Ministry of Justice

Surveying Prisoner Crime Reduction (SPCR)

Adjusting for Missing Data
Technical Report

Ian Brunton-Smith, University of Surrey

James Carpenter and Mike Kenward,
London School of Hygiene and Tropical Medicine

Roger Tarling, University of Surrey

Ministry of Justice Research Series
2014

Analytical Services exists to improve policy making, decision taking and practice by the Ministry of Justice. It does this by providing robust, timely and relevant data and advice drawn from research and analysis undertaken by the department's analysts and by the wider research community.

Disclaimer

The views expressed are those of the authors and are not necessarily shared by the Ministry of Justice (nor do they represent Government policy).

First published 2014

Any enquiries regarding this publication should be sent to us at mojanalyticalservices@justice.gsi.gov.uk

This publication is available for download at http://www.justice.gov.uk/publications/research-and-analysis/moj

ISBN 978-1-84099-630-2

Acknowledgements

We would like to thank Kathryn Hopkins, Rik Van de Kerckhove, and Nisha de Silva from the Ministry of Justice for their guidance and support throughout this project. We would also like to thank Andrew Cleary from Ipsos MORI for his openness and willingness to explain the complexities of the fieldwork for the project, which proved essential to understanding the nature of missingness.

Disclaimer

The views expressed are those of the authors and are not necessarily shared by the Ministry of Justice (nor do they represent Government policy).

Contents

List of tables

List of figures

1. Introduction

This report is concerned with a large longitudinal survey, Surveying Prisoner Crime Reduction (SPCR), and the recovery of 'missing' data from the survey. SPCR involved interviewing a large group of prisoners during and after their prison sentences. In some cases, the interviews were not conducted, as the prisoner could not be contacted or did not want to be interviewed. This meant that the prisoners' answers were 'missing' from the dataset, meaning that the dataset was smaller than planned, and potentially biased. This report explains how statistical techniques were used to 'recover' the missing data, where possible, allowing more robust analysis of the survey data to be conducted, and more rigorous findings to be produced.

1.1 Surveying Prisoner Crime Reduction

Surveying Prisoner Crime Reduction (SPCR) is a longitudinal cohort study which aimed to track the progress of approximately 4,000 newly sentenced prisoners in England and Wales over the period from 2005 to 2010. SPCR consists of a sample of prisoners which was broadly representative of most prison receptions (Sample1) and a sample of longer-sentenced prisoners (Sample2). At the time it was the largest survey of prisoners ever undertaken in Britain. Ipsos MORI was commissioned to carry out the survey by the Research Development and Statistics Directorate (RDS) National Offender Management Service (NOMS) at the Home Office, now Offender Management and Sentencing Analytical Services (OMSAS) at the Ministry of Justice (MoJ).

The broad aim of SPCR was to explore how interventions might work in combination to address a range of prisoners' needs. More specifically, SPCR aimed to record prisoners' problems and needs on reception into prison and how these are addressed during and after custody. Then, taking into account prisoners' background characteristics, SPCR aimed to assess the combined effect of any interventions on subsequent offending and other outcomes after release from prison.

SPCR was designed as a cohort study: the objective was to interview the same prisoners on three occasions, or 'waves' and a subset of prisoners on a fourth occasion. At Wave 1 prisoners were selected for the study and interviewed on reception into prison. Prisoners were interviewed a second time (Wave 2) just prior to being released from prison. Thus interviews at Waves 1 and 2 were conducted while the prisoners were in prison. Waves 3 and 4 occurred in the community after the prisoners had been released from prison. The intention at Wave 3 was to interview the entire sample two months following release from

prison. Attempts were made to re-interview a subsample of those prisoners successfully interviewed at Wave 3 four months later; that is, six months following release from prison (Wave 4). However, as only a subset of prisoners were eligible for interview at Wave 4 (longer-sentenced prisoners only), this final stage of SPCR is not considered further in this report. This report focuses on Waves 1, 2 and 3, where all prisoners qualified for interview.

In a representative cohort (or panel) study, which SPCR was designed to be, the aim is to randomly select participants at the outset, and ensure that the same participants are interviewed throughout each stage (or wave) of the survey, to monitor changes to individuals over time. However, it is generally difficult to secure interviews with the same participants again and again over time, and as participants cannot be replaced (due to the nature of the study), their answers to later interviews become 'missing'. This is also called 'survey attrition'. Missing data also occurs even when a subject is interviewed – it may not be possible to obtain answers to every single question, through error, or because the interviewee does not want to answer particular questions.

In summary therefore, there are two main reasons why data can be missing from such studies:

1. The interview participant is not interviewed as per the interview schedule, and therefore no responses to any questions are recorded for that person for that particular interview. In missing data terminology, this is called '**unit nonresponse**', as the interviewee is considered a 'unit' of the study.
2. The interview participant is interviewed, but some answers to some questions are not provided by the interviewee. This is called '**item nonresponse**' as each question in the interview is considered an 'item'.

In the first case, where the participant is not interviewed (unit nonresponse) there are two reasons why the interview may not have taken place. The first is when the participant could not be contacted, and therefore could not be asked if they wish to be interviewed. The second occurs when the subject is successfully contacted, but they do not wish to be interviewed. Therefore the main issues associated with missing survey data are unit nonresponse (as a result of either contact or compliance issues) and item nonresponse. In SPCR there was little item nonresponse at each wave. Thus if prisoners were interviewed, they largely answered all the questions. The problem for SPCR was unit nonresponse; at Waves 2 and 3 many prisoners could either not be contacted, or did not wish to participate further in the study.

Unit nonresponse, resulting in attrition of the sample at Waves 2 and 3 led the sample after Wave 1 to be smaller than originally intended and potentially biased if certain types of prisoners were more likely to be interviewed than others. For example, prisoners with more chaotic lifestyles may have been more difficult to contact in the community at Wave 3. Any resulting findings at this wave will therefore not be representative of these types of prisoners. The smaller sample size also leads to less efficient estimates (as the standard errors will be larger[1]). Thus it is important to investigate the nature and causes of missing data in any study and to take steps where possible to rectify the potential problems that may result, both in terms of biased and therefore unrepresentative samples, and samples which are smaller than they could be, and therefore less robust.

The aim of this study was to address the problem of missing data in SPCR. The extent of missing data was explored and a feasibility study conducted to determine whether the missing data could be recovered using statistical proxies. There are a number of methods which can be used to adjust for missing data and the one selected for SPCR was Multiple Imputation (MI), which is described later in this report. Also provided are instructions on how MI can be used to adjust survey estimates from Wave 2 and Wave 3.

A short, non-technical summary of the procedures for undertaking MI with SPCR can also be found in Brunton-Smith et al., (2014).

Drawing on the missing data analysis undertaken on SPCR and explained in this report, MI procedures have been used to allow robust analysis using data from Wave 2 and 3. Not all missing data were recovered, particularly where sample sizes were small, or where MI was unfeasible for other reasons. Descriptive results from Waves 2 and 3, and statistical analysis using Waves 1 to 3 are available on the gov.uk website (Hopkins and Brunton-Smith, 2014, Brunton-Smith and Hopkins, 2013, Brunton-Smith and Hopkins, 2014), along with Wave 1 results, which contain no imputed data (MoJ, 2010, Williams et al., 2012a, 2012b, Hopkins, 2012, Boorman and Hopkins, 2012, Cunniffe et al., 2012, Light et al., 2013).

1.2 Structure of the technical report

The initial stage of this study explored the nature of missing data in SPCR. The complex design of SPCR, involving interviewing prisoners at up to four waves, is described in **Chapter 2**. This chapter also includes details of the extent of missingness across Waves 2 and 3 of the sample, as well as an examination of the reasons why data were missing.

[1] Standard errors are a function of the sample size.

Chapter 3 provides a discussion of the theory underpinning missingness and its potential impact on the precision and accuracy of survey estimates, before outlining the range of statistically principled methods available to adjust for missing data. Multiple Imputation (MI) is identified as the most appropriate method to deal with missing data in the SPCR, therefore a detailed explanation of MI is provided, including an annotated description of the steps to be taken when undertaking MI using the software package Stata.

Procedures for conducting specific analyses in SPCR when data are missing are discussed in **Chapter 4**. This includes examples where the aim is to estimate descriptive statistics, as well as more complex regression analyses. Examples are provided for analyses using Sample1 and Sample2 of SPCR. Given that the most important finding of interest was whether prisoners in the survey reoffended, proven re-offending based on Police National Computer (PNC) records was chosen as the dependent variable for many of the examples of regression analyses. Appropriate sensitivity analyses are proposed and described to test the assumptions underpinning multiple imputation.

In **Chapter 5** the main conclusions from the assessment of missing data in SPCR are summarised together with further guidance when applying multiple imputation in practice.

For ease of reference, auxiliary variables suitable for missing data adjustments at Wave 2 and Wave 3 are listed at **Appendix I**.

Appendix II presents the full Stata code, with accompanying instructions, to implement the examples and the sensitivity analyses set out in Chapter 4.

2. The extent and nature of missing data in SPCR

2.1 Sample structure of SPCR

Two samples[2] of prisoners were selected for SPCR from 117 prisons in England and Wales, starting in 2005. The first (Sample1), comprising 1,435 prisoners, was selected to be representative of all prisoners received into prison and was first interviewed between November 2005 and November 2006.The majority of prisoners received into prison in England and Wales are serving short-term sentences (one year or less). In order to obtain more information pertaining to longer-term prisoners, a second sample of 2,414[3] prisoners was selected to be representative of those received into prison serving a sentence between 18 months and four years (Sample2), and was first interviewed between June 2006 and November 2007.

Prisoners were sampled to be interviewed at Wave 1 in prison from 2005 to 2007 (depending on when they were sentenced), at Wave 2 in prison from 2005 to 2010 (depending on their sentence length) and at Wave 3 in the community (or back in prison, if they had returned) from 2006 to 2010 (also depending on their sentence length). Prisoners were matched to the Police National Computer (PNC) and their offending histories and reoffending data one and two years after release were added to the dataset (see Boorman & Hopkins, 2012 for details). Data contained in SPCR therefore spans from the beginning of the prisoners' criminal careers (frequently in childhood) through to mid 2011 (when the last reoffending snapshot was taken) with the bulk of the data comprising survey answers collected from 2006 to 2010, but which also asked about the prisoners' backgrounds (including childhood) and future plans.

Sample1 and Sample2 were designed to be examined separately from one another, with Sample1 representing the experiences of all prisoners, and Sample2 representing those prisoners serving longer sentences. However, together the total number of prisoners in SPCR was 3,849 and all were interviewed at Wave 1, interviews being conducted as soon as practicable after the prisoner was first received into prison.

A second interview with all respondents (Wave 2) was planned to be conducted in the two weeks prior to release from prison. However, for those serving very short sentences (all of whom were in Sample1) it was not practical to arrange a second interview shortly after the

[2] See the Wave 1 Technical Report (Cleary et al., 2012) for details of sampling and methodology.
[3] Sample2 was also designed to over-sample female offenders, who are also likely to be under-represented in a simple random sample.

first interview (especially if for any reason the first interview had been delayed). For this group, only one interview took place in prison and a selection of 'Wave 2 questions' were asked at the first interview. This subgroup of Sample1, comprising 737 prisoners, is referred to as 'Sample1combW12' (because Waves 1 and 2 were combined). The other subgroup of Sample1, comprising 698 prisoners, was interviewed on two separate occasions while in prison and will be referred to as Sample1sepW12 (because Waves 1 and 2 were separate). The intention was to interview all prisoners, in whichever sample or subgroup, at Wave 3, ideally 2 months after release from prison.

