultaneous regressions with common coefficients on the data from the two levels. Their method, which is termed hierarchical related regression, is extended in Jackson *et al.* (2008a) and further examples are given in Jackson *et al.* (2008b). The aims of combining data in this context are to reduce ‘ecological bias’ and generally improve inferences from using a single dataset. Molitor *et al.* (2009) combine multiple data sources using Bayesian graphical models to inform about different aspects of a research question, linking various sub-models by shared parameters. A further example of combining information from two surveys using a hierarchical Bayesian approach is provided by Raghunathan *et al.* (2007), who find that their combined estimation procedure helps adjust for both non-response and non-coverage errors.

Our motivation is to improve our covariate imputations by supplementing sparse data, and our approach is to fit simultaneous regressions with shared coefficients to related datasets from different sources. We now explore the benefits of using data from another longitudinal survey to help with the imputation of the missing covariates in our MCS income example.

7.1.1 BCS70 educational level data

The additional data is taken from a different British birth cohort study, the BCS70, which was briefly introduced in Section 3.1. Sweep 5 of the BCS70 was carried out in 1999/2000 when the cohort members were aged 30, and sweep 6 took place in 2004/5 at age 34. The timings of these two sweeps are not very different to those for the first two sweeps of the MCS, and the information collected on the BCS70 cohort members covers similar domains of interest to those used for the parents of the MCS cohort members, including income, education and work status details. Given the ages of the BCS70 cohort members, we are able to select a subset who have similar characteristics to the MCS main respondents whose income we are modelling. This BCS70 subset is restricted to female cohort members who are in a household with children, single in sweep 5, and in work and not self-employed in sweeps 5 and 6. The inclusion criteria selects on households with any children aged 0-16, rather than being limited to younger children. This has the advantage of increasing the sample size. Further detail on these criteria can be found in Appendix A.3, and we find that there are 157 fully observed BCS70 cohort members who meet these criteria.

For these 157 individuals, we create a set of data that includes educational level at sweeps 5 and 6, and whether the individual gained a partner by sweep 6. Educational variables are not recorded directly for the highest level of national vocational qualification equivalence, and instead are calculated using a number of academic and vocational qualification variables. Details of these calculations and the source of the variables are provided in Appendix A.3.

The BCS70 cohort members who are all aged 30 at sweep 5 are on average a little older than the mothers of the MCS cohort, who have a mean age of 26.9 years and 95% interval, calculated from the 2.5% and 97.5% quantiles, of (19,39) at sweep 1. 41.4% of the BCS70 individuals gain a partner between sweeps compared to 33.6% of the MCS individuals, which may in part reflect the slightly longer time of 4 years between the two sweeps for the BCS70 data, compared to a mean of 3 years for the MCS data. The number of individuals gaining higher educational levels between sweeps is shown in Table 7.1. The percentage of individuals increasing their educational level is similar for the BCS70 and MCS individuals who have educational level 0 at sweep 1, but higher for the BCS70 individuals if they have level 1 educational qualifications at sweep 1 (see Table 7.2).

Table 7.1: MCS income example: sweep 5 by sweep 6 counts of educational level for 157 selected BCS70 individuals

	<i>BCSedu</i> ₆ =0	<i>BCSedu</i> ₆ =1	<i>BCSedu</i> ₆ =2
<i>BCSedu</i> ₅ =0	48	6	1
<i>BCSedu</i> ₅ =1	0	65	6
<i>BCSedu</i> ₅ =2	0	0	31

Table 7.2: MCS income example: comparison of educational level data from MCS and BCS70
observed *edu* from MCS (% of row total) *BCSedu* from BCS70 (% of row total)

	<i>edu</i> ₂ =0	<i>edu</i> ₂ =1	<i>edu</i> ₂ =2	total		<i>edu</i> ₆ =0	<i>edu</i> ₆ =1	<i>edu</i> ₆ =2	total
<i>edu</i> ₁ =0	89.5	8.8	1.8	100.0	<i>edu</i> ₅ =0	87.3	10.9	1.8	100.0
<i>edu</i> ₁ =1	0.0	98.0	2.0	100.0	<i>edu</i> ₅ =1	0.0	91.5	8.5	100.0
<i>edu</i> ₁ =2	0.0	0.0	100.0	100.0	<i>edu</i> ₅ =2	0.0	0.0	100.0	100.0

The subscript on *edu* indicates sweep.

An examination of the educational level data, split into individuals who gain a partner between sweeps and those who do not, reveals that those who gain a partner are more likely to increase their educational level from 0 to 1. 19% and 18% of the individuals who have educational level 0 in the first sweep and gain a partner between sweeps also increase their educational level for the MCS and BCS70 data respectively. The proportion gaining a degree is similar for those remaining single and those gaining a partner in both studies.

7.1.2 Use in the covariate model of missingness

We now incorporate BCS70 data into the covariate model of missingness part of our joint MCS income model, with the aim of increasing the precision of the imputations. This is done by adding an extra set of equations to model the BCS70 data, which have the same parameters as the equations for imputing the missing MCS covariates.

Recall that in JMA, the covariate model of missingness models the joint distribution of the MCS covariates *edu*, *sing* and *age* using a sequence of conditional univariate models as given by Equation 6.7. Using a parallel set of equations of exactly this form to model the equivalent BCS70 variables is undesirable on two counts. Firstly, the variance and covariances of *BCSage* are 0, so *BCSage* will provide no information about *BCSedu* or *BCSsing* and our three variable joint model collapses to a two variable joint model. Secondly, the time between the two modelled sweeps is approximately a year longer for BCS70 than MCS, which provides more opportunity for individuals to gain additional educational qualifications or acquire a partner.

We deal with the first issue by only setting up equations for *BCSedu* and *BCSsing*. Note that this solution is implementable because we are using a sequence of conditional univariate models, but would not work with a joint model. To take account of the time difference, we add an additional explanatory variable, sweep gap (*sgap*), to the equations which calculate the means of *edu*^{*} and *sing*^{*}. Hence the equations for $\nu_{i,1}$ and $\nu_{i,2}$ in Equation 6.7 are replaced by

$$\begin{aligned}\nu_{i,1} &= edu.\eta_0 + \sum_{k=2}^3 (edu.\eta_{edu[k]} \times edu_{[k]i1}) + (edu.\eta_{sgap} \times MCSsgap) \\ \nu_{i,2} &= sing.\eta_0 + (sing.\eta_{sgap} \times MCSsgap)\end{aligned}\tag{7.1}$$

where *MCSsgap* is set to 3. This model could be elaborated to allow *MCSsgap* to vary by individual, in line with their varying differences in age between sweeps, but implementing this would be complicated by missing data.

The new equations for modelling the BCS70 data are

$$\begin{aligned}eduB_i^* &\sim N(\zeta_{i,1}, \varrho)I(LB_{i,1}, RB_{i,1}) \\ singB_i^* | eduB_i^* &\sim N(\zeta_{i,2} + \varrho^{-1}\kappa_{12}(eduB_i^* - \zeta_{i,1}), 1 - \varrho^{-1}\kappa_{12}^2)I(LB_{i,2}, RB_{i,2}) \\ \zeta_{i,1} &= edu.\eta_0 + \sum_{k=2}^3 (edu.\eta_{edu[k]} \times BCSedu_{[k]i1}) + (edu.\eta_{sgap} \times BCSsgap) \\ \zeta_{i,2} &= sing.\eta_0 + (sing.\eta_{sgap} \times BCSsgap)\end{aligned}\tag{7.2}$$

where *BCSsgap* is set to 4 and boundary conditions are imposed as for the MCS data. Whereas *edu* and *sing* in our MCS data have observed and missing values, only fully observed individuals have been selected for inclusion in our BCS70 data, so *BCSedu* and *BCSsing* have no missing values. The key point is that parameters ϱ , κ_{12} , *edu*. η_0 , *edu*. $\eta_{edu[k]}$, *edu*. η_{sgap} , *sing*. η_0 and *sing*. η_{sgap} are common to both sets of equations and so the BCS70 data helps with their estimation.

We make these changes to the covariate model of missingness in JMA, forming a new joint model, which we call JMB. An examination of the trace plots and the Gelman-Rubin convergence statistic suggests that this model converges. The parameter estimates for JMB are shown in Table 7.3, which includes a comparison with JMA.

Table 7.3: MCS income example: comparison of parameter estimates with (JMB) and without (JMA) including external data from BCS70

	JMA		JMB		% diff ^a
σ	0.36	(0.33,0.40)	0.36	(0.33,0.40)	-0.1
ς	0.23	(0.19,0.26)	0.23	(0.19,0.26)	0.1
γ_1	1.96	(1.81,2.11)	1.96	(1.81,2.12)	0.2
γ_2	1.92	(1.78,2.06)	1.92	(1.78,2.06)	0.2
γ_3	1.89	(1.71,2.07)	1.90	(1.72,2.07)	0.2
γ_4	1.95	(1.73,2.16)	1.95	(1.73,2.17)	0.2
γ_5	1.88	(1.72,2.04)	1.89	(1.72,2.05)	0.2
γ_6	1.92	(1.75,2.10)	1.92	(1.75,2.11)	0.1
γ_7	1.82	(1.66,1.98)	1.83	(1.67,2.00)	0.2
γ_8	1.88	(1.66,2.10)	1.88	(1.66,2.10)	0.1
γ_9	1.93	(1.77,2.08)	1.93	(1.77,2.09)	0.1
β_{age}	0.11	(0.08,0.14)	0.11	(0.08,0.14)	0.3
$\beta_{edu[NVQ2\&3]}$	0.16	(0.08,0.24)	0.16	(0.08,0.24)	0.1
$\beta_{edu[NVQ4\&5]}$	0.32	(0.22,0.41)	0.31	(0.22,0.41)	-1.0
β_{eth}	-0.05	(-0.18,0.07)	-0.06	(-0.18,0.07)	3.0
β_{reg}	-0.12	(-0.24,-0.01)	-0.13	(-0.25,0.00)	1.9
β_{sing}	-0.16	(-0.25,-0.07)	-0.16	(-0.25,-0.07)	-1.3
θ_0	-1.45	(-3.16,0.13)	-1.39	(-3.06,0.32)	-4.4
θ_{level}	0.33	(0.13,0.56)	0.32	(0.11,0.55)	-2.6
δ_{change}	0.40	(0.17,0.67)	0.39	(0.15,0.65)	-2.5
$\theta_{ctry[Wales]}$	-0.14	(-0.77,0.49)	-0.14	(-0.77,0.49)	0.1
$\theta_{ctry[Scotland]}$	0.25	(-0.36,0.87)	0.25	(-0.36,0.86)	1.7
$\theta_{ctry[Northern Ireland]}$	0.20	(-0.41,0.84)	0.21	(-0.42,0.83)	1.6
θ_{eth}	-1.17	(-1.86,-0.52)	-1.16	(-1.85,-0.50)	-1.0
$\theta_{sc[intermediate]}$	-0.07	(-0.77,0.66)	-0.09	(-0.80,0.65)	24.6
$\theta_{sc[lower supervisory \& technical]}$	0.04	(-0.93,1.07)	0.01	(-0.98,1.03)	-79.2
$\theta_{sc[semi-routine \& routine]}$	0.04	(-0.68,0.80)	0.02	(-0.73,0.78)	-56.2
$edu.\eta_0$	-0.27	(-0.38,-0.17)	-0.54	(-0.87,-0.18)	98.8
$edu.\eta_{edu[NVQ2\&3]}$	0.83	(0.72,0.96)	0.88	(0.78,0.97)	5.0
$edu.\eta_{edu[NVQ4\&5]}$	7.56	(2.92,9.63)	6.65	(2.84,9.44)	-12.0
$edu.\eta_{sgap}$			0.08	(-0.02,0.17)	
$sing.\eta_0$	-0.36	(-0.51,-0.21)	-0.77	(-1.65,0.12)	112.9
$sing.\eta_{sgap}$			0.14	(-0.13,0.39)	
$age.\eta_0$	0.35	(0.34,0.36)	0.35	(0.34,0.36)	0.0
$age.\eta_{age}$	1.00	(0.99,1.00)	1.00	(0.99,1.00)	0.0

Table shows the posterior mean, with the 95% interval in brackets.

^a % difference in parameter estimates from JMA to JMB

The additional data has little impact on the model of interest and response model of missingness parameters, with large percentage changes only occurring for parameters with posterior means close to zero. However, there are substantial changes to the parameters in the covariate model of missingness equations for calculating the means of the latent variables edu^* and $sing^*$, resulting from the addition of the $sgap$ covariate. In line with our expectation that the longer the gap between sweeps, the more likely an individual is to gain higher educational qualifications and acquire a partner, the signs of the posterior means of $edu.\eta_{sgap}$ and $sing.\eta_{sgap}$ are positive. There is evidence that the effect is stronger for $sing$ than edu , although there is also greater uncertainty in the parameter estimate.

Our real interest is in the covariate imputations rather than the covariate model of missingness parameter estimates. A comparison of these between JMB and JMA is presented in Table 7.4. Only those elements which can vary are shown (by design edu_{21} , edu_{31} and edu_{32} are 0 and edu_{33} is 26). The differences are slight, and the inadequacies in the imputations such as no individual increasing their educational level by two categories remains. While we had hoped that the additional data would increase precision, the 95% intervals on these summary statistics have not reduced and in fact there is increased uncertainty for edu_{22} and edu_{23} .

Table 7.4: MCS income example: comparison of covariate imputations with (JMB) and without (JMA) including external data from BCS70

	JMA		JMB		% diff ^c
edu_{11}^a	34.6	(29,39)	35.2	(30,39)	1.7
edu_{12}^a	4.4	(0,10)	3.8	(0,9)	-13.1
edu_{13}^a	0.0	(0,0)	0.0	(0,0)	
edu_{22}^a	98.2	(94,101)	97.1	(92,101)	-1.1
edu_{23}^a	2.8	(0,7)	3.9	(0,9)	37.0
$sing_1^b$	98.5	(79,116)	98.5	(80,116)	0.0
$sing_2^b$	67.5	(50,87)	67.5	(50,86)	0.0

Table shows posterior means, and 95% intervals in brackets

^a The first index indicates the value of sweep 1 edu , and the second index the value of sweep 2 edu .

^b The index indicates the value of sweep 2 $sing$.

^c % difference in parameter estimates from JMA to JMB

To understand what is happening, we re-run JMB, but this time we increase the number of records from BCS70 tenfold by using 10 replicates of each record (JMBx10). This has little impact on the 95% interval widths for the summaries of the covariate imputations. However, as the form of the covariate model of missingness is unchanged between JMB and JMBx10, we can now compare the precision of the estimates of its parameters. We find that the 95% interval widths for the covariate model of missingness parameter estimates relating to edu and $sing$ are reduced by about a third, as shown in Table 7.5. By contrast, the 95% interval widths for the age related parameters whose estimation is not helped by the addition of the BCS70 data are unchanged. However, the estimation of the latent

variables is determined by the mean and the variance of a Normal distribution. While the means are now being estimated with greater precision, the estimated variances are not shrinking (see σ_{edu^*} and σ_{sing^*} in Table 7.5). The true variability of a latent variable is estimated more precisely with extra data, but is not expected to go down as it is not influenced by sample size. Hence the increased precision in the covariate model of missingness parameter estimates does not filter through to the estimates of individual edu_i^* and $sing_i^*$, or the missing edu_i and $sing_i$.

Table 7.5: MCS income example: comparison of interval widths of covariate model of missingness parameters for joint models run with 157 BCS70 records (JMB) and 10 replicates of the BCS70 records (JMBx10)

	JMB			JMBx10			interval width % diff ^a
	posterior mean	95% interval	interval width	posterior mean	95% interval	interval width	
$edu.\eta_0$	-0.54	(-0.87,-0.18)	0.70	-0.57	(-0.80,-0.35)	0.45	35.0
$edu.\eta_{edu[NVQ2\&3]}$	0.88	(0.78,0.97)	0.19	0.92	(0.88,0.97)	0.10	49.3
$edu.\eta_{edu[NVQ4\&5]}$	6.65	(2.84,9.44)	6.60	6.43	(3.21,8.87)	5.66	14.2
$edu.\eta_{sgap}$	0.08	(-0.02,0.17)	0.19	0.07	(0.01,0.13)	0.12	40.1
$sing.\eta_0$	-0.77	(-1.65,0.12)	1.77	-0.78	(-1.44,-0.13)	1.31	26.3
$sing.\eta_{sgap}$	0.14	(-0.13,0.39)	0.52	0.14	(-0.03,0.31)	0.34	35.8
$age.\eta_0$	0.35	(0.34,0.36)	0.02	0.35	(0.34,0.36)	0.02	-0.6
$age.\eta_{age}$	1.00	(0.99,1.00)	0.02	1.00	(0.99,1.00)	0.02	0.0
$\sigma_{edu^*}^b$	0.26	(0.23,0.28)	0.06	0.30	(0.28,0.31)	0.04	36.3
$\sigma_{sing^*}^b$	0.99	(0.96,1.00)	0.04	1.00	(0.99,1.00)	0.01	79.0
$\sigma_{age^*}^b$	0.07	(0.07,0.08)	0.01	0.07	(0.07,0.08)	0.01	-1.2

^a % difference in interval widths from JMB to JMBx10

^b σ_{edu^*} , σ_{sing^*} , σ_{age^*} are the standard deviations of the latent variables for *edu*, *sing* and *age* respectively

We also run the covariate model of missingness of JMB in isolation (JMBcmom), and find that there are changes in both the covariate model of missingness parameter estimates and covariate imputations, particularly for those relating to *sing*. For example, the posterior mean of the percentage of individuals with missing partnership status who remain single in sweep 2 increases from 59% (48%,72%) for JMB to 66% (59%,78%) for JMBcmom (95% intervals shown in brackets). This demonstrates that information is being drawn from other parts of the joint model to help with the covariate imputations.

Using additional BCS70 data has not changed the posterior means of the covariate imputations substantially, reassuring us that it is reasonable to combine the MCS and BCS70 data and providing extra confidence in the results. Although we can estimate the parameters in the covariate model of missingness with more accuracy, we do not get increased precision in the covariate imputations.

7.2 Expert knowledge

So far, all our models have been set up with non-informative or mildly informative priors. However, one of the advantages of the Bayesian approach is the ability to incorporate expert knowledge through informative priors (Best *et al.*, 1996; Dunn *et al.*, 2003; White *et al.*, 2004), and this has been an area of considerable research. An example in a missing data context is provided by Scharfstein *et al.* (2003) who incorporate prior beliefs about selection bias parameters, elicited by asking an expert about the odds of drop-out for a proportional change in response. Their motivation is to allow weak distributional assumptions to be placed on a continuous outcome in a selection model. Working with binary outcomes, White *et al.* (2008a) discuss eliciting prior beliefs about the odds ratio between success and missingness for use in a pattern-mixture model.

O'Hagan *et al.* (2006) review the psychological and statistical aspects of eliciting priors and discuss a selection of published examples. The basic technique for eliciting a distribution involves eliciting summaries of an expert's distribution, and then fitting a distribution to match the elicited summaries. Typically these summaries are quantiles or central credible intervals. The literature suggests that there is a tendency for credible intervals to be too short, so eliciting credible intervals with high probabilities should be avoided. Overfitting (eliciting more summaries than are needed) and feedback are two recommended ways of improving an elicitation. Most of the literature suggests that it is usually better to ask people about observable quantities (possibly conditioned on other observable quantities), rather than unobservable quantities such as regression coefficients (Kadane and Wolfson, 1998).

Eliciting a distribution for more than one variable adds substantial complexity to the process, unless the variables can be considered independent. In the case of assumed independence, the expert's beliefs can be elicited separately for each variable. Otherwise measures of association between the variables are required, which may be more accurately elicited through conditional or joint probabilities, rather than asking questions directly about correlations. Chaloner *et al.* (1993) present a method of graphical elicitation for a parametric bivariate distribution, in which the expert specifies the quantiles of the marginal distributions for two variables and the probability that both variables are larger than their median value. However, research into the elicitation of multivariate distributions is not yet well developed.

In discussing Bayesian approaches to randomised trials, Spiegelhalter *et al.* (1994) recommend using a *community* of priors to cover the perspectives of a range of individuals. These may include a *reference* prior (represents minimal prior information), a *clinical* prior (formalised opinion of well-informed specific individuals or derived from meta-analyses of similar studies), a *sceptical* prior and an *enthusiastic* prior. The sceptical and enthusiastic priors may be thought to provide reasonable

bounds to the community of priors. Further examples of constructing a community of priors are provided by Greenhouse and Seltman (2005), who create two priors based on historical information from previous clinical trials in addition to sceptical and enthusiastic priors, and White *et al.* (2005), who create informative priors for treatment $\times$ covariate interactions through expert elicitation.

A number of specialist software packages are now available to help with the process of elicitation. These include the SHEffield ELicitation Framework (SHELF) and ELICITOR.

SHELF is a framework for eliciting the beliefs of a group of experts, developed by Tony O’Hagan and Jeremy Oakley (available from <http://www.tonyohagan.co.uk/shelf/>). It consists of a number of components, including briefing notes, templates for the elicitation and software for fitting distributions in the R language. SHELF accommodates several different protocols for eliciting a distribution, allowing specific probabilities, lower and upper quartiles or lower and upper tertiles to be elicited.

ELICITOR is graphical elicitation software, written as a WinBUGS add-on by Mary Kynn, to provide informative priors for Bayesian generalised linear models, for ecological applications (available from <http://silmaril.math.sci.qut.edu.au/~whateley/>). The method is a modification of that developed by Garthwaite and Al-Awadhi (2006), which allows the use of piecewise linear functions, and elicits conditional and unconditional medians and quartiles from an expert using interactive graphical software to allow covariances to be assessed. The original version of ELICITOR only allowed for the elicitation of normal prior distributions for logistic regression models, but has subsequently been extended to encompass different link functions and prior distributions. However, ELICITOR does not currently allow covariances to be elicited, instead assuming that each variable is independent. It uses the method of conditional mean priors (Bedrick *et al.*, 1996) on the scale of the response variable. This method is applicable to generalised linear models, and involves eliciting information about mean responses at fixed values of the covariates, which is then used to derive an informative prior distribution on the regression coefficients.

Three methods of eliciting a credible interval are given in Mary Kynn’s ELICITOR Version 3.3 User Manual (available from <http://silmaril.math.sci.qut.edu.au/~whateley/UserManual.pdf>). In the *direct fractile elicitation method*, the expert is invited to put an interval around their initial estimate, such that they are $x\%$ sure of their estimate. A typical interval is 90%, so the 5% and 95% fractiles are elicited. An alternative traditional method is the *bisection method*, in which the expert is told that the true value is below their original estimate and asked to choose another value which they believe the true value is equally likely to be above or below. This value is the elicited lower quartile, and an upper quartile can be elicited in the same way, dividing the range into four equally likely intervals. In the third method, developed for ELICITOR, the expert is asked to provide upper, lower and middle estimates and then asked how sure they are that the true value lies in this range.

We now investigate the elicitation of expert priors by creating informative priors for the parameters of the response model of missingness for the MCS income example. This involves eliciting normal prior distributions for Bayesian logistic regression models. We choose ELICITOR to help with this process, as it was created for exactly this type of elicitation and automatically generates appropriate code for WinBUGS.

7.2.1 Elicitation

Although it is clear from the literature that it is preferable to elicit the views of several experts, as this exercise is for illustrative purposes we interview a single expert. Our recruited expert has general knowledge about missing data in longitudinal studies and specific knowledge about missing MCS family income. We now describe the interview, which follows the recommendations in the ELICITOR user manual. Our expert is not shown any of our modelling results before the completion of the elicitation exercise.

Selection of variables to explain income missingness

To begin with, we ask our expert whether the five variables we have been using in our response model of missingness, *level*, *change*, *sc*, *eth* and *ctry*, are suitable for explaining missingness of income variables in sweep 2. He confirms that they are and does not suggest additional explanatory variables.

Eliciting optimum values and design points

Next we ask our expert which category or level of each variable he thinks would maximise the probability of an individual responding to the income question in sweep 2. This is called the *optimum value*, and his belief for each variable is shown in Table 7.6.

Table 7.6: MSC income example: explanatory variables for income missingness in sweep 2

name	description	design points	optimum value
level	level of hourly pay (sweep 1)	£4, £10, £25 and £50	£10
change	change in hourly pay (sweep 2-1)	-£5, £0 and £5	£0
sc	social class	NS-SEC 1, NS-SEC 2 and NS-SEC 4/5 ^b	NS-SEC 2
eth	ethnicity	white and non-white	white
ctry ^a	country	EWS and NI	EWS

^a EWS=England/Wales/Scotland; NI=Northern Ireland

^b NS-SEC 1=managerial and professional occupations; NS-SEC 2=intermediate occupations; NS-SEC 4=lower supervisory and technical occupations; NS-SEC 5=semi-routine and routine occupations

Then we ask which other values of the variables should be used for eliciting the response probability. These and the optimum value are known as the *design points* of the variable, and are also given in Table 7.6. Our expert's choices necessitate three changes to our response model of missingness.

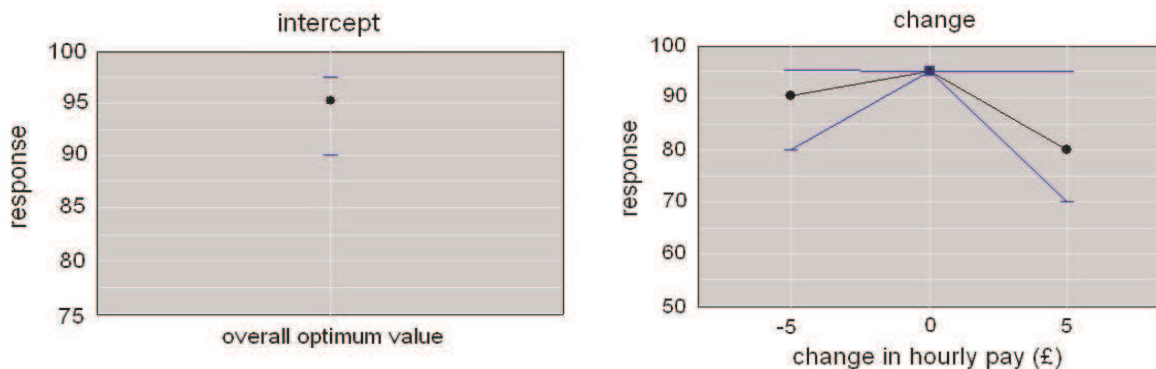
- 1 Our expert believes that the response probability would be the same for individuals with social class classification NS-SEC4 and NS-SEC5. So we combine categories 3 and 4, and use a three category rather than four category *sc*.
- 2 Our expert does not expect any difference in the probability of response between individuals residing in England, Scotland and Wales. Hence we combine these three countries to form a two category rather than four category *ctry*.
- 3 Our expert's opinion is that an individual is most likely to respond if their hourly pay is £10 (a little higher than the median sweep 1 hourly pay of £7) and their hourly pay rate does not change between sweeps. This means that the linear functional form which we have been using for variables *level* and *change* is no longer suitable. Following discussion with our expert, we decide to switch to a piecewise linear form.

Eliciting the response probability at the overall optimum value

The *overall optimum value* occurs when all the covariates are set to the optimum value of their design points. This equates to the equation intercept and is the highest probability of observing income. First we elicit the median value for the intercept by asking our expert how many individuals out of a sample of 100 he would expect to respond, if they all had optimum values for the five covariates. Our expert expects that 95% would respond. We plot this value using the ELICITOR software, and then ask our expert to confirm that he feels that the true value was equally likely to be above or below 95%, which he does.

We then consider the confidence that our expert has in his estimate, eliciting lower and upper quartiles by the bisection method. For the lower quartile, we ask our expert to assume that the true value is actually below his estimate, and then to choose a value such that the true value is equally likely to be higher or lower. We elicit the upper quartile similarly. We graph these values using ELICITOR (see left graph in Figure 7.1) and as a check, we ask our expert to confirm that he feels that the true value is equally likely to be inside or outside this interval.

Figure 7.1: MCS income example: screen shots from ELICITOR of the original elicitation



Eliciting the response probability at the other design points

The five explanatory variables are now considered in turn. Each variable is assumed independent, so covariances do not need to be assessed, and response probabilities for the design points are elicited assuming all the other variables are at optimum level. The probability for the optimum value is the same as the intercept graph, and the probabilities (median and 50% interval) for the remaining design points are elicited as before, again using the ELICITOR software. These elicited values are shown in the “original elicited %” columns in Table 7.7. The right graph in Figure 7.1 shows a screen shot from ELICITOR of the elicitation for variable *change*.

Table 7.7: MSC income example: elicited values of response percentage for sweep 2 income

explanatory variable	design point	original elicited %			adjusted elicited %		
		lq ^b	median	uq ^b	lq ^b	median	uq ^b
intercept		90	95	98			
hourly pay (sweep 1)	£4	80	85	90			
	£10	optimum value					
	£25	80	85	90			
	£50	65	75	85	75	80	
hourly pay change ^a	-£5	80	90	95			
	£0	optimum value					
	£5	70	80	95	80	90	
social class	NS-SEC 1	85	90	95			
	NS-SEC 2	optimum value					
	NS-SEC 4/5	85	90	95			
ethnicity	white	optimum value					
	non-white	70	80	90	80	85	
country	England/Wales/Scotland	optimum value					
	Northern Ireland	80	85	90	85	90	95

^a change in hourly pay from sweep 1 to sweep 2

^b lq=lower quartile; uq=upper quartile

Providing feedback on the elicitation

Providing feedback is an important part of the elicitation process, as it allows the expert to reconsider his assessments. ELICITOR enables feedback during the elicitation, including the display of alternative intervals. For example, if a 50% interval is elicited, then we can provide our expert with more information on the implications of his chosen values by asking ELICITOR to display a 90% interval. Any variable can be revisited at any stage in the elicitation. We use these features throughout the interview. Our expert feels that it would be useful to see the implied median for the *worst case*, i.e. when all the variables are set to their minimum design points. He expects a value of about 60%. This cannot be generated without running WinBUGS to forward sample from our expert’s prior, so we agree to provide this feedback after the meeting.

During discussion of the elicitation process, our expert comments that he had found questions about the *change* variable the most difficult part. This is the variable which allows informative missingness in our model, and about which there is greatest uncertainty on account of the missingness.

Converting the elicited values into WinBUGS code

ELICITOR automatically converts the probabilities of response at different design points into WinBUGS code with informative priors, and Equation 7.3 shows the resulting equations for this elicitation. Note that in this equation the parameters of the Normal distributions are written as the mean and variance, whereas WinBUGS uses the mean and precision.

$$\begin{aligned}
\text{logit}(p_i) &= \theta_0 + \text{Piecewise}(\text{level}_i) + \text{Piecewise}(\text{change}_i) \\
&\quad + (\theta_{\text{ctry}} \times \text{ctry}_{i1}) + (\theta_{\text{eth}} \times \text{eth}_{i1}) + \sum_{k=1}^3 (\theta_{\text{sc}[k]} \times \text{sc}_{[k]i1}) \\
\text{level}_i &= \text{hpay}_{i1} \\
\text{change}_i &= \text{hpay}_{i2} - \text{hpay}_{i1} \\
\text{Piecewise}(\text{level}_i) &= \begin{cases} \theta_{\text{level}[1]} \times (\text{level}_i - 10) : & \text{level}_i < 10 \\ \theta_{\text{level}[2]} \times (\text{level}_i - 10) : & 10 \leq \text{level}_i < 25 \\ \theta_{\text{level}[3]} \times (\text{level}_i - 25) + \theta_{\text{level}[2]} \times 15 : & \text{level}_i \geq 25 \end{cases} \\
\text{Piecewise}(\text{change}_i) &= \begin{cases} \delta_1 \times \text{change}_i : & \text{change}_i < 0 \\ \delta_2 \times \text{change}_i : & \text{change}_i \geq 0 \end{cases} \tag{7.3} \\
\theta_0 &\sim N(3.0, 1.64) \\
\theta_{\text{level}[1]} &\sim N(0.21, 0.01); \theta_{\text{level}[2]} \sim N(-0.085, 0.002); \theta_{\text{level}[3]} \sim N(-0.026, 0.001) \\
\delta_1 &\sim N(0.15, 0.06); \delta_2 \sim N(-0.32, 0.10) \\
\theta_{\text{ctry}} &\sim N(-1.3, 0.36) \\
\theta_{\text{eth}} &\sim N(-1.6, 1.04) \\
\theta_{\text{sc}[1]} &\sim N(-0.8, 0.83); \theta_{\text{sc}[2]} = 0; \theta_{\text{sc}[3]} \sim N(-0.8, 0.83)
\end{aligned}$$

where the second index on the variables indicates sweep, and $\text{sc}_{[k]}$ is a binary indicator for *sc* category k (1 if sc_i is category k , 0 otherwise).