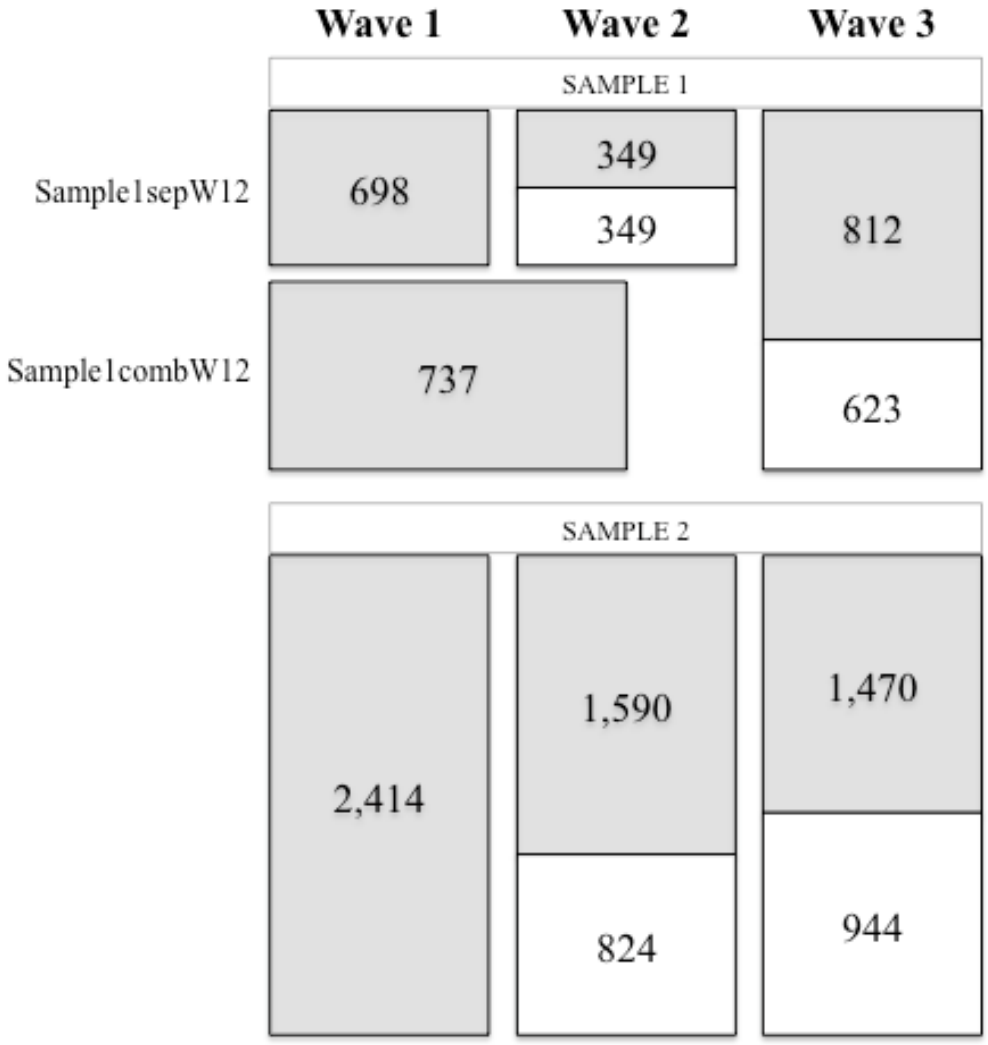
Further details about the sample design of the Waves 1 to 4 interviews, including information on the sampling frame, overall response rates, and representativeness of the sample can be found on the MoJ website (Cleary et al., 2012a; 2012b; 2014).

2.2 The extent of missing data in SPCR

Despite achieving a representative sample of prisoners at Wave 1, the degree of cohort attrition (respondents not being re-interviewed) after Wave 1 was notable. The degree of attrition is summarised in Figure 2.1. At Wave 2, exactly half the Sample1sepW12 (349) were not interviewed. There was no attrition at Wave 2 for Sample1combW12 (because Wave 2 questions were asked at the Wave 1 interview). In any strategy to adjust for missing data at Wave 2, Sample1combW12 should not be included, as being missing was not possible for this group.

At Wave 3, where all prisoners were intended to be interviewed two months following release from prison, 623 or 43.4% of the 1,435 prisoners in Sample1 were not interviewed.

Figure 2.1: SPCR prisoners interviewed (shaded boxes) and not interviewed (unshaded) across Waves 1 to 3



For the sample of longer-term prisoners, Sample2, 824 (or 34.1%) were not interviewed at Wave 2 and attrition had risen to 39.1% (944 prisoners) at Wave 3.

2.3 The impact of missing data on sample representativeness

Missing data can be particularly problematic when the sample achieved at later waves differs systematically from the original sample, leading to potentially biased estimates. With fully observed (collected) data at Wave 1 of SPCR, it is possible to see the extent that the Wave 2 and Wave 3 samples may lead to biased estimates by comparing the distribution of key variables (prisoner characteristics) across survey waves. Table 2.1 includes details of the observed data at Wave 1, 2 and 3, with comparisons of prison sentence length, age, gender, ethnicity, offence type, and drug use (as measured at Wave 1).

Table 2.1 Characteristics[4] of SPCR Prisoners at Waves 1, 2 and 3: percentages

	Sample1			Sample2		
	W1	W2	W3	W 1	W2	W3
Gender						
Male	90.8	90.7	92.0	83.4	86.2	84.8
Age						
18-20	12.1	13.4	14.0	15.5	16.0	16.3
21-24	20.7	20.0	21.1	20.8	20.3	21.5
25-29	22.9	21.9	23.0	18.3	18.6	17.7
30-39	29.1	30.1	27.8	26.1	26.5	26.7
40-49	11.9	11.3	10.8	14.1	13.8	13.3
50+	3.3	3.3	3.2	5.3	4.7	4.6
Ethnicity						
White	84.4	84.6	86.8	81.8	82.3	84.6
Mixed	3.8	3.9	4.3	4.2	4.3	3.7
Asian	3.6	3.0	4.2	5.4	4.5	5.2
Black	7.4	7.4	4.3	8.0	8.3	6.1
Chinese or Other	0.8	1.1	0.4	0.7	0.6	0.3
Offence group						
Violence	16.5	15.8	18.1	21.4	21.8	21.6
Sexual offences	0.8	0.7	1.0	5.6	6.9	5.0
Robbery	1.8	2.1	1.9	8.8	8.9	9.0
Burglary	6.3	6.3	7.9	13.0	14.1	14.0
Theft and handling	20.6	20.8	19.8	7.5	7.0	7.3
Fraud and forgery	3.0	2.4	1.9	3.3	2.5	3.2
Drug offences	5.3	5.4	5.3	25.7	23.7	26.1
Motoring offences	17.6	18.1	16.3	2.7	2.9	2.5
Other offences	20.9	21.6	21.1	11.5	11.9	11.1
Offence not recorded	7.2	6.7	6.9	0.5	0.4	0.3
Sentence length						
<= 6 months	66.2	69.7	65.5	-	-	-
> 6 months <= 1 year	15.2	12.4	14.4	-	-	-
> 1 year <= 18 months	5.7	5.3	6.2	14.8	12.3	14.3
> 18 months <= 2 years	5.0	5.1	4.9	28.3	28.1	28.4
> 2 years <= 3 years	5.6	5.7	6.8	38.5	42.9	39.9
> 3 years <= 4 years	2.3	1.8	2.2	18.4	16.8	17.4
Drug use in last month						
Class A drugs	44.7	46.0	45.0	37.0	38.9	38.0
Class B/C drugs	52.3	54.0	53.8	45.1	47.7	45.9
Total	**1,435**	**1,086**	**812**	**2,414**	**1,590**	**1,470**

In general there was a high degree of consistency in the observed data across the three waves of the survey, despite the missing data at Waves 2 and 3. For Sample1 there was some suggestion of an over-representation of those serving shorter sentences at Wave 2, as 69.7% of the Sample1 Wave 2 sample was sentenced to less than or equal to six months, compared with 66.2% of the Sample1 Wave 1 sample. There also appeared to be an under-representation of black offenders at Wave 3 (7.4% of Sample1 had reported their ethnicity as 'black' at Wave 1, but only 4.3% were observed at Wave 3) and a slight over-representation of those imprisoned for burglary or violence in the Wave 3 observed (interviewed) sample. For Sample2 there was a slight under-representation of those serving sentences between six

[4] Some of these characteristics have been published on the MoJ website. Estimates reported here may use different classifications/groupings to other published SPCR results and/or may not be representative of the SPCR samples. Findings reported here were used in the missing data analysis process and are not intended as population estimates.

to 18 months and those serving over 3 years at Wave 2, and slightly fewer black respondents at Wave 3. A more detailed assessment of potential bias across the full list of variables from Wave 1 confirmed there were relatively few differences between the observed samples at each wave.

2.4 Exploring the reasons for missing data across Waves 2 and 3

Once the extent of missing data in SPCR was understood, the next step was to try to understand the reasons for its occurrence. This analysis was used to inform decisions on the appropriateness of assumptions regarding 'missing mechanisms' and also to identify 'auxiliary variables' which can help to adjust for missing data (topics which are defined and discussed later in this report).

Despite a high degree of unit nonresponse (prisoners not being interviewed) across the waves of the survey, the levels of nonresponse to particular survey questions (item nonresponse) amongst those interviewed was low – typically fewer than 10 cases were missing from any given question/item. This meant the problem of missing data was mainly restricted to groups of prisoners' answers being missing at a particular wave or waves (e.g. prisoners who reported their ethnicity as 'black' being more likely to be missing from Wave 3 interviews). As a result, this analysis of missingness did not look further at the issue of item nonresponse, as it was expected to have little impact on any survey findings.

Cohort attrition usually occurs in a uniform fashion, with those respondents opting out of an earlier wave of the survey remaining absent at subsequent waves. This would mean that Wave 3 of SPCR should suffer equal, or more attrition than Wave 2. In contrast, however, some prisoners who were missing at Wave 2 of SPCR (where the interviews were held in prison) were interviewed in the community at Wave 3. This suggested that the reason for absence at Wave 2 was not solely down to the prisoner opting out of the survey, but was also because of practical restrictions on the ability of interviewers to access prisoners for interview. These included the prisoner having been transferred to a different prison, or there being insufficient prison staff to allow interviewers access to the prison or the prisoner.

To better understand the extent that missing data in SPCR may be a product of these two different mechanisms – one reflecting practical constraints on the interview process and one resulting from individual decisions not to participate – interviews with Ipsos MORI (the data collection agency) were conducted about the data collection process at each wave. Ipsos MORI provided detailed information including: information on the interviewer-recorded reason for unit nonresponse; details of each prison at which an attempted interview was

made (with interviews attempted as many as seven times in up to four different prisons if a prisoner was moved between prison establishments); and when the interview was conducted.

Analysis of the interview data confirmed the existence of two distinct, but related, processes leading to unit missing data at Waves 2 and 3 of the survey: **non-contact**, and **noncompliance** amongst those where successful contact was made. Non-contact refers to the inability of the interviewer to establish direct contact with a prisoner, whilst noncompliance was an active decision made by the prisoner not to take part in the survey. Noncompliance was only possible amongst the prisoners who were successfully contacted, thus the two categories of nonresponse represent two distinct phases of the data collection process. First the interviewer had to contact the prisoner, and then amongst those contacted, the prisoner chose whether to comply or decline to be interviewed.

Wave 2: Interviews conducted in prisons before release

Table 2.2 provides details of how missing data was split between non-contact and noncompliance for the two samples and also provides details of the reason for the non-contact or noncompliance. Sample1combW12 were omitted from this analysis as non-contact was not possible as they were asked Wave 2 questions at the same time as they were asked Wave 1 questions. The table shows that half of Sample1sepW12 was interviewed at Wave 2, as was nearly two-thirds (65.9%) of Sample2. For those prisoners who were not interviewed, non-contact accounted for most of the nonresponse: 41.4% of those in Sample1sepW12 and just over a quarter (26.8%) of Sample2. Noncompliance was a relatively small problem: only 8.6% of those in Sample1sepW12 and 7.3% of Sample2 did not wish to be interviewed at Wave 2, once contacted.

Table 2.2: Wave 2: number of SPCR prisoners missing by reason (sample1combW12 omitted) for each sample

	Sample1sepW12		Sample2	
	Frequency	%	Frequency	%
Interviewed	**349**	**50**	**1590**	**65.9**
Nonresponse (not interviewed)				
Noncontact	**289**	**41.4**	**648**	**26.8**
• Access not provided by prison	79	11.3	125	5.2
• Access – insufficient time	159	22.8	329	13.6
• No contact	26	3.7	142	5.9
• Other (absconded, deceased, unavailable, no match)	25	3.6	52	2.2
Noncompliance	**60**	**8.6**	**176**	**7.3**
• Arranged but not brought forward (no reason obtained)	38	5.4	63	2.6
• Refused	22	3.2	113	4.7
Total	**698**	**100**	**2414**	**100**

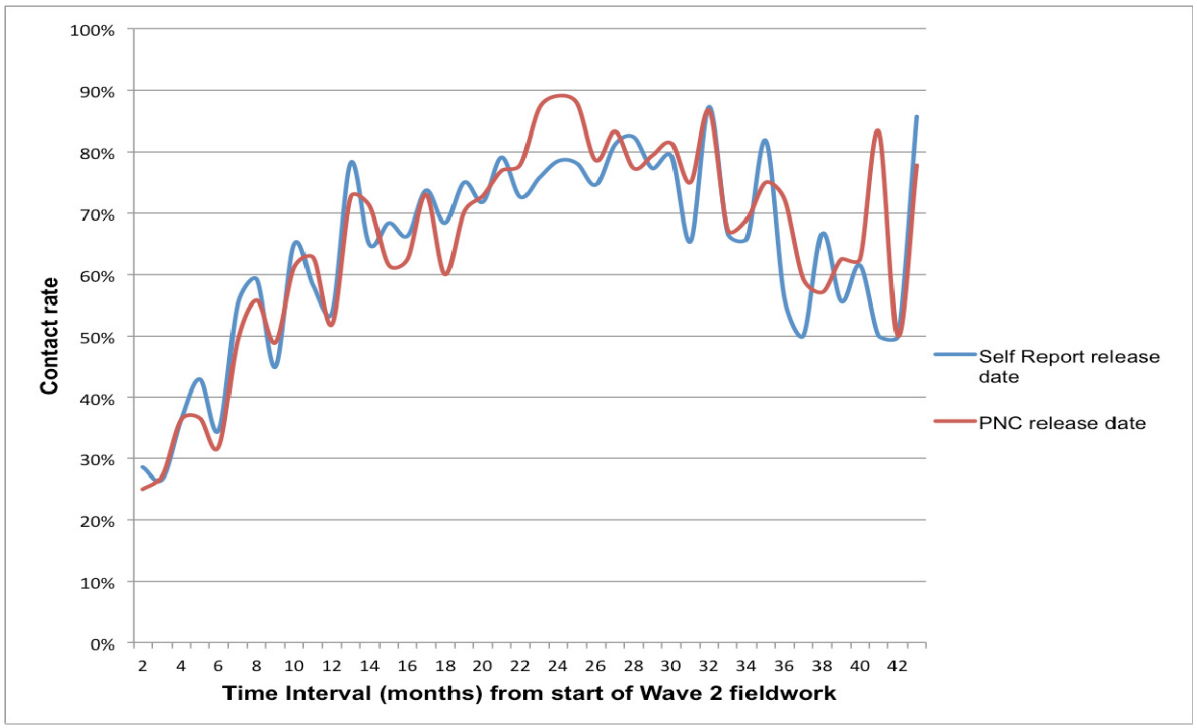
Source: Ipsos MORI

Non-contact

Non-contact was largely due to insufficient time given to establish contact with prisoners (22.8% and 13.6% respectively). Ipsos MORI highlighted the importance of the lead-in time allocated to secure re-interview at Wave 2, with insufficient time allocated to this process during the early stages of SPCR. As a result many potential respondents had already been released (frequently on Home Detention Curfew) before Ipsos MORI was able to identify which prison they were in. As this problem became evident, Ipsos MORI extended the lead-in period to allow more time to establish contact, which led to an improvement in contact rates. The contact rates were higher for Sample2 because fieldwork for Sample2 began seven months after the fieldwork for Sample1, and was thus less affected by the problems experienced during the early stages of data collection.

The result of the change in procedures is evident in Figure 2.2, which plots monthly contact rates against the time interval (in months) between the start of Wave 2 fieldwork and the prisoners' date of release for all those eligible for re-interview prior to release from prison (Sample1sepW12, plus Sample2). Both self-reported release date as collected at Wave 1 (used to schedule re-interview), and release date as recorded from PNC data are shown, although in practice, both track each other closely.

Figure 2.2: Wave 2 contact rate by time interval (in months) between start of Wave 2 fieldwork and SPCR prisoners' release date



During the first 12 months of data collection the contact rate gradually increased, but remained lower than during subsequent months. This was largely a result of improved contact rates in Sample1, with most of the data collection for Sample2 occurring later in the fieldwork. The improvement in contact rates over time matches Ipsos MORI's report of gradually extending the lead-in time between the sixth and 12th month of data collection. The increased fluctuation towards the end of the data collection period reflects the comparatively small number of interviews attempted during this time. This was because there were relatively few prisoners serving longer (e.g. three to four years) sentences.

The second issue raised by Ipsos MORI was the inability of some prisons to provide access for interviews. As a result of the high degree of inter-prison mobility between Wave 1 and 2 (31% of respondents had moved prisons between the two waves), contact rates were based on the final prison at which an interview was attempted. Considerable variability was evident in the ability of prisons to allow interviewers access to the prison site for re-interview. Ten of the 117 sampled prisons did not contact any prisoner for interview (although for most of these the target number was very small) and a further 12 had contact rates below 50%. At the other extreme, eleven prisons had 100% contact rates.

Noncompliance

In contrast, noncompliance was less of a problem at Wave 2, with only around eight per cent in both samples actively opting not to take part in the survey (see Table 2.2). Unlike issues of non-contact, which were more likely to reflect practical problems associated with gaining access to prisoners, refusal might be expected to be more closely influenced by individual motivational factors. However, some variability in refusal rates was also evident across prisons. Whilst compliance was high across the majority of prisons (with no refusals in 34 prisons), a small group of prisons were identified with particularly low levels of compliance (although in these prisons the target number was very small). The reason for this is not known, but could be due to staff/prisoner influence or other factors which might lead to many prisoners in one location not wishing to be interviewed.

Wave 3: Interviews conducted in the community after release (or in prison for those who had returned)

Despite Wave 3 being intended as an interview in the community two months after leaving prison, a considerable minority (n=707) of prisoners were back in prison at the time of re-interview. This group is referred to as 'Prison Returnees', and interviews with 507 of these prisoners were secured (see Table 2.3). Ipsos MORI did not record details of why 200 of these returnees were not interviewed, but based on Wave 2 responses, it is expected that the majority of this group would be noncontacts rather than refusals. Table 2.3 provides details of the Wave 3 interview results.

Table 2.3: Wave 3: number of SPCR prisoners missing by reason

	Sample1 Frequency	%	Sample2 Frequency	%
COMMUNITY SAMPLE				
Interviewed	**587**	**52.0**	**1,188**	**59.0**
Nonresponse (not interviewed)				
Noncontact	**487**	**43.1**	**693**	**34.4**
• Property ineligible: vacant, derelict, demolished, not found, or other	29	2.6	38	1.9
• No contact	141	12.5	189	9.4
• Other (too ill, moved, died, never lived at address)	196	17.4	253	12.6
• No details recorded	121	10.7	213	10.6
Noncompliance	**55**	**4.9**	**132**	**6.5**
• Prisoner declined interview	49	4.3	118	5.9
• Other household member declined on behalf of prisoner	6	0.5	11	0.5
• Entry to block/scheme refused by warden	0	0.0	3	0.1
Total	**1,129**	**100**	**2,012**	**100**
PRISON RETURNEES				
Interviewed	**225**	**73.5**	**282**	**70.3**
Nonresponse (not interviewed)	**81**	**26.5**	**119**	**29.7**
Total	**306**	**100**	**401**	**100**

Source: Ipsos MORI

Non-contact

Where interviews were attempted in the community, the majority of nonresponse was again identified as noncontact: 43.1% of Sample1 and 34.4% of Sample2 were identified as non-contact. This included around 2% where the property was identified as 'ineligible' (vacant, derelict or demolished), and 12.5% and 9.4% respectively where no contact was made at the household. A further 10.7% of Sample1 and 10.6% of Sample2 were classified as non-contact by Ipsos MORI, but no details were accurately recorded about the specific reason for non-contact.

Amongst Prison Returnees, 26.5% of Sample1 and 29.7% of Sample2 could not be contacted. The non-contact rate for Sample1 Prison Returnees was noticeably lower while the non-contact rate for Sample2 Prison Returnees was slightly higher than the corresponding non-contact rates for these samples at Wave 2. The reasons for this were not recorded.

Noncompliance

Noncompliance rates were again lower than levels of non-contact, with only 4.9% of Sample1 and 6.5% of Sample2 actively declining to be re-interviewed (see Table 2.3).

Across both Waves 2 and 3 of SPCR only a small proportion of the total sample was actively opting themselves out of the survey. Instead, the problem of missing data was primarily limited to types of non-contact.

The low levels of noncompliance at Wave 2 and 3 gave some confidence that it was possible to make suitable missing data adjustments across SPCR. Identification of two separate processes leading to missing data at each wave (non-contact and noncompliance) also allowed for a more nuanced assessment of the potential drivers of missing data, with separate assessments of each process identifying those variables most closely associated with each.

2.5 Empirical models predicting missingness

To identify the factors associated with a prisoner being missing at Wave 2 or Wave 3, a series of logistic regression models were specified and fitted. The intention was to explore systematic differences in the likelihood of being interviewed at Waves 2 and 3 of the survey based on the range of individual characteristics available from the initial interview (Wave 1) of SPCR, fieldwork information provided by Ipsos MORI, and data from the Police National Computer (PNC), which contained reoffending and criminal history data for each prisoner.

Two dependent variables were constructed to help explore missing data in SPCR at each wave. 'Contact' identified whether a respondent included in the sample at Wave 1 was successfully contacted at subsequent waves (contact=1, non-contact=0[5]). 'Compliance' identified whether a Wave 1 prisoner agreed to participate at later waves (compliance=1, noncompliance=0).