The mean of the prior for θ_0 , is the logit of the elicited median for the intercept (as a proportion), i.e.

$$\text{prior mean of } \theta_0 = \text{logit}(0.95) = \log\left(\frac{0.95}{1 - 0.95}\right) = 2.9.$$

The small discrepancies between the results of this and subsequent calculations and the values shown in Equation 7.3, are due to rounding errors and the inexact entering of the elicited values into ELICITOR through its “click and point” mechanism. ELICITOR makes use of an approximation from Walpole and Myers (1993), p.211, for calculating the quantile, $q(f)$, of a $\text{Normal}(\mu, \sigma^2)$ distribution, as given

by

$$q_{\mu,\sigma}(f) \approx \mu + \sigma \left(4.91 \{ f^{0.14} - (1-f)^{0.14} \} \right). \quad (7.4)$$

Using Equation 7.4, the standard deviations can be approximated as

$$\sigma \approx \frac{q(f) - \mu}{4.91 \{ f^{0.14} - (1-f)^{0.14} \}}. \quad (7.5)$$

Hence to calculate an approximate standard deviation, we need the elicited probability from one quantile in addition to the median. As two additional quantiles are available, ELICITOR uses the average of the approximated standard deviation from both quantiles. So the variance, σ^2 , of the prior for θ_0 , is estimated as follows:

$$\begin{aligned} \sigma_{lq} &= \frac{\text{logit}(0.90) - \text{logit}(0.95)}{4.91 \{ 0.25^{0.14} - (1 - 0.25)^{0.14} \}} \\ \sigma_{uq} &= \frac{\text{logit}(0.98) - \text{logit}(0.95)}{4.91 \{ 0.75^{0.14} - (1 - 0.75)^{0.14} \}} \\ \sigma^2 &= \frac{1}{2} (\sigma_{lq}^2 + \sigma_{uq}^2) = 1.61 \end{aligned}$$

where σ_{lq} and σ_{uq} are the approximations based on the lower and upper quantiles respectively.

For the parameters associated with the binary and categorical variables, the prior mean is calculated by subtracting the logit of the elicited median for the intercept from the logit of the elicited mean of its associated explanatory variable. For example,

$$\text{prior mean of } \theta_{eth} = \text{logit}(0.80) - \text{logit}(0.95) = -1.6.$$

The remaining prior variances are calculated in a similar way to the prior variance for θ_0 , for example the prior variance of θ_{eth} is

$$\frac{1}{2} \left(\left(\frac{\text{logit}(0.70) - \text{logit}(0.80)}{4.91 \{ 0.25^{0.14} - (1 - 0.25)^{0.14} \}} \right)^2 + \left(\frac{\text{logit}(0.90) - \text{logit}(0.80)}{4.91 \{ 0.75^{0.14} - (1 - 0.75)^{0.14} \}} \right)^2 \right) = 1.05.$$

The only difference in converting the elicitation for the continuous variables, is that the distance between the design points must be taken into account in calculating the prior mean to place it on a unit scale. As an example, we consider the calculation of the prior means for the two parameters in the piecewise linear equation for *change*,

$$\begin{aligned} \text{prior mean of } \delta_1 &= \frac{\text{logit}(0.90) - \text{logit}(0.95)}{-5 - 0} = 0.15 \\ \text{prior mean of } \delta_2 &= \frac{\text{logit}(0.80) - \text{logit}(0.95)}{5 - 0} = -0.31. \end{aligned}$$

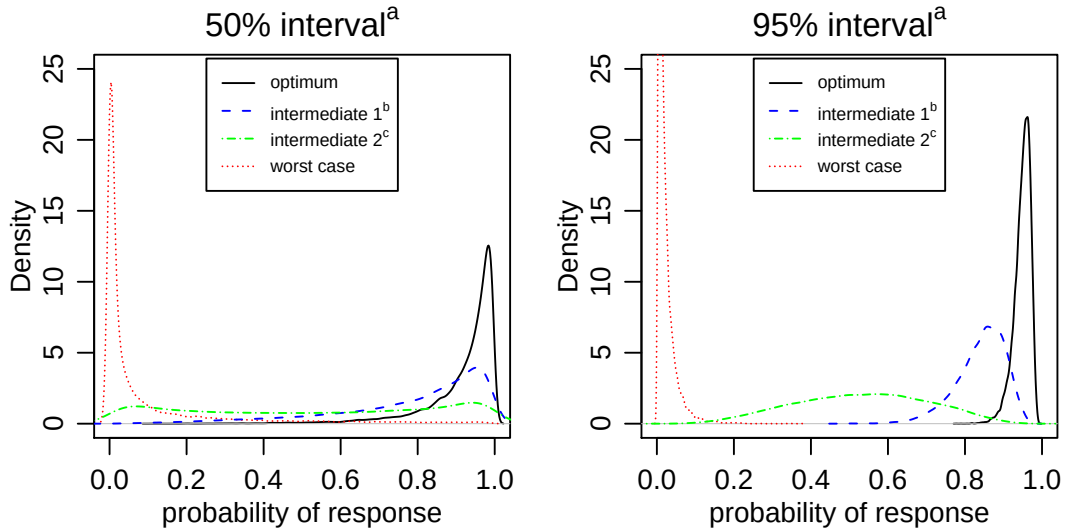
The prior variance calculations are the same as for the categorical variables.

The WinBUGS code generated by ELICITOR is run for a number of design points, including those elicited during the interview and the minimum as requested by our expert. Based on a sample of

20,000 iterations, the median for the probability of response in the worst case turns out to be 1%, very different from the 60% expected by our expert. The 95% interval, based on the 2.5 and 97.5 percentiles of the prior distribution is (0.00,0.75). The density of the probability of response in this worst case is shown as the dotted red line in the left plot of Figure 7.2. The densities for the optimum case (solid black line) and two intermediate cases are also shown. The median response probabilities for the design points tie up with the elicited values, confirming that the model is correctly implemented.

In order to see the effect of eliciting narrower intervals, we also forward sample from WinBUGS using code that is generated assuming the expert’s uncertainty was elicited via a 95% interval (i.e. the lower and upper quartiles become the 2.5 and 97.5 percentiles respectively). The impact can be seen in the right plot of Figure 7.2. Note that the medians are unchanged.

Figure 7.2: MCS income example: prior densities generated by forward sampling using WinBUGS, based on the original elicitation for selected design points



^a In the interview a 50% interval was elicited. To enable a comparison with narrower intervals, we also forward sample from WinBUGS using code that is generated assuming the expert’s uncertainty was elicited via a 95% interval (i.e. the lower and upper quartiles become the 2.5 and 97.5 percentiles respectively).

^b Intermediate design point 1: *level* = £4, all other variables at optimum

^c Intermediate design point 2: *level* = £4, *change* = £5, all other variables at optimum

Revisiting the elicited values

In the light of this, our expert adjusts some of his medians and intervals as shown under the “adjusted elicited %” heading in Table 7.7. We generate new WinBUGS code using ELICITOR and rerun our model. The worst case response probability median is still only 9%. This worst case is very extreme, as it assumes that the individual is paid at an hourly rate of £50 in sweep 1, has an increase in pay of £5 an hour between sweeps, is in social class 1, is non-white and lives in Northern Ireland. To put this into context, out of 505 individuals, only one individual is paid £50 or more in sweep 1, and only one has an hourly rate between £25 and £50. However, it does also reveal a difficulty with this approach. The rate of response rapidly decreases as probabilities are multiplied and having good intuition about

probabilities that are combined is difficult.

Part of the problem is that the probabilities are not really independent as assumed in the elicitation. Hindsight suggests that it would have been better to focus on eliciting information on fewer variables and allow for correlation between these variables. This would require modifying ELICITOR or developing an alternative method.

7.2.2 Use in the response model of missingness

To explore the effect of including this expert knowledge, we run two more joint models based on JMB. Changes are only made to the response model of missingness. In the first model, JMC, we change the functional form of covariates *level* and *change* to piecewise linear and aggregate *sc* and *ctry* into three and two categories respectively (*sc* now has categories NS-SEC 1, NS-SEC 2 and NS-SEC 4/5, and *ctry* is split England/Wales/Scotland (EWS) and Northern Ireland). However, we continue to use uninformative priors on the response model of missingness parameters. Our second new joint model, JMD, is the same as JMC but has informative priors on the response model of missingness parameters, generated from the adjusted elicitation. Both these models include the BCS70 educational data (the findings discussed below are replicated if the BCS70 data is excluded), and their convergence is satisfactory.

Table 7.8 shows the parameter estimates for these new models, including comparisons of JMC with JMB and JMD with JMC. For JMC, which has the more complex form of the response model of missingness but not the informative priors, we find that there is insufficient information in the data for $\theta_{level[3]}$ to be estimated with any accuracy (recall there are only two individuals with sweep1 *hpay* above £25). In JMD, which has the informative priors, this is resolved.

Impact of switching to a piecewise linear functional form

There are some differences in the model of interest parameter estimates for JMC compared to JMB, in particular for β_{sing} , whose posterior mean falls 30%. Also, there are some small changes to the covariate model of missingness parameter estimates, in particular for those associated with *sing*, which have a small impact on the covariate imputations with fewer individuals gaining a partner between sweeps (see Table 7.9). It is difficult to compare the response model of missingness parameters directly because of the change in functional form. However, *level*, *change* and *eth* remain important predictors of missingness.

To see which of the functional form changes are impacting the estimate of β_{sing} , we run three variants of JMC in which the only changes from JMB are 1) the functional form of *level*, 2) the functional form of *change* and 3) the aggregation of *sc* and *ctry*. These show that it is the switch from linear to

Table 7.8: MCS income example: comparison of parameter estimates using a response model of missingness with linear functional form (JMB) and piecewise linear functional form with (JMD) and without informative priors (JMC)

	JMB		JMC		% diff ^a	JMD		% diff ^b
σ	0.36	(0.33,0.40)	0.45	(0.39,0.50)	22.4	0.39	(0.34,0.45)	-11.9
ς	0.23	(0.19,0.26)	0.13	(0.05,0.20)	-41.1	0.20	(0.12,0.25)	47.4
γ_1	1.96	(1.81,2.12)	1.98	(1.83,2.13)	0.7	2.00	(1.85,2.15)	0.9
γ_2	1.92	(1.78,2.06)	1.95	(1.81,2.08)	1.4	1.97	(1.83,2.11)	0.9
γ_3	1.90	(1.72,2.07)	1.92	(1.74,2.09)	1.1	1.95	(1.77,2.13)	1.7
γ_4	1.95	(1.73,2.17)	1.99	(1.76,2.21)	1.8	2.02	(1.80,2.24)	1.7
γ_5	1.89	(1.72,2.05)	1.90	(1.75,2.06)	0.9	1.92	(1.77,2.09)	1.1
γ_6	1.92	(1.75,2.11)	1.92	(1.75,2.09)	0.0	1.94	(1.76,2.11)	0.7
γ_7	1.83	(1.67,2.00)	1.88	(1.71,2.04)	2.6	1.88	(1.72,2.04)	0.1
γ_8	1.88	(1.66,2.10)	1.90	(1.69,2.12)	1.3	1.91	(1.69,2.13)	0.3
γ_9	1.93	(1.77,2.09)	1.95	(1.79,2.10)	0.9	1.97	(1.81,2.12)	1.1
β_{age}	0.11	(0.08,0.14)	0.12	(0.09,0.15)	5.8	0.12	(0.09,0.15)	0.5
$\beta_{edu[NVQ2\&3]}$	0.16	(0.08,0.24)	0.17	(0.09,0.25)	6.2	0.17	(0.09,0.25)	-2.7
$\beta_{edu[NVQ4\&5]}$	0.31	(0.22,0.41)	0.35	(0.25,0.45)	11.2	0.32	(0.22,0.42)	-7.5
β_{eth}	-0.06	(-0.18,0.07)	-0.06	(-0.18,0.06)	10.2	-0.07	(-0.19,0.06)	7.9
β_{reg}	-0.13	(-0.25,0.00)	-0.14	(-0.26,-0.03)	14.4	-0.14	(-0.27,-0.03)	-0.1
β_{sing}	-0.16	(-0.25,-0.07)	-0.11	(-0.20,-0.01)	-30.3	-0.07	(-0.15,0.02)	-34.6
θ_0	-1.39	(-3.06,0.32)	2.56	(1.52,3.73)	-284.5	1.70	(1.03,2.41)	-33.7
$\theta_{level[1]}$	0.32	(0.11,0.55)	0.25	(0.07,0.46)	-20.7	0.14	(0.03,0.26)	-43.1
$\theta_{level[2]}$			0.68	(0.35,1.05)		-0.02	(-0.09,0.06)	-102.4
$\theta_{level[3]}$			-8067	(-22450,-315)		-0.01	(-0.05,0.02)	-100.0
$\delta_{change[1]}$	0.39	(0.15,0.65)	0.65	(0.39,0.94)	69.2	0.19	(0.05,0.35)	-70.3
$\delta_{change[2]}$			-0.21	(-0.35,-0.06)		-0.11	(-0.25,0.10)	-45.9
$\theta_{ctry[Wales]}$	-0.14	(-0.77,0.49)						
$\theta_{ctry[Scotland]}$	0.25	(-0.36,0.86)						
$\theta_{ctry[NI]}$	0.21	(-0.42,0.83)	0.31	(-0.29,0.93)	47.4	0.05	(-0.47,0.58)	-82.8
θ_{eth}	-1.16	(-1.85,-0.50)	-1.12	(-1.75,-0.50)	-3.2	-0.96	(-1.45,-0.48)	-13.9
$\theta_{sc[NS-SEC1]}$			0.08	(-0.68,0.83)		0.27	(-0.31,0.88)	245.2
$\theta_{sc[NS-SEC2]}$	-0.09	(-0.80,0.65)						
$\theta_{sc[NS-SEC4/4\&5]}^c$	0.01	(-0.98,1.03)	0.06	(-0.46,0.60)	790.5	-0.18	(-0.64,0.27)	-383.3
$\theta_{sc[NS-SEC5]}$	0.02	(-0.73,0.78)						
$edu.\eta_0$	-0.54	(-0.87,-0.18)	-0.52	(-0.84,-0.18)	-3.5	-0.49	(-0.81,-0.15)	-5.2
$edu.\eta_{edu[NVQ2\&3]}$	0.88	(0.78,0.97)	0.88	(0.78,0.97)	-0.2	0.87	(0.78,0.97)	-0.2
$edu.\eta_{edu[NVQ4\&5]}$	6.65	(2.84,9.44)	4.17	(2.01,7.93)	-37.3	4.97	(2.22,7.51)	19.3
$edu.\eta_{sgap}$	0.08	(-0.02,0.17)	0.07	(-0.02,0.16)	-5.7	0.06	(-0.03,0.15)	-10.3
$sing.\eta_0$	-0.77	(-1.65,0.12)	-0.95	(-1.75,-0.13)	24.0	-0.99	(-1.82,-0.15)	4.2
$sing.\eta_{sgap}$	0.14	(-0.13,0.39)	0.18	(-0.06,0.42)	33.6	0.19	(-0.06,0.44)	5.2
$age.\eta_0$	0.35	(0.34,0.36)	0.35	(0.34,0.36)	0.1	0.35	(0.34,0.36)	0.0
$age.\eta_{age}$	1.00	(0.99,1.00)	1.00	(0.99,1.00)	0.0	1.00	(0.99,1.00)	0.0

Table shows the posterior mean, with the 95% interval in brackets.

^a % difference in parameter estimates from JMB to JMC

^b % difference in parameter estimates from JMC to JMD

^c first definition applies to model JMB (4 category *sc*); the second definition applies to models JMC and JMD (3 category *sc*).

Table 7.9: MCS income example: comparison of covariate imputations using a response model of missingness with linear functional form (JMB) and piecewise linear functional form with (JMD) and without informative priors (JMC)

	JMB		JMC		% diff ^c	JMD		% diff ^d
edu_{11}^a	35.2	(30,39)	34.5	(29,38)	-2.1	34.0	(28,38)	-1.2
edu_{12}^a	3.8	(0,9)	4.5	(1,10)	19.2	5.0	(1,11)	9.3
edu_{13}^a	0.0	(0,0)	0.0	(0,0)		0.0	(0,0)	
edu_{22}^a	97.1	(92,101)	96.3	(91,100)	-0.8	95.8	(90,100)	-0.5
edu_{23}^a	3.9	(0,9)	4.7	(1,10)	20.5	5.2	(1,11)	10.5
$sing_1^b$	98.5	(80,116)	107.1	(91,122)	8.7	109.5	(95,124)	2.2
$sing_2^b$	67.5	(50,86)	58.9	(44,75)	-12.7	56.5	(42,71)	-4.2

Table shows posterior means, and 95% intervals in brackets

^a The first index indicates the value of sweep 1 *edu*, and the second index the value of sweep 2 *edu*.

^b The index indicates the value of sweep 2 *sing*.

^c % difference in parameter estimates from JMB to JMC

^d % difference in parameter estimates from JMC to JMD

piecewise linear functional form for *change* which is making the difference.

Impact of using informative priors on the response model of missingness parameters

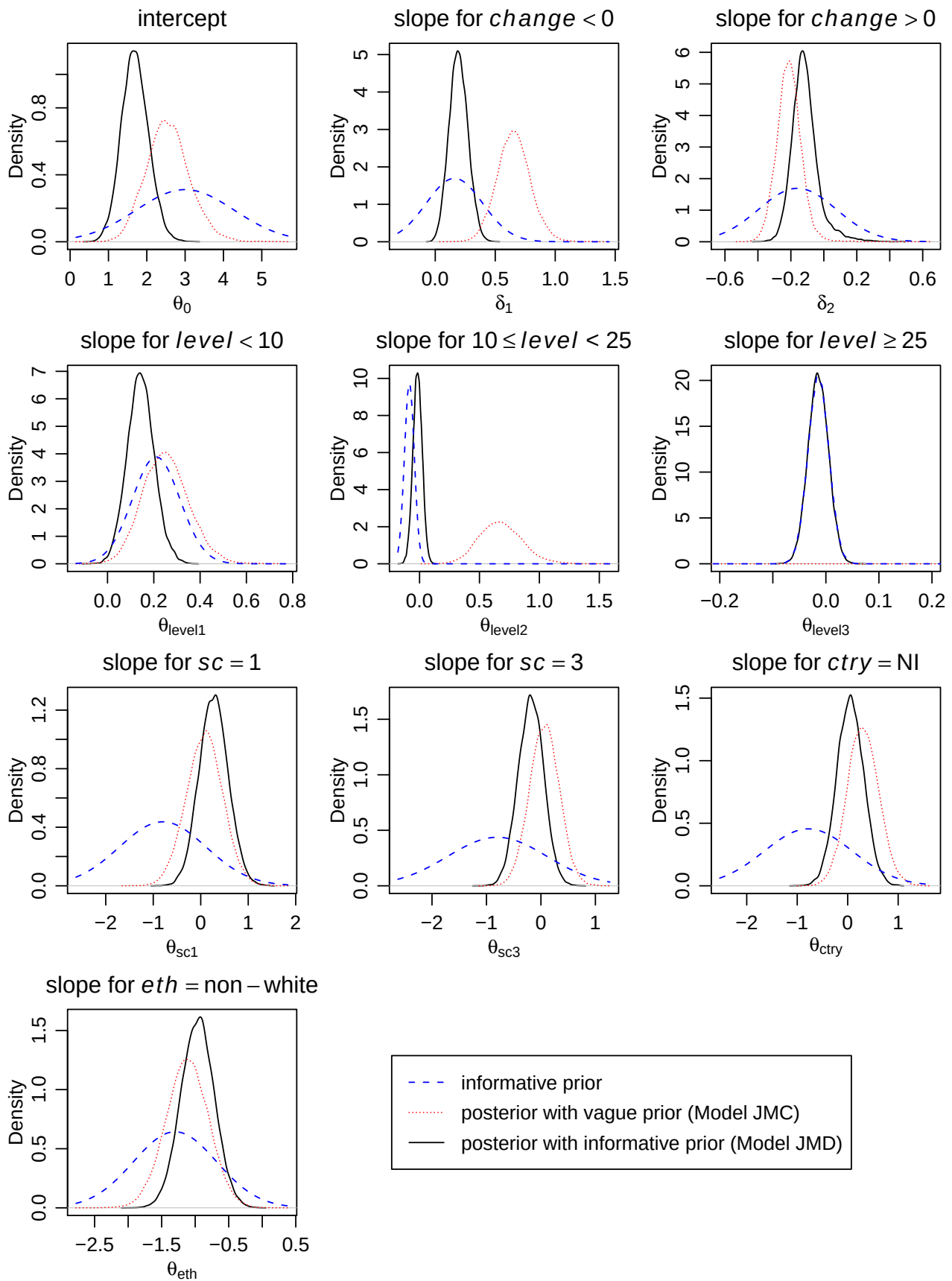
Comparing JMD to JMC, the most notable difference in the model of interest parameter estimates is a further fall in the posterior mean of β_{sing} . We find small further changes to the covariate model of missingness parameter estimates and little change in the covariate imputations. We will comment on these differences more fully in discussing sensitivity analysis in Chapter 8, and for now restrict ourselves to discussing the impact of the informative priors on the response model of missingness parameters, which is substantial.

The effects of the informative priors are easier to see from Figure 7.3, which shows for each parameter the densities of the posterior for JMC (red dotted line), the posterior for JMD (black solid line) and the informative priors from the adjusted elicitation used in JMD (dashed blue line). The posterior for JMC is a proxy for the likelihood since vague priors are assumed for all parameters. Note that for $\theta_{level[3]}$ the JMC posterior is very flat, and displays as a horizontal line at zero.

There are substantial differences between some of the priors and likelihoods. For θ_{eth} they are in approximate agreement. For θ_{ctry} the modes of the prior and likelihood have different signs. For $\theta_{sc[NVQ1]}$ and $\theta_{sc[NVQ4\&5]}$ the likelihood does not support our expert's view that social class NS-SEC 1, 4 and 5 individuals are less likely to respond than those categorised as NS-SEC 2. With the optimum values for the covariates at different values for the JMD posterior distribution compared to its prior distribution, we also find a sizeable difference for the constant, θ_0 .

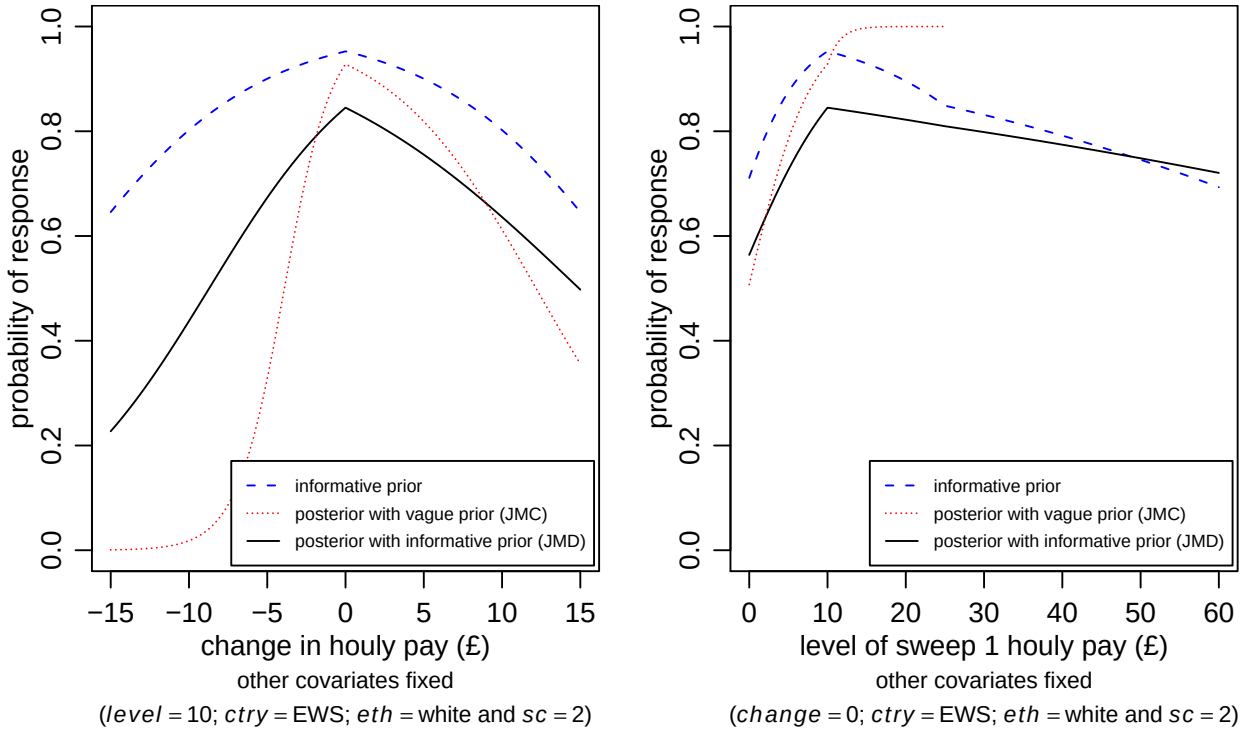
Of particular interest are the parameters for *change*, whose modes for the priors and the likelihoods have the same signs. For δ_1 the magnitude of the likelihood is four times that of the prior, but the

Figure 7.3: MCS income example: prior and posterior distributions for parameters in the response model of missingness



posterior is much closer to the prior. An alternative view of the combined effects of these priors and likelihoods is shown in the left graph in Figure 7.4. This plots the probability of response against change in hourly pay for the informative prior and posteriors for JMC and JMD, assuming that all the other covariates are fixed at their optimum values (i.e. $level=10$, $ctry=EWS$, $eth=white$ and $sc=NS-SEC\ 2$). We find that the shape of the JMD posterior is similar to its informative prior, with individuals whose hourly pay does not change between sweeps most likely to respond. Of course, the shape of the JMC posterior is partly predetermined because the change-point in the piecewise linear function is fixed at our expert's suggestion. However, we know from our work with synthetic data, Section 4.5, that fitting a response model of missingness which estimates the change point in a piecewise linear regression is very difficult. We do not attempt to do this, as we consider that it is only sensible to use the piecewise linear form if we can fix the change-point based on an expert's advice. Without an expert, we would stick to the simpler linear form.

Figure 7.4: MCS income example: prior and posterior probability of response as the change in pay varies and as the level of pay in sweep 1 varies, with the other parameters fixed



The JMC posterior line is not plotted for $level$ beyond £25 because of the problems with the calculation of $\theta_{level[3]}$.

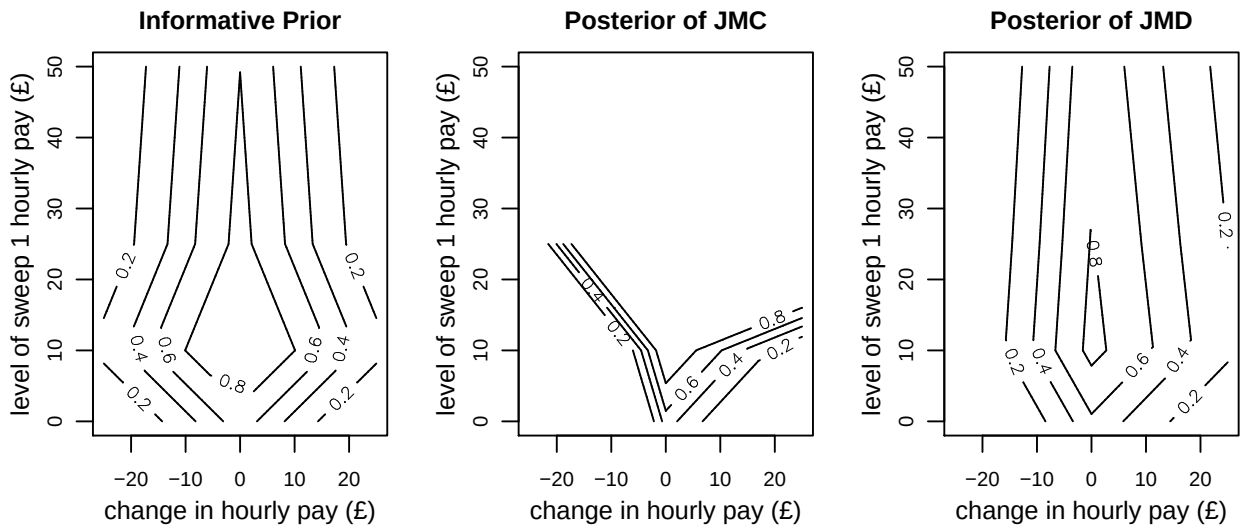
The right graph in Figure 7.4 shows the probability of response plotted against the level of hourly pay in sweep 1. Again the JMD posterior and informative prior have a similar shape. The JMC posterior line is not plotted beyond £25 because of the problems with the calculation of $\theta_{level[3]}$. From Figure 7.3 we see that virtually all the information for the JMD posterior of $\theta_{level[3]}$ comes from its prior, with the two lines almost coincident. Figure 7.3 also shows conflict in the likelihood and prior for $\theta_{level[2]}$.

From Figure 7.3 we see that the expert priors are quite uncertain, and for the parameters associated

with *ctry*, *eth* and *sc* dominated by the data. However the priors are much more prominent in the estimation of the *level* parameters, where less information can be drawn from the data, and the *change* parameters whose estimation is otherwise dependent on model assumptions. We also notice that for some parameters (notably $\theta_{level[1]}$, but also δ_2 , $\theta_{sc[NVQ1]}$ and θ_{eth}) the mode of the JMD posterior is not between the modes of its likelihood and prior. This shows that information is being drawn from different parts of the model in their calculation. In particular, we suspect that the estimation of the parameters *change* and *level* are entwined and that ideally we should have considered an interaction between them.

Figure 7.5 displays contour lines showing how the probability of response changes for the informative prior and posteriors for JMC and JMD, as the level of sweep 1 hourly pay and the change in hourly pay between sweeps vary, with all the other covariates assumed fixed at their optimum values. We note that the probability of response decreases more slowly for positive than negative change in pay for the JMC posterior and to a lesser extent for the JMD posterior, while the prior is symmetric.

Figure 7.5: MCS income example: probability of response for varying *change* and *level*



All other regressors are fixed s.t. *eth*=white; *sc*=NS-SEC2 and *ctry*=EWS.

The probabilities for JMC are not plotted for *level* beyond £25 because of the problems with the calculation of $\theta_{level[3]}$.