A key benefit of SPCR was the considerable amount of information available about prisoners who were missing after Wave 1, from which could be derived a large number of potential explanatory variables. Demographic information was included for each offender, covering: age; gender; ethnicity; marital status; socio-economic status; educational qualifications; and whether the respondent had dependent children. Information was also included about the length of the current sentence, with those serving shorter sentences expected to be more difficult to contact as a result of the difficulties of setting up subsequent interviews. The type of offence that prisoners were serving their sentence for was also recorded, as was self reported details of whether the respondent had received any prior prison sentences.

[5] The convention in missing data analysis is to predict whether an individual is observed, not whether they are missing.

Information was collected about the type of accommodation lived in prior to the SPCR prison sentence, which was expected to be closely aligned with contact rates at Wave 3. A distinction was made between those that reported living in 'unstable' accommodation prior to the sentence (those sleeping rough, in a hostel, in probationary accommodation, or in a squat), those living in 'short-term' accommodation (resident at previous address for less than six months), and those that were living with family. Medical information was also included: whether the respondent reported having a long-standing illness, whether they were registered with a family doctor (GP) and whether they were in receipt of benefit or income support in the year prior to their sentence. In addition, information on drug and alcohol use prior to the SPCR sentence was recorded, as well as whether the prisoner was previously employed, and whether English was a Foreign Language for each prisoner.

As contact rates improved over the length of the data collection period, a variable was constructed to indicate whether the prisoner had been released during the first 12 months of the data collection period or at a later time point. Whether a respondent agreed to have their address traced (as recorded at Wave 1), and whether they agreed to have their record linked to data from the Department for Work and Pensions (DWP) or Her Majesty's Revenue and Customs (HMRC) was included to capture possible contact problems during the community interviews. Finally, information on whether each eligible sample member was contacted and complied with the Wave 2 sample was incorporated in the Wave 3 analysis.

As well as individual data derived from the Wave 1 interviews and information from Ipsos MORI, data from the PNC about each eligible sample member were also examined. This allowed for the inclusion of information about the offending history of each respondent, Information about whether the prisoner was a first time offender, as well as the number of convictions a prisoner had received in the 12 months prior to the SPCR sentence was included. Also recorded was whether the prisoner had been convicted in the last 12 months for a burglary, robbery, theft, or violent offence and, finally, whether the prisoner went on to be reconvicted in the year after their release from prison for the SPCR offence.

To ensure that all characteristics associated with missing data were identified, an initial bivariate analysis was undertaken examining the associations between each of the dependent variables and the full range of explanatory variables. Those explanatory variables that were found to be significantly associated with a dependent variable were incorporated into logistic regression models in order to isolate those that were independently associated with missingness. The models were estimated using a backward stepwise regression

strategy, with all the variables identified as significant in the bivariate models included in the first instance.

Models that predicted whether an individual included at Wave 1 was successfully contacted at Wave 2 (or Wave 3) were explored first. Amongst those that were successfully contacted, a second model that predicted whether a contacted individual chose to comply with the survey request was created. This two-stage process accurately reflected the structure of missing data in SPCR.

The analysis revealed substantial variations in contact and compliance rates across prisons at Wave 2 (and amongst the Prison Returnee sample at Wave 3). To incorporate these influences on likelihood of being interviewed, a multilevel logistic modelling approach was adopted (Goldstein, 2003). This enabled the differential contact and compliance rates across prisons to be accounted for, identifying those prisons with lower levels of contact and compliance. Incorporating the multilevel structure also ensured that significant individual predictors of contact and compliance were correctly estimated.

The purpose was to identify all factors that were associated with data being missing, and because the data for each sample were collected jointly by Ipsos MORI, Sample1 and Sample2 were amalgamated. However, as a sensitivity check, similar models were explored separately for each sample to determine whether additional processes specific to these samples were evident. No additional predictors of missing data were identified, therefore only the general models are presented here.

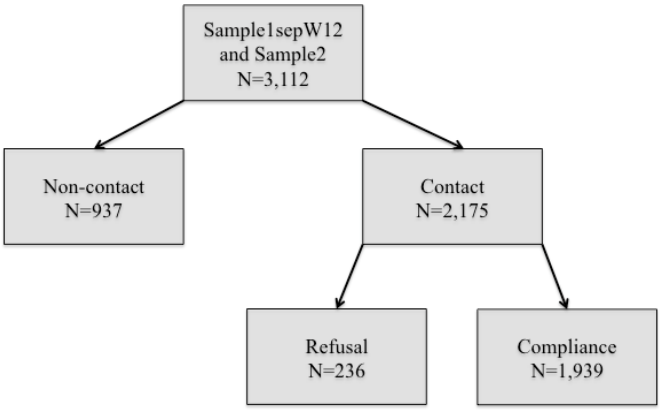
2.6 The logistic regression models

The models presented below show those predictors significantly associated with contact and compliance ($p<.05$). Results are presented as odds ratios. For example, in models predicting contact, odds ratios greater than one indicated a higher chance of being successfully contacted when compared against the reference group, and odds ratios below one represent a lower chance of being successfully contacted.

Wave 2

At Wave 2 the analysis was restricted to the 3,112 respondents from Sample1sepW12 and Sample2 (as it was not necessary to re-contact those in ‘Sample1combW12). When securing an interview, an attempt was made to contact each survey prisoner and then those that were contacted then had the opportunity to agree to be interviewed or refuse to participate further in the survey. This two stage model structure is demonstrated graphically in Figure 2.3.

Figure 2.3: Structure of missing data at Wave 2



Contact

Table 2.4 details the final multilevel model that shows significant predictors of a prisoner being successfully contacted (but not necessarily interviewed) for second interviews at Wave 2 (in-prison).

Table 2.4: Predicting contact with SPCR prisoners for second interview at Wave 2[1]

	B	S.E	Odds Ratio	Sig
Early release date (first 12 months)	-1.03	0.18	0.36	<0.01
Sentence length (ref: >18 months <= 3 years)				
<= 6 months	-0.58	0.21	0.56	0.01
> 6 months <= 1 year	-0.32	0.23	0.72	0.17
> 1 year <= 18 months	-0.63	0.13	0.53	<0.01
> 3 years <= 4 years	-0.53	0.12	0.59	<0.01
SPCR offence (ref: acquisitive, violence, breach or other)				
Theft	-0.48	0.14	0.62	<0.01
Drug offence	-0.32	0.11	0.73	<0.01
Prior prison sentence	0.36	0.1	1.43	<0.01
Prior sentence for burglary	0.37	0.1	1.45	<0.01
Prison non-contact	-1.8	0.26	0.16	<0.01
(constant)	1.21	0.1		
Prison variance	0.22	0.07		
Sample size	3,110			

[1] With the exception of sentence length and SPCR offence, all other explanatory variables are binary and the reference category is the absence of the attribute indicated.

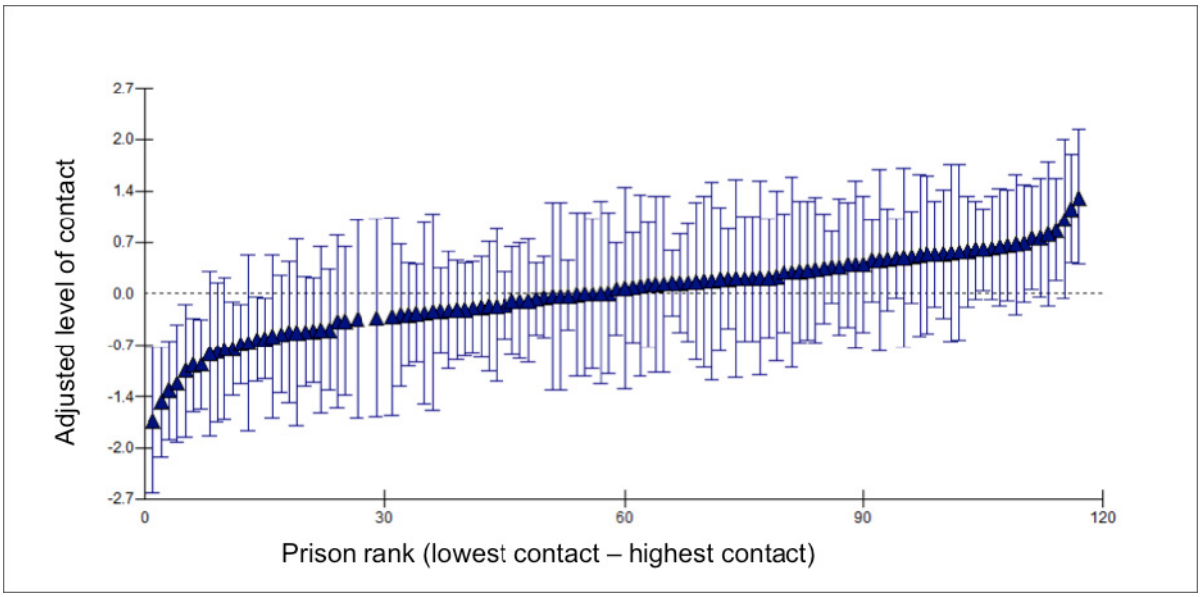
This confirms the effect of an early release date from prison, with the odds of being contacted amongst those released in the first 12 months of the survey approximately one third (0.36) the odds of those released after the first 12 months (whilst all other factors remain constant). This supports reports from Ipsos MORI that extending the re-interview lead-in time over the first year of data collection resulted in significantly higher contact rates thereafter.

A number of explanatory variables reflecting differences in the characteristics of sentenced prisoners were also identified. Sentence length was found to be significantly related to contact at Wave 2, confirming the apparent bias evident in Table 2.1. Those serving shorter sentences were less likely to be re-contacted. This finding further reflects the problem of an insufficient lead-in time to secure a re-interview, with less available time for re-interview amongst those serving shorter sentences. Those serving sentences of over three years were also less likely to be re-contacted, perhaps reflecting difficulties of maintaining accurate records for these prisoners. Sample2 was restricted to those serving sentences of between 18 months and four years, therefore the lower odds of being contacted evident for those serving sentences of less than or equal to six months are isolated to Sample1.

Those serving a sentence for a theft or drug offence were particularly likely not to be re-contacted at Wave 2. In contrast, those who had served prior prison sentences were more likely to be re-contacted at Wave 2, as were those who had received a prior sentence for burglary. It was difficult to draw too many concrete conclusions about why prior sentences were associated with a higher or lower likelihood of being re-contacted, however it is possible that this reflected systematic differences in the types of prison these offenders were sent to. The lower likelihood of contact amongst those serving sentences for drug offences might be because these individuals were involved in treatment programmes during their prison sentence, perhaps reducing their availability for re-interview.

Having adjusted for systematic differences in the likelihood of being contacted across the sample, significant variability in contact rates remained between prisons. This clearly demonstrated that some prisons were less able to allow access to interviewers. Figure 2.4 includes details of the variations in contact rates across all 117 prisons included in the sample. Adjustment was also made for the number of prisoners within each prison by adopting the standard practice of ‘shrinking’ contact rates for prisons with few prisoners towards the overall mean in order to reduce their undue influence on the estimated contact rate (Goldstein, 2003).

Figure 2.4: Variation in rates of contact with prisoners achieved by interviewers across the 117 prisons participating in SPCR, at Wave 2



The horizontal line across the centre of the graph reflects the average adjusted contact rate across all prisons, with each prison's difference from this average represented as a triangle. These differences have been ranked from lowest on the left (representing those prisons with contact rates lower than the average) to highest on the right (representing those prisons with contact rates higher than the average). 95% confidence intervals were also included around each estimate, allowing identification of those prisons that had significantly higher or lower contact rates (where prisons have the same rank, the triangles have been superimposed on one another).

From the graph seven prisons were identified with contact rates that were significantly lower than average. Including an identifier for these prisons in the logistic models (Prison Non-contact) accounted for more than half of the residual variation in contact rates between prisons. Taken alongside the systematic variations associated with the characteristics of prisoners within each prison, more than 60% of the variance initially attributed to prison differences was explained.

Compliance amongst prisoners who were successfully contacted

Table 2.5 includes details of the final multilevel model predicting compliance (agreeing to be interviewed) at Wave 2 amongst the eligible sample of 2,175 that were successfully re-contacted by Ipsos MORI.

Table 2.5: Predicting SPCR prisoners' compliance (agreeing to be interviewed) at Wave 2[1]

	B	S.E	Odds Ratio	Sig
Young respondent (18-20)	0.51	0.26	1.67	0.05
Sentence length (ref: > 18 months <= 3 years)				
<= 6 months	-1.17	0.23	0.31	0.05
> 6 months <= 1 year	-0.18	0.38	0.83	<0.01
> 1 year <= 18 months	-0.40	0.22	0.67	0.63
> 3 years <= 4 years	-0.49	0.20	0.62	0.07
Prison noncompliance	-1.48	0.27	0.23	0.02
(constant)	2.53	0.13		
Prison variance	0.17	0.10		
Sample size	2,175			

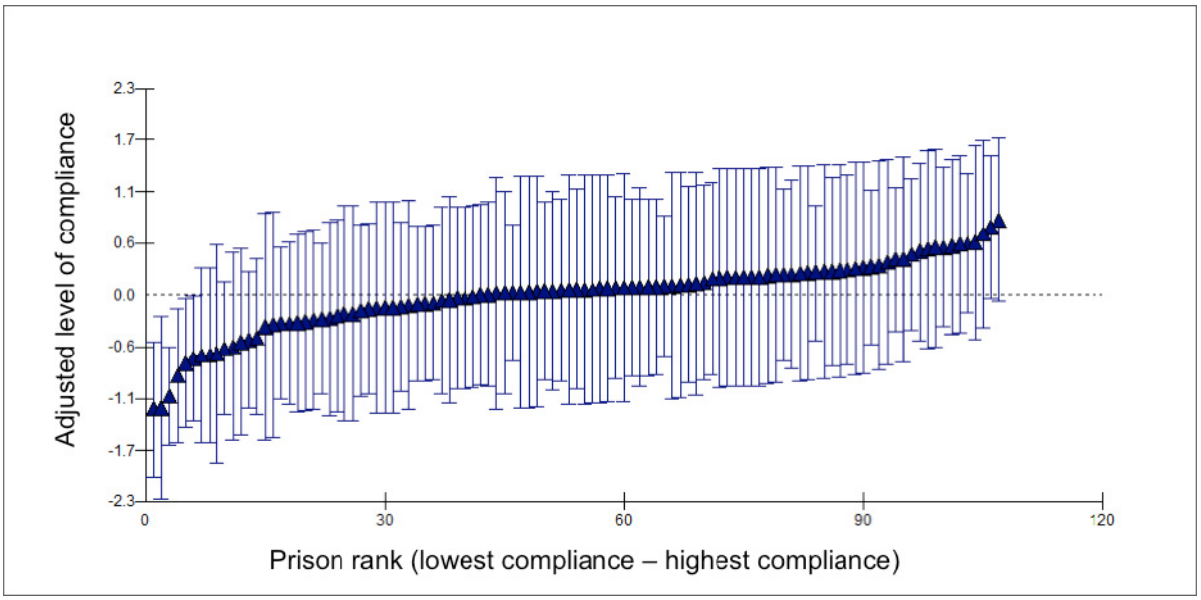
[1] With the exception of sentence length, all other explanatory variables are binary and the reference category is the absence of the attribute indicated.

Fewer significant predictors of compliance were evident when compared to levels of contact. This reflected the comparatively small group of prisoners who refused to participate in the survey once contact had been successfully made. Given that a decision to refuse to be interviewed was more likely to be driven by individual prisoners' motivations, the apparent lack of systematic differences was reassuring, suggesting that refusal to participate in the Wave 2 interview was essentially random.

Prisoners serving particularly short sentences (less than or equal to 6 months in Sample1, and 18 months in Sample2) were significantly less likely to agree to re-interview, as were those serving particularly long sentences. In contrast, the youngest prisoners were significantly more likely than other age groups to agree to be interviewed.

Systematic differences between prisons with respect to compliance rates of prisoners were also evident. Figure 2.5 shows variations in compliance rates across the 104 prisons where contact was made.

Figure 2.5: Variation in compliance rates (agreement to be interviewed) across the 104 prisons where contact was made with prisoners, at Wave 2

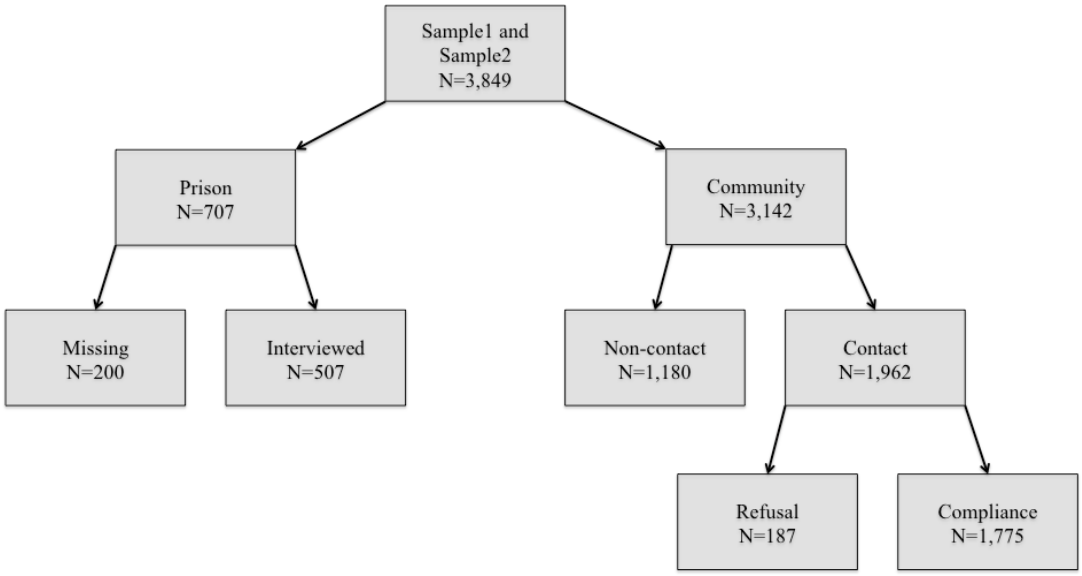


A small group of prisons was identified where compliance rates were significantly lower than average. Information on these five prisons was included as a binary explanatory variable (Prison refusal) in the model. The inclusion of this information accounted for all remaining significant variability in compliance between prisons. This means that in these five prisons, prisoners tended to refuse to participate in the Wave 2 interview. The reason for this is not known, and may be due to staff/prisoner influence, or other factors.

Wave 3: Interviews conducted in the community after release (or back in prison for Prison Returnees)

The data structure at Wave 3 was more complex than Wave 2. The majority of Wave 3 interviews were conducted in the community but 707 prisoners were back in prison at the time of re-interview (Figure 2.6). Amongst the community sample, an attempt was made to contact each prisoner. Those contacted then had the opportunity to agree to be interviewed (comply) or refuse to take part in the survey. Amongst the Prison Returnee sample, however, only whether an interview was achieved was recorded. Ipsos MORI was not able to provide information about whether lack of success was attributable to non-contact or noncompliance.

Figure 2.6: Structure of missing data at Wave 3



Community sample

The 3,142 eligible respondents who were resident in the community at the time of interview were considered first.

Contact

Table 2.6 presents details of the final logistic model identifying significant predictors of contact amongst the community sample. In common with Wave 2, a number of significant predictors of contact at Wave 3 were evident.

Table 2.6: Predicting contact with SPCR prisoners at Wave 3 (in the community)[1]

	B	S.E	Odds Ratio	Sig
Young respondent (18–20)	0.27	0.12	1.30	0.02
Ethnicity (black)	-0.67	0.14	0.51	<0.01
English a Foreign Language	-0.65	0.17	0.52	<0.01
SPCR offence (ref: acquisitive, violence, theft or other)				
Drug offence	0.28	0.10	1.33	<0.01
Breach	-0.39	0.16	0.68	0.02
No consent to DWP/HMRC matching	-0.38	0.08	0.69	<0.01
Never had a full time job	-0.27	0.13	0.76	0.03
Has educational qualifications	0.18	0.08	1.20	0.02
Lived with family prior to SPCR	0.44	0.08	1.55	<0.01
Not registered with GP	-0.46	0.13	0.63	<0.01
Crack cocaine dependency (daily user)	-0.52	0.13	0.60	<0.01
(constant)	0.38	0.09		
Sample size	3,103			

[1] With the exception of SPCR offence, all other explanatory variables are binary and the reference category is the absence of the attribute indicated.

Black respondents and those reporting speaking English as a Foreign Language were significantly less likely to be contacted in the community at Wave 3. Also less likely to be contacted were those identified as daily users of crack cocaine prior to their prison sentence, those who never worked before their sentence and those not registered with a GP[6] - characteristics one might expect to be associated with an unstable lifestyle. Prisoners who served an SPCR sentence for breach of an existing sentence were also less likely to be contacted at Wave 3, perhaps reflecting an unwillingness to provide up-to-date details of their address upon release. Those who did not agree to have their records linked with DWP and HMRC records at the Wave 1 interview were also identified as less likely to be contacted.

In contrast, those who reported they were living with family prior to their SPCR sentence, and those with educational qualifications were more likely to be contacted. Younger respondents (aged 18-20) and those who served an SPCR sentence for a drug offence were also more likely to be contacted.

Compliance

Once contacted, respondents decided whether or not to be interviewed at Wave 3. Table 2.7 presents estimates from the final logistic model based on the 1,962 prisoners who were contacted and asked whether they wanted to be interviewed.

Table 2.7: Predicting agreement to be interviewed at Wave 3 (in the community) by SPCR prisoners, once contacted [1]

	B	S.E	Odds Ratio	Sig
No consent to address matching	-0.57	0.22	0.56	0.01
Non-contact at Wave 2	-0.52	0.17	0.60	<0.01
Refused interview at Wave 2	-0.80	0.28	0.45	0.00
(constant)	2.53	0.11		
Sample size	1,962			

[1] All explanatory variables are binary and the reference category is the absence of the attribute indicated.

Like the Wave 2 compliance model, a significantly reduced range of factors associated with compliance were evident (again suggesting that refusal was not a large problem in the survey). The strongest predictor of compliance at Wave 3 was having been interviewed at Wave 2 (those who refused the interview or could not be contacted at Wave 2 had significantly lower odds of being interviewed at Wave 3). This reflects the attrition process common to longitudinal studies, emphasising the importance of prior survey involvement to ensure compliance at later survey waves. A small group actively opted out of the survey

[6] General Practitioner (Family doctor).

early and remained absent at subsequent waves, with a further group who chose to opt out of the survey at the later stage. Agreement to be re-interviewed was also less likely amongst those that did not give consent for their address to be matched at Wave 1, suggesting that one reason for opting not to provide contact details was a desire not to be re-contacted.