In view of our findings, we propose that our elicitation strategy should be revisited as part of future research in this area. More emphasis needs to be placed on identifying the variables which are not well identified by the data, in this case *level* and *change*, and the elicitation effort should concentrate on these key variables. Although allowing for correlation between variables will complicate the elicitation process substantially, it would be beneficial for these variables. It is also clear that the functional forms of these parameters are very important, and this should be reflected in the elicitation process. Plots similar to those in Figure 7.5 may be helpful in this revised strategy. A way of eliciting uncertainty regarding these functional forms should be investigated, and in the case of a piecewise linear functional form, this includes uncertainty about the position of the knots. It would also be better to elicit views

from more than one expert.

In the next chapter, as part of our sensitivity analysis, we investigate the effect of different priors on the response model of missingness parameters on the results given by JMD.

Summary: incorporating information from additional sources

The estimation of some parameters in the models of missingness which form part of a joint model can be difficult, particularly where information is limited. Such problems can be mitigated by incorporating additional information into the model. Sources of additional information include data from other studies and expert knowledge, and each of these has been investigated using our MCS income example.

Additional data from BCS70 is included by expanding the covariate model of missingness to simultaneously fit two sets of equations with common parameters, where one set comprises the original equations for imputing the missing MCS covariates and the other set consists of new equations for modelling the BCS70 data. The posterior means of the covariate imputations do not change substantially, providing extra confidence in the results. The extra BCS70 data allows the parameters in the covariate model of missingness to be estimated with greater accuracy, but does not give increased precision in the covariate imputations. It has little impact on the model of interest parameters and so does not affect any conclusions relating to our original substantive questions.

Using the ELICITOR software, expert priors for all the parameters of the response model of missingness were elicited assuming that each variable is independent. As a result of the elicitation, we also switched from a linear to piecewise linear functional form for the *level* and *change* parameters in the response model of missingness. This change and the use of the informative priors had a substantial impact on the estimate of β_{sing} , and so affects our conclusions regarding the association of partnership status with income. Unsurprisingly, there is also a large impact on the response model of missingness parameters. Eliciting independent priors resulted in some implausibly extreme probabilities of missingness for certain variable combinations, suggesting that the priors should be jointly elicited. Based on this elicitation exercise, a good elicitation strategy in a missing data context would identify and concentrate on weakly identified variables, focussing on the choice of functional form and allowing for correlation between these variables.

Chapter 8

Designing sensitivity analyses

In this chapter we investigate designing sensitivity analyses for data with missing values, using our MCS income example. Molenberghs and Kenward (2007) suggest that one approach is to fit a number of plausible MNAR models, or carry out a primary analysis supplemented by a number of modifications. We follow this procedure, selecting a *base model* and fitting alternatives. Other approaches to sensitivity analysis are discussed in Section 2.4.6.

We take JMD, which was introduced in the previous chapter, as our base model and compare it to a number of alternative models in which some of the main assumptions are varied. To recap, the model of interest of JMD uses the set of covariates $\{age, edu, eth, reg, sing\}$, a log transform of the response and a t_4 error distribution, its covariate model of missingness includes BCS70 data and the key features of its response model of missingness are a piecewise linear functional form for *level* and *change* and informative priors based on the final elicitation. The WinBUGS code for JMD is given in Appendix E.1. The sensitivities discussed in this chapter are by no means exhaustive, but are chosen to demonstrate the types of analyses that can be performed. Alternative assumptions are considered for each part of the joint model in turn, focussing on how they interact with the assumption that the response is MNAR. We also explore the robustness of conclusions regarding our original motivating questions to choice of model.

Our assessment of each model includes an examination of the fit of the seven re-issued individuals discussed in Section 3.1.2, whose sweep 2 hourly pay is modelled as missing, although it is actually known. We also compare the DIC_W for the response model of missingness and the associated scaled p_D . We do not attempt to calculate any DIC based on the observed data likelihood, as the models we are using are too complex for this to be done easily. Instead we use the re-issues to examine the fit of the model of interest part of the joint model.

8.1 Choice of sensitivity analyses for the MCS income example

We start by considering possible sensitivities for our MCS example, and select a range for implementation and analysis. Each of our chosen sensitivities varies from our base model, JMD, in a single aspect so their individual effects can be assessed. A second stage of sensitivity analysis could combine several sensitivities which are shown to have a sizeable impact on results.

8.1.1 Sensitivity to model of interest assumptions

We revisit the assumptions made in setting up the original model of interest, and decide that the areas of chief concern, given our analysis of the complete cases (Section 5.3.1) and the insights gained from the simulations undertaken in Chapter 4, are as follows:

- choice of error distribution;
- choice of explanatory variables;
- choice of transform for the response.

For each of these, there are a number of alternatives to the choice incorporated in JMD, but we limit our sensitivities to the alternative we consider most plausible.

Varying the error distribution

In selecting our initial model of interest we considered using Normal or t_4 errors, and chose t_4 errors for robustness to outliers. As a sensitivity we run a model, JMG, which is the same as JMD but uses a Normal error distribution instead of a t_4 error distribution. (JME and JMF will be introduced shortly.)

Kenward (1998) provides an example where changing from a Normal to a heavy-tailed t distribution removed the evidence for MNAR, because the outliers were better accommodated. We started with a t_4 distribution, but found some evidence of MNAR, and are interested in whether the amount of evidence changes if a Normal distribution is used.

Incorporating additional explanatory variables

In Section 5.3.1 we found evidence for the inclusion of an age^2 term and $age \times edu$ interaction terms, but excluded them to avoid adding complexity at an early stage of model formation. We form JMH by adding both these terms to the model of interest in JMD. Their inclusion is complicated by the need to ensure consistency in the imputation of the missing values of age , age^2 and $age \times edu$. This is achieved by calculating the missing values of age^2 and $age \times edu$ using the imputations of age . As with age , age^2 and $age \times edu$ are standardised to improve convergence using fixed values of their means and standard deviations based on the observed data. In calculating age^2 , we square the original

(unstandardised) age and then standardise.

An alternative transform

From our exploration of selection models with simulated data in Chapter 4, we know that the choice of response transform is a key assumption, since distributional skewness can be confused with informative missingness. As an alternative to the log transformation, we use a cube root transformation in setting up model JMI, which is otherwise identical to JMD.

8.1.2 Sensitivity to response model of missingness assumptions

As with the model of interest, we could set up sensitivities in which the explanatory variables in the response model of missingness are varied. However, we restrict our attention to sensitivities involving the functional form and priors of these variables.

In building a joint model which incorporated expert knowledge in Chapter 7, we set up two models, JMB and JMC, which can be regarded as sensitivities to JMD for assessing the impact of incorporating expert knowledge. JMB excludes all aspects of the elicited information, while JMC uses the functional form of the response model of missingness parameters suggested by our expert but non-informative priors on these parameters. As further sensitivities to the choice of response model of missingness priors, we run two other models, JME and JMF. JME has response model of missingness priors generated from the initial elicitation, and in JMF all the response model of missingness parameters are fixed to their final elicitation median values and not estimated.

If the response missingness is MAR then $\delta_1 = \delta_2 = 0$. We know that these two parameters are difficult for the model to estimate, and that our expert found the part of elicitation involving their associated variable, change in income between sweeps, the most difficult. Verbeke *et al.* (2001) envisage a sensitivity analysis in which the changes in the parameters or functions of interest are studied for different values of δ . Therefore, we also carry out a sensitivity analysis in which a series of models are run with these two parameters fixed. We refer to this group of models as JMJ, and it contains forty-nine variants which are formed by combining seven values of δ_1 with each of seven values of δ_2 . We use the same set of values, namely $\{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$, for both δ_1 and δ_2 . This range encompasses the values elicited from our expert and those estimated by the models we have fitted to date. The design includes seven variants in which the functional form of *change* is linear, i.e. $\delta_1 = \delta_2$, with the $\delta_1 = \delta_2 = 0$ variant equivalent to assuming the response is MAR.

8.1.3 Sensitivity to covariate model of missingness assumptions

A range of sensitivities can also be constructed by varying the assumptions in the covariate model of missingness. Most obviously, we could include additional covariates in the equations for ν (Equation 6.7). One variant was explored in imputing the missing values of two correlated binary variables, Section 6.2.1, and found to have little impact on the model of interest parameter estimates. We might also consider expanding the covariate model of missingness to include imputing the missing values of *reg*. However, as only four observed individuals moved between London and the regions between sweeps, this is a very low priority, which is not pursued.

A further possibility is to allow the covariates to be MNAR, rather than MAR as assumed to date. This raises a number of questions, for example should we use separate missingness indicators for the covariates and the response or should we use an overall missingness indicator for attrition? If we use separate indicators, a new sub-model linked to the existing covariate model of missingness is required. In implementing this we would need a different indicator for each covariate pattern of missingness. Alternatively, if we use an overall missingness indicator for attrition, we then also require a method for dealing with any item missingness that occurs in the response or covariates. Although in theory a model allowing MNAR covariates could be designed, it may currently be computationally prohibitive in WinBUGS. Attempting to implement such a model is beyond the scope of this thesis.

8.1.4 Sensitivity to dataset

Verbeke *et al.* (2001) point out that one or a few influential subjects can lead to non-random dropout being found in an analysis, and Kenward (1998) provides an example where discarding two subjects with anomalous profiles removes the evidence for MNAR. In setting up our MCS income dataset, we excluded four individuals with suspicious looking extreme pay values (Section 3.1.2). A possible sensitivity analysis might investigate whether any further exclusions are justified and investigate the impact. However, as we have used t_4 errors for robustness to outliers, we consider this unlikely to have a large impact.

One of the criteria used in forming our dataset excluded individuals who stopped working in sweep 2. An additional sensitivity might fit a model to an expanded dataset that includes those who moved out of the workforce by sweep 2, treating their sweep 2 hourly pay as missing. The rationale for this is consistency, in that we do not know whether the individuals with missing sweep 2 pay were working or had moved out of the workforce. However, we do not pursue this idea.

8.2 Results of sensitivity analyses for the MCS income example

For convenience we discuss the findings of our selected sensitivity analyses in two parts, which we call *Assumption Sensitivity* (JMB-JMI) and *Parameter Sensitivity* (the forty-nine variants of JMJ). All the models are run with 2 chains for 100,000 iterations of which 50,000 are burn-in and a thinning rate of 5, and their convergence is satisfactory.

8.2.1 Assumption sensitivity

The differences between each of the models JMB-JMI and JMD are summarised in Table 8.1, and fuller details can be found in Table 5 in the model glossary. The first four (JMB, JMC, JME and JMF) allow us to explore sensitivity to the choice of functional form and priors in the response model of missingness, while the last three (JMG, JMH and JMI) are sensitivities to different model of interest assumptions.

Table 8.1: MSC income example: summary of differences of joint models for sensitivity analysis from base model (JMD)

model	difference from JMD ^a
JMB	linear functional form for <i>level</i> and <i>change</i> ; non-informative priors on RMoM parameters
JMC	non-informative priors on RMoM parameters
JME	informative priors on RMoM parameters based on original elicitation
JMF	RMoM parameters fixed to their final elicitation median values and not estimated
JMG	Normal error distribution for model of interest
JMH	additional covariates age^2 and $age \times edu(3)$ for model of interest
JMI	cube root transform of response for model of interest

All models run with 2 chains for 100,000 iterations of which 50,000 are burn-in and a thinning rate of 5.

^a The key features of JMD are: model of interest - covariates $\{age, edu, eth, reg, sing\}$, a log transform of the response and a t_4 error distribution; response model of missingness - piecewise linear functional form for *level* and *change* and informative priors based on the final elicitation.

A full set of parameter estimates is given for these seven models along with JMD in Table 8.2. Non-negligible differences in the parameter estimates of JMB-JMC and JME-JMH from JMD are highlighted in bold, where for this purpose a non-negligible difference is defined as an absolute difference greater than 0.02 and a percentage difference greater than 10%. We consider each part of the model in turn.

Table 8.2: MSC income example: parameter estimates for joint models JMB-JMI (with non-negligible differences^c from JMD highlighted in bold)

	JMD	JMB	JMC	JME	JMF	JMG	JMH	JMI ^a
σ	0.39 (0.34,0.45)	0.36 (0.33,0.40)	0.45 (0.39,0.50)	0.39 (0.34,0.44)	0.40 (0.36,0.45)	0.41 (0.36,0.45)	0.39 (0.34,0.45)	0.25 (0.22,0.28)
ς	0.20 (0.12,0.25)	0.23 (0.19,0.26)	0.13 (0.05,0.20)	0.20 (0.14,0.25)	0.19 (0.14,0.24)	0.11 (0.00,0.23)	0.19 (0.12,0.25)	0.13 (0.09,0.16)
γ_1	2.00 (1.85,2.15)	1.96 (1.81,2.12)	1.98 (1.83,2.13)	2.00 (1.84,2.15)	2.00 (1.85,2.16)	2.02 (1.85,2.16)	1.99 (1.83,2.14)	1.96 (1.86,2.06)
γ_2	1.97 (1.83,2.11)	1.92 (1.78,2.06)	1.95 (1.81,2.08)	1.97 (1.82,2.11)	1.97 (1.83,2.11)	1.96 (1.82,2.11)	1.96 (1.82,2.10)	1.94 (1.85,2.03)
γ_3	1.95 (1.77,2.13)	1.90 (1.72,2.07)	1.92 (1.74,2.09)	1.95 (1.77,2.13)	1.96 (1.78,2.14)	1.90 (1.72,2.10)	1.94 (1.76,2.11)	1.93 (1.81,2.04)
γ_4	2.02 (1.80,2.24)	1.95 (1.73,2.17)	1.99 (1.76,2.21)	2.02 (1.80,2.25)	2.03 (1.80,2.27)	2.07 (1.83,2.32)	2.01 (1.78,2.23)	1.97 (1.83,2.12)
γ_5	1.92 (1.77,2.09)	1.89 (1.72,2.05)	1.90 (1.75,2.06)	1.93 (1.76,2.09)	1.94 (1.77,2.10)	1.96 (1.79,2.13)	1.93 (1.77,2.10)	1.91 (1.81,2.02)
γ_6	1.94 (1.76,2.11)	1.92 (1.75,2.11)	1.92 (1.75,2.09)	1.94 (1.76,2.12)	1.94 (1.76,2.12)	1.96 (1.77,2.15)	1.95 (1.78,2.13)	1.92 (1.80,2.03)
γ_7	1.88 (1.72,2.04)	1.83 (1.67,2.00)	1.88 (1.71,2.04)	1.88 (1.71,2.04)	1.89 (1.72,2.06)	1.92 (1.74,2.10)	1.88 (1.72,2.05)	1.88 (1.77,1.99)
γ_8	1.91 (1.69,2.13)	1.88 (1.66,2.10)	1.90 (1.69,2.12)	1.91 (1.69,2.13)	1.92 (1.69,2.14)	1.97 (1.75,2.21)	1.90 (1.67,2.12)	1.90 (1.76,2.04)
γ_9	1.97 (1.81,2.12)	1.93 (1.77,2.09)	1.95 (1.79,2.10)	1.97 (1.81,2.13)	1.98 (1.81,2.13)	1.99 (1.82,2.16)	1.97 (1.81,2.13)	1.94 (1.84,2.04)
β_{age}	0.12 (0.09,0.15)	0.11 (0.08,0.14)	0.12 (0.09,0.15)	0.12 (0.08,0.15)	0.12 (0.09,0.15)	0.11 (0.07,0.14)	0.36 (0.09,0.63)	0.07 (0.05,0.10)
β_{age}^2							-0.27 (-0.53,0.00)	
$\beta_{edu}[\text{NVQ2\&3}]$	0.17 (0.09,0.25)	0.16 (0.08,0.24)	0.17 (0.09,0.25)	0.17 (0.08,0.25)	0.17 (0.09,0.25)	0.13 (0.05,0.22)	0.16 (0.08,0.25)	0.10 (0.05,0.15)
$\beta_{edu}[\text{NVQ4\&5}]$	0.32 (0.22,0.42)	0.31 (0.22,0.41)	0.35 (0.25,0.45)	0.32 (0.22,0.42)	0.32 (0.23,0.42)	0.32 (0.22,0.43)	0.27 (0.17,0.37)	0.21 (0.14,0.27)
$\beta_{age \times edu}[\text{NVQ2\&3}]$							0.00 (-0.08,0.08)	
$\beta_{age \times edu}[\text{NVQ4\&5}]$							0.11 (0.01,0.21)	
β_{eth}	-0.07 (-0.19,0.06)	-0.06 (-0.18,0.07)	-0.06 (-0.18,0.06)	-0.07 (-0.19,0.06)	-0.07 (-0.20,0.06)	0.01 (-0.13,0.14)	-0.05 (-0.18,0.08)	-0.04 (-0.12,0.04)
β_{reg}	-0.14 (-0.27,-0.03)	-0.13 (-0.25,0.00)	-0.14 (-0.26,-0.03)	-0.15 (-0.26,-0.02)	-0.15 (-0.28,-0.02)	-0.14 (-0.27,-0.01)	-0.15 (-0.27,-0.02)	-0.09 (-0.17,-0.02)
β_{sing}	-0.07 (-0.15,0.02)	-0.16 (-0.25,-0.07)	-0.11 (-0.20,-0.01)	-0.07 (-0.15,0.01)	-0.06 (-0.14,0.03)	-0.06 (-0.16,0.04)	-0.07 (-0.15,0.01)	-0.05 (-0.11,0.00)
θ_0	1.70 (1.03,2.41)	-1.39 (-3.06,0.32)	2.56 (1.52,3.73)	1.72 (1.05,2.44)	3 (1.41,2.88)	2.13 (1.41,2.88)	1.70 (1.04,2.39)	1.66 (0.99,2.38)
$\theta_{level[1]}$	0.14 (0.03,0.26)	0.32 (0.11,0.55)	0.25 (0.07,0.46)	0.14 (0.03,0.26)	0.21 (0.04,0.29)	0.16 (0.04,0.26)	0.14 (0.03,0.26)	0.16 (0.05,0.28)
$\theta_{level[2]}$	-0.02 (-0.09,0.06)		0.68 (0.35,1.05)	-0.02 (-0.09,0.05)	-0.085 (-0.07,0.07)	0.00 (-0.07,0.07)	-0.02 (-0.09,0.06)	-0.01 (-0.09,0.06)
$\theta_{level[3]}$	-0.01 (-0.05,0.02)		-8067 (-22450,-315)	-0.03 (-0.09,0.04)	-0.014 (-0.05,0.02)	-0.01 (-0.05,0.02)	-0.01 (-0.05,0.02)	-0.01 (-0.05,0.02)
$\delta_{change[1]}$	0.19 (0.05,0.35)	0.39 (0.15,0.65)	0.65 (0.39,0.94)	0.19 (0.05,0.34)	0.16 (0.14,0.44)	0.29 (0.14,0.44)	0.19 (0.05,0.34)	0.20 (0.05,0.36)
$\delta_{change[2]}$	-0.11 (-0.25,0.10)		-0.21 (-0.35,-0.06)	-0.12 (-0.25,0.09)	-0.16 (-0.34,-0.07)	-0.22 (-0.34,-0.07)	-0.12 (-0.25,0.07)	-0.06 (-0.22,0.18)
$\theta_{ctry[\text{Wales}]}$		-0.14 (-0.77,0.49)						
$\theta_{ctry[\text{Scotland}]}$		0.25 (-0.36,0.86)						
$\theta_{ctry[\text{NI}]}$	0.05 (-0.47,0.58)	0.21 (-0.42,0.83)	0.31 (-0.29,0.93)	-0.12 (-0.60,0.37)	-0.78 (-1.48,-0.46)	0.06 (-0.47,0.60)	0.05 (-0.46,0.58)	0.05 (-0.47,0.57)
θ_{eth}	-0.96 (-1.45,-0.48)	-1.16 (-1.85,-0.50)	-1.12 (-1.75,-0.50)	-0.97 (-1.48,-0.46)	-1.3 (-2.33,-0.87)	-0.96 (-1.47,-0.46)	-0.96 (-1.45,-0.47)	-0.99 (-1.48,-0.50)
$\theta_{sc}[\text{NS-SEC1}]$	0.27 (-0.31,0.88)		0.08 (-0.68,0.83)	0.27 (-0.33,0.87)	-0.8 (-1.82,-0.23)	0.24 (-0.38,0.87)	0.28 (-0.32,0.89)	0.23 (-0.36,0.82)
$\theta_{sc}[\text{NS-SEC2}]$		-0.09 (-0.80,0.65)						
$\theta_{sc}[\text{NS-SEC4/4\&5}]^b$	-0.18 (-0.64,0.27)	0.01 (-0.98,1.03)	0.06 (-0.46,0.60)	-0.18 (-0.64,0.27)	-0.8 (-1.82,-0.23)	-0.17 (-0.65,0.31)	-0.19 (-0.65,0.26)	-0.16 (-0.60,0.29)
$\theta_{sc}[\text{NS-SEC5}]$		0.02 (-0.73,0.78)						
$edu.\eta_0$	-0.49 (-0.81,-0.15)	-0.54 (-0.87,-0.18)	-0.52 (-0.84,-0.18)	-0.51 (-0.85,-0.15)	-0.51 (-0.83,-0.18)	-0.50 (-0.83,-0.18)	-0.51 (-0.87,-0.17)	-0.50 (-0.83,-0.14)
$edu.\eta_{edu}[\text{NVQ2\&3}]$	0.87 (0.78,0.97)	0.88 (0.78,0.97)	0.88 (0.78,0.97)	0.88 (0.78,0.98)	0.88 (0.78,0.97)	0.88 (0.78,0.97)	0.88 (0.78,0.98)	0.87 (0.78,0.97)
$edu.\eta_{edu}[\text{NVQ4\&5}]$	4.97 (2.22,7.51)	6.65 (2.84,9.44)	4.17 (2.01,7.93)	4.33 (2.04,9.27)	6.7 (2.57,9.52)	5.69 (2.11,9.38)	5.58 (2.27,8.86)	5.20 (2.03,9.39)
$edu.\eta_{sgap}$	0.06 (-0.03,0.15)	0.08 (-0.02,0.17)	0.07 (-0.02,0.16)	0.07 (-0.03,0.16)	0.07 (-0.02,0.16)	0.07 (-0.03,0.16)	0.07 (-0.03,0.17)	0.07 (-0.04,0.16)
$sing.\eta_0$	-0.99 (-1.82,-0.15)	-0.77 (-1.65,0.12)	-0.95 (-1.75,-0.13)	-1.06 (-1.87,-0.28)	-1.03 (-1.86,-0.24)	-1.03 (-1.82,-0.23)	-1.04 (-1.88,-0.21)	-1.01 (-1.88,-0.18)
$sing.\eta_{sgap}$	0.19 (-0.06,0.44)	0.14 (-0.13,0.39)	0.18 (-0.06,0.42)	0.21 (-0.02,0.45)	0.20 (-0.03,0.45)	0.20 (-0.04,0.44)	0.21 (-0.04,0.45)	0.20 (-0.05,0.46)
$age.\eta_0$	0.35 (0.34,0.36)	0.35 (0.34,0.36)	0.35 (0.34,0.36)	0.35 (0.34,0.36)	0.35 (0.34,0.36)	0.35 (0.34,0.36)	0.35 (0.34,0.36)	0.35 (0.34,0.36)
$age.\eta_{age}$	1.00 (0.99,1.00)	1.00 (0.99,1.00)	1.00 (0.99,1.00)	1.00 (0.99,1.00)	1.00 (0.99,1.00)	1.00 (0.99,1.00)	1.00 (0.99,1.00)	1.00 (0.99,1.00)

Table shows the posterior mean, with the 95% interval in brackets.

^a The parameters for JMI are not directly comparable with the other models, since JMI uses a $\sqrt{\cdot}$ rather than $\log$ transform of the response, so differences are not highlighted.

^b The first definition applies to model JMB (4 category sc); the second definition applies to models JMC to JMI (3 category sc).

^c Absolute difference > 0.02 and percentage difference $> 10\%$.

Robustness of the conclusions to substantive questions

Recall that we are particularly interested in the parameters of the model of interest relating to educational level, ethnicity and partnership status to answer our original motivating questions. From JMD, our base model, we conclude that having qualifications equivalent to NVQ levels 2 or 3 is clearly associated with a higher hourly rate of pay, and estimate the associated increase to be 18%. However, there is considerable uncertainty surrounding the magnitude of this increase, as the 95% interval from JMD is (9%,28%). Further, having degree level qualifications (NVQ levels 4 or 5) is associated with approximately double this increase. Regarding ethnicity, we find little evidence to suggest that being non-white is associated with lower levels of pay. There is weak evidence that gaining a partner between sweeps is associated with lower pay, though the magnitude of the resulting change is much smaller than that associated with educational qualifications.

Looking at JMB, which excludes all expert knowledge relating to the response model of missingness functional form and priors, we find that the estimates of the parameters associated with educational level and ethnicity are virtually unchanged. However, JMB provides much stronger evidence that gaining a partner is associated with lower pay (the magnitude of the posterior mean of β_{sing} more than doubles for JMB compared to JMD). As regards the model of interest parameter estimates of JMC (which excludes the expert priors but keeps the piecewise linear functional form of the missingness model), we find that the β_{sing} posterior mean is about halfway between those from JMD and JMB, the $\beta_{edu[NVQ4\&5]}$ is a little higher than for JMD, but the other parameters are very similar to JMD. We also find that our other two sensitivities to choice of prior (JME and JMF) give similar model of interest parameter estimates to JMD. In our example, our conclusions to two of the questions are fairly robust to our choice of priors and functional form for the response model of missingness, but there is some sensitivity to these choices for the question relating to partnership status.

Turning to our sensitivities regarding different aspects of the specification of the model of interest, we find that assuming Normal rather than t_4 errors makes little difference to the β_{sing} estimate, slightly lowers the estimate of the parameter associated with educational level NVQ2/3 but not the parameter associated with educational level NVQ4/5, and removes all evidence of hourly pay being associated with ethnicity. However, these differences do not change the conclusions based on JMD.

The additional covariates make comparisons for JMH harder. The parameter estimates and 95% credible intervals support the inclusion of the age^2 term and an interaction between age and having a degree. However, there is no evidence of an interaction between age and educational level NVQ2/3. More importantly, it provides no evidence for altering the conclusions regarding our three substantive questions. JMI is not strictly comparable with the other models, because of its change of response scale, but the message is unaltered.

Robustness of the response model of missingness parameter estimates

The estimates of the response model of missingness parameters for JMD were discussed in Section 7.2.2. To recap, these suggest that sweep 2 pay is more likely to be missing for individuals who are non-white, but do not provide much evidence that social class or country make a difference. Those with low or high levels of hourly pay in sweep 1 are more likely to be missing, as are those whose pay changes substantially between sweeps.

These findings were compared with JMB and JMC in Section 7.2.2 in the discussion of the impact of the expert information. The main points to emerge were that for the parameters associated with *sc*, *eth* and *ctry* the data dominates the prior, but the situation is more complex for the parameters associated with *level* and *change*. The informative priors are required to enable all the parameters in the more complex piecewise linear functional form to be estimated with any accuracy, but regardless of functional form and choice of priors, the *change* parameter estimates (δ) are statistically significant, providing evidence of informative missingness given the model assumptions.

Using the original elicitation (JME) or adding extra covariates to the model of interest (JMH) makes little difference to the response model of missingness parameters. However, assuming Normal rather than t_4 errors for the model of interest (JMG) increases the magnitude of the posterior means of the δ , but has little impact on the other response model of missingness parameters. The evidence of informative missingness remains when using a cube root transform for the response in the model of interest (JMI).

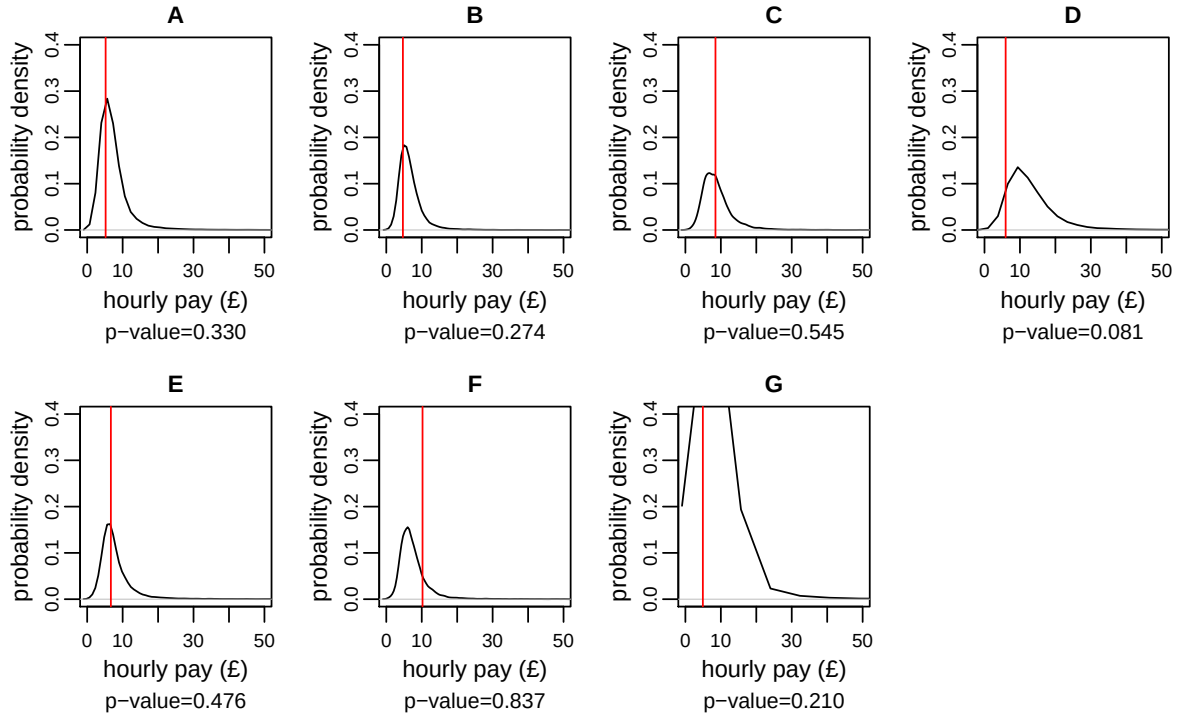
Robustness of the imputations of the missing covariates

The covariate model of missingness parameter estimates for all the main sensitivities, apart from JMB, are similar to those for JMD. The main differences between JMB and JMD are for *sing. η_0* and *sing. η_{sgap}* . This feeds into small differences in the summaries of the covariate imputations for JMB compared to JMD (Table 7.9). Any differences between the other models and JMD are slight.

Fit of re-issued individuals

In Section 3.1.2 we described how data that was collected from seven individuals, as a result of re-issues, was set to missing. We label these individuals A to G and use these data for checking the fit of our models. Figure 8.1 displays a density plot of the posterior predictive distribution of hourly pay for each individual imputed by JMD. The observed value of hourly pay is indicated by a vertical red line. A Bayesian p-value is displayed beneath each graph, calculated in WinBUGS as the proportion of iterations in which a higher value than the observed value was imputed (see Section 2.4.3 for further details on Bayesian p-values). Bayesian p-values are not changed by monotonic transformations, so the same values are obtained by working with the log of hourly pay.

Figure 8.1: MCS income example: posterior predictive distribution and observed value of hourly pay to show the fit of re-issued individuals using base model (JMD)



The observed value of hourly pay is indicated by the red line.

Hourly pay is estimated reasonably well for all seven individuals. It is clear from the plots that the predictions for all the individuals are subject to considerable uncertainty. Taken together, the plots for the seven individuals do not suggest any great conflict with the proposed model.