Prison Returnee sample

The same two stage process was anticipated for the Prison Returnee sample, with prison access restrictions accounting for the largest share of missing data. However, records of the type of missing data were not maintained by Ipsos MORI for the Prison Returnee sample, with information restricted to a single code indicating the respondent was back in prison. As a result, a distinction was not made between types of nonresponse amongst this sample. Nevertheless, details of the prison where the final attempted contact was made were retained, allowing systematic differences between prisons to be explored in a multilevel logistic regression model. The final model is shown in Table 2.8.

Table 2.8: Predicting response at Wave 3 (in prison) by SPCR prisoners who had returned to prison[1]

	B	S.E	Odds Ratio	Sig
Young respondent (18–20)	0.78	0.39	2.19	0.04
SPCR burglary	0.80	0.37	2.22	0.03
Prior sentence for robbery	0.74	0.35	2.10	0.03
(constant)	0.85	0.31		
Prison variance	6.23	1.84		
Sample size	707			

[1] All explanatory variables are binary and the reference category is the absence of the attribute indicated.

Only three factors were significantly associated with the likelihood of being interviewed across this sample. Younger respondents were significantly more likely to respond, similarly, those who served their original SPCR sentence for burglary, and those who had previously been convicted of a robbery were more likely to be re-interviewed. There was also considerable variability between prisons, with a number of prisons identified where no respondents were successfully re-interviewed.

2.7 Summary

This chapter has provided a detailed assessment of the reasons for missing data in SPCR, identifying a number of important issues that guide the strategy to deal with missingness. Firstly, the problem of missing data was largely restricted to unit missing data (prisoners were not contactable, or did not want to be interviewed), with item missing data (where prisoners were interviewed, but did not provide answers to some questions) comparatively

low (typically fewer than 10 answers were missing from any question). This meant that any missing data adjustment could be framed in relation to prisoners not being interviewed at Wave 2 or Wave 3, rather than adopting an item specific approach. This also meant information from Wave 1 that was available for all prisoners (whether or not they were interviewed at Wave 2 or Wave 3) could be incorporated as 'auxiliary' information.

Information from the interviewing organisation, Ipsos MORI, showed that at Wave 2 and Wave 3 the majority of missing data was the result of the inability on the part of the interviewers to re-contact prisoners. In contrast, comparatively few prisoners refused to participate in the Waves 2 and 3 interview if they were successfully contacted, with less than 10% of prisoners actively opting out at each wave. This had important implications for missing data adjustments, allowing assumptions to be made about missing data primarily in relation to contact rates, such as prisons not being able to grant access, rather than factors associated with prisoners not agreeing to be interviewed.

Exploratory analysis also revealed that despite a reasonably high degree of missing data across the survey, analysis of basic prisoner demographics and characteristics from Waves 1 to Wave 3 showed that there was little change across the survey. This does not mean that particular survey answers were unbiased, because prisoners who answered might differ systematically in other ways (not related to the characteristics investigated) from those who did not at Wave 2 and Wave 3. However, when assessed in conjunction with the evidence that the majority of missing data was the result of non-contact, rather than noncompliance, it does support the theory that the data was missing at random (MAR, Rubin 1987). Information on this assumption, and the implications for missing data adjustments are provided in Chapter 3.

Treating the missing data problem as a two-stage process – interviewers first attempting to make contact with each prisoner, with those successfully contacted able to opt in or out of the survey – confirmed the importance of the data collection process, with a range of characteristics identified that were associated with non-contact. At Wave 2, prison access arrangements were identified as particularly important, with a number of prisons particularly difficult to access. Insufficient time to secure re-interview, and the time between initial interview and Wave 2 interview (represented by sentence length) were also associated with non-contact at Wave 2. At Wave 3 non-contact was more closely related to individual prisoner characteristics, with regular drug users, those living in unstable accommodation prior to their sentence, no registered GP, and English as a Foreign Language, amongst those that were less likely to be re-contacted.

Chapter 3 provides a theoretical summary of the problem of missing data, before showing how the variables identified as predictive of missing data can, under appropriate assumptions, be used as ‘auxiliary’ variables in missing data adjustment strategies to improve survey estimates from incomplete data. These ‘auxiliary’ variables are listed in Appendix I.

3. Adjusting for missing data and the use of Multiple Imputation for analysing SPCR in the software package Stata

In this chapter the process of using associations between auxiliary variables and missingness (identified in the previous chapter) to make adjustments for missing data is explained. In addition some of the theoretical background and the alternative approaches to recovering missing data are described. A more detailed summary of the theory and methodology of missing data can be found in Enders (2010): a recent, comprehensive and accessible text, and Carpenter and Kenward (2013).

3.1 Missing data: Issues and assumptions

When data are missing, inference inevitably rests on assumptions that cannot be definitively verified from the data at hand. Broadly speaking, these assumptions can either be framed in terms of the so-called *missing data mechanism*; i.e. the probabilistic mechanism giving rise to the missing data, or in terms of the conditional distribution of partially observed variables given fully observed variables.

The conceptual framework for handling missing data was introduced by Rubin (1976), who classified missing data mechanisms into separate classes. Within a class, certain analyses are valid, and others are not. The key issue is dependence among missingness and unobserved variables, because it is the existence of such relationships that potentially undermines the validity of analyses.

Missing Completely at Random (MCAR)

Broadly speaking, MCAR matches intuitively the idea of random missingness. The chance that a unit is missing on an occasion does not depend on any of the missing values, or on values obtained on other occasions. The observed data are a simple random sample of the complete data and the probability of data being missing on Y is unrelated to Y or Xs. Under MCAR a completers' analysis is valid, if potentially inefficient (the standard errors of the estimates are larger).

Missing at Random (MAR)

In spite of the name, MAR does not correspond to the intuitive notion of randomness. Under MAR, missingness (i.e. the chance of data being missing) may be associated with any variables (factors such as sentence length, ethnicity, etc.), observed or not. However,

conditional on the values taken by observed variables (i.e. once the factors associated with missingness are known) there is no residual association with unobserved ones (the missing values can be predicted from information in the dataset alone). This is a subtle distinction, and intuitively it corresponds to the idea that any link between missingness and unobserved variables can be 'explained' by the variables that have been observed. Thus the probability of missing data on a variable Y is related to some other measured variable(s), X, but not the values of Y itself once these other variables have been included. An important consequence of the MAR assumption is that relationships that are seen among variables for units that are observed hold as well for units that are not fully observed, but have similar values on observed variables. While simple completers analyses are generally invalid under MAR, adjustments can be made to correct for this that are based on the observed data. These adjustments exploit the fact that relationships among variables for missing units can be 'borrowed', in an appropriate way, from those that are observed.

Missing Not at Random (MNAR)

A situation which is not MCAR or MAR is MNAR. Even adjusting for observed variables, there remain associations between missingness and unobserved data. Thus the probability of missing data on Y is related to values of Y, even after controlling for X.

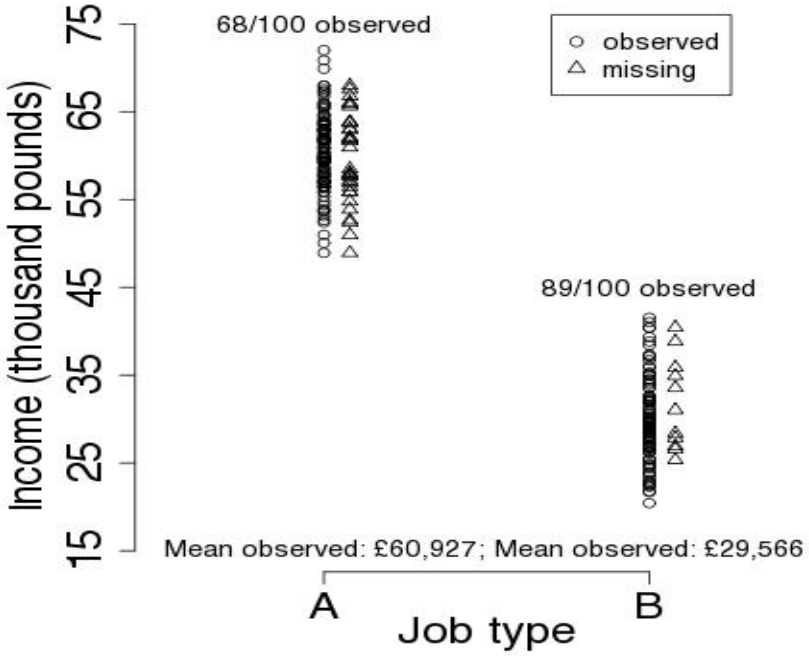
These different missing data mechanisms are illustrated by way of an example set out in Figure 3.1. Suppose 200 people were surveyed to estimate average income. We know 100 of them have job type A and 100 job type B. Only 157 out of 200 answer the question on income. Three possible assumptions that could be made are:

1. The missing incomes are a random selection of the data that were intended to be observed (MCAR);

2. Within job type, the missing incomes are a random selection of the data that were intended to be observed (MAR), and

3. Within job type, missing incomes are systematically different from those that were observed (MNAR).

Figure 3.1: Illustration of role of assumptions in analysis of partially observed data



The first assumption implies that the probability of observing the income data is unrelated to any inference about income that may be made. In terms of the classes of missing data mechanisms, the data are *missing completely at random (MCAR).* In this case, a valid estimate of mean income can be obtained by simply averaging the observed data. In general, when data are MCAR, analysis restricted to individuals with no missing data – known as *complete cases* or *complete records* analysis – will be unbiased, but may be imprecise (that is, has greater standard errors).

In this example, the key point that rejects MCAR is that there was more missing data in the group with higher income. In other words, it is evident from Figure 3.1 that the chance of observing income varies with job type (68% of those in job type A answered the question whereas 89% of those in job type B answered the question). Furthermore, the observed incomes are higher for those with job type A than for those with job type B. Thus merely averaging the observed incomes will not produce the correct result. Put another way, the assumption of MCAR is implausible.

The second assumption implies that the probability of observing income data depends on income (those in job type B with lower incomes are more likely to answer the question than those in job type A with higher incomes), but critically, within job type *the chance of seeing income is independent of the income value* (that is, these are a random sample of incomes within each job type). This is an example of the assumption known generically as *missing at*

random (MAR). MAR is a conditional independence statement: the chance of observing a variable depends on its value, but given other observed variables, this association is broken.

In the example in Figure 3.1, MAR implies that a valid estimate of the mean income within each job type can be obtained by averaging the observed incomes within that job type, giving £60,927 and £29,566 for job type A and B respectively. A valid estimate of the overall average income assuming missing incomes are MAR given job type is then (100 × £60,927 + 100 × £29,566) / 200 = £45,246.

Notice that the MAR assumption has been explained as the 'probability of observing income, given job type no longer depends on income'. As Figure 3.1 illustrates, this means that within strata of job type, the distribution of seen and unseen incomes is the same. This is an example of the correspondence – referred to above – between assumptions about the missingness mechanism on the one hand and assumptions about the distribution of partially observed variables given fully observed variables on the other.