We also calculate the mean square error (MSE) of the fit of hourly pay for the seven individuals, and use this as a summary measure of the performance of our models in predicting their sweep 2 hourly pay. Table 8.3 shows the mean, median and 95% interval MSE of the fit of hourly pay for the seven individuals for our base model, JMD, and the seven models run as sensitivities. The posterior distribution of this MSE is somewhat skewed, so the median is a better measure than the mean, which is unstable due to outlying values of hourly pay for some individuals. The model excluding the expert knowledge (JMB) fits the seven re-issued individuals best, and there are slight improvements over JMD in this respect by adding the extra covariates or using the cube root transform.

Insight from the response model of missingness DIC_W and scaled p_D

Table 8.4 shows the DIC_W for the response model of missingness, calculated using *logitp* plug-ins, for models JMB to JMI. Only $\bar{D}$ is shown for JMC, because the two individuals with the highest level of pay in sweep 1 generate a very large *logitp*, preventing the calculation of $\hat{D}$. ($\hat{D}$ calculated excluding these two individuals is 479.6 for JMC, compared with 531.4 for JMB and 599.8 for JMD.) For JMF in which the response model of missingness parameters are fixed, the *logitp* for the observed responses are also fixed and consequently not monitored in WinBUGS. We calculate these *logitp* outside WinBUGS using the fixed values, in order to be able to calculate $\hat{D}$. So, for JMF, the only

Table 8.3: MCS income example: MSE of imputed hourly pay for seven re-issued individuals for joint models JMB-JMI

	MSE for hourly pay		
	mean	median	95% interval
JMB	9.7	8.8	(2.9,22.1)
JMC	94.0	18.4	(3.1,345.5)
JMD	201.6	19.1	(3.7,257.6)
JME	121.4	19.5	(3.8,273.2)
JMF ^a	866.4	26.4	(4.3,406.6)
JMG	32.5	23.0	(3.8,113.1)
JMH	158.6	15.0	(2.9,236.5)
JMI	28.4	14.7	(3.6,95.2)

^a Calculation of mean excludes 4 outlying values.

uncertainty contributing to the p_D comes from the missing responses.

Table 8.4: MCS income example: response model of missingness DIC_W calculated using *logitp* plug-ins for joint models JMB-JMI

	$\bar{D}$	<i>logitp</i> plug-ins			
		$\hat{D}$	p_D	DIC_W	scaled p_D
JMB	570.2	531.7	38.5	608.7	0.154
JMC ^a	543.9				
JMD	625.1	599.8	25.3	650.4	0.083
JME	626.6	601.2	25.4	652.0	0.083
JMF	657.0	633.3	23.7	680.7	0.074
JMG	592.1	557.1	35.0	627.1	0.135
JMH	624.8	599.3	25.6	650.4	0.084
JMI	629.4	611.2	18.2	647.7	0.045

^a $\hat{D}$ could not be calculated using *logitp* plug-ins.

Comparing JMB-JMF, the DIC_W suggests that the model excluding expert knowledge (JMB) explains the missingness best. However, the mean deviance suggests that the piecewise linear functional form (JMC) does better than the linear form (JMB), although this does not penalise for the additional complexity in JMC. Amongst the models with piecewise linear functional form and informative priors in their response model of missingness, our base model (JMD) does best. There is little to choose between JMD and JME, but the model with fixed priors (JMF) has a higher DIC_W , which we expect, as it does not use the data to learn anything about the missingness parameters. (We do not include JMG-JMI in our comparison as they have different models of interest, and our strategy is to use DIC_W to compare joint models with the same model of interest but different models of missingness.)

The scaled p_D provides evidence of informative missingness for all the models, but it is greater if

Normal errors are assumed for the model of interest (JMG) and less if a cube root transform is used (JMI). From our research with simulated data, these are precisely the type of changes we expect to make a difference.

8.2.2 Parameter sensitivity

We now examine the robustness of our conclusions regarding the substantive questions to changes in the δ parameters, by analysing the forty-nine variants of JMJ. The range of results, in terms of the proportional increase in hourly pay associated with educational level, ethnicity and partnership status are shown in Table 8.5. As with our assumption sensitivity analysis, the effect of gaining a partner between sweeps is most sensitive. If all the JMJ variants are plausible then we cannot even be sure about the direction of this effect, as the models suggest a range of conclusions from strong evidence of a positive effect to strong evidence of a negative effect. The results from the MAR analysis lie between the two extremes, close to the base model (JMD). We now look at ways of presenting the JMJ results in more detail.

Table 8.5: MCS income example: estimates of proportional change in pay associated with selected covariates for JMJ variants compared with base model (JMD)

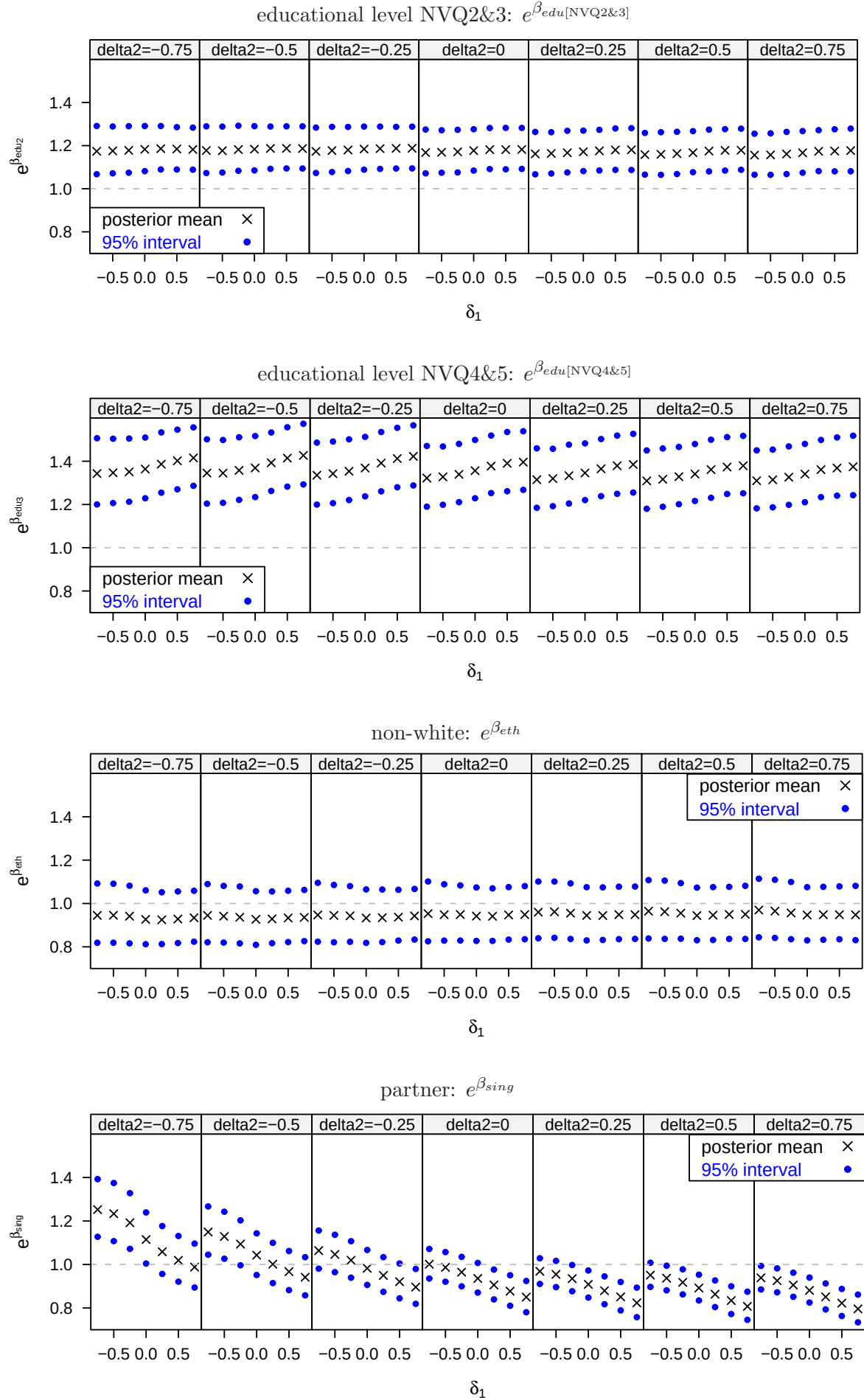
	JMJ			JMD
	minimum	maximum	MAR ^a	
edu[NVQ2&3]	1.16 (1.06,1.26)	1.19 (1.09,1.29)	1.18 (1.08,1.28)	1.18 (1.09,1.28)
edu[NVQ4&5]	1.31 (1.18,1.45)	1.43 (1.29,1.57)	1.36 (1.23,1.50)	1.38 (1.25,1.52)
eth	0.92 (0.81,1.05)	0.97 (0.84,1.11)	0.94 (0.83,1.07)	0.94 (0.82,1.06)
sing	0.80 (0.73,0.86)	1.25 (1.13,1.39)	0.94 (0.87,1.01)	0.93 (0.86,1.02)

Table shows the posterior mean, with the 95% interval in brackets.

^a $\delta_1 = 0$ and $\delta_2 = 0$ is MAR

Yun *et al.* (2007) plot a quantity of interest against a sensitivity parameter, for a set of models in which this sensitivity parameter is fixed to a range of values, and use the resulting sensitivity plot to show the dependence of their conclusions to assumptions about the missingness. We extend this idea to two sensitivity parameters, using trellis graphs to demonstrate the level of robustness of our quantities of interest to changes in δ_1 and δ_2 . Figure 8.2 shows the posterior mean and 95% interval of the proportional increase in pay associated with covariates relating to our questions of interest. Regardless of the δ values, there is no evidence that being non-white is associated with lower pay, and strong evidence that a higher educational level is associated with higher pay, although the magnitude of the associated increase for degree level education is less robust. However, the effect of gaining a

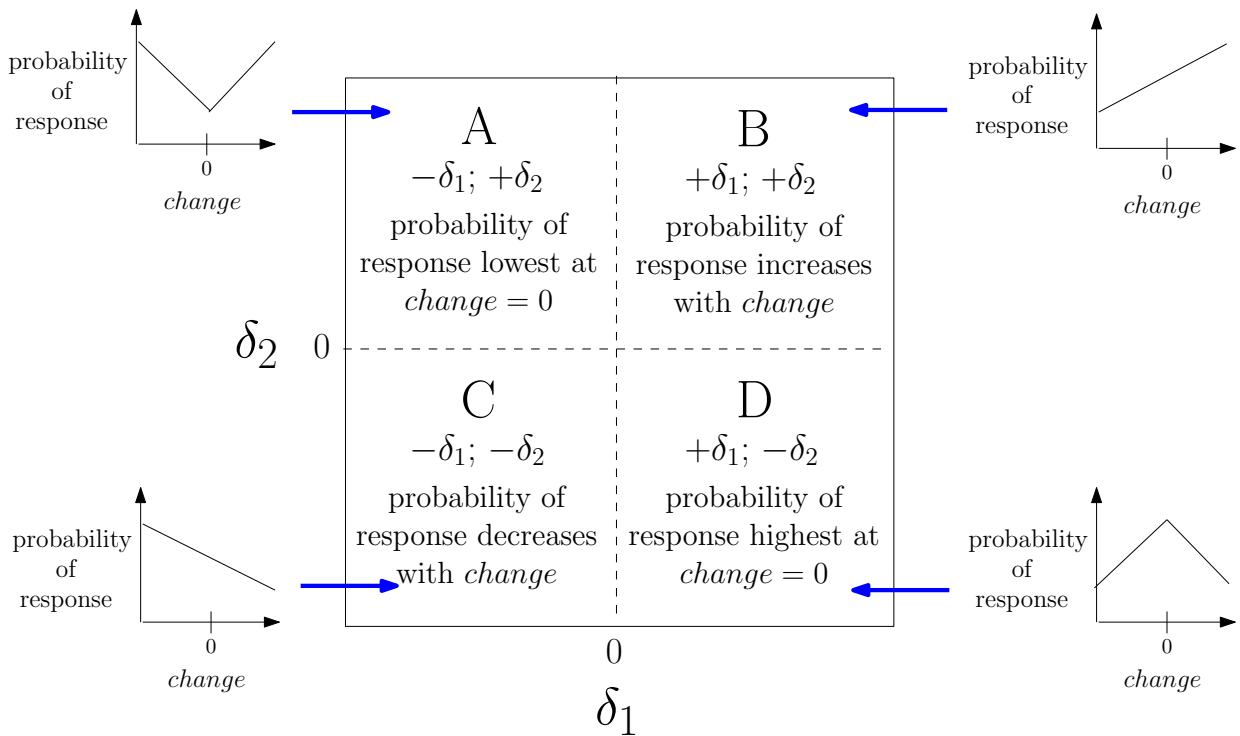
Figure 8.2: MCS income example: proportional increase in pay associated with selected covariates versus δ_1 , conditional on δ_2 from JMJ variants



partner is very unclear.

To better understand the plausibility of the different δ values, we use an alternative presentation of the results, plotting the posterior means as a series of contour lines. The resulting graph has four quadrants, determined by the signs of δ_1 and δ_2 . Figure 8.3 summarises how the probability of response will vary with the change in level of pay between sweeps in each of the quadrants, labelled A-D. It is the direction of the lines in the sample graphs for each quadrant which is important. (These mini graphs assume all other covariates are held constant, and assume δ_1 and δ_2 have the same magnitude although this need not be the case.) The views of our expert suggest that δ values in quadrant D are most plausible.

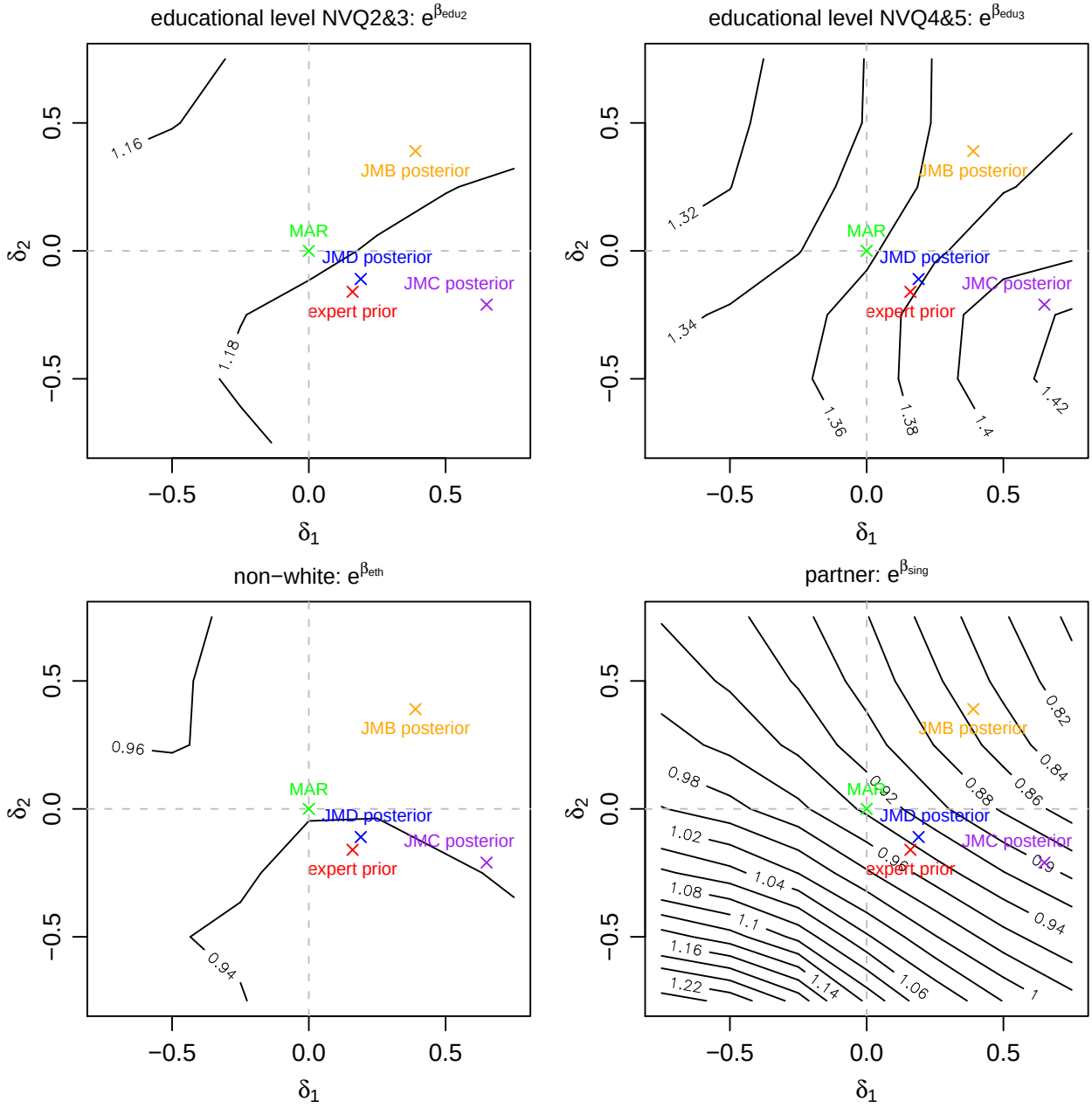
Figure 8.3: MCS income example: the four quadrants of δ_1 and δ_2



Our alternative plots are shown in Figure 8.4, with a number of points of interest marked on each. These points are at the δ values relating to MAR missingness (MAR), our expert's prior (expert prior) and the posteriors for model JMB which assumes a linear functional form for the response model of missingness, JMC which is a proxy for the likelihood and our base model JMD (JMB posterior, JMC posterior and JMD posterior respectively). By definition, the JMB posterior is constrained to lie on a diagonal line passing through quadrants B and C such that $\delta_1 = \delta_2$, and is actually in quadrant B. Our expert prior, JMC posterior and JMD posterior all lie in the top part of quadrant D.

The closer the contour lines, the greater the variation in the proportional change in pay associated with a selected covariate as δ_1 and δ_2 change. So the sparsity of lines in the two left plots of Figure 8.4 indicates the robustness of the results relating to educational level NVQ2&3 and being non-white.

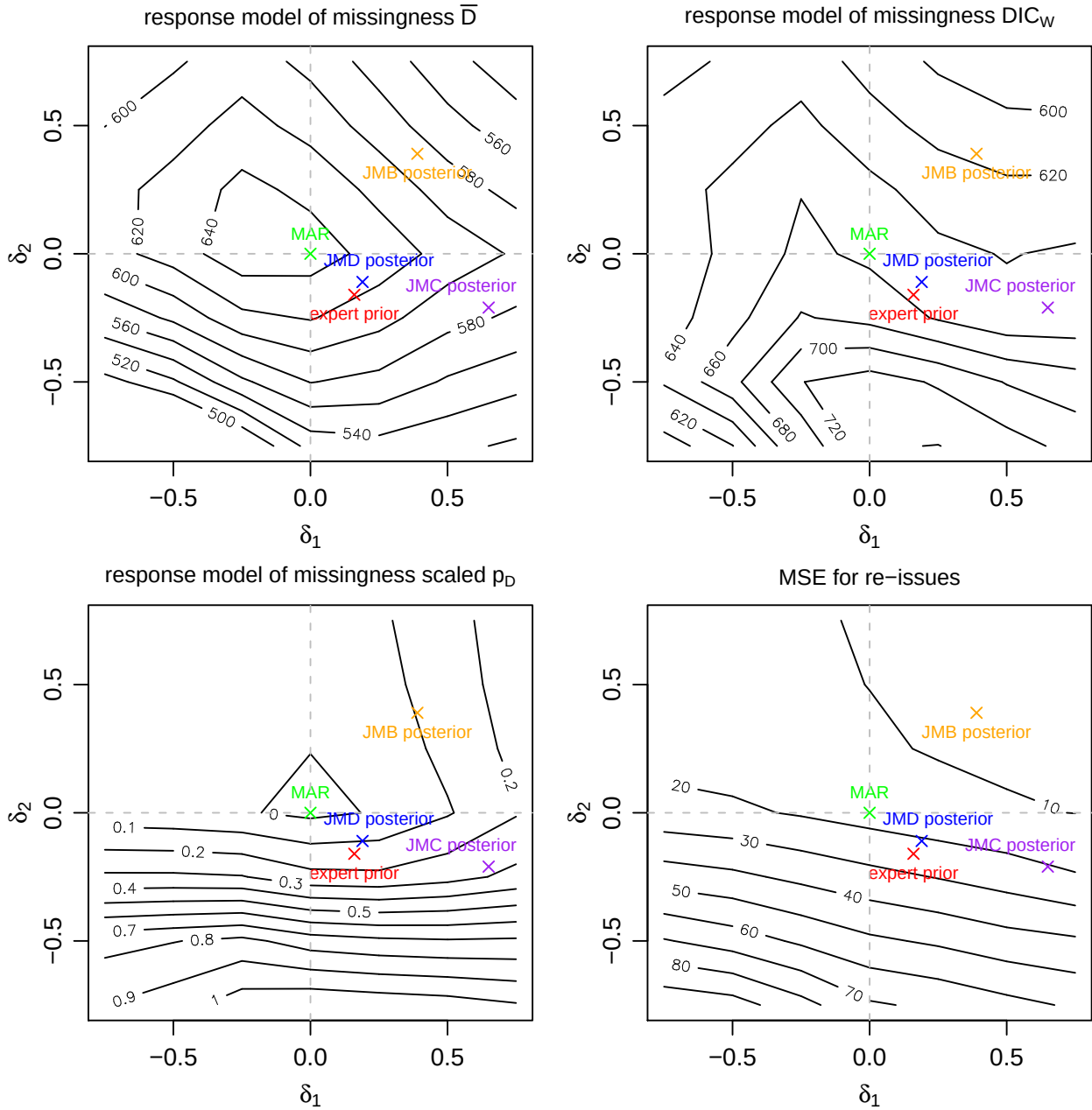
Figure 8.4: MCS income example: posterior mean of proportional change in pay associated with selected covariates versus δ_1 and δ_2 from JMJ variants



By contrast, the dense contours in the bottom right plot show that the proportional increase of 1.24 in pay associated with gaining a partner when δ_1 and δ_2 are at their most negative ($\delta_1 = \delta_2 = -0.75$), reduces as either δ_1 or δ_2 increases, until there is a substantial proportional decrease (0.8) when δ_1 and δ_2 are at their most positive ($\delta_1 = \delta_2 = 0.75$).

As with the assumption sensitivity analysis, we can gain some insight about these models by looking at the DIC_W for the response model of missingness and its associated scaled p_D . We can also use the mean square error of the fit of hourly pay for the seven re-issued individuals as a measure of model fit. These, along with the mean deviance for the response model of missingness are shown in Figure 8.5, which uses the same format as Figure 8.4.

Figure 8.5: MCS income example: response model of missingness mean deviance, DIC_W and scaled p_D , and the mean square error of the fit of hourly pay for the re-issued individuals versus δ_1 and δ_2 from JMJ variants



As expected, the mean deviance, $\bar{D}$, decreases as JMJ departs from MAR in any direction. The DIC_W plot suggests that models in the lower half of quadrant D do not explain the missingness as well as other models. The message from the scaled p_D plot, is that the departures from MAR have a much greater impact on the model of interest in quadrants C and D as δ_2 becomes increasingly negative, than in quadrants A and B. This is consistent with the closer contours in the bottom quadrants in Figure 8.4. The mean square error (MSE) for the seven re-issued individuals provides little support for the models in quadrant C and those in the lower half of quadrant D. Of our fitted models, JMB is the best based on the measures of fit shown in Figure 8.5.

Combining this information suggests that a plausible range of values for our quantities of interest

should be based on models which fall in quadrant B, possibly extending downwards into the upper part of quadrant D. This leads us to conclude that there is weak evidence that gaining a partner is associated with lower pay, in line with the assumption sensitivity.

In all these analyses, the estimation of the parameter associated with *sing* has been most sensitive to model changes. β_{sing} differs from the other parameters in that it can only draw information from sweep 2 values (by design it always takes value 1 in sweep 1). To see if the other parameters exhibit similar sensitivity if sweep 1 data cannot help with their estimation, we run a reduced version of JMJ in which the random effects are removed from the model of interest, and its parameters are estimated using sweep 2 data only. We run four variants only, using the four combinations of the extreme values of δ_1 and δ_2 (i.e. $\delta_1 = \delta_2 = -0.75$; $\delta_1 = -0.75, \delta_2 = 0.75$; $\delta_1 = \delta_2 = 0.75$ and $\delta_1 = 0.75, \delta_2 = -0.75$). We now find a lot more variation in the estimates of the other parameters. This suggests that longitudinal models may robustify parameter estimates through the inclusion of fully observed data from another sweep in some circumstances.

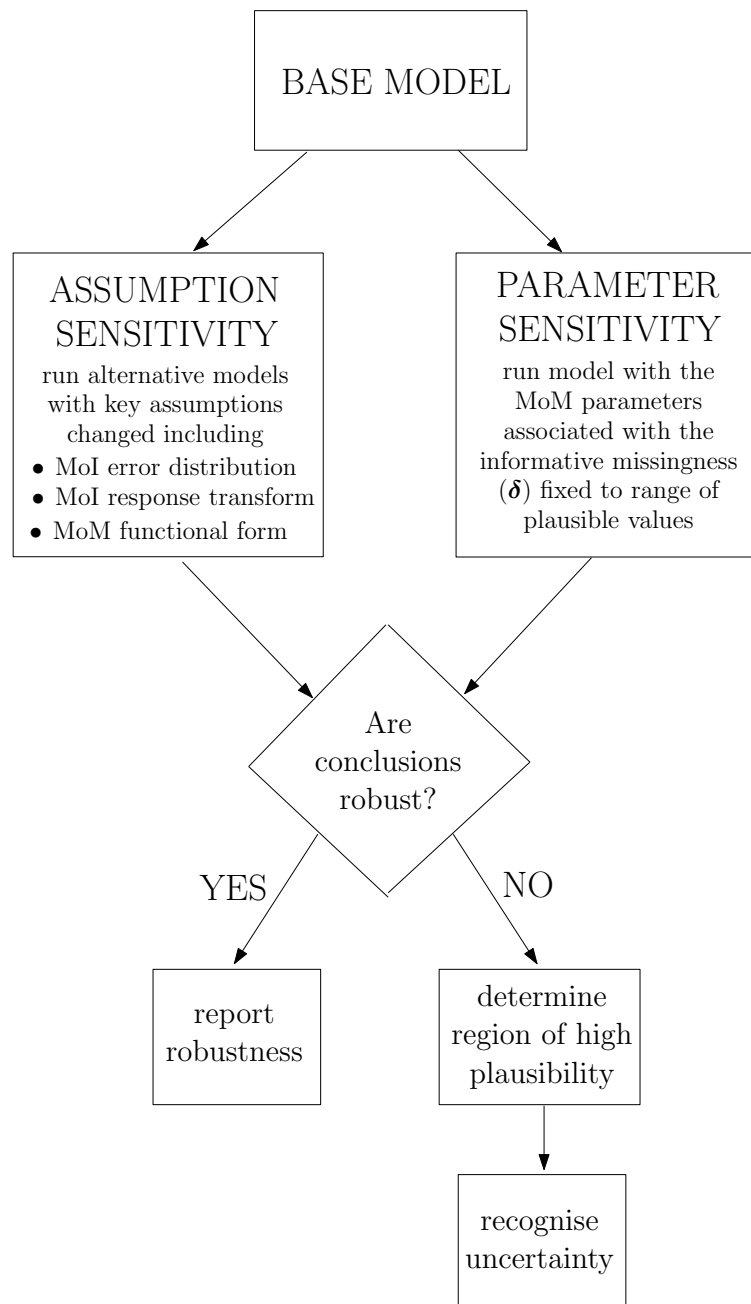
8.3 Strategy for sensitivity analyses

When modelling missing data, some form of sensitivity analysis is essential because we are forced to make assumptions which are untestable from the data. As discussed, there are many possible options, and at least to some extent the choice will be determined by the problem at hand. However, based on our experience, we propose a general strategy for use when a joint model with a selection model parameterisation is used for analysis of a dataset with suspected non-ignorable missingness.

This strategy, outlined in Figure 8.6, assumes that a reasonable base model has already been formed. Two types of sensitivity analysis should then be carried out, an assumption sensitivity and a parameter sensitivity. For the assumption sensitivity, a number of alternative models should be run, formed from the base model by changing key assumptions. These should include, but not be limited to, changes in the model of interest error distribution, the transformation of the model of interest response and the functional form of the model of missingness. The parameter sensitivity involves running the base model with the parameters controlling the extent of the departure from MAR fixed to values in a plausible range.

The results of both sets of sensitivities should then be analysed to establish how much the quantities of interest vary. A range of plots, providing complementary views of the analysis is recommended. Plots like those in Figure 8.2 are particularly useful. If the conclusions are robust, this should be reported. Otherwise a range of diagnostics (e.g. fit of a hold-out sample and model of missingness DIC_W) should be used to determine a region of high plausibility, and the uncertainty recognised. The sensitivity analysis may also suggest that the base model should be reconsidered, or more external

Figure 8.6: Proposed strategy for sensitivity analyses for models with missingness



information sought from experts or related studies to help assess the sensitivities.

Summary: designing sensitivity analyses

Sensitivity analysis is crucial when modelling missing data, to allow the robustness of conclusions about the questions of interest to be investigated. We propose a dual approach in which (1) a base model is compared to a number of alternatives (assumption sensitivity) and (2) the base model is rerun a number of times with the parameters controlling the extent of the departure from MAR fixed to a range of plausible values (parameter sensitivity). An outline of our proposed strategy is shown diagrammatically in Figure 8.6.

We have demonstrated this strategy using our MCS income example, and find that our conclusions regarding the substantive questions about educational level and ethnicity are robust. Higher levels of education are associated with higher hourly pay, but ethnicity makes little difference. However, there is considerable uncertainty about the effect of change in partnership status on income. An investigation of the fit of different aspects of our joint models, using the hold-out sample of re-issued individuals and the model of missingness DIC_W , provides greatest support for JMB and suggests that gaining a partner is associated with lower pay.

Chapter 9

Bayesian modelling of missing data: conclusions and extensions

Each chapter has already been summarised, so here we draw together our findings and suggest avenues of future research.

9.1 Strengths and limitations of Bayesian joint models

Bayesian full probability modelling (indicated hereafter by BJM - Bayesian Joint Modelling) provides a flexible and ‘statistically principled’ way of modelling non-random missing data, using a selection model factorisation of a joint model. This type of joint model comprises a model of interest and model(s) of missingness. By estimating the unknown parameters and the missing data simultaneously, this modelling approach ensures that their estimation is coherent. Since the required joint models are built in a modular way, they are easy to adapt, facilitating sensitivity analysis which is crucial when the missing data mechanism is unknown.

Provided the different parts of the joint model are correctly specified, BJM performs much better than simplistic alternative methods, such as complete case analysis, in terms of bias in the model of interest parameter estimation and general model fit. However, the results are not robust to incorrect specification of the different parts of the joint model. In particular, the performance of BJM can be adversely affected if the error distribution of the model of interest is misspecified or the functional form of the model of missingness is incorrectly specified. Unfortunately, we can never be sure about the correct choices, so sensitivity analysis should always be carried out on these key assumptions. We recommend that expert knowledge or other external information is used to help select a range of plausible options.