Looking at Figure 3.1 it is clear that the MAR assumption cannot be verified from the observed data alone (circles); to do this the unseen data (triangles) would be needed. In this hypothetical example the unobserved data (triangles) have been added in and assumed to be broadly similar in value to the observed data, thus making MAR plausible. However, in practice the values of the triangles are not known, which could display a very different pattern and a different relationship to the observed data (circles). If that were the case, the data mechanism would be *missing not at random (MNAR)* – assumption 3 above.

The above illustrates the key point that the analysis requires assumptions which (though they may be shown to be plausible on the basis of the observed data and other information) can nevertheless not be definitively verified. In general, the procedure to deal with missing data is as follows:

1. Perform a preliminary analysis of the data to explore the pattern and extent of missing data, and identify plausible predictors of data being missing and plausible predictors of the unseen values;

2. Taking the information in Step 1 into account, formalise an assumption about the missing data mechanism and relate this to the generic assumptions of MCAR, MAR, MNAR;

3. Using an appropriate statistical method, impute missing data and perform an analysis under the assumption in step 2, and

4. Explore the robustness of inferences in Step 3 to alternative, plausible assumptions about the missingness mechanism. This is referred to as 'sensitivity analysis'.

In practice, step 1 will give insights into why data are missing and the factors associated with missingness (for example, is response related to gender – are men less likely to respond than women)? If the exploration at step 1 reveals no discernible pattern and no factors can be identified to explain missing data, one may assume at step 2 that data are missing completely randomly with respect to our inferential questions – the data mechanism is MCAR. It is then appropriate to perform the analysis on complete cases.

Most likely a pattern will be revealed at step 1. For example in Figure 3.1, job type is predictive of income being missing, and likely very predictive of income (step 1). As a result the data mechanism is not MCAR but could be MAR or MNAR (step 2). It is not certain which, as there is no information on the values of the missing data (triangles – in Figure 3.1) and hence their relationship to the observed data (circles). In this situation MAR is invariably assumed as the starting point for an analysis. Assuming MAR in Figure 3.1 (missing data depends on job type, but conditional on this there is no association between income and missing values), the distribution of observed income within job type is representative. This leads to the estimate of mean income of £45,246 above (step 3). However, it may be appropriate to explore the robustness of this estimate to possible differences – within job type – between the mean observed and unobserved income (step 4).

In SPCR, detailed analysis of the reasons for missing data highlighted that most missing data was a result of failure to contact respondents, with refusal comparatively rare. At Wave 2, non-contact was clearly linked with the process of data collection, with lower contact rates early in the process, some prisons particularly prone to non-contact, and those prisoners serving shorter sentences less likely to be contacted. As a result, the assumption that conditional on these observed process-related differences in contact, prisoners were missing at random is judged to be plausible.

Refusal was less clearly linked to the data collection process. However, comparatively few prisoners refused to be interviewed if successfully contacted (fewer than 10%), and there was little evidence of systematic differences in the types of prisoner that refused to comply with the survey request. As a result, the MAR assumption was again plausible, with the relatively small amount of refusers making it unlikely that departures from this assumption would have a large impact on results. The sensitivity of results to different assumptions regarding contact and compliance is discussed in Chapter 4.

At Wave 3 non-contact was again identified as the dominant source of missing data. Here a number of individual characteristics were associated with the propensity to be contacted,

acting as auxiliary variables to support the MAR assumption. However, because of the move in interview location from prison to the community and the associated increase in the potential explanations for respondents not being contacted, as well as evidence that sample members with more 'risky' profiles (e.g. no stable address prior to Wave 1, regular drug users prior to Wave 1) are less likely to be contacted, the MAR assumption is judged stronger, that is, less plausible at Wave 3. As a result, more care should be taken when assessing the significance of results from models using Wave 3 data, and the sensitivity of results to alternative assumptions about the missingness mechanism should be explored.

3.2 Options for missing data adjustment

There is now a very large literature on the statistical problem of handling missing data (e.g. Rubin 1987, Little and Rubin 2002, Schafer 1997, Molenberghs and Kenward 2007) and as a result a wide range of approaches have been developed to recover lost information (e.g. Diggle and Kenward, 1994; Scharfstein et al, 1999, Carpenter et al, 2006; Carpenter and Plewis, 2010, Carpenter and Kenward, 2013). The approaches that have been developed to deal with missing data can be broadly divided into ad hoc ones and those that can be described as statistically principled. The distinction between these largely rests on the underlying rationale.

Ad hoc approaches are usually defined in terms of the procedure itself, commonly some form of simple single imputation, such as replacing a missing value by the mean of the observed values for that variable. The resulting completed data set is then typically analysed as though complete. Unless this is done in a carefully structured way, with proper account taken of the fact that some data have been 'created', something that is rarely done in practice, such a procedure is invalid. Such a method can be biased, but more importantly, the measures of precision and resulting inferences will be wrong, the analyst mistakenly acting as though there is more data than there really is. Unless the proportion of missing data is very small, such methods are best avoided.

In contrast, principled methods are derived in a coherent way from appropriate and clearly stated assumptions about the mechanisms underpinning missingness. The aim is typically to maximise the efficiency of analyses and/or to correct for biases due to missingness. Such analyses can be framed in a variety of ways, but three general methodological approaches can be identified: Inverse Probability Weighting; Full Information Maximum Likelihood and Multiple Imputation. These various approaches differ greatly in their flexibility, technical complexity and current practical applicability.

Inverse Probability Weighting (IPW)

IPW has a long history in survey sampling, and is associated with the original paper by Horvitz and Thompson (1952). Here a model for the probability of a unit being missing is constructed from the predictors identified at the earlier exploratory phase. The inverse of the fitted probabilities of observing the units are then used to weight the units, and in so doing produce an analysis that represents the original sampled population. For example, if explorations of missingness identified that women were half as likely to remain in the sample as men, then the observed responses from women would each be given twice as much emphasis in the final analysis (a weight of 2 would be applied). The result of these adjustments is that the analytic survey appears, at least at face value, to accurately represent the original population. However, whilst this approach is conceptually clear, and represents a straightforward and general adjustment strategy in its most basic form such adjustments can be shown to be inefficient (e.g. Scharfstein et al 1999).

A growing, rather technical literature, now exists that develops more efficient estimators based on Inverse Probability Weighting, which have potentially useful robustness properties (e.g. Scharfstein et al, 1999). However, there currently exists no sufficiently general and flexible implementation for these models, and as such they have been largely restricted to bespoke analyses in very specific research settings.

Full Information Maximum Likelihood (FIML)

In the FIML approach a joint multivariate model is constructed for all the relevant outcome, and partially incomplete variables. The substantive model is subsumed within this and the overall model is then fitted to the data. If the model is correct then the resulting inferences will be valid. Many such models have been applied in a wide variety of settings. However, such models can be very elaborate and tend to be bespoke, with large modelling changes required when moving among substantive problems. Further, FIML approaches require a generally high level of technical knowledge to correctly specify the joint model (typically constructed within a structural equation modelling framework). The high dimensionality and complexity of the required joint models in FIML also increase the chances that there will be numerical and computational problems, such as lack of convergence.

Multiple Imputation (MI)

Multiple Imputation provides an alternative and very flexible approach to adjusting for missing data (Rubin, 1987, Kenward and Carpenter, 2007, Carpenter and Kenward, 2013). In this approach an imputation model is constructed which relates the potentially missing observations to those variables predictive both of missingness and of the incomplete variable itself. The multiple imputation model consists of two main stages. First, using random draws from the fitted imputation model, the data set is completed (i.e the missing data are replaced by randomly chosen imputations). The substantive model (the model being used for analysis) is fitted to the completed data set, and the analysis is undertaken. This whole process is repeated, using new random imputations on each occasion, thus generating sets of estimates and their corresponding variances. At the second stage, using rules developed by Rubin (e.g. see Rubin 1987), the sets of estimates and variances are combined to give one overall estimate and a valid measure of precision. Because the imputation process is repeated for each research question, a new imputed dataset is created each time a new analysis is undertaken. For this reason, a ‘final’ imputed dataset has not been produced, and the process described above may need to be repeated each time a new anaysis is undertaken. Most importantly the precision estimates take account of the process of imputation, so there is no sense of data being "created" or "made up", in contrast to the ad hoc single imputation approaches.

An important advantage of the MI approach is that it is straightforward to include variables in the imputation model that need not be in the substantive model. For such an analysis the MAR assumption may be far more plausible than for the more restrictive MCAR specification. Further, the same imputation model, and hence imputed data sets, may be appropriate for several different substantive analyses. MI also has the conceptual advantage that the same substantive model is retained that would have been used had there been no missing values. Hence there is a sense in which the MI analysis is "closer" conceptually to the originally conceived substantive analysis.

Multiple Imputation was judged to be the most appropriate method for dealing with the problem of missing data in SPCR. Not only are the results from MI effectively equivalent to those from FIML procedures in particular settings under MAR, the approach is simpler to implement and is more flexible and generalisable across a range of different research questions with comparatively few adjustments required.

3.3 Performing Multiple Imputation under Missing at Random (MAR)

This section sets out the rationale behind Multiple Imputation and outlines in more detail the steps taken in order to conduct MI. The intention here is not to provide a theoretical account, which can be found in Enders (2010) or Carpenter and Kenward, (2013). Furthermore, in practice these steps are performed by appropriate statistical software, which is discussed in the next section. The purpose here is to provide the reader with an intuitive understanding of the method, sufficient to specify the analysis to be undertaken by the software.

Suppose we have a substantive model of research interest with dependent variable *Y* and explanatory variables *X*. Data can be missing on any of the variables. In addition we have a set of auxiliary variables *Z* which are predictive of missingness. Of course, depending on the research question and substantive model of interest, any of the auxiliary variables could be regarded as an explanatory variable *X*. For example, drug use in custody (X) may be predictive of outcomes on release, such as reoffending (Y). Each of these variables may have missing data. Ethnicity, age, and sentence length (auxiliary variables, Z) of the respondent may predict whether the data is missing or not, and the value of that missing data. This information can be used, therefore, to predict the missing values of interest.

Assuming missing data are MAR

Step 1. A multivariate response model is fitted where the partially observed variables are treated as response variables and the fully observed variables are the explanatory variables. Include in this model (a) all the variables in the substantive model of interest (*Y* and *X*) and (b) any auxiliary variables, *Z*, which improve prediction of the missing values, especially if they are also predictive of values being missing. This is the *imputation model* which assumes multivariate normality for the variables in the dataset.

An alternative to specifying an explicit multivariate normal distribution of the partially observed variables given the fully observed variables (which is a requirement in the approach above) is to instead specify a series of conditional distributions of each variable on all the others. This is known as the *Full Conditional Specification (FCS)* algorithm, often also referred to as the *Chained Equations* approach to multiple imputation. The attraction of this approach is that no joint model has to be explicitly specified and linear regression can be replaced by an appropriate generalised linear model for discrete and categorical variables.

Step 2 creates *K* ‘completed’ datasets by drawing the missing data from the imputation model fitted in 1 above, taking account of both the uncertainty in the parameter estimates of

the imputation model, and – given the parameter estimates – the variability of the missing data implied by the model. In doing this, the missing data are being drawn from the estimated distribution of the missing data given the observed data. Creating the *K* imputed datasets provides a computationally convenient way of representing this distribution.[7]

Step 3. The substantive model of interest is fitted to each imputed dataset, giving *K* sets of parameter estimates and their standard errors.