If no knowledge about the shape of the missingness is available from external sources or expert opinion, the safest strategy is to use a linear logit for the model of missingness. If the missingness is linear

or close to linear, we have shown that there are potential benefits, and if not, its effect appears to be reasonably benign at least in the examples we have studied. Other functional forms should only be used if supported by external evidence, and informative priors can be derived to help with the estimation of the model of missingness parameters.

A drawback to this approach is that these types of complex statistical models do not run quickly for large datasets in WinBUGS, the readily available software typically used for Bayesian analysis. Additionally, we found that the current capability of the WinBUGS software limits the scope for easily implementing complex joint models which incorporate combinations of certain types of correlated covariates with missingness, as these may require more specialised MCMC sampling algorithms.

An alternative approach to modelling missing data, multiple imputation, avoids some of the speed and computational issues associated with BJM, because it is a two stage approach which imputes the missing data and analyses the model of interest separately, and so divides the problem making it more tractable. However, the trade-off is the need to ensure that the imputation and analysis models are compatible, which is not trivial. Additionally, unless you are prepared to generate imputations using a series of conditional univariate models, the problems associated with specifying a multivariate imputation model for correlated variables remain. Also, while multiple imputation can be applied if data are MNAR, this is not entirely straightforward.

One of the advantages of the Bayesian approach is that it allows expert knowledge to be incorporated through informative priors. This is particularly useful for parameters in the model of missingness that are difficult to estimate. From our experience, we recommend that the elicitation process focusses on weakly identified variables, particularly those allowing informative missingness. It should allow for correlation between these variables and pay particular attention to their functional form.

9.2 Substantive questions

Three real examples were used in our investigation of BJM. Findings for the two smaller examples, NCDS educational test scores and the antidepressant trial, are summarised at the end of Chapter 5. Conclusions regarding our main example, MCS income, can be found in the Chapter 8 summary. Our approach allowed the robustness of the conclusions to the substantive questions to be tested and any uncertainty reported. These examples also enabled us to develop a strategy for modelling longitudinal data with missing values using a Bayesian approach, which was one of the main objectives of this thesis. We discuss this in the next section.

9.3 Strategy for Bayesian modelling of non-ignorable missing data

We start by proposing a general strategy for building appropriate models in Section 9.3.1, that can be used by social scientists and other researchers. Although designed for analysing data from longitudinal studies, it is also applicable for cross sectional data. This approach allows the uncertainty from the missing data to be taken into account, and all relevant sources of information relating to the question under investigation to be utilised. In Section 9.3.2 we discuss how the DIC can be used for evaluating such models.

9.3.1 Proposed model building strategy

Based on the research carried out for this thesis, our proposed strategy for building Bayesian models for data with missing covariates and responses is encapsulated by the following steps.

- 1 Select an initial model of interest based on complete cases. This will include choosing a transform for the response, model structure and a set of explanatory variables. It is worth noting plausible alternatives at this stage, which can then be incorporated into the sensitivity analysis in step 5.
- 2 Add a covariate model of missingness to produce realistic imputations of any missing covariates, taking account of possible correlation between these covariates as necessary. A latent variable approach can be used for binary or categorical variables. Check the reasonableness of this model by comparing the pattern of the imputed values with the observed values. Crucially this allows incomplete cases to be included in the model estimation.
- 3 Add a response model of missingness to allow informative missingness in the response. In the absence of any prior knowledge, the safest strategy is to assume a linear relationship between the probability of missingness and the response (or change in response) in the response model of missingness. More complex models of missingness with vague priors are difficult to estimate, as they are reliant on limited information from assumptions about other parts of the model.
- 4 To complete the formation of a base model, add additional data and/or expert knowledge into the various sub-models to help with parameter estimation, if available. Information relating to the response model of missingness has the potential to make the biggest impact, in particular regarding its functional form. Additional data may come from other, similar studies or be provided by earlier/later sweeps not under investigation.
- 5 Carry out a series of sensitivities to determine the robustness of any conclusions to different plausible assumptions for all parts of the model. A particularly key sensitivity is the choice of transform of the response variable in the model of interest. Expert knowledge can help with setting up the parameter sensitivity.

This approach was followed for our MCS income example, except that we carried out steps 2 and 3 in reverse, because of our greater focus on missing response data as opposed to missing covariates. However, when modelling using WinBUGS, the ordering proposed above seems more natural, since a combined model of interest and covariate model of missingness will run with missing responses assuming MAR, whereas a combined model of interest and response model of missingness will not run with missing covariates. So dealing with the covariate model of missingness first, avoids the necessity of working with a restricted dataset which excludes records with missing covariates or “filling in” the missing covariates using simplistic assumptions prior to running the model.

Our proposed strategy assumes that the covariates are MAR, but in principle step 2 can be elaborated to allow MNAR covariates. This is a possible extension that is discussed in Section 9.4. Also, step 3 is only relevant for MNAR responses, since it is not necessary to explicitly model missing response if it is MAR.

At each step, checks of model fit should be carried out to ensure that the models are plausible. Although this strategy forms a base model, the aim is not to select a single model, but a range of models encompassing different realistic assumptions. These are then used to determine the robustness of the conclusions about the substantive questions and examine the uncertainty as necessary.

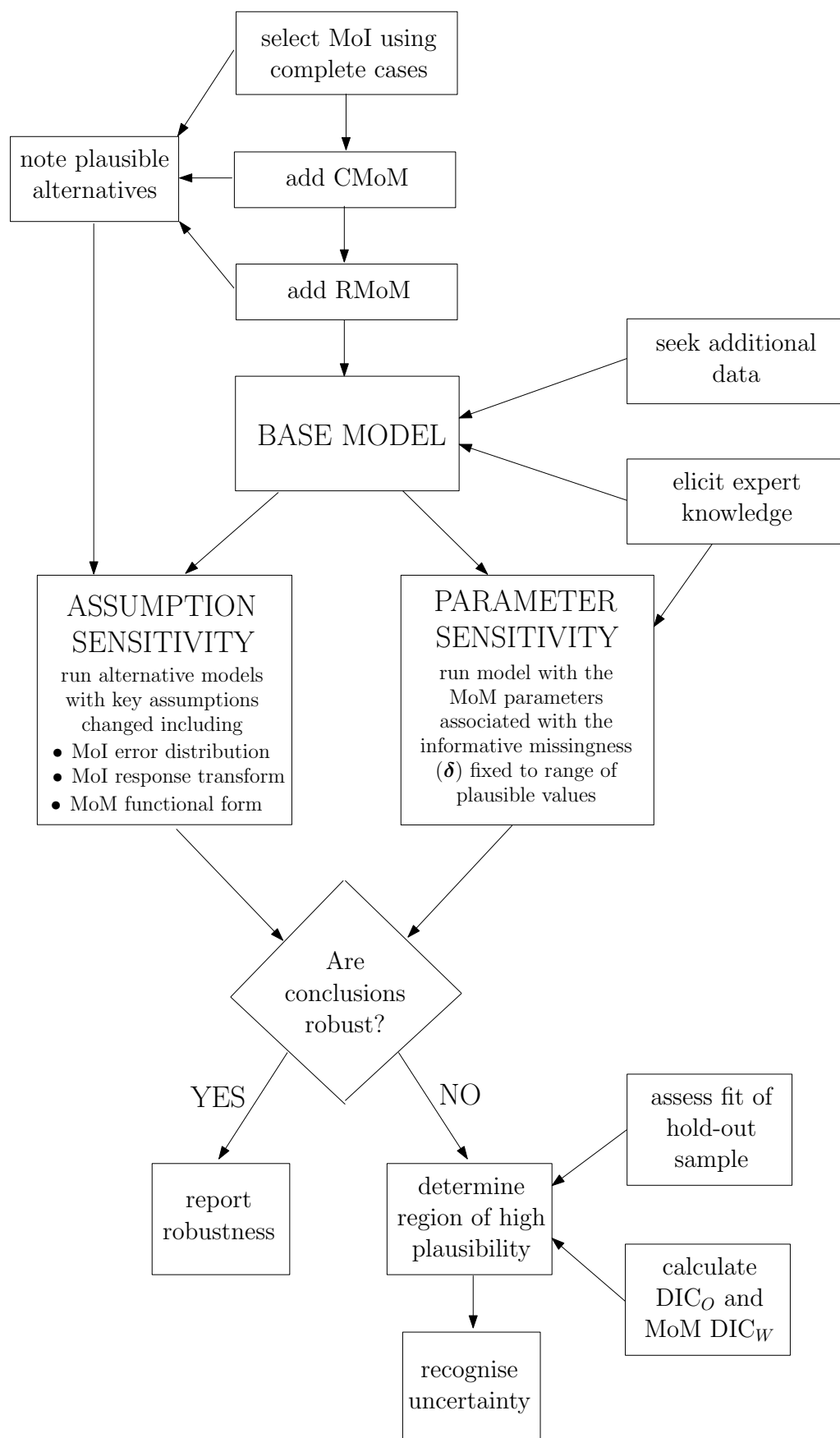
This strategy is summarised diagrammatically in Figure 9.1. Note that the fit of a hold-out sample and the DIC can be used to help determine which models are plausible. Uses of the latter are described in the next section.

There are situations where it may be necessary to adapt this strategy. For example, if the dataset to be modelled is very large, or there are large numbers of covariates with missingness, then running times may be prohibitive or computational issues encountered. In these circumstances, one option is to take a two stage approach and impute some or all of the covariates prior to running the Bayesian joint model. In this case the issues surrounding multiple imputation regarding compatibility will apply. It may be possible to identify covariates where using simplistic assumptions to impute their missingness is acceptable, as we did with the region covariate, *reg*, in the MCS income example. If not all the covariates are correlated, another option is to split the covariate model of missingness into several smaller sub-models.

9.3.2 Proposed use of DIC in the presence of missing data

For complete data, DIC is routinely used by Bayesian statisticians to compare models, a practice facilitated by its automatic generation in WinBUGS. However, using DIC in the presence of missing data is far from straightforward. The usual issues surrounding the choice of plug-ins are heightened, and in addition we must ensure that its construction is sensible. It would be a mistake to use it

Figure 9.1: Proposed strategy for Bayesian modelling of non-ignorable missing data



to select a single model, but we have identified three ways in which it has the potential to provide insight into proposed models. These may give conflicting messages, and deciding how much emphasis is placed on each requires judgement.

- 1 A DIC based on the observed data likelihood, DIC_O , can help with the choice of the model of interest. It should be used to compare joint models built with the same model of missingness but different models of interest. This DIC_O cannot be generated by WinBUGS, but can be calculated from WinBUGS output using other software. Daniels and Hogan (2008) provide an algorithm for its calculation, which we have adapted and implemented for both simulated and real examples. We recommend performing two sets of checks: (1) that the plug-ins are reasonable (i.e. if posterior means are used, they should come from symmetric, unimodal posterior distributions) and (2) that the size of the samples generated from the likelihoods is adequate (we suggest plotting deviance against sample length and checking for stability, as in Figure 4.8). For DIC_O , we must choose plug-ins that ensure consistency in the calculation of the posterior mean deviance and the plug-in deviance, so that missing values are integrated out in both parts of the DIC. This consideration makes *logitp* plug-ins for the selection model of missingness unsuitable.
- 2 The model of missingness scaled p_D can indicate the level of departure from missing at random, given the model assumptions. It comes from the conditional DIC generated by WinBUGS, DIC_W . However, we have found that it is not always robust to the choice of plug-ins used, and recommend caution in its interpretation if the plug-ins are suspect.
- 3 The model of missingness part of DIC_W allows comparison of the fit of the model of missingness for joint models with the same model of interest.

Although we propose using two component measures from the conditional DIC routinely generated by WinBUGS, the total DIC_W has asymptotic and coherency difficulties for selection models. We do not recommend using either the WinBUGS total DIC or the WinBUGS model of interest DIC. An alternative to using DIC to compare models, is to assess model fit using a set of data not used in the model estimation, if available.

9.4 Extensions

There is scope for extending the MCS income example in a number of directions. One possibility is to allow for sweep 1 missingness. If the sweep 1 missingness is restricted to the response, this would require extending the missing value indicator to incorporate sweep 1 pay and adapting the response model of missingness accordingly. If the sweep 1 missingness also included covariates, the covariate model of missingness would also need revisiting. In particular, the modified covariate model of missingness would need to account for within-person correlation between covariate values at sweeps 1 and

2 and include individual random effects (Carpenter and Goldstein, 2004). Another possibility is to distinguish between item and wave missingness, or between types of non-response such as refusals and non-contacts, which are known to have different predictors (Plewis *et al.*, 2008). Clearly more information is available for individuals with item rather than wave missingness, and this extra information could be utilised by building separate sub-models for the two types of missingness. Distinguishing between different types of non-response leads to a missingness indicator with three or more categories, which could be modelled using multinomial logistic regression.

Sweep 3 MCS data is now available and could be used for validation. Either sweep 3 hourly pay could be predicted using the models fitted to data from the first two sweeps and compared with the observed values, or the model parameters re-estimated using the three sweeps of data and the results compared with their previous estimation based on two sweeps. The missingness pattern for sweep 3 is non-monotonic, with 1,444 families who were missing in sweep 2 responding at sweep 3 (Ketende, 2008). Whether this extra data can be used to robustify the effect of gaining a partner is of particular interest. It would also allow us to test whether adding extra sweeps reduces the necessity of allowing for informative missingness in longitudinal models.

Our models could also be developed to allow informative missingness in the covariates. Some of the issues surrounding this were outlined in Section 8.1.3, but all the options entail adding substantial complexity to existing joint models through additional sub-models. Implementing models with this level of complexity may not be feasible in WinBUGS, in which case alternative software and/or approximations to the full joint model would need to be sought or developed.

Another area of future work is the use of the DIC for models with missing data. We implemented a DIC based on the observed data likelihood, DIC_O , for some of our simpler examples with response missingness, but questions concerning its implementation with more complex models remain. In particular, we did not explore its implementation and practical use when there are missing covariates as well as missing responses. Our implementation for this thesis was tailored to specific examples, so a first stage to this work would be to generalise the R functions that we developed, so that a DIC_O can be evaluated routinely for a variety of models allowing for informative missingness in the response. Extensions for missing covariates would require adapting the underlying algorithms before tackling the problem of implementation.

Appendix A

Datasets

In Appendix A we provide additional details about the datasets used in this thesis. These include information on which variables we extracted from the UK data archive files, and how they were processed.

A.1 Details of NCDS educational test scores data

The NCDS educational test score dataset is introduced in Section 3.1.1, and analysed in Section 5.1. Details of the source of the variables in this dataset are given in Table A.1.

Table A.1: Details of the source of NCDS educational test scores variables

name	description	variable code
<i>maths7</i>	mathematics test score at age 7	N90
<i>maths11</i>	mathematics test score at age 11	N926
<i>maths16</i>	mathematics test score at age 16	N2930
<i>sc11</i>	social class at age 11	N1687

All variables extracted from rec1.tab (PMS & NCDS1-3 variables - most “useful”), included in the SN:3148 NCDS Perinatal Data and Sweeps One to Five 1958-1991 download from the UK data archive.

A.2 Details of MCS income data

An overview of the MCS income dataset is given in Section 3.1.2, and this dataset is analysed extensively in Chapters 5 to 8. Details of the criteria for a main respondent’s inclusion in this dataset are given in Table A.2, and Table A.3 provides the source of the variables selected for this dataset. The main respondent is predominantly the cohort member’s natural mother, and there are just three exceptions in our subset of the data (two are the natural father and one the foster mother). Not all the required variables are in the UK data archive downloads, but are calculated from a number of variables which are available. Details of these calculations are also given in this section. In addition,

Tables A.4 and A.5 provide descriptions of the NVQ levels which determine the *edu* categories and the NS-SEC classifications behind the *sc* categories respectively.

Table A.2: Criteria for inclusion of record in MCS income dataset

criteria	condition used ^a
is single in sweep 1	<i>adresp00</i> = 1
has paid work in sweep 1	<i>admwrk00</i> = 1
is not self-employed in sweep 1	<i>amsepa00</i> < 0 and <i>adm05s00</i> ≠ 3
has not stopped working in sweep 2	<i>admwrk00</i> ≠ 2 and <i>admwrk00</i> ≠ -1
is not self-employed in sweep 2	<i>bmsepa00</i> < 0 and <i>bdm05s00</i> ≠ 3

^a shown in terms of the MCS eight character codes

Table A.3: Details of the source of MCS income variables

name	description	variable type	sweep 1 code ^a	sweep 2 code ^b
<i>hpay</i>	hourly net pay	calculated	see below	see below
<i>age</i>	main respondent's age at interview	derived	<i>admagi00</i>	<i>bdmagi00</i>
<i>hsize</i>	number of people in household	derived	<i>adtotp00</i>	<i>bdtotp00</i>
<i>kids</i>	number of siblings in household plus number of cohort babies	derived	<i>adtots00</i>	<i>bdtots00</i>
<i>edu</i>	main respondent's educational level: the level of National Vocational Qualification (NVQ) equivalence of the individual's highest academic or vocational educational qualification	combination ^c	<i>adnvqm00</i>	see below
<i>sc</i>	main respondent's social class: National Statistics Socio-Economic Classification (NS-SEC) grouped into 5 categories	derived	<i>adm05s00</i>	<i>bdm05s00</i>
<i>eth</i>	main respondent's ethnic group: 2 levels (white/non-white) aggregated from 6 category census classification	derived	<i>adm06e00</i> ^e	<i>bdm06e00</i> ^e
<i>sing</i>	single/partner: from parent interview response summary	derived	1 by selection	<i>bdresp00</i> ^f
<i>reg</i>	residing region: 2 levels (London/other) aggregated from 12 category classification	admin ^d	<i>aaregn00</i> ^g	<i>baregn00</i> ^g
<i>ctry</i>	country	admin ^d	<i>aactry00</i>	n/a
<i>stratum</i>	stratum within country: 9 category classification	admin ^d	<i>aaptty00</i>	n/a
<i>ward</i>	fieldwork point number incorporating super-wards (created by combing small wards, as described in Plewis (2007a), p.12)	admin ^d	<i>aasptn00</i>	n/a

^a All sweep 1 variables extracted from *mcs1_household_grid_june_2006.tab*, included in the SN:4683 Millennium Cohort Study First Survey, 2001-2003 (MCS1) download from the UK data archive.

^b All sweep 2 variables extracted from *mcs2_household_grid_june_2006.tab*, included in the SN:5350 Millennium Cohort Study Second Survey, 2003-2005 (MCS2) download from the UK data archive. All sweep 2 variables are set to missing if the main respondent changed between sweeps (*bmpnum00* ≠ *ampnum00*).

^c derived variable for sweep 1 and calculated for sweep 2

^d survey administrative data variable

^e aggregated into 2 levels: white (code 1) and non-white (codes 2-6)

^f aggregated into 2 levels: single (codes 1 and 7) and partner (codes 2-6)

^g aggregated into 2 levels: London (code 3) and other (codes 1-2 and 4-12)

Calculation of *hpay*

Hourly net pay, *hpay*, is calculated using main respondent variables taken from Module J: Employment and Education as follows:

- 1 Calculate annual net pay, *apay*, using the amount of take-home pay last time respondent paid (sweep 1: *amneta00*; sweep 2: *bmneta00*) and the time period covered by last take-home pay (sweep 1: *amnetp00*; sweep 2: *bmnetm00*).
- 2 Calculate the number of hours worked in a week excluding overtime, *hrs*, by taking the minimum of the number of hours usually worked in a week including usual overtime (sweep 1: *amwkdir00*; sweep 2: *bmwkdir00*) and 35 (assumed to be the length of a standard working week excluding lunch breaks).
- 3 Calculate hourly net pay assuming 46 working weeks in a year and excluding overtime, according to the formula

$$hpay = \frac{apay}{hrs \times 46}.$$

Overtime pay is included in *apay*, so ideally we would exclude only unpaid overtime from *hrs*. Unfortunately, we cannot distinguish between unpaid and paid overtime from the available data, so we exclude all overtime hours. Despite this inconsistency in the treatment of overtime, we consider the method adequate for the purpose of this methodological research.

Calculation of *edu* for sweep 2

There is no derived educational variable for sweep 2, similar to *adnvqm00* for sweep 1, therefore we create one by updating *adnvqm00* as follows:

- 1 Create a single variable, *newac*, indicating the highest level of new academic qualifications acquired since sweep 1 using variables *bmnnacq0a* – *bmnnacq0h*.
- 2 Create a single variable, *newvo*, indicating the highest level of new vocational qualifications acquired since sweep 1 using variables *bmnnvcq0a* – *bmnnvcq0h*.
- 3 Combine *newac* and *newvo* into a single variable, *newqual*, indicating the highest level of new academic or vocational qualifications acquired since sweep 1.
- 4 Take the maximum of *adnvqm00* and *newqual*.

Note that the calculation outlined above is only valid for creating a sweep 2 educational variable for existing families where the main respondent is unchanged between sweeps. By definition our datasets only include such individuals.

Table A.4: Levels of National Vocational Qualification (NVQ) equivalence

<i>edu</i> category	NVQ level	description
5	NVQ5	Higher degree
4	NVQ4	Diploma in higher education; First degree; Nursing/other medical qualification; Professional qualification at degree level
3	NVQ3	A/AS/S levels; NVQ/SVQ/GSVG level 3
2	NVQ2	O level/GSCE grades A-C; NVQ/SVQ/GSVG level 2; Trade apprenticeships
1	NVQ1	GSCE grades D-G; NVQ/SVQ/GSVG level 1
0	None	None of these qualifications

Table A.5: National Statistics Socio-Economic Classification (NS-SEC)

<i>sc</i> category	NS-SEC classification	description
1	1	Managerial and Professional occupations
2	2	Intermediate occupations
n/a	3	Small employers and own account workers
3	4	Lower supervisory and technical occupations
4	5	Semi-routine and routine occupations

A.3 Details of BCS70 data

In Chapter 7 we incorporate a subset of data from BCS70 into our MCS income model, to help with the imputation of the missing covariates. Details of the criteria used to determine if a BCS70 cohort member should be included in this subset are given in Table A.6. We require data relating to educational level, *BCSedu*, and whether an individual is single or has a partner, *BCSsing*, as described in Table A.7.

Table A.6: Criteria for inclusion of record in BCS70 dataset

criteria	condition used ^a
is female	$dmsex = 2$
is single in sweep 5	$dm\text{spart} = 2$
is full-time or part-time paid employee in sweep 5	$econact = 1$ or 2
children (0-16 years) in household in sweep 5	$anychd = 1$
is full-time or part-time paid employee in sweep 6	$bd7ecact = 1$ or 2

^a shown in terms of the BCS70 variable codes

Calculation of *BCSedu*

There are no derived educational level variables giving the individual's highest academic or vocational educational qualification for sweeps 5 or 6 of the BCS70 data. We create these variables from

Table A.7: Details of the source of BCS70 variables

name	description	variable type	sweep 1 code ^a	sweep 2 code ^b
<i>BCSedu</i>	cohort member's educational level ^c	calculated	see below	see below
<i>BCSsing</i>	single/partner	derived	single by selection	<i>bd7spphh</i> ^d

^a All sweep 5 variables extracted from bcs2000.dta, included in the SN:5558 1970 British Cohort Study: Twenty-Nine-Year Follow-up, 1999-2000 download from the UK data archive.

^b All sweep 6 variables extracted from bsc.2004.followup.dta, included in the SN:5585 1970 British Cohort Study: Thirty-Four-Year Follow-up, 2004-2005 download from the UK data archive.

^c The level of National Vocational Qualification (NVQ) equivalence of the individual's highest academic or vocational educational qualification.

^d Codes whether cohort member has a partner in household (Y/N).

scratch, using the original variables coded from answers to detailed questions about qualifications. Our algorithms are implemented using the R software (R Development Core Team, 2007).

The algorithm for calculating the sweep 5 *BCSedu* is based on the SPSS syntax to derive highest academic and vocational qualifications, provided by Andrew Jenkins, Gerald Makepeace and Samantha Parsons in Appendix 7 of the report on the NCDS/BCS70 1999-2000 sweep compiled by Shepherd (2002). The questions asked about academic and vocational qualifications in sweep 5 are given in Section 7 (Lifelong Learning) of the NCDS/BCS70 1999-2000 Follow-ups CAPI Documentation prepared by the National Centre for Social Research NCDS/BCS70 Team. This documentation is part of the SN:5558 1970 British Cohort Study: Twenty-Nine-Year Follow-up, 1999-2000 download from the UK data archive. The qualifications used in creating *BCSedu* are as follows:

- GCSE, O levels and CSE;
- A/S and A levels;
- Scottish school qualifications;
- Degrees;
- Nursing qualifications, diplomas and PGCE;
- ONC, OND, HNC, HND;
- BATES;
- City and Guilds;
- ROSA qualification;
- Pitmans qualifications;
- NVQ and GNVQ;
- other vocational qualifications (including trade apprenticeships and HGV licence);

and the mapping of these qualifications into the national qualifications framework is tabulated on page A138 of Shepherd (2002).

For sweep 6, the questions about education are asked in terms of qualifications gained since the previous interview. Those relating to academic qualifications are given in Section 14 (Lifelong Learning) of the

British Birth Cohort Study 1970 (BCS70) 2004 Follow up Questionnaire documentation prepared by the National Centre for Social Research, while those for vocational qualifications are in Section 15 (Vocational Qualifications). We adapt the algorithm for sweep 5, calculate the highest additional academic and vocational qualifications, and then set sweep 6 *BCSedu* to be the maximum of the level of the additional qualifications and sweep 5 *BCSedu*.

Appendix B

Simulated non-linear missingness

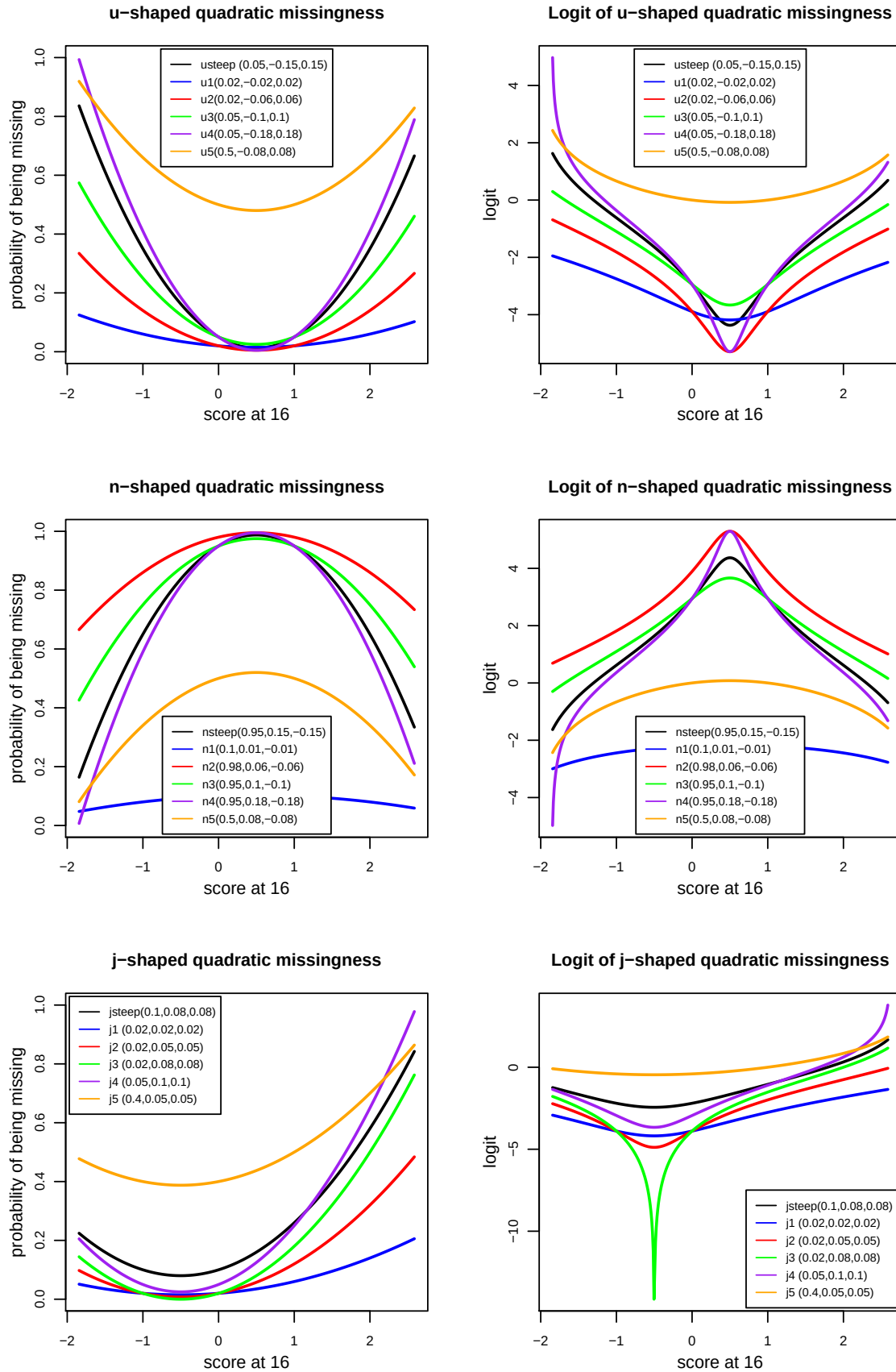
Appendix B provides details of our exploration of Bayesian full probability modelling using synthetic data scenarios with non-linear missingness, which are discussed in Section 4.5.

B.1 Quadratic missingness

To investigate the performance of joint models when the shape of the missingness is quadratic, we create 18 datasets for scenario Scores.quad by imposing quadratic missingness on the Scores data using Equation 4.5. We form six quadratic data sets with u-shaped quadratic missingness (*usteepest* and *u1-u5*), six with n-shaped quadratic missingness (*nsteep* and *n1-n5*) and six with j-shaped quadratic missingness (*jsteep* and *j1-j5*), as shown in Figure B.1. In a similar way to a linear logit being approximately linear for linear missingness, a quadratic logit is approximately quadratic for quadratic missingness if the probabilities are neither very low or very high. However, as can be seen from the plots, if the probabilities approach 0 or 1, then the logit becomes pinched. These shapes become increasingly extreme as the probability range nears its end points.

We start by assessing the performance of a linear logit missingness model, and run models TRUE, CC and JM on each dataset. All the models converge except for the n-shaped quadratic, *n2*, which has an extremely high percentage of missingness. The underlying shape of the distribution of the score at age 16 which has far fewer observations in its tails than the middle of the range, see Figure 3.1, ensures that the n-shaped quadratics have the potential to result in much higher proportions of missingness. The results of these runs are summarised in Table B.1. Clearly it is not possible for a linear logit to approximate a quadratic missingness shape closely. For the u-shaped and j-shaped quadratics there are small reductions in the β_1 bias, but for the n-shaped quadratics there are small increases. From the MSE, we find that the overall fit of JM is slightly better than CC for the u-shaped and j-shaped quadratics, but there is a slight deterioration for two of the n-shaped quadratics, when the β_0 bias also increases.

Figure B.1: Shapes of the quadratic missingness imposed on the datasets in scenario Scores.quad and their associated logits



The figures in brackets in the graph legends are the values of the ϕ parameters in Equation 4.5.

Table B.1: Scores.quad: modelling quadratic shaped missingness with a linear logit

	% β_1 bias			absolute β_0 bias			% MSE increase ^a			% of missingness
	CC	JM	%red. ^b	CC	JM	red. ^c	CC	JM	diff. ^d	
uSteep	-11.8	-8.9	3.0	0.05	-0.02	0.03	2.4	1.2	1.2	19.6
u1	-1.5	-1.5	0.1	0.00	0.00	0.00	0.0	0.0	0.0	4.5
u2	-2.8	-2.1	0.8	0.03	0.01	0.02	0.4	0.1	0.3	9.1
u3	-6.9	-6.6	0.3	0.02	0.01	0.01	0.7	0.6	0.1	14.3
u4	-16.4	-14.2	2.1	0.04	-0.01	0.03	4.4	3.1	1.3	22.9
u5	-9.4	-8.8	0.5	0.06	0.01	0.05	2.1	1.1	1.0	56.6
nSteep	27.4	28.5	-1.1	-0.11	0.04	0.07	13.9	12.0	1.9	80.8
n1	0.9	1.0	-0.1	0.00	0.00	0.00	0.0	0.0	0.0	8.9
n2*	-	-	-	-	-	-	-	-	-	91.0
n3	25.0	26.0	-1.0	-0.08	0.11	-0.03	9.7	12.0	-2.3	85.4
n4	19.5	22.4	-3.0	-0.17	0.04	0.13	13.2	7.9	5.3	77.5
n5	8.3	8.6	-0.3	0.01	0.04	-0.03	1.0	1.6	-0.5	43.1
jSteep	-9.2	-5.8	3.5	-0.05	0.00	0.05	1.9	0.5	1.5	16.9
j1	-1.2	-1.0	0.3	-0.01	-0.01	0.01	0.1	0.0	0.0	4.4
j2	-5.2	-2.5	2.7	-0.04	-0.01	0.03	0.7	0.1	0.6	7.5
j3	-8.6	-5.4	3.2	-0.06	-0.03	0.03	1.8	0.6	1.2	8.7
j4	-8.9	-3.1	5.8	-0.07	-0.02	0.06	2.4	0.2	2.2	13.3
j5	-10.3	-10.1	0.2	-0.06	-0.02	0.04	2.6	1.7	0.9	48.7

^a % increase in MSE from TRUE^b % reduction in absolute bias from CC to JM^c reduction in absolute bias from CC to JM^d difference between CC and JM columns

* run did not converge

Given the limitations of a linear logit model when the true missing data mechanism is not linear, we explore the use of a quadratic logit model, using Equation 4.7, with a logistic(0,1) prior for θ_0 and weakly informative $N(0,0.68)$ priors for θ_1 and θ_2 . We expect a quadratic logit model to be more flexible and to better approximate quadratic shaped missingness.

Models JM and TRUE are fitted using a quadratic logit for the missingness model to all the datasets in Scores.quad. All the models converged, apart from the j-shape with the highest percentage of missingness, $j5$, and the results are summarised in Table B.2.

Few datasets have reduced β_1 bias compared to the complete case analysis, and only half have any improvement in fit as measured by the MSE. We find that there are difficulties in approximating the quadratic missingness shapes. The tails of the n-shapes were poorly mimicked in all cases. However j-shapes were approximated for $j2$ and $j4$, which may account for the reduced β_1 bias for these two runs. We also note that there is a reversal in the direction of the bias for JM and CC for the u-shaped and n-shaped quadratics, possibly caused by over-adjusting due to poor fit in the tails of the missingness model.

Table B.2: Scores.quad: modelling quadratic shaped missingness with a quadratic logit

	% β_1 bias			absolute β_0 bias			% MSE increase ^a			% of missingness
	CC	JM	%red. ^b	CC	JM	red. ^c	CC	JM	diff. ^d	
uSteep	-11.8	9.6	2.2	0.05	-0.06	-0.01	2.4	2.0	0.4	19.6
u1	-1.5	3.5	-2.0	0.00	0.00	0.00	0.0	0.2	-0.1	4.5
u2	-2.8	3.4	-0.6	0.03	0.00	0.03	0.4	0.2	0.2	9.1
u3	-6.9	8.6	-1.6	0.02	-0.02	0.00	0.7	1.1	-0.3	14.3
u4	-16.4	9.8	6.6	0.04	-0.05	-0.01	4.4	2.0	2.4	22.9
u5	-9.4	11.9	-2.6	0.06	-0.03	0.03	2.1	2.2	-0.1	56.6
nSteep	27.4	-26.6	0.8	-0.11	0.09	0.01	13.9	12.5	1.4	80.8
n1	0.9	-3.3	-2.4	0.00	0.01	-0.01	0.0	0.2	-0.2	8.9
n2	17.7	-46.2	-28.5	-0.12	0.20	-0.08	7.5	36.7	-29.3	91.0
n3	25.0	-28.5	-3.5	-0.08	0.14	-0.05	9.7	14.8	-5.1	85.4
n4	19.5	-21.6	-2.2	-0.17	0.07	0.11	13.2	8.1	5.0	77.5
n5	8.3	-10.3	-2.0	0.01	0.07	-0.06	1.0	2.7	-1.7	43.1
jSteep	-9.2	-8.8	0.4	-0.05	-0.01	0.05	1.9	1.1	0.8	16.9
j1	-1.2	-3.0	-1.7	-0.01	-0.01	0.01	0.1	0.1	-0.1	4.4
j2	-5.2	0.3	4.9	-0.04	-0.01	0.03	0.7	0.0	0.7	7.5
j3	-8.6	-8.9	-0.3	-0.06	-0.04	0.02	1.8	1.3	0.5	8.7
j4	-8.9	-2.5	6.4	-0.07	-0.02	0.06	2.4	0.1	2.3	13.3
j5*	-	-	-	-	-	-	-	-	-	48.7

^a % increase in MSE from TRUE^b % reduction in absolute bias from CC to JM^c reduction in absolute bias from CC to JM^d difference between CC and JM columns

* run did not converge

The difficulties in approximating quadratic shaped missingness are illustrated by a set of three plots for *usteep*, *nsteep* and *jsteep* in Figure B.2, which are described below.

Model of Interest Plot: The regression lines for the model of interest produced by models JM (labelled *Missing Model*), TRUE (labelled *True Line*) and CC (labelled *Complete Cases*) are superimposed on a plot of the response against the covariate. Observed and missing individuals are represented by different symbols: a grey open circle for observed and a yellow cross for missing. This graph allows a visual assessment of the impact of the inclusion of the chosen model of missingness. An improvement over a simple complete case analysis is an indication of an appropriately chosen model of missingness.

Model of Missingness Plot: The true probability of missingness (labelled *True Probability*) and the fitted probability of missingness produced by the two models, JM (labelled *Missing Model*) and TRUE (labelled *Full Response*), are plotted against the response. TRUE indicates the best possible estimate of the true probability of missingness, conditional on the selected form of the missingness model and allowing for sampling variability. We are looking for the *Missing Model* line to be as close to the *Full Response* line as possible.

Link Plot: Similar to the model of missingness plot, but plots the logit link, i.e. transformed probability of missingness, against the response.

Figure B.2: Examples of modelling quadratic shaped missingness with a quadratic logit

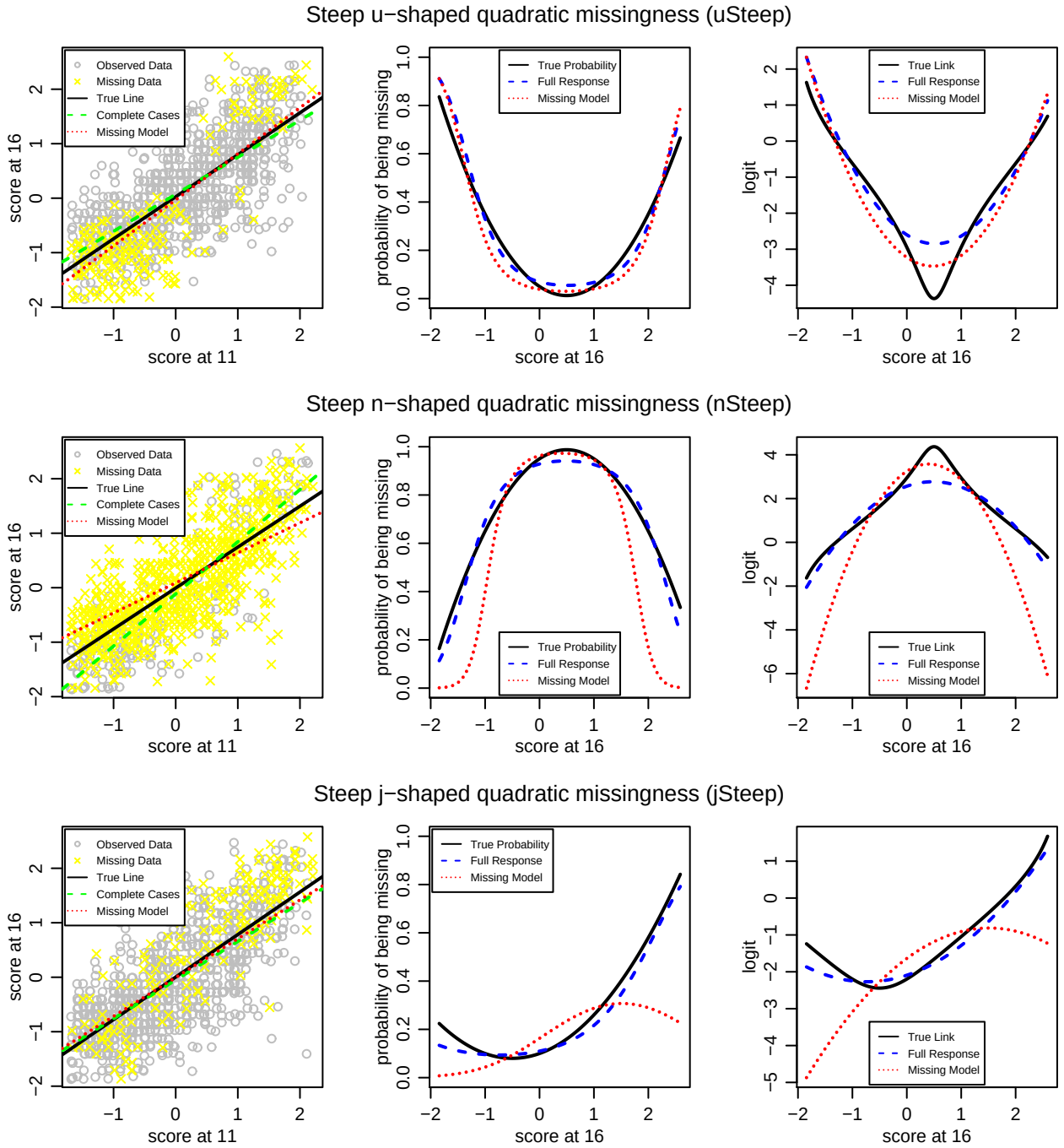
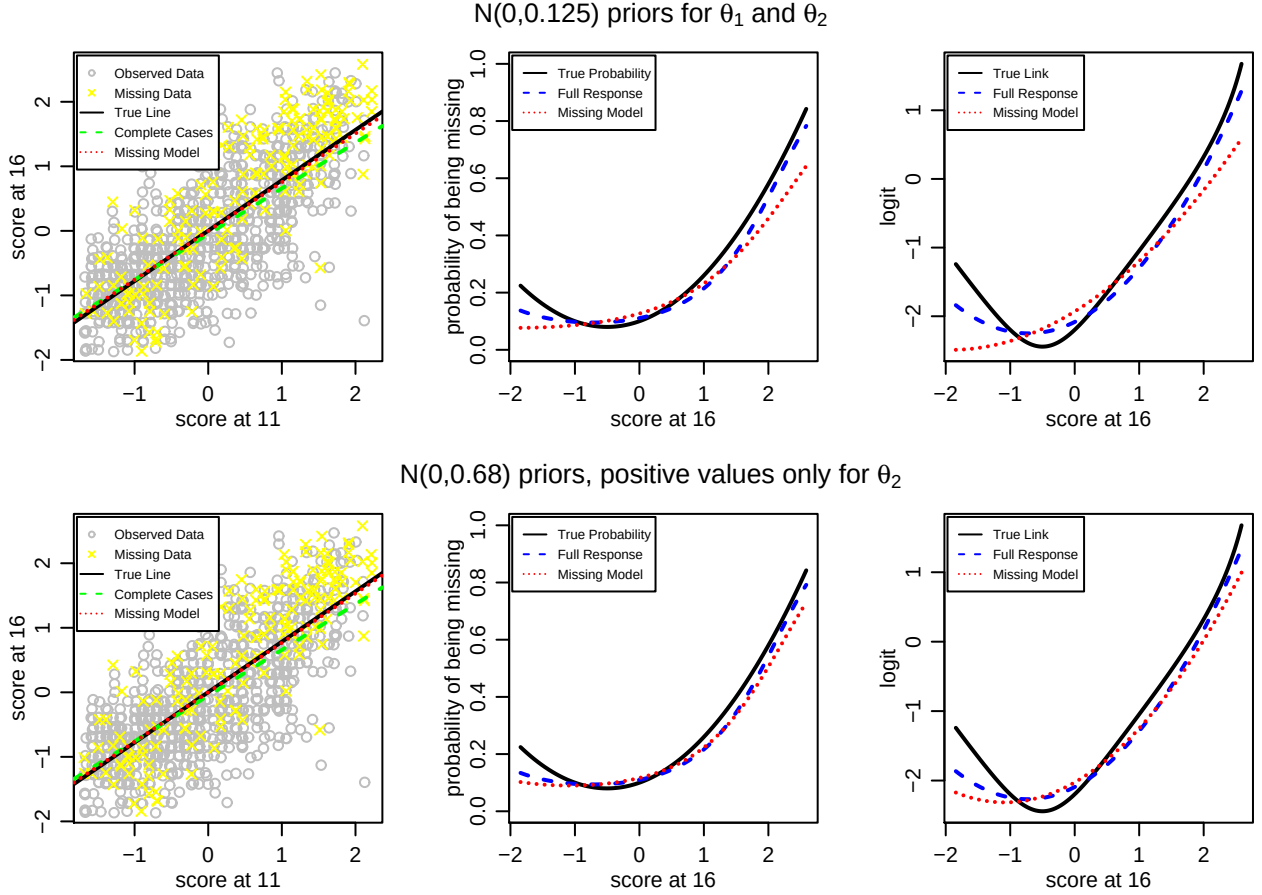


Figure B.2 shows that the quadratic logit approximates the u-shape best, struggles with the tails of the n-shape and fails to even approximate the j-shape. To see if we can improve the fit, we rerun the models using two sets of more informative priors for the model of missingness parameters: $\theta_1, \theta_2 \sim N(0, 0.125)$ and θ_2 restricted to positive (u-shapes and j-shapes only) or negative (n-shapes) values only. There are problems running the models with θ_2 restricted to negative values for the datasets with n-shaped missingness, with error messages suggesting that the models are not well defined. Otherwise, there are generally gains from using these priors. As an example, Figure B.3 shows that the j-shape is now

approximated for *jsteep* for both sets of priors. This demonstrates that there are potential modelling gains if we can justify the use of more informative priors.

Figure B.3: The effects of different model of missingness priors on *jsteep*



We also consider the performance of a piecewise linear logit model with one knot, such that

$$\begin{aligned} \text{logit}(p_i) &= \theta_0 + \theta_1 y A_i + \theta_2 y B_i \\ y A_i &= \min(y_i, \text{ChangePt}) \\ y B_i &= y_i - y A_i \end{aligned}$$

where *ChangePt* indicates the position of the knot. We use a logistic(0,1) prior for θ_0 , weakly informative $N(0,0.68)$ priors for θ_1 and θ_2 , and a uniform prior over the range of response values for *ChangePt*.

This model is fitted to the Scores.quad datasets, and the results are shown in Table B.3. All the u-shaped datasets converge and result in a reduction in the β_1 bias compared to CC. The n-shaped and j-shaped quadratics converge for the lower percentages of missingness, and either improve the β_1 bias or lead to a very small deterioration. There is a small improvement in fit according to the MSE for all datasets.

The fit is shown graphically for the *usteep*, *nsteep* and *jsteep* test datasets in Figure B.4. The performance of TRUE suggests that a piecewise linear logit model with one knot has the potential to mimic

Table B.3: Scores.quad: modelling quadratic shaped missingness with a piecewise linear logit

	% β_1 bias			absolute β_0 bias			% MSE increase ^a			% of missingness
	CC	JM	%red. ^b	CC	JM	red. ^c	CC	JM	diff. ^d	
uSteep	-11.8	4.3	7.6	0.05	-0.05	0.00	2.4	0.9	1.5	19.6
u1	-1.5	0.4	1.1	0.00	0.00	0.00	0.0	0.0	0.0	4.5
u2	-2.8	1.0	1.8	0.03	0.00	0.03	0.4	0.0	0.4	9.1
u3	-6.9	3.4	3.5	0.02	-0.02	0.00	0.7	0.3	0.5	14.3
u4	-16.4	4.1	12.2	0.04	-0.05	0.00	4.4	0.8	3.6	22.9
u5	-9.4	2.3	7.0	0.06	-0.02	0.04	2.1	0.2	2.0	56.6
n1	0.9	-1.4	-0.4	0.00	0.01	0.00	0.0	0.0	0.0	8.9
n5	8.3	-4.7	3.6	0.01	0.05	-0.04	1.0	1.0	0.0	43.1
j1	-1.2	-1.8	-0.5	-0.01	-0.01	0.01	0.1	0.1	0.0	4.4
j2	-5.2	-2.4	2.7	-0.04	-0.01	0.03	0.7	0.1	0.6	7.5
j3	-8.6	-7.5	1.1	-0.06	-0.04	0.03	1.8	1.0	0.8	8.7
j4	-8.9	-3.3	5.6	-0.07	-0.02	0.06	2.4	0.2	2.2	13.3

nsteep, *n2* – *n4*, *jsteep* and *j5* did not converge

^a % increase in MSE from TRUE

^b % reduction in absolute bias from CC to JM

^c reduction in absolute bias from CC to JM

^d difference between CC and JM columns

some quadratic shapes quite closely, but with four parameters to estimate for the missing model, convergence problems are encountered with JM. Of the three examples, only *usteep* converged, and even then the shape of the probability of being missing is not well approximated.

B.2 Step missingness

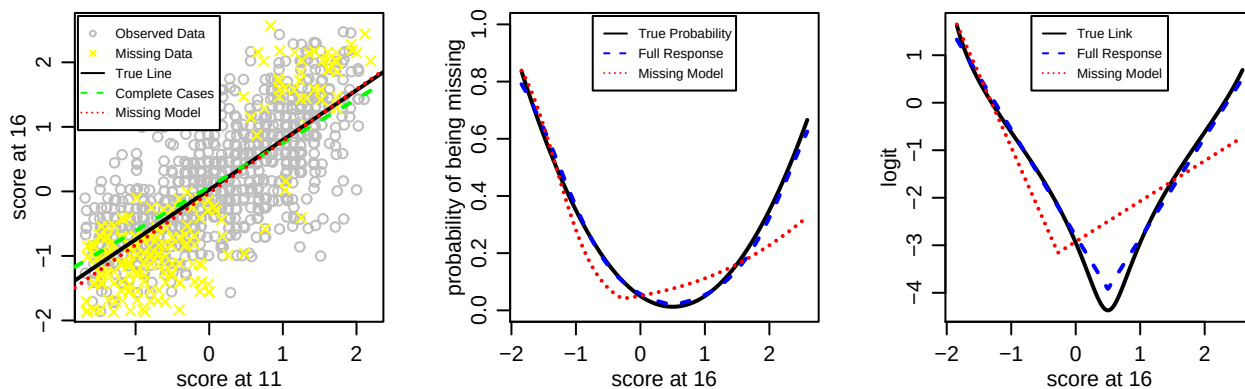
We also create six datasets with step shaped missingness using Equation 4.6, with the step at different positions (scenario Scores.step). Once again we run models TRUE, CC and JM on each dataset, using a linear logit model of missingness in TRUE and JM. In all cases where there is convergence, we find that the linear logit reduces the bias in the β parameters in the model of interest and lowers the MSE (Table B.4).

From the example in Figure B.5 we see that the step missingness is modelled quite accurately by the linear logit model. Using less informative priors for the model of missingness parameters, allows a closer approximation of the step shape to be obtained, but does not result in substantial further reductions in β_1 bias.

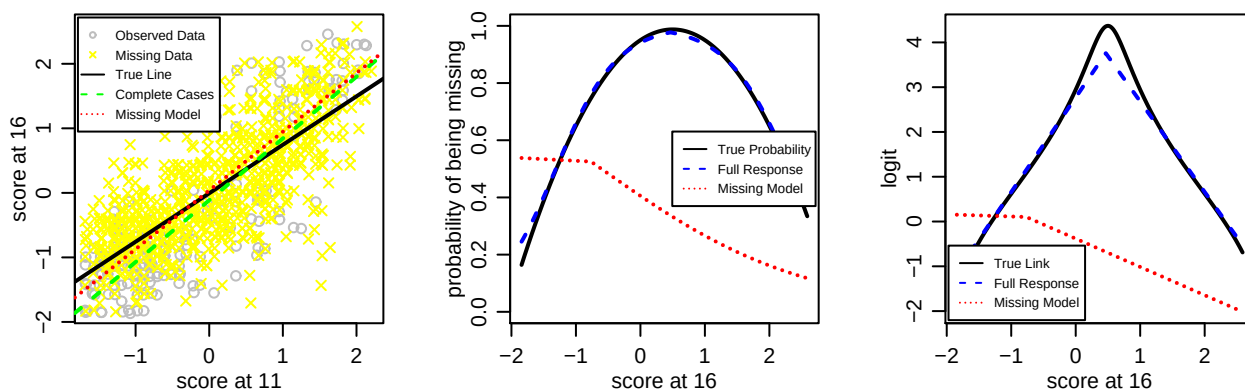
Further, we try modelling the Scores.step datasets using a step function for the missingness model. This is implemented using the WinBUGS *step* function, with a $N(0,100)$ prior on the *step* parameter. Sometimes the true position of the step is modelled exactly, and sometimes it failed to locate the step. Even when the step is positioned correctly in the missingness model, there is only limited improvement

Figure B.4: Examples of modelling quadratic shaped missingness with a piecewise linear logit

Steep u-shaped quadratic missingness (uSteep)



Steep n-shaped quadratic missingness (nSteep)



Steep j-shaped quadratic missingness (jSteep)

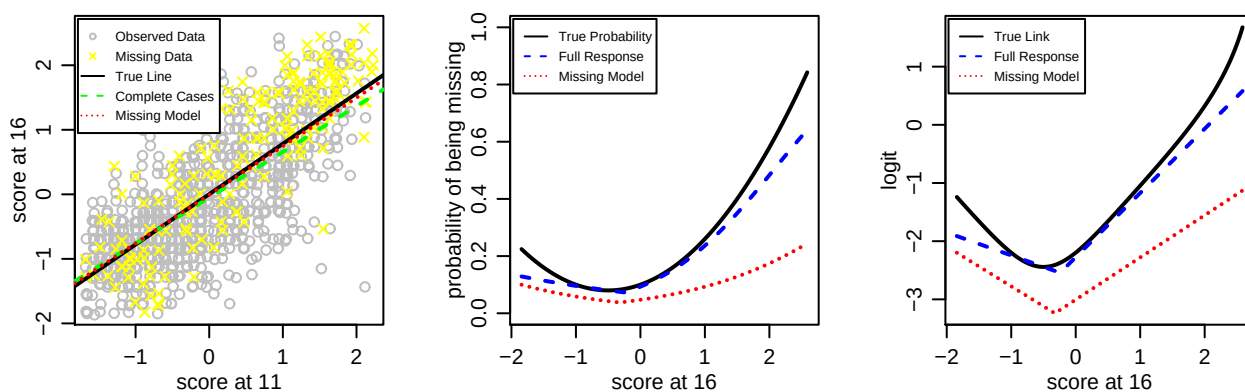
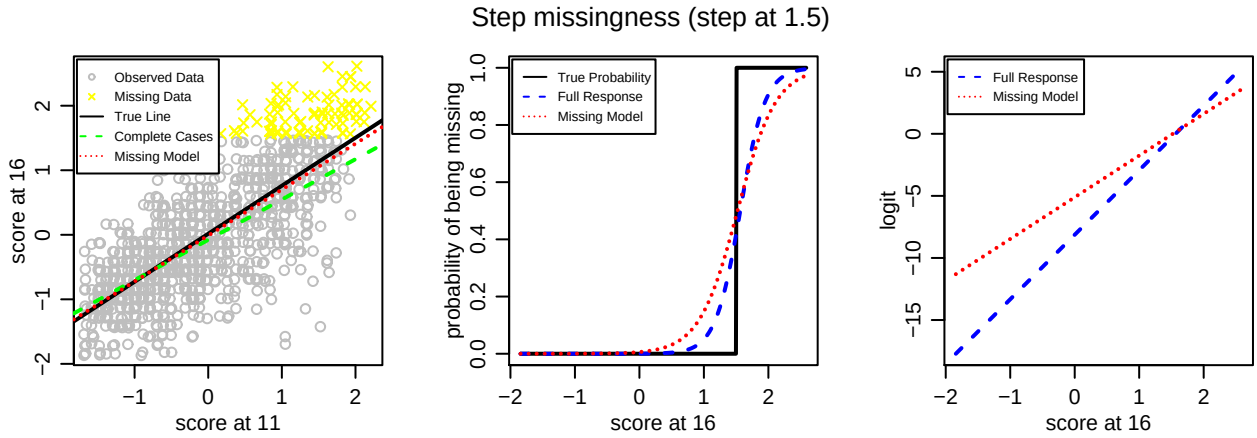


Table B.4: Scores.step: modelling step shaped missingness with a linear logit

step position	% β_1 bias			absolute β_0 bias			% MSE increase ^a			% of missingness
	CC	JM	%red. ^b	CC	JM	red. ^c	CC	JM	diff. ^d	
-1 [*]	-	-	-	-	-	-	-	-	-	86.5
-0.5	-83.1	-45.0	38.1	-0.83	-0.39	0.44	249.4	61.8	187.7	59.5
0	-66.5	-31.5	35.0	-0.60	-0.25	0.34	157.4	31.4	125.9	44.9
1	-36.4	-15.3	21.1	-0.26	-0.11	0.15	32.0	5.8	26.2	19.2
1.5	-16.0	-4.3	11.7	-0.09	-0.03	0.07	5.4	0.4	5.0	8.1
2	-7.6	-1.3	6.2	-0.04	-0.01	0.03	1.3	0.0	1.3	3.9

^a % increase in MSE from TRUE^b % reduction in absolute bias from CC to JM^c reduction in absolute bias from CC to JM^d difference between CC and JM columns^{*} run did not converge

Figure B.5: Example of modelling step shaped missingness with a linear logit



in the fit of the model of interest. The standard deviation on the step parameter is small when the step is well positioned and very large otherwise. Not surprisingly, when the step function is used to model other missingness shapes it often results in a deterioration in the fit of the model of interest, sometimes alarmingly so.

Appendix C

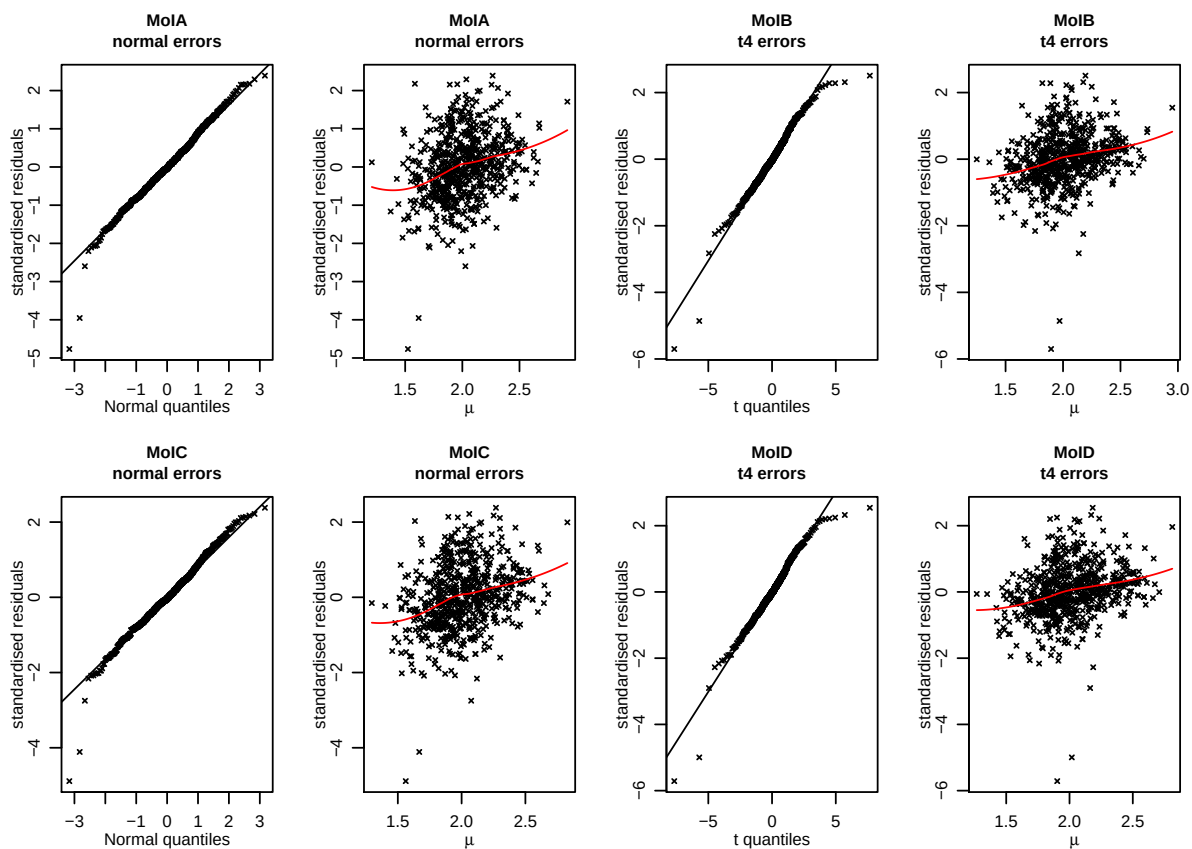
MCS income example

This appendix contains detailed additional information associated with the increasingly complex models which are fitted to our MCS income example in Chapters 5 to 8.

C.1 Models of interest fitted to complete cases

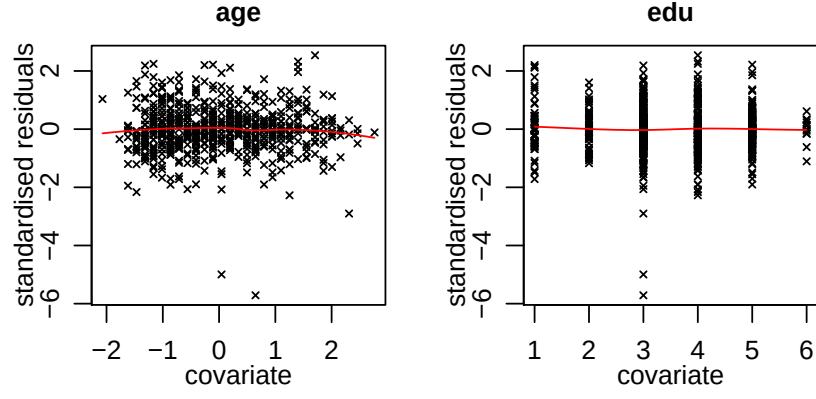
Figures C.1 and C.2 are examples of residual plots for the models of interest that we consider.

Figure C.1: MCS income example: residual plots for models of interest MoIA-MoID



The red lines are fitted splines

Figure C.2: MCS income example: residual plots v selected covariates for MoID



The red lines are fitted splines

Table C.1 shows the full set of parameter estimates for the five models of interest, MoIA-MoIE, described in Section 5.3.1.

Table C.1: MCS income example: parameter estimates for models of interest MoIA-MoIE

	MoIA		MoIB		MoIC		MoID		MoIE	
σ	0.32	(0.30,0.35)	0.33	(0.30,0.37)	0.32	(0.30,0.35)	0.33	(0.30,0.37)	0.33	(0.30,0.37)
ς	0.16	(0.10,0.21)	0.19	(0.15,0.23)	0.19	(0.15,0.23)	0.21	(0.17,0.25)	0.21	(0.17,0.25)
ρ_1	0.07	(0.00,0.18)	0.07	(0.00,0.19)						
ρ_2	0.13	(0.01,0.25)	0.07	(0.00,0.18)						
ρ_3	0.17	(0.01,0.38)	0.13	(0.01,0.34)						
ρ_4	0.24	(0.01,0.70)	0.24	(0.01,0.71)						
ρ_5	0.12	(0.01,0.26)	0.11	(0.01,0.26)						
ρ_6	0.17	(0.01,0.40)	0.17	(0.01,0.41)						
ρ_7	0.06	(0.00,0.17)	0.06	(0.00,0.18)						
ρ_8	0.14	(0.01,0.43)	0.16	(0.01,0.48)						
ρ_9	0.07	(0.00,0.20)	0.08	(0.00,0.22)						
γ_1	2.19	(1.98,2.40)	2.13	(1.93,2.33)	2.17	(1.98,2.37)	2.13	(1.94,2.33)	2.14	(1.95,2.33)
γ_2	2.13	(1.93,2.34)	2.09	(1.91,2.29)	2.11	(1.93,2.29)	2.10	(1.91,2.27)	2.10	(1.93,2.28)
γ_3	2.10	(1.84,2.36)	2.08	(1.84,2.33)	2.06	(1.82,2.29)	2.08	(1.85,2.30)	2.09	(1.86,2.31)
γ_4	2.23	(1.85,2.59)	2.15	(1.80,2.51)	2.22	(1.94,2.50)	2.17	(1.90,2.44)	2.18	(1.91,2.45)
γ_5	2.12	(1.90,2.34)	2.05	(1.84,2.26)	2.11	(1.91,2.30)	2.06	(1.86,2.26)	2.07	(1.87,2.27)
γ_6	2.11	(1.86,2.36)	2.05	(1.81,2.28)	2.10	(1.89,2.30)	2.05	(1.84,2.25)	2.06	(1.85,2.26)
γ_7	2.13	(1.90,2.35)	2.07	(1.85,2.28)	2.11	(1.90,2.32)	2.07	(1.86,2.28)	2.08	(1.88,2.29)
γ_8	2.14	(1.85,2.42)	2.05	(1.76,2.34)	2.14	(1.90,2.38)	2.06	(1.81,2.31)	2.06	(1.81,2.32)
γ_9	2.13	(1.90,2.35)	2.08	(1.87,2.29)	2.12	(1.92,2.32)	2.09	(1.89,2.29)	2.10	(1.90,2.30)
β_{age}	0.11	(0.07,0.15)	0.12	(0.08,0.16)	0.11	(0.07,0.15)	0.12	(0.08,0.16)	0.12	(0.09,0.16)
β_{hsize}	0.00	(-0.04,0.04)	0.00	(-0.04,0.03)	0.00	(-0.03,0.04)	0.00	(-0.04,0.03)		
β_{kids}	0.00	(-0.04,0.05)	0.00	(-0.04,0.05)	0.01	(-0.04,0.05)	0.01	(-0.04,0.05)		
$\beta_{edu[1]}$	-0.17	(-0.33,0.00)	-0.14	(-0.31,0.03)	-0.17	(-0.33,0.00)	-0.15	(-0.31,0.02)	-0.16	(-0.32,0.00)
$\beta_{edu[2]}$	0.04	(-0.09,0.17)	0.07	(-0.07,0.21)	0.03	(-0.10,0.16)	0.06	(-0.07,0.20)	0.05	(-0.08,0.18)
$\beta_{edu[3]}$	0.08	(-0.06,0.22)	0.10	(-0.05,0.25)	0.08	(-0.06,0.22)	0.09	(-0.05,0.24)	0.09	(-0.06,0.23)
$\beta_{edu[4]}$	0.23	(0.09,0.38)	0.26	(0.11,0.40)	0.24	(0.10,0.38)	0.25	(0.10,0.40)	0.24	(0.10,0.38)
$\beta_{edu[5]}$	0.42	(0.15,0.69)	0.47	(0.20,0.74)	0.42	(0.15,0.69)	0.45	(0.19,0.72)	0.44	(0.18,0.70)
β_{eth}	-0.09	(-0.25,0.07)	-0.09	(-0.24,0.06)	-0.05	(-0.19,0.11)	-0.07	(-0.21,0.08)	-0.07	(-0.21,0.08)
β_{reg}	-0.21	(-0.37,-0.05)	-0.18	(-0.32,-0.03)	-0.20	(-0.34,-0.06)	-0.18	(-0.31,-0.05)	-0.18	(-0.31,-0.05)
β_{sing}	-0.07	(-0.15,0.00)	-0.08	(-0.15,0.00)	-0.07	(-0.15,0.00)	-0.07	(-0.15,0.00)	-0.07	(-0.14,0.00)

Table shows the posterior mean, with the 95% interval in brackets

C.2 Effective sample size

Table C.2 shows the effective sample size (Section 2.4.2) for estimating the posterior means for a selection of MCS income models. The actual MCMC sample size for all the models shown is 20,000.

Table C.2: MCS income example: effective sample size for estimating posterior means

	MoIE	JM.R8	JM.C1	JMA	JMB	JMC	JMD
β_{age}	13058	6403	13039	12005	11477	12746	11719
$\beta_{edu[NVQ1]}$	1931	1966					
$\beta_{edu[NVQ2/2\&3]}^a$	1340	1332	3050	2785	2838	3966	2964
$\beta_{edu[NVQ3]}$	1565	1576					
$\beta_{edu[NVQ4/4\&5]}^a$	1372	1389	3706	3392	3427	4281	3158
$\beta_{edu[:NVQ5]}$	4091	4470					
β_{eth}	3559	2196	2616	2633	2362	3953	2815
β_{reg}	1018	884	1046	990	923	1449	1117
β_{sing}	17422	12787	16856	1452	1140	3162	5564
γ_1	730	699	1048	1052	974	1515	1179
γ_2	765	651	1083	1018	934	1520	1127
γ_3	1189	896	1627	1569	1353	2272	1654
γ_4	1365	1155	2045	1863	1901	2715	2105
γ_5	819	728	1164	1119	1073	1719	1262
γ_6	871	931	1296	1356	1334	1950	1431
γ_7	859	766	1183	1154	1067	1610	1214
γ_8	1318	1462	2160	2374	2082	3046	2444
γ_9	817	746	1158	1144	1067	1579	1252
σ	11172	4485	7137	2640	2735	502	909
ς	7213	3556	5147	3834	4304	348	790
θ_0		508		538	473	575	1705
$\theta_{level[1]}$		483		500	461	611	2093
$\theta_{level[2]}$						869	6933
$\theta_{level[3]}$						11521	15603
$\delta_{change[1]}$		495		510	481	542	1518
$\delta_{change[2]}$						1131	1031
$\theta_{ctry[Wales]}$		16792		16428	16011		
$\theta_{ctry[Scotland]}$		17496		16085	16825		
$\theta_{ctry[NI]}$		17133		16458	15378	17313	18947
θ_{eth}		5143		5204	5419	6258	14744
$\theta_{sc[NS-SEC1]}$						3128	7307
$\theta_{sc[NS-SEC2]}$		1335		2009	1878		
$\theta_{sc[NS-SEC4/4\&5]}^b$		1597		2196	1936	9456	11779
$\theta_{sc[NS-SEC5]}$		870		1114	943		
$edu.\eta_0$			1898	1818	160	167	170
$edu.\eta_{edu[NVQ2\&3]}$			2138	2171	1838	2320	1915
$edu.\eta_{edu[NVQ4\&5]}$			9	12	9	18	14
$edu.\eta_{sgap}$					163	189	179
$sing.\eta_0$			15866	4392	385	466	427
$sing.\eta_{sgap}$					403	445	425
$age.\eta_0$			20000	20000	20486	20000	20892
$age.\eta_{age}$			20000	19065	18416	19363	20000

^a first definition applies to models including MoIE (6 category *edu*); the second definition applies to models including MoIF (3 category *edu*)

^b first definition applies to models with 4 category *sc*; the second definition applies to models with 3 category *sc*.

C.3 Additional details on the JM.C1 covariate model of missingness

Section 6.2.1 discusses a covariate model of missingness for dichotomised *edu* and *sing*, concentrating on the imputations of the missing covariates which are of primary interest. Here, for completeness, we consider the parameter estimates for the covariate model of missingness for the five variants of JM.C1 described in Section 6.2.1, which are shown in Table C.3. The posterior means and intervals for the η_i parameters are very similar for JM.C1a-c which are fully identifiable models, but there are substantial differences between these three and JM.C1d. For JM.C1a, the diagonal elements of Σ are 1 as required, and the posterior mean of κ suggests very little correlation is being found between *edu*^{*} and *sing*^{*}. The JM.C1b and JM.C1c κ estimates are the same as the JM.C1a estimate, so we are satisfied that JM.C1a-c are effectively the same model, albeit with different specifications, after allowing for Monte Carlo error. The 95% intervals on the elements in the unrestricted covariance matrix for JM.C1d are large, reflecting the identifiability issues discussed in Section 6.1.1.

Table C.3: MCS income example: covariate model of missingness parameter estimates for five variants of JM.C1, which impute the missing values for dichotomised educational level and single/partner allowing for correlation between these two covariates

	JM.C1a		JM.C1b		JM.C1c		JM.C1d		JM.C1e	
<i>edu</i> . η_0	-2.73	(-3.67,-2.02)	-2.75	(-3.72,-2.02)	-2.71	(-3.69,-2.01)	-2.46	(-3.97,-1.14)	-2.65	(-3.53,-1.96)
<i>edu</i> . η_{edu}	8.00	(5.28,10.33)	8.13	(5.59,10.31)	8.12	(5.30,10.33)	7.70	(4.05,10.27)	8.54	(5.91,10.46)
<i>edu</i> . η_{age}	-0.80	(-1.58,-0.16)	-0.82	(-1.63,-0.17)	-0.80	(-1.61,-0.17)	-0.72	(-1.49,-0.13)	-0.74	(-1.49,-0.11)
<i>edu</i> . η_{eth}									-1.26	(-4.00,0.80)
<i>sing</i> . η_0	-0.44	(-0.58,-0.30)	-0.44	(-0.58,-0.30)	-0.44	(-0.58,-0.30)	-0.70	(-1.38,-0.22)	-0.48	(-0.63,-0.32)
<i>sing</i> . η_{age}	-0.15	(-0.30,-0.01)	-0.15	(-0.29,-0.01)	-0.15	(-0.29,-0.01)	-0.24	(-0.63,-0.01)	-0.17	(-0.32,-0.02)
<i>sing</i> . η_{eth}									0.25	(-0.18,0.67)
κ	0.04	(-0.39,0.45)	0.04	(-0.39,0.45)	0.04	(-0.40,0.46)			0.04	(-0.39,0.46)
Σ_{11}	1.00						0.91	(0.18,2.35)	1.00	
Σ_{12}	0.04	(-0.39,0.45)					0.06	(-0.82,0.98)	0.04	(-0.39,0.46)
Σ_{22}	1.00						3.00	(0.28,8.60)	1.00	

Table shows the posterior mean, with the 95% interval in brackets

Turning to the more complicated variant, JM.C1e, we find that the 95% intervals for the additional parameters contain 0. This supports our decision to use the simpler variants as a basis for extending JM.C1. We also find that there is very little difference in the parameter estimates for the model of interest between all five models (estimates not shown).

Appendix D

Deviance Information Criteria (DIC)

We illustrate the use of different formulations of the DIC using the Scores data in Section 4.6 and the antidepressant trial data in Section 5.2. Two of these formulations, DIC_O which is based on the observed data likelihood and DIC_F , based on the full data likelihood, cannot be computed using WinBUGS alone. Instead we implement algorithms based on suggestions by Daniels and Hogan (2008) (DH), using the R software with calls to WinBUGS to carry out MCMC runs where necessary. Details of these algorithms are given in this Appendix, along with some notes on their implementation. The equations underpinning these DIC can be found in Section 2.4.5. As usual $f(\mathbf{y}|\boldsymbol{\beta}, \sigma)$ is the model of interest, typically Normal or t in our applications, and $f(\mathbf{m}|\mathbf{y}, \boldsymbol{\theta})$ is a Bernoulli model of missingness in a selection model.

D.1 Algorithm for DIC based on the observed data likelihood (DIC_O)

Two versions of the algorithm for calculating DIC_O are implemented, the first uses reweighting as described by DH and the second avoids this reweighting by carrying out a “nested” MCMC sampler.

D.1.1 A version using the DH reweighting method

The algorithm for calculating DIC_O using reweighting proceeds as follows:

- 1 Carry out a standard MCMC run on the joint model $f(\mathbf{y}, \mathbf{m}|\boldsymbol{\beta}, \sigma, \boldsymbol{\theta})$. Save samples of $\boldsymbol{\beta}$, σ and $\boldsymbol{\theta}$, denoted by $\boldsymbol{\beta}^{(k)}$, $\sigma^{(k)}$ and $\boldsymbol{\theta}^{(k)}$, $k = 1, \dots, K$, which we shall call the ‘Ksample’.
- 2 Evaluate the posterior means of $\boldsymbol{\beta}$, σ and $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}$ and $\hat{\boldsymbol{\theta}}$. (Evaluate $\hat{\sigma}$ on the log scale and then back transform, see Section 4.6.2 for rationale.)
- 3 Generate a sample $\mathbf{y}_{mis}^{(q)}$, $q = 1, \dots, Q$, from the appropriate likelihood evaluated at the posterior means, e.g. $y_{mis} \sim N(\mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$. We shall call this the ‘Qsample’.

- 4 For each value of $(\beta^{(k)}, \sigma^{(k)})$ in the Ksample, and each value of $\mathbf{y}_{mis}^{(q)}$ from the Qsample, calculate the weight

$$w_q^{(k)} = \frac{f(\mathbf{y}_{mis}^{(q)} | \mathbf{y}_{obs}, \beta^{(k)}, \sigma^{(k)})}{f(\mathbf{y}_{mis}^{(q)} | \hat{\beta}, \hat{\sigma})}.$$

For some applications this calculation simplifies to

$$w_q^{(k)} = \frac{f(\mathbf{y}_{mis}^{(q)} | \beta^{(k)}, \sigma^{(k)})}{f(\mathbf{y}_{mis}^{(q)} | \hat{\beta}, \hat{\sigma})}.$$

Then evaluate

$$E_{\mathbf{y}_{mis} | \mathbf{y}_{obs}, \beta^{(k)}, \sigma^{(k)}} \{f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta}^{(k)})\} \approx \frac{\sum_{q=1}^Q w_q^{(k)} f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}^{(q)}, \boldsymbol{\theta}^{(k)})}{\sum_{q=1}^Q w_q^{(k)}}.$$

Denote this expectation by $h^{(k)}$.

- 5 Calculate the posterior expectation of the observed data log likelihood as

$$\log L(\beta, \sigma, \boldsymbol{\theta} | \mathbf{y}_{obs}, \mathbf{m}) \approx \frac{1}{K} \sum_{k=1}^K \left[\log L(\beta^{(k)}, \sigma^{(k)} | \mathbf{y}_{obs}) + \log h^{(k)} \right].$$

Multiply this by -2 to get the posterior mean of the deviance, denoted $\bar{D}$.

- 6 Evaluate the plug-in observed data log likelihood using the posterior means from the Ksample as

$$\log L(\hat{\beta}, \hat{\sigma} | \mathbf{y}_{obs}) + \log \left(E_{\mathbf{y}_{mis} | \mathbf{y}_{obs}, \hat{\beta}, \hat{\sigma}} \{f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \hat{\boldsymbol{\theta}})\} \right)$$

where

$$E_{\mathbf{y}_{mis} | \mathbf{y}_{obs}, \hat{\beta}, \hat{\sigma}} \{f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \hat{\boldsymbol{\theta}})\} \approx \frac{1}{Q} \sum_{q=1}^Q f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}^{(q)}, \hat{\boldsymbol{\theta}}).$$

The required plug-in likelihood can be calculated by using plug-ins on different scales. In a regression framework these can be the stochastic parameters used to calculate the linear predictor (which we shall call the $\boldsymbol{\theta}$ plug-ins) or the linear predictor itself ($\text{logit}(p_i)$ for our model of missingness, which we shall refer to as *logitp* plug-ins). We explore both options. Using the *logitp* plug-ins, the dependence on $\mathbf{y}$ in the expectation disappears, and

$$E_{\mathbf{y}_{mis} | \mathbf{y}_{obs}, \hat{\beta}, \hat{\sigma}} \{f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \hat{\boldsymbol{\theta}})\} \approx f(\mathbf{m} | \text{logitp}).$$

In this case, the plug-in log likelihood does not use the Qsample.

Multiply this plug-in log likelihood by -2 to get the plug-in deviance, denoted $\hat{D}$.

- 7 Finally, calculate $\text{DIC}_O = 2\bar{D} - \hat{D}$.

D.1.2 A version without reweighting (Gold Standard)

We compare the above algorithm, with a version which avoids reweighting and generates a Qsample for every member of the Ksample. We will refer to this as our *Gold Standard* DIC_O algorithm. The required changes are:

- 1 Unchanged.
- 2 Unchanged.
- 3 For each member of the Ksample, generate a sample $\mathbf{y}_{mis}^{(kq)}$, $q = 1, \dots, Q$, from the appropriate likelihood evaluated at $\boldsymbol{\beta}^{(k)}$ and $\sigma^{(k)}$, e.g. $y_{mis}^k \sim N(\mathbf{X}\boldsymbol{\beta}^{(k)}, \sigma^{(k)2})$. We denote the sample associated with member k of the Ksample as ‘Qsample^(k)’.
- 4 Then evaluate

$$h^{(k)} = E_{\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\beta}^{(k)}, \sigma^{(k)}} \{f(\mathbf{m}|\mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta}^{(k)})\} \approx \frac{1}{Q} \sum_{q=1}^Q f(\mathbf{m}|\mathbf{y}_{obs}, \mathbf{y}_{mis}^{(kq)}, \boldsymbol{\theta}^{(k)}).$$

- 5 Unchanged.
- 6 Generate a new Qsample, $\mathbf{y}_{mis}^{(q)}$, $q = 1, \dots, Q$, using $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ (as in step 3 of the reweighted algorithm). Evaluate the equations in step 6 of the reweighted algorithm using these $\mathbf{y}_{mis}^{(q)}$.
- 7 Unchanged.

D.2 Algorithm for DIC based on the full data likelihood (DIC_F)

The algorithm to evaluate DIC_F using the DH reweighting method is as follows:

- 1 Carry out a standard MCMC run on the joint model $f(\mathbf{y}, \mathbf{m}|\boldsymbol{\beta}, \sigma, \boldsymbol{\theta})$. Save samples of $\boldsymbol{\beta}$, σ , $\boldsymbol{\theta}$ and $\mathbf{y}_{mis}$, denoted by $\boldsymbol{\beta}^{(k)}$, $\sigma^{(k)}$, $\boldsymbol{\theta}^{(k)}$ and $\mathbf{y}_{mis}^{(k)}$, $k = 1, \dots, K$, which we shall call the ‘Ksample’.
- 2 Calculate the posterior mean of the full data log likelihood in the standard way as

$$\frac{1}{K} \sum_{k=1}^K \left[\log L(\boldsymbol{\beta}^{(k)}, \sigma^{(k)}|\mathbf{y}_{obs}, \mathbf{y}_{mis}^{(k)}) + \log L(\boldsymbol{\theta}^{(k)}, \mathbf{y}_{obs}, \mathbf{y}_{mis}^{(k)}|\mathbf{m}) \right].$$

Multiply this by -2 to get the posterior mean of the deviance, denoted $\bar{D}$.

- 3 Evaluate the posterior means of $\mathbf{y}_{mis}$, denoted by $\hat{\mathbf{y}}_{mis}$.
- 4 Carry out another MCMC run on the joint model, this time fixing the values of $\mathbf{y}_{mis}$ at $\hat{\mathbf{y}}_{mis}$. Save samples of $\boldsymbol{\beta}, \sigma$ and $\boldsymbol{\theta}$, denoted by $\boldsymbol{\beta}^{(q)}$, $\sigma^{(q)}$ and $\boldsymbol{\theta}^{(q)}$, $q = 1, \dots, Q$. We shall call this the ‘Qsample’.

- 5 For each value of $\mathbf{y}_{mis}^{(k)}$ in the original MCMC sample, and each value of $(\boldsymbol{\beta}^{(q)}, \sigma^{(q)}, \boldsymbol{\theta}^{(q)})$, calculate the weight

$$w_q^{(k)} = \frac{f(\mathbf{y}_{obs}, \mathbf{y}_{mis}^{(k)} | \boldsymbol{\beta}^{(q)}, \sigma^{(q)}) f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}^{(k)}, \boldsymbol{\theta}^{(q)})}{f(\mathbf{y}_{obs}, \hat{\mathbf{y}}_{mis} | \boldsymbol{\beta}^{(q)}, \sigma^{(q)}) f(\mathbf{m} | \mathbf{y}_{obs}, \hat{\mathbf{y}}_{mis}, \boldsymbol{\theta}^{(q)})}. \quad (\text{D.1})$$

Then evaluate the plug-in estimates for DIC_F as

$$\begin{aligned} \hat{\beta}(\mathbf{y}_{mis}^{(k)}) &= \frac{\sum_{q=1}^Q w_q^{(k)} \beta^{(q)}}{\sum_{q=1}^Q w_q^{(k)}} \\ \hat{\sigma}(\mathbf{y}_{mis}^{(k)}) &= \frac{\sum_{q=1}^Q w_q^{(k)} \sigma^{(q)}}{\sum_{q=1}^Q w_q^{(k)}} \\ \hat{\theta}(\mathbf{y}_{mis}^{(k)}) &= \frac{\sum_{q=1}^Q w_q^{(k)} \theta^{(q)}}{\sum_{q=1}^Q w_q^{(k)}}. \end{aligned}$$

Note that there are K plug-in estimates for each parameter, one corresponding to each member of the Ksample. To calculate $\hat{\sigma}(\mathbf{y}_{mis}^{(k)})$, work with $\log(\sigma)$ and back transform, as recommended by Spiegelhalter *et al.* (2002). As with DIC_O we explore two sets of plug-in parameters for the model of missingness, $\boldsymbol{\theta}$ plug-ins and *logitp* plug-ins.

- 6 Evaluate the plug-in log likelihood as

$$\frac{1}{K} \sum_{k=1}^K \left[\log L \left(\hat{\beta}(\mathbf{y}_{mis}^{(k)}), \hat{\sigma}(\mathbf{y}_{mis}^{(k)}) | \mathbf{y}_{obs}, \mathbf{y}_{mis}^{(k)} \right) + \log L \left(\hat{\theta}(\mathbf{y}_{mis}^{(k)}), \mathbf{y}_{obs}, \mathbf{y}_{mis}^{(k)} | \mathbf{m} \right) \right].$$

Multiply this plug-in log likelihood by -2 to get the plug-in deviance, denoted $\hat{D}$.

- 7 Finally, calculate $\text{DIC}_F = 2\bar{D} - \hat{D}$.

D.3 Implementation

We implement these algorithms for simulated and real examples, using the R software with calls to WinBUGS at step 1 of the DIC_O algorithm and steps 1 and 4 of the DIC_F algorithm, to carry out the MCMC runs. There are both speed and computation issues.

Working with likelihoods causes computation problems in calculating the weights for some examples, which can be avoided by working with log likelihoods and exponentiating later. The implementation for DIC_O is as follows:

$$w_q^{(k)} = \exp \left[\log \left(f(\mathbf{y}_{mis}^{(q)} | \boldsymbol{\beta}^{(k)}, \sigma^{(k)}) \right) - \log \left(f(\mathbf{y}_{mis}^{(q)} | \hat{\beta}, \hat{\sigma}) \right) \right].$$

The equation for DIC_F is similar, but we also need to add a constant to each weight before exponentiating to avoid underflow errors. This constant cancels when the weights are used to evaluate the plug-in estimates, and so has no impact on the calculated DIC_F . Similar tricks are required to successfully evaluate the expectation $h^{(k)}$ in step 4 of the DIC_O algorithm.

In our examples where the likelihood factorises into products over individuals, the performance of the code in terms of speed is improved by calculating any

$$\sum_{q=1}^Q f(\mathbf{m}|\mathbf{y}_{obs}, \mathbf{y}_{mis}^{(q)}, \boldsymbol{\theta}) = \sum_{q=1}^Q \left\{ \prod_{i=1}^{Nobs} f(m_i|y_i, \boldsymbol{\theta}) \prod_{j=1}^{Nmis} f(m_j|y_j^{(q)}, \boldsymbol{\theta}) \right\}$$

term as

$$\prod_{i=1}^{Nobs} f(m_i|y_i, \boldsymbol{\theta}) \sum_{q=1}^Q \left\{ \prod_{j=1}^{Nmis} f(m_j|y_j^{(q)}, \boldsymbol{\theta}) \right\} = f(\mathbf{m}_{obs}|\mathbf{y}_{obs}, \boldsymbol{\theta}) \sum_{q=1}^Q f(\mathbf{m}_{mis}|\mathbf{y}_{mis}^{(q)}, \boldsymbol{\theta}),$$

where $Nobs$ and $Nmis$ are the number of observed and missing responses respectively, and $\mathbf{m}_{obs}$ and $\mathbf{m}_{mis}$ are the missing value indicators for $\mathbf{y}_{obs}$ and $\mathbf{y}_{mis}$ respectively. This factorisation also enables us to avoid evaluating some of the contributions from $\mathbf{y}_{obs}$ in Equation D.1, which will cancel.

The Gold Standard DIC_O algorithm is an adaption of the reweighted DIC_O algorithm, using the methods outlined above for reducing computation and speed problems where necessary.

Appendix E

WinBUGS code

This appendix contains the WinBUGS code for running the joint model JMD and the model of interest MoIRJ.

JMD is selected as our main example because it is one of the most complex models, containing a model of interest, a covariate model of missingness, a response model of missingness, data from an additional data set and expert priors. The code required for running all the other models used for the MCS income example discussed in this thesis can be formed by making minor modifications to appropriate extracts from the JMD code. The only exceptions are the two models that include a reversible jump variable selection sub-model. So one of these, MoIRJ, is shown as a further example.

Note on JMD code:

Function *elicitor.piecewise*, written by Mary Kynn, can be downloaded from the WinBUGS development website (<http://www.winbugs-development.org.uk/>).

Note on MoIRJ code:

Functions *jump.lin.pred* and *jump.lin.pred.pred* are part of the WinBUGS Jump interface (Lunn *et al.*, 2005; Lunn *et al.*, 2006), and can also be downloaded from the WinBUGS development website.

E.1 Code for running joint model JMD

```
# WinBUGS code for running JMD
# Model of Interest: hpay logged, covariates {age, edu(3), eth(2), reg(2), sing(2)}, individual random effects,
# stratum specific intercepts and t4 errors.
# Covariate Model of Missingness: imputes correlated 3 category edu, binary sing and continuous age using BCS70 data
# Response Model of Missingness: MNAR - logit(pi) regressed on {change, level, ctry(2), eth(2), sc(2)},
# using piecewise linear functions for change and level, with informative priors from final elicitation.

model
{
  for (i in 1:N) { # N individuals

    # Model of Interest
    for (t in 1:2) { # 2 sweeps
      hpay[i,t]~dt(mu[i,t],tau,4)
      mu[i,t]<-beta0[i]+beta0.stratum[stratum[i]]+beta.eth[eth[i]]+
        beta.age*age[i,t]+beta.edu[edu[i,t]]+beta.sing[sing[i,t]]+beta.reg[reg[i,t]]
      e.hpay[i,t]<-exp(hpay[i,t]) # unlog hpay for response missingness model
      resid[i,t]<-(hpay[i,t]-mu[i,t])/sigma # calculate residuals
    }
    beta0[i]~dnorm(0,beta0.tau) # individual random effects

    # Response Model of Missingness - sweep 2 only
    paid[i]~dbern(p[i])
    logit(p[i]) <- theta0+theta.eth[eth[i]]+theta.ctry[ctry[i]]+theta.sc[sc[i]]+levelPW[i]+changePW[i]
    linkp[i]<- theta0+theta.eth[eth[i]]+theta.ctry[ctry[i]]+theta.sc[sc[i]]+levelPW[i]+changePW[i]
    levelPW[i] <- elicitor.piecewise(e.hpay[i,1],4,1,level.slope[],level.knot[])
    change[i]<-e.hpay[i,2]-e.hpay[i,1]
    changePW[i] <- elicitor.piecewise(change[i],3,1,change.slope[],change.knot[])

    # log likelihood contribution of individuals
    logL.RMM[i]<-(paid[i]*log(p[i]))+((1-paid[i])*log(1-p[i]))

    # Covariate Model of Missingness - MCS data, sweep 2 only
    edu.star[i]~dnorm(nu[i,1],edu.prec)I(left[i,1],right[i,1]) # latent variable for MCS edu
    sing.star[i]~dnorm(nu[i,2],sing.prec)I(left[i,2],right[i,2]) # latent variable for MCS sing
    age[i,2]~dnorm(nu[i,3],age.prec) # age does not require a latent variable as it is continuous
    nu[i,1]<-edu.eta0+edu.eta1[edu[i,1]]+(edu.eta2*MCS.sgap)
    nu[i,2]<-sing.eta0+(sing.eta2*MCS.sgap)+(kappa[1]*(edu.star[i]-nu[i,1])/varrho)
    nu[i,3]<-age.eta0+age.eta1*age[i,1]+(A1*(edu.star[i]-nu[i,1]))+(A2*(sing.star[i]-nu[i,2]))

  }

  # Covariate Model of Missingness - extra data from BCS70
  for (i in 1:Nbcs70) { # loop through BCS70 data - all observed
    edu.starB[i]~dnorm(nuB[i,1],edu.prec)I(leftB[i,1],rightB[i,1]) # latent variable for BCS70 edu
    sing.starB[i]~dnorm(nuB[i,2],sing.prec)I(leftB[i,2],rightB[i,2]) # latent variable for BCS70 sing
    nuB[i,1]<-edu.eta0+edu.eta1[BCSedu[i,1]]+(edu.eta2*BCS70.sgap)
    nuB[i,2]<-sing.eta0+(sing.eta2*BCS70.sgap)+(kappa[1]*(edu.starB[i]-nuB[i,1])/varrho)
    leftB[i,1]<-1*equals(BCSedu[i,2],3)-limit*equals(BCSedu[i,2],1) # -limit if edu=1, 0 if edu=2 and 1 if edu=3
    rightB[i,1]<-1*equals(BCSedu[i,2],2)+limit*equals(BCSedu[i,2],3) # 0 if edu=1, 1 if edu=2, limit if edu=3
    leftB[i,2]<- -limit*equals(BCSsing[i,2],1) # -limit if sing=1, 0 if sing=2
    rightB[i,2]<-limit*equals(BCSsing[i,2],2) # 0 if sing=1, limit if sing=2
  }

  limit<-10
  MCS.sgap<-3 # number of years between MCS sweeps
  BCS70.sgap<-4 # number of years between BCS70 sweeps
}
```

```

for (j in 1:Nobs.edu) { # loop through observed covariate edu
  # -limit if edu=1, 0 if edu=2 and 1 if edu=3
  left[edu.olst[j],1]<-1*equals(edu[edu.olst[j],2],3)-limit*equals(edu[edu.olst[j],2],1)
  # 0 if edu=1, 1 if edu=2, limit if edu=3
  right[edu.olst[j],1]<-1*equals(edu[edu.olst[j],2],2)+limit*equals(edu[edu.olst[j],2],3)
}

for (j in 1:Nmiss.edu) { # loop through missing covariate edu
  # set lower bound to prevent highest educational level going down between sweeps
  left[edu.mlst[j],1]<-1*equals(edu[edu.mlst[j],1],3)-limit*equals(edu[edu.mlst[j],1],1)
  right[edu.mlst[j],1]<-limit
  # impute missing sweep 2 edu
  edu[edu.mlst[j],2]<-step(edu.star[edu.mlst[j]])+step(edu.star[edu.mlst[j]]-1)+1
}

for (j in 1:Nobs.sing) { # loop through observed covariate sing
  left[sing.olst[j],2]<- -limit*equals(sing[sing.olst[j],2],1) # -limit if sing=1, 0 if sing=2
  right[sing.olst[j],2]<-limit*equals(sing[sing.olst[j],2],2) # 0 if sing=1, limit if sing=2
}

for (j in 1:Nmiss.sing) { # loop through missing covariate sing
  left[sing.mlst[j],2]<- -limit
  right[sing.mlst[j],2]<-limit
  # impute missing sweep 2 sing
  sing[sing.mlst[j],2]<-step(sing.star[sing.mlst[j]])+1
}

for (i in 1:Nri) { # loop through re-issues who provided data
  ri.hpay[i]<-exp(hpay[ri.lst[i],2]) # unlog hourly pay
  ri.pval[i]<-step(Tpay[i]-ri.hpay[i])
  ri.MSEcontrib[i]<-pow(Tpay[i]-ri.hpay[i],2)
  ril.hpay[i]<-hpay[ri.lst[i],2] # logged hourly pay
  ril.pval[i]<-step(log(Tpay[i])-ril.hpay[i])
  ril.MSEcontrib[i]<-pow(log(Tpay[i])-ril.hpay[i],2)
}
ri.MSE<-sum(ri.MSEcontrib[])/Nri
ril.MSE<-sum(ril.MSEcontrib[])/Nri

D.RMM<- -2*sum(logL.RMM[]) # calculate Bernoulli Deviance (response missingness model)

# Priors for model of interest

for (st in 1:9) { beta0.stratum[st]~dnorm(0,0.00000001) } # 9 stratum specific intercepts
beta.age~dnorm(0,0.00000001)
beta.edu[1]<-0 # alias first level of edu beta
for (e in 2:3) { beta.edu[e]~dnorm(0,0.00000001) } # 3 levels of education
beta.eth[1]<-0 # alias first level of eth (ethnicity) beta
beta.eth[2]~dnorm(0,0.00000001)
beta.sing[1]<-0 # alias first level of sing beta
beta.sing[2]~dnorm(0,0.00000001)
beta.reg[1]<-0 # alias first level of reg beta
beta.reg[2]~dnorm(0,0.00000001)

beta0.sigma~dnorm(0,0.00000001)I(0,)
beta0.tau<-1/(beta0.sigma*beta0.sigma)

tau~dgamma(0.001,0.001)
sigma<-sqrt(2 / tau) # t errors on 4 degrees of freedom

```

Priors for Response Model of Missingness: informative priors from final elicitation

theta0 ~ dnorm(3.0, 0.61)

level.knot[1] <- 0

level.slope[1] ~ dnorm (0.21, 95)

level.knot[2] <- 10.0

level.slope[2] ~ dnorm(-0.085, 600)

level.knot[3] <- 25.0

level.slope[3] ~ dnorm (-0.014, 2700)

level.knot[4] <- 1000

change.knot[1] <- -100

change.slope[1] ~ dnorm(0.16, 18.0)

change.knot[2] <- 0.0

change.slope[2] ~ dnorm (-0.16, 18.0)

change.knot[3] <- 100

theta.eth[1] <- 0

theta.eth[2] ~ dnorm(-1.3, 2.6)

theta.sc[1] ~ dnorm(-0.8, 1.2)

theta.sc[2] <- 0

theta.sc[3] ~ dnorm(-0.8, 1.2)

theta.ctr[1] <- 0

theta.ctr[2] ~ dnorm(-0.78, 1.3)

Priors for Covariate Model of Missingness

edu.eta0~dnorm(0,0.00000001)

edu.eta1[1]<-0 # alias first level

edu.eta1[2]~dnorm(0,0.00000001)

edu.eta1[3]~dnorm(0,0.00000001)

edu.eta2~dnorm(0,0.00000001)

sing.eta0~dnorm(0,0.00000001)

sing.eta2~dnorm(0,0.00000001)

age.eta0~dnorm(0,0.00000001)

age.eta1~dnorm(0,0.00000001)

kappa[1]~dunif(-1,1)

kappa[2]~dunif(-1,1)

kappa[3]~dunif(-1,1)

edu.prec~dgamma(0.001,0.001)

varrho<-1/edu.prec # edu variance

inv.varphi~dgamma(0.001,0.001)

varphi<-1/inv.varphi # joint model age variance

sing.prec<-1/(1-(pow(kappa[1],2)/varrho))

sing.sigma<-1/sqrt(sing.prec)

divisor<-varrho-pow(kappa[1],2)

age.prec<-1/(varphi-(pow(kappa[2],2)-(2*kappa[1]*kappa[2]*kappa[3])+varrho*pow(kappa[3],2))/divisor)

age.sigma<-1/sqrt(age.prec)

A1<-(kappa[2]-kappa[1]*kappa[3])/divisor

A2<-(kappa[3]-kappa[1]*kappa[2])/divisor

```
# Covariate contingency tables
```

```
for (i in 1:3) { # loop through sweep 1 categories
  for (j in 1:3) { # loop through sweep 2 categories
    for (k in 1:Nmiss.edu) { edu.count[k,i,j]<-equals(edu[edu.mlst[k],1],i)*equals(edu[edu.mlst[k],2],j) }
    edu.tab[i,j]<-sum(edu.count[,i,j]) # counts individuals with missing sweep 2 edu only
  }
}
```

```
for (i in 1:3) { # loop through sweep 1 categories
  for (j in 1:3) { # loop through sweep 2 categories
    edu.pctab[i,j]<-edu.tab[i,j]/sum(edu.tab[i,]) # percentages of individuals with missing sweep 2 edu only
  }
}
```

```
for (j in 1:2) { # loop through sweep 2 categories
  for (k in 1:Nmiss.sing) { sing.count[k,j]<-equals(sing[sing.mlst[k],2],j) }
  sing.tab[j]<-sum(sing.count[,j]) # counts individuals with missing sweep 2 sing only
}
```

```
for (i in 1:3) { # loop through edu categories
  for (j in 1:2) { # loop through sing categories
    for (k in 1:N) { edusing.count[k,i,j]<-equals(edu[k,2],i)*equals(sing[k,2],j) }
    edusing.tab[i,j]<-sum(edusing.count[,i,j]) # counts all individuals
  }
}
```

```
}
```

E.2 Code for running model of interest MoIRJ

```
# WinBUGS code for running MoIRJ
# Includes a reversible jump variable selection sub-model
# The binary components of the categorical covariates can be selected separately

model
{
  for (i in 1:N) { # N individuals
    for (t in 1:2) { # 2 sweeps
      # Model of Interest
      hpay[i,t]~dt(mu[i,t],tau,4)
      mu[i,t]<-beta0[i]+beta0.stratum[stratum[i]]+jump.mu[i,t]
      # Create full expanded covariate matrix, X
      X[i,t,1]<-age[i,t]
      X[i,t,2]<-sq.age[i,t]
      X[i,t,3]<-hsize[i,t]
      X[i,t,4]<-kids[i,t]
      # 5 binary variables for 6 category education
      X[i,t,5]<-edu1[i,t]
      X[i,t,6]<-edu2[i,t]
      X[i,t,7]<-edu3[i,t]
      X[i,t,8]<-edu4[i,t]
      X[i,t,9]<-edu5[i,t]
      # 1 binary variable for 2 category ethnicity
      X[i,t,10]<-eth2[i,t]
      # 1 binary variable for 2 category single/partner
      X[i,t,11]<-sing2[i,t]
      # 1 binary variable for 2 category region
      X[i,t,12]<-reg2[i,t]
      # 5 variables for age.education interaction
      X[i,t,13]<-age.edu1[i,t]
      X[i,t,14]<-age.edu2[i,t]
      X[i,t,15]<-age.edu3[i,t]
      X[i,t,16]<-age.edu4[i,t]
      X[i,t,17]<-age.edu5[i,t]
    }
    beta0[i]~dnorm(0,beta0.tau) # individual random effects
  }

  jump.mu[1:N,1:2]<-jump.lin.pred(X[1:N,1:2,1:Q],k,beta.prec)
  id[1]<-jump.model.id(jump.mu[1:N,1:2])
  k~dbin(0.5,Q) # Q is set to 17
  beta.prec~dgamma(0.001,0.001)

  for (st in 1:9) { beta0.stratum[st]~dnorm(0,0.00000001) } # 9 stratum specific intercepts
  beta0.sigma~dnorm(0,0.00000001)I(0,)
  beta0.tau<-1/(beta0.sigma*beta0.sigma)
  tau~dgamma(0.001,0.001)
  sigma<-sqrt(2 / tau ) # t errors on 4 degrees of freedom

  pred[1:(Q + 1)] <- jump.lin.pred.pred(jump.mu[1:N,1:2], X.pred[1:(Q + 1), 1:Q])
  for (i in 1:Q) {
    # create an identity matrix
    X.pred[i, i] <- 1
    for (j in 1:(i - 1)) { X.pred[i, j] <- 0 }
    for (j in (i + 1):Q) { X.pred[i, j] <- 0 }
    X.pred[(Q + 1), i] <- 0 # Q+1 row, a vector of zeros, will enable calculation of constant
    effect[i] <- pred[i] - pred[Q + 1]
    included[i]<-step(abs(effect[i])-0.000000000000000000000001)
  }
}
```

Bibliography

- Abayomi, K., Gelman, A., and Levy, M. (2008). Diagnostics for Multivariate Imputations. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **57**, (3), 273–91.
- Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, **88**, (422), 669–79.
- Allison, P. D. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological Methods and Research*, **28**, 301–9.
- Barnard, J. and Meng, X.-L. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research*, **8**, 17–36.
- Bedrick, E. J., Christensen, R., and Johnson, W. (1996). A New Perspective on Priors for Generalized Linear Models. *Journal of the American Statistical Association*, **91**, (436), 1450–60.
- Behr, A., Bellgardt, E., and Rendtel, U. (2005). Extent and Determinants of Panel Attrition in the European Community Household Panel. *European Sociological Review*, **21**, (5), 489–512.
- Best, N. G., Spiegelhalter, D. J., Thomas, A., and Brayne, C. E. G. (1996). Bayesian Analysis of Realistically Complex Models. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **159**, (2), 323–42.
- Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **26**, (2), 211–52.
- Brooks, S. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–55.
- Brooks, S. P. and Roberts, G. O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, **8**, 319–35.
- Browne, W. J. (2003). *MCMC estimation in MLwiN Version 2.0 User Manual*. Centre for Multilevel Modelling, Institute of Education, University of London, Available from <http://seis.bris.ac.uk/~frwjb/materials/mcmcman2.pdf>.
- Browne, W. J. and Draper, D. (2000). Implementation and Performance Issues in the Bayesian and Likelihood Fitting of Multilevel Models. *Computational Statistics*, **15**, (3), 391–420.
- Bunch, D. S. (1991). Estimability in the multinomial probit model. *Transportation Research*, **25B**, (1), 1–12.

- Calderwood, L., Ketende, S., and MacDonald, J. (2008). Patterns of longitudinal participation in the Millennium Cohort Study. Prepared for the Panel Surveys Workshop in Essex July 2008 and posted on www.iser.essex.ac.uk.
- Carpenter, J. and Kenward, M. (2005). Missing Data. www.missingdata.org.uk. accessed on 15 March 2006.
- Carpenter, J., Pocock, S., and Lamm, C. J. (2002). Coping with missing data in clinical trials: A model-based approach applied to asthma trials. *Statistics in Medicine*, **21**, 1043–66.
- Carpenter, J. R. and Goldstein, H. (2004). Multiple imputation in MLwiN. *Multilevel modelling newsletter*, **16**, (2).
- Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **169**, (3), 571–84.
- Carpenter, J. R., Kenward, M. G., and White, I. R. (2007). Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research*, **16**, (3), 259–75.
- Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). Deviance Information Criteria for Missing Data Models. *Bayesian Analysis*, **1**, (4), 651–74.
- Chaloner, K., Church, T., Louis, T. A., and Matts, J. P. (1993). Graphical Elicitation of a Prior Distribution for a Clinical Trial. *The Statistician*, **42**, (4), 341–53.
- Chen, M.-H. and Dey, D. K. (2000). *Generalised Linear Models: A Bayesian Perspective*, chapter 8: Bayesian Analysis for Correlated Ordinal Data Models, pp. 133–55. Marcel Dekker Inc.
- Chib, S. (2000). *Generalised Linear Models: A Bayesian Perspective*, chapter 7: Bayesian Methods for Correlated Binary Data, pp. 113–31. Marcel Dekker Inc.
- Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, **85**, (2), 347–61.
- Clayton, D., Spiegelhalter, D., Dunn, G., and Pickles, A. (1998). Analysis of Longitudinal Binary Data from Multi-Phase Sampling (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **60**, (1), 71–87.
- Collett, D. (2003). *Modelling Binary Data*, (2nd edn). Chapman & Hall.
- Collins, L. M., Schafer, J. L., and Kam, C.-M. (2001). A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods*, **6**, (4), 330–51.
- Congdon, P. (2001). *Bayesian Statistical Modelling*. John Wiley and Sons.
- Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, **91**, (434), 883–904.

- Daniels, M. J. and Hogan, J. W. (2008). *Missing Data In Longitudinal Studies Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman & Hall.
- Dansie, B. R. (1985). Parameter estimability in the multinomial probit model. *Transportation Research*, **19B**, (6), 526–8.
- Demirtas, H. and Schafer, J. L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, **22**, 2553–75.
- Dempster, A. P. (1973). The direct use of likelihood for significance testing. In *Proceedings of Conference on Foundational Questions in Statistical Inference*, (ed. O. Barndorff-Nielsen, P. Blaesild, and G. Schou), pp. 335–54. University of Aarhus.
- Dempster, A. P. (1997). The direct use of likelihood for significance testing. *Statistics and Computing*, **7**, 247–52.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **39**, (1), 1–38.
- Diggle, P., Farewell, D., and Henderson, R. (2007). Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal (with discussion). *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **56**, (5), 499–550.
- Diggle, P. and Kenward, M. G. (1994). Informative Drop-out in Longitudinal Data Analysis (with discussion). *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **43**, (1), 49–93.
- Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, (2nd edn). Oxford University Press.
- Draper, D. (1995). Assessment and Propagation of Model Uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **57**, (1), 45–97.
- Dunn, M. C., Kadane, J. B., and Garrow, J. R. (2003). Comparing Harm Done by Mobility and Class Absence: Missing Students and Missing Data. *Journal of Educational and Behavioral Statistics*, **28**, (3), 269–88.
- Dunson, D. B. (2000). Bayesian Latent Variable Models for Clustered Mixed Outcomes. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **62**, (2), 355–66.
- Durrant, G. B. and Skinner, C. (2006). Using data augmentation to correct for non-ignorable non-response when surrogate data are available: an application to the distribution of hourly pay. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **169**, (3), 605–23.
- Everitt, B. S. (ed.) (1999). *Statistical Methods in Medical Research*, **8**, (1). Hodder Arnold.
- Faris, P. D., Ghali, W. A., Brant, R., Norris, C. M., Galbraith, P. D., and Knudston, M. L. (2002). Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *Journal of Clinical Epidemiology*, **54**, 184–91.

- Feinstein, L. (2003). Inequality in the Early Cognitive Development of British Children in the 1970 Cohort. *Economica*, **70**, 73–97.
- Fitzmaurice, G. M. (2003). Methods for Handling Dropouts in Longitudinal Clinical Trials. *Statistica Neerlandica*, **57**, (1), 75–99.
- Fogelman, K. R. and Goldstein, H. (1976). Social Factors associated with changes in Educational Attainment between 7 and 11 Years of Age. *Educational Studies*, **2**, (2), 95–109.
- Fogelman, K. R., Goldstein, H., Essen, J., and Ghodsian, M. (1978). Patterns of Attainment. *Educational Studies*, **4**, (2), 121–30.
- Forster, J. J. and Smith, P. W. F. (1998). Model-Based Inference for Categorical Survey Data Subject to Non-Ignorable Non-Response (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **60**, (1), 57–70.
- Garthwaite, P. H. and Al-Awadhi, S. A. (2006). Quantifying Opinion about Logistic Regression using Interactive Graphics. Technical report, Department of Statistics, Open University.
- Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models. *Journal of the American Statistical Association*, **94**, (445), 247–53.
- Gelman, A. (2004). Exploratory Data Analysis for Complex Models. *Journal of Computational and Graphical Statistics*, **13**, (4), 755–79.
- Gelman, A. (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, (2), 1–19.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*, (2nd edn). Chapman & Hall.
- Gelman, A. and Raghunathan, T. E. (2001). Comment on “Conditionally Specified Distributions: An Introduction”. *Statistical Science*, **16**, (3), 268–9.
- Gelman, A., van Mechelen, I., Verbeke, G., Heitjan, D. F., and Meulders, M. (2005). Multiple Imputation for Model Checking: Completed-Data Plots with Missing and Latent Data. *Biometrics*, **61**, 74–85.
- Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*, (1st edn). Chapman & Hall.
- Goldstein, H. (1979). Some Models for Analysing Longitudinal Data on Educational Attainment (with discussion). *Journal of the Royal Statistical Society, Series A (General)*, **142**, (4), 407–42.
- Goldstein, H., Carpenter, J., Kenward, M. G., and Levin, K. A. (2008). Multilevel models with multivariate mixed response types. *Statistical Modelling*. To appear.
- Greenhouse, J. B. and Seltman, H. (2005). Using prior distributions to synthesize historical evidence: comments on the Goodman-Sladky case study of IVIg in Guillain-Barre syndrome. *Clinical Trials*, **2**, 311–8.

- Greenland, S. (1996). Basic Methods for Sensitivity Analysis of Biases. *International Journal of Epidemiology*, **25**, (6), 1107–16.
- Greenland, S. (2005). Multiple-bias modelling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **168**, (2), 267–306.
- Hawkes, D. and Plewis, I. (2006). Modelling non-response in the National Child Development Study. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **169**, (3), 479–91.
- Hawkes, D. and Plewis, I. (2008). Missing Income Data in the Millenium Cohort Study: Evidence from the First Two Sweeps. CLS cohort studies, working paper 2008/10, Institute of Education, University of London.
- Hogan, J. W. and Laird, N. M. (1997). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, **16**, 259–72.
- Horton, N. J. and Kleinman, K. P. (2007). Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. *The American Statistician*, **61**, (1), 79–90.
- Horton, N. J. and Laird, N. M. (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research*, **8**, (1), 37–50.
- Horton, N. J. and Lipsitz, S. R. (2001). Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables. *The American Statistician*, **55**, (3), 244–54.
- Huang, L., Chen, M.-H., and Ibrahim, J. G. (2005). Bayesian Analysis for Generalized Linear Models with Nonignorably Missing Covariates. *Biometrics*, **61**, 767–80.
- Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-Data Methods for Generalized Linear Models: A Comparative Review. *Journal of the American Statistical Association*, **100**, (469), 332–46.
- Jackson, C., Best, N., and Richardson, S. (2006). Improving ecological inference using individual-level data. *Statistics in Medicine*, **25**, 2136–59.
- Jackson, C., Best, N., and Richardson, S. (2008a). Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **171**, (1), 159–78.
- Jackson, C., Richardson, S., and Best, N. (2008b). Studying place effects on health by synthesising individual and area-level outcomes. *Social Science and Medicine*, **67**, 1995–2006.
- Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G., and Kenward, M. G. (2006). The nature of sensitivity in montone missing not at random models. *Computational Statistics and Data Analysis*, **50**, 830–58.
- Jones, A. M., Koolman, X., and Rice, N. (2006). Health-related non-response in the British Household

- Panel Survey and European Community Household Panel: using inverse-probability-weighted estimators in non-linear models. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **169**, (3), 543–69.
- Kadane, J. B. (1993). Subjective Bayesian Analysis for Surveys with Missing Data. *The Statistician*, **42**, (4), 415–26.
- Kadane, J. B. and Wolfson, L. J. (1998). Experiences in Elicitation. *The Statistician*, **47**, (1), 3–19.
- Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). Markov Chain Monte Carlo in Practice: A Roundtable Discussion. *The American Statistician*, **52**, (2), 93–100.
- Kass, R. E. and Natarajan, R. (2006). A Default Conjugate Prior for Variance Components in Generalized Linear Mixed Models (Comment on Article by Browne and Draper). *Bayesian Analysis*, **1**, (3), 535–42.
- Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in Medicine*, **17**, 2723–32.
- Kenward, M. G. and Carpenter, J. (2007). Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, **16**, (3), 199–218.
- Kenward, M. G. and Molenberghs, G. (1999). Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research*, **8**, (1), 51–83.
- Kerman, J., Gelman, A., Zheng, T., and Ding, Y. (2007). *Handbook of Data Visualization*, chapter 16: Visualization in Bayesian Data Analysis, pp. 709–24. Springer.
- Ketende, S. (2008). Millennium Cohort Study: Technical Report on Response. Technical report, 2nd edition, Institute of Education, University of London.
- King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review*, **95**, (1), 49–69.
- Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential Imputations and Bayesian Missing Data Problems. *Journal of the American Statistical Association*, **89**, (425), 278–88.
- Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Society*, **88**, (421), 125–34.
- Little, R. J. A. (1995). Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Society*, **90**, (431), 1112–21.
- Little, R. J. A. and Rubin, D. B. (1983). On Jointly Estimating Parameters and Missing Data by Maximizing the Complete-Data Likelihood. *The American Statistician*, **37**, (3), 218–20.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, (2nd edn). John Wiley and Sons.

- Liu, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, **6**, 113–9.
- Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*, (1st edn). Springer-Verlag.
- Longford, N. T., Tyrer, P., Nur, U. A. M., and Seivewright, H. (2006). Analysis of a long-term study of neurotic disorder, with insights into the process of non-response. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **169**, (3), 507–23.
- Lunn, D. J., Best, N., and Whittaker, J. C. (2005). Generic reversible jump MCMC using graphical models. Technical report EPH-2005-01, Imperial College London.
- Lunn, D. J., Whittaker, J. C., and Best, N. (2006). A Bayesian Toolkit for Genetic Association Studies. *Genetic Epidemiology*, **30**, 231–47.
- Lynn, P. (ed.) (2006). *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **169**, (3). Blackwell Publishing Limited.
- Ma, G., Troxel, A. B., and Heitjan, D. F. (2005). An index of local sensitivity to nonignorable drop-out in longitudinal modelling. *Statistics in Medicine*, **24**, 2129–50.
- Marshall, E. C. and Spiegelhalter, D. J. (2003). Approximate cross-validators predictive checks in disease mapping models. *Statistics in Medicine*, **22**, 1649–60.
- McCulloch, R. and Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, **64**, 207–40.
- McCulloch, R. E., Polson, N. G., and Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, **99**, 173–93.
- Michiels, B., Molenberghs, G., Bijmens, L., Vangeneugden, T., and Thijs, H. (2002). Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Statistics in Medicine*, **21**, 1023–41.
- Molenberghs, G., Beunckens, C., Sotito, C., and Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **70**, (2), 371–88.
- Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*, (1st edn). John Wiley and Sons.
- Molenberghs, G., Kenward, M. G., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, **84**, (1), 33–44.
- Molenberghs, G., Michiels, B., Kenward, M. G., and Diggle, P. J. (1998). Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, **52**, (2), 153–61.
- Molenberghs, G., Thijs, H., Kenward, M. G., and Verbeke, G. (2003). Sensitivity Analysis of Continuous Incomplete Longitudinal Outcomes. *Statistica Neerlandica*, **57**, (1), 112–35.
- Molitor, N.-T., Best, N., Jackson, C., and Richardson, S. (2009). Using Bayesian graphical models to

- model biases in observational studies and to combine multiple data sources: Application to low birth-weight and water disinfection by-products. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*. To appear.
- Nicoletti, C. and Peracchi, F. (2005). Survey response and survey characteristics: microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **168**, (4), 763–81.
- O’Hagan, A. (2003). *Highly Structured Stochastic Systems*, chapter 14, pp. 423–53. Oxford University Press.
- O’Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts’ Probabilities*, (1st edn). John Wiley and Sons.
- Oudshoorn, K., van Buuren, S., and van Rijkevorsel, J. (1999). Flexible multiple imputation by chained equations of the avo-95 survey. Technical report PG/VGZ/99.045, TNO Prevention and Health.
- Paddock, S. M. (2007). Bayesian variable selection for longitudinal substance abuse treatment data subject to informative censoring. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **56**, (3), 293–311.
- Peruggia, M. (1997). On the Variability of Case-Deletion Importance Sampling Weights in the Bayesian Linear Model. *Journal of the American Statistical Association*, **437**, (92), 199–207.
- Pettitt, A. N., Tran, T. T., Haynes, M. A., and Hay, J. L. (2006). A Bayesian hierarchical model for categorical longitudinal data from a social survey of immigrants. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **169**, (1), 97–114.
- Plewis, I. (2007a). The Millennium Cohort Study: Technical Report on Sampling. Technical report, 4th edition, Institute of Education, University of London.
- Plewis, I. (2007b). Non-Response in a Birth Cohort Study: The Case of the Millennium Cohort Study. *International Journal of Social Research Methodology*, **10**, (5), 325–34.
- Plewis, I., Calderwood, L., Hawkes, D., and Nathan, G. (2004). National Child Development Study and 1970 British Cohort Study Technical Report: Changes in the NCDS and BCS70 Populations and Samples over Time. Technical report, 1st edition, Institute of Education, University of London.
- Plewis, I., Ketende, S. C., Joshi, H., and Hughes, G. (2008). The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First Two Waves of the UK Millennium Cohort Study. *Journal of Official Statistics*, **24**, (3), 1–22.
- Plummer, M., Best, N., Cowles, K., and Vines, K. (2007). *coda: Output analysis and diagnostics for MCMC*. R package version 0.13-1.
- R Development Core Team (2007). *R: A Language and Environment for Statistical Comput-*

ing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

- Raghunathan, T. E. and Siscovick, D. S. (1998). Combining exposure information from various sources in an analysis of a case-control study. *The Statistician*, **47**, (2), 333–47.
- Raghunathan, T. E., Xie, D., Schenker, N., Parsons, V. L., Davis, W. W., Dodd, K. W., and Feuer, E. J. (2007). Combining Information from Two Surveys to Estimate County-Level Prevalence Rates of Cancer Risk Factors and Screening. *Journal of the American Statistical Association*, **102**, (478), 474–86.
- Rassler, S. (2003). A Non-Iterative Bayesian Approach to Statistical Matching. *Statistica Neerlandica*, **57**, (1), 58–74.
- Rotnitzky, A. and Robins, J. (1997). Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in Medicine*, **16**, 81–102.
- Rousseeuw, P. J. and Molenberghs, G. (1994). The Shape of Correlation Matrices. *The American Statistician*, **48**, (4), 276–9.
- Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, **63**, (3), 581–92.
- Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Society*, **91**, (434), 473–89.
- Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, **57**, (1), 3–18.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, (1st edn). Chapman & Hall.
- Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, **8**, (1), 3–15.
- Schafer, J. L. (2003). Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ. *Statistica Neerlandica*, **57**, (1), 19–35.
- Schafer, J. L. and Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, **7**, (2), 147–77.
- Scharfstein, D. O., Daniels, M. J., and Robins, J. M. (2003). Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes. *Biostatistics*, **4**, (4), 495–512.
- Schenker, N. and Raghunathan, T. E. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine*, **26**, 1802–11.
- Shepherd, P. (2002). NCDS/BCS70 1999-2000 Follow-ups: Guide to the Combined Dataset. CLS Cohort Studies Working Paper, Centre for Longitudinal Studies, Institute of Education, University of London.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures

- of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **64**, (4), 583–639.
- Spiegelhalter, D. J., Thomas, A., and Best, N. G. (1994). Computation on Bayesian Graphical Models. In *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting*, (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith), pp. 407–25. Oxford University Press.
- Spiegelhalter, D. J., Thomas, A., Best, N. G., and Lunn, D. (2003). *WinBUGS Version 1.4 User Manual*. MRC Biostatistics Unit, Cambridge, Available from www.mrc-bsu.cam.ac.uk/bugs.
- Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*, **12**, (3), 1–16.
- Troxel, A. B., Ma, G., and Heitjan, D. F. (2004). An index of local sensitivity to nonignorability. *Statistica Sinica*, **14**, 1221–37.
- van Bureen, S. (ed.) (2003). *Statistica neerlandica*, **57**, (1). Blackwell Publishing Limited.
- van Buuren, S. (2005). Multiple Imputation Online. www.multiple-imputation.com. accessed on 16 March 2006.
- van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. *Statistics in Medicine*, **18**, 681–94.
- van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, **76**, (12), 1049–64.
- van Buuren, S. and Oudshoorn, C. G. M. (2005). Multivariate imputation by chained equations, mice v1.14 user’s manual. Technical report PG/VGZ/00.038, TNO Prevention and Health.
- van Buuren, S. and Oudshoorn, K. (1999). Flexible multiple imputation by MICE. Technical report PG/VGZ/99.054, TNO Prevention and Health.
- Vella, F. (1998). Estimating Models with Sample Selection Bias: A Survey. *Journal of Human Resources*, **33**, (1), 127–69.
- Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., and Kenward, M. G. (2001). Sensitivity Analysis for Nonrandom Dropout: A Local Influence Approach. *Biometrics*, **57**, 7–14.
- Wakefield, J. (2004). Ecological inference for 2 x 2 tables. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **167**, (3), 385–445.
- Walpole, R. E. and Myers, R. H. (1993). *Probability and Statistics for Engineers and Scientists*, (5th edn). MacMillan.
- White, I. R., Carpenter, J., Evans, S., and Schroter, S. (2004). Eliciting and using expert opinions about dropout bias in randomised controlled trials. Technical report, London School of Hygiene and Tropical Medicine.

- White, I. R., Higgins, J. P. T., and Wood, A. M. (2008a). Allowing for uncertainty due to missing data in meta-analysis - Part 1: Two-stage methods. *Statistics in Medicine*, **27**, 711–27.
- White, I. R., Pocock, S. J., and Wang, D. (2005). Eliciting and using expert opinions about influence of patient characteristics on treatment effects: a Bayesian analysis of the CHARM trials. *Statistics in Medicine*, **24**, 3805–21.
- White, I. R., Welton, N. J., Wood, A. M., Ades, A. E., and Higgins, J. P. T. (2008b). Allowing for uncertainty due to missing data in meta-analysis - Part 2: Hierarchical models. *Statistics in Medicine*, **27**, 728–45.
- Wood, A. M., White, I. R., and Hotopf, M. (2006). Using number of failed contact attempts to adjust for non-ignorable non-response. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **169**, (3), 525–42.
- Yun, S.-C., Lee, Y., and Kenward, M. G. (2007). Using hierarchical likelihood for missing data problems. *Biometrika*, **94**, (4), 905–19.