Step 4. Following rules developed by Rubin (Rubin, 1987) and commonly referred to as 'Rubin's rules', combine the results of the *K* sets to produce one Multiple Imputation parameter (essentially the average of the *K* parameters) and an estimate of its variance.

In practice, the analyst will have identified the variables they wish to include in the substantive model of interest (*Y* and *X*). Once the analyst has further specified the imputation model (to include variables *Y*, *X* and *Z*) and the *K* number of imputations to be performed, the rest proceeds automatically by the appropriate software to which we now turn.

3.4 Software

Different software packages adopt different numerical algorithms to carry out MI. However, for the analysis of SPCR Stata is recommended, because of its practical utility and flexibility (and it performs all calculations on the appropriate scale automatically). Even within Stata there are two routines, MI (adopting the multivariate normal approach mentioned in Step 1 above) and ICE (the alternative chained equation approach mentioned in Step 1). The most flexible package is ICE, and this is adopted for all of the analyses presented later in this report.

ICE (Imputation by Chained Equations) was developed by Patrick Royston from UCL, and can be added as a routine to Stata. The routine can be accessed via his home page: http://www.homepages.ucl.ac.uk/~ucakjpr/. Alternatively, it can be accessed direct from Stata; in Stata, type: net from http://www.homepages.ucl.ac.uk/~ucakjpr/stata and follow the instructions on the screen to select, download and install the program(s) of your choice. The programs are mi_ice and mim2 (although mim2 is less necessary).

[7] It is for the analyst to choose the number of imputations, K. The choice should reflect a balance between ensuring sufficient imputations to generate good estimates of uncertainty, and unnecessary computational burden. For the examples that follow 40 imputations were chosen, which will be sufficient in most applications. In practice, the researcher should explore the sensitivity of results to differing numbers of imputations by varying K.

Undertaking Multiple Imputation for SPCR in Stata

This section provides a brief description of the key Stata code required. In setting out the generic code, the section also augments the steps of the analysis outlined above.

The nature of missingness within SPCR was explored in Chapter 2 and a series of auxiliary variables identified that were associated with and predictive of missingness for Waves 2 and 3. The auxiliary variables are listed in Appendix I, they are all binary variables. For use in Multiple Imputation models, all auxiliary variables have been coded as if they had no item missing values to ensure that ICE does not try to impute values for these variables. Any missing responses were recoded as 0 to represent the absence of the attribute in question.

The imputation model is generally robust to 'over-fitting' of auxiliary variables; therefore all significant predictors of missing data should be included in the imputation model, irrespective of the source or type of missing data they were associated with (non-contact or refusal to co-operate once contacted). However, to ensure that the model is not over-complicated, it is recommended only to use auxiliary variables predictive of missing data at Wave 2 when interested in recovering data for variables at Wave 2, and auxiliary variables predictive of missing data at Wave 3 when imputing Wave 3 variables. When data are missing from more than one wave, the auxiliary variables from each wave can be combined.

An analysis using the generic Stata code given in the box below can now begin. Each line is explained in turn.

```
use "{file path}", clear

preserve

keep {varlist to be recovered, varlist of model of interest, auxiliary varlist}

drop if {conditional command}

{command}{varlist of model of interest}

mi set wide

mi ice {varlist to be recovered}{varlist of model of interest}{auxiliary varlist}{conditional command},
                conditional(var: conditional command) add(K) seed(n)

mi estimate {, esampvarok}: {command}{varlist of model of interest}

restore
```

use simply identifies the data file. Adding ‘clear’ is advisable as it ensures that any data file analysed previously is removed.

preserve is used to save an image of the data file in its current form in Stata’s working memory, and allows the user to quickly return to a ‘clean’ dataset following any analysis. This is useful because the imputation process appends all imputed data on the end of the working data file, quickly increasing the size of the dataset if more than one MI is undertaken. Any variables and data that are deleted before the analysis is undertaken can also be quickly recovered. The restore command at the end of the code is used to return to the original data file.

keep is used to restrict the working file to those variables used in the analysis (the list of variables in the model of interest and the auxiliary variables associated with missing data). When used in conjunction with preserve/restore, this is a convenient strategy for running MI that avoids the generation of overly complex and large datasets.

drop if will often be needed to omit certain cases from the analysis. In SPCR this is sometimes used to omit respondents from a particular sample (for example, Sample2 representative of longer term prisoners) if the inclusion of that sample is not relevant to the proposed analysis. The command is also used to omit cases with any item missing data, ensuring that ICE does not try to impute these values (in practice ICE makes no distinction between item and unit missing data).

{command} Before embarking on multiple imputation, undertaking a ‘complete case’ analysis which can later be compared with any analysis performed on imputed data is recommended. A command is issued; for example, regress, logit, depending on the analysis required. MI is compatible with a wide range of existing Stata commands, listed in the Stata manual.

mi set wide Before setting up the imputation model, the data file is set to ‘wide’ format (the Stata default). This is a memory efficient structure to deal with imputed data, which appends the imputed data from the variables with missing values on to the end of the data file as a series of new variables. The imputed variables can be identified in the dataset with the extension _1_, _2_, _3_... included before the variable name.

mi ice is the basic command used by Stata to indicate the specification of an imputation model. The three sets of variables which need to be included are set out; the variables to be recovered, the variables of substantive interest in the analysis at hand, and the appropriate

auxiliary variables. ICE automatically recognises when a variable to be recovered is binary or continuous and uses the appropriate estimation procedure in the imputation model. For ordinal and categorical variables it is necessary to explicitly identify the level of measurement by adding the prefix o. if the variable is ordinal and m. if the variable is categorical. The ordering of included variables does not matter as ICE will automatically identify the variables with missing data, but it is clearest to start with imputed variables, followed by model of interest, followed by auxiliary. It is also possible to restrict imputation to a subset of observed cases here, by adding a conditional statement before the comma.

conditional(var: conditional command) is required when attempting to impute data from variables with complex routing. var identifies the variable that is to be imputed only for a subset of respondents from the main data file that satisfy a given routing criteria. The given routing criteria is identified with the conditional command, and can include a variable that is itself imputed from the data. This allows for a wide range of conditional imputations to be specified. When using this conditional command, it is also necessary to add the esampvaryok command to mi estimate (described below). This allows for the fact that the imputed sample size at each imputation can vary depending on how many missing cases are classified that satisfy the conditional specification.

add(K) is used to identify the number of separate imputed sets of data to generate. In the examples to follow 40 separate imputations are specified, although it is possible to draw more than this if required. Thus 40 separately imputed sets of the variables with missing data are produced, with each containing complete information on all respondents (these are appended to the end of the dataset).

mi estimate Having performed the imputations, the substantive model of interest is then re-estimated on each of the imputed sets of data independently. Parameter estimates from each imputed analysis are then combined along with a combined measure of the precision of the estimates (this incorporates the variance within each imputed dataset, as well as the variance across imputed datasets). When including conditional impute commands in the imputation model, esampvaryok needs to be added before the model of interest is specified.

These are the main steps in undertaking all of the imputation models presented in this report. There are, however, several modifications, qualifications and additions which are required when conducting specific analyses. These are discussed in Appendix II, which includes all Stata code from the examples in Chapter 4. This includes code to compare the imputed data with the observed data as an initial check for unexpected, implausible differences in the

distributions that might indicate potential specification problems with the imputation model. This also provides details of the sensitivity analyses undertaken to assess the appropriateness of the missing data models.

4. Multiple Imputation examples from SPCR

In this chapter a number of examples are provided showing how the imputation procedures (outlined in Chapter 3) could be implemented with SPCR. First, situations are considered when interest lies in descriptive information from variables measured at Wave 2 or 3 which contain missing data. Second, imputation procedures suitable for multivariate analyses are shown, where data are partially observed for some of the included dependent or explanatory variables.

When interest is in variables with missing data at Wave 2, the imputation procedures should include those variables that were significant predictors of contact and compliance at Wave 2 as 'auxiliary' information. When interest is in variables with missing data at Wave 3, the predictors of contact, compliance, and contact in prison should be used. Appendix I lists these auxiliary variables, and the sources of missing data they are associated with. The imputation model is generally robust to 'over-fitting' of auxiliary variables, therefore all significant predictors of missing data will be included in the imputation model, irrespective of the type of missing data they were associated with (non-contact or noncompliance).

SPCR is comprised of two separate samples. Sample1 was designed to be representative of all receptions into prison, whilst Sample2 was an oversample of those serving longer prison sentences (between 18 months and 4 years). Examples are included that use each of these samples. Note also, however, that not all Wave 2 questions were asked of every prisoner. Only a subset of Wave 2 questions were put to those in Sample1 who were only interviewed once in prison (Sample1combW12). It was felt that imputation procedures were only appropriate for questions asked of all prisoners. For the other questions, not asked of every prisoner, it would not make conceptual sense to use imputation procedures to attempt to adjust estimates since many individuals were not given the opportunity to opt in or out of the survey at this time point. Furthermore, any analysis based on the subset of 698 prisoners (Sample1sepW12) who were asked a Wave 2 question not asked of all would be biased as Sample1sepW12 is not representative of Sample1 and hence not representative of prisoners received into prison.

Full Stata code for all the examples is set out in Appendix II.

4.1 Imputation for descriptive statistics

The first examples are concerned with producing descriptive information about prisoners' experiences in prison, derived from variables that were measured at Wave 2 of SPCR. Specifically, whether prisoners reported doing any paid work during their sentence (J2Work), a question asked of all prisoners at Wave 2.

4.1.1 Sample1

Based on complete cases, the analysis began by calculating details of the total number of respondents from Sample1 who reported undertaking any paid work whilst in prison (see Table 4.1). This indicates that approximately 49% of prisoners reported being in paid employment during their prison sentence, based on the 1,078 prisoners with observed data. This includes 729 respondents with observed data measured at Wave 1 from the combined questionnaire (Sample1combW12 – omitting the 8 respondents from the sample who did not have valid data for this item) and 349 with observed data collected at Wave 2 from Sample1sepW12.

Table 4.1: Percentage of respondents having undertaken paid work during their sentence (Sample1)

Have you undertaken paid work during your sentence?	Complete cases	MI Results		
			95% confidence Interval	
	Proportion	Proportion		
Yes	49.1%	53.4%	50.2%	56.7%
No	50.9%	46.6%	43.3%	49.8%
Sample size	1,078	1,427		

The imputation model must include the model of interest – in this case just the variable with missing data (J2Work) and the list of auxiliary variables linked to the probability of missing data. In the current example missing data is from a Wave 2 variable, so the list of auxiliary variables is those linked to the probability of missing data at Wave 2. In general, including any other variables from other waves that are associated with the variable with missing data (here J2Work) should also be considered, whether or not they are predictive of the values of that variable being missing.

In this example a total of 40 sets of imputations were produced, with the missing data imputed from the 349 complete cases in Sample1sepW12. The imputed data was then merged with the observed data from the 729 respondents in Sample1combW12 who responded to the question as part of their Wave 1 and Wave 2 combined interview to create 40 'complete' datasets. The descriptive information from each 'complete' dataset was then

combined using Rubin's rules to produce an overall estimate for the full sample. In addition, 95% confidence intervals have been included alongside this estimate to reflect the additional uncertainty that is associated with the imputed data.

In contrast to the complete case analysis, the imputed analysis (Table 4.1, 'MI Results') estimates that approximately 53% of prisoners were in paid work during their sentence, suggesting that the complete case analysis slightly underestimates the proportion of prisoners that work.

To explore the difference between the complete case and the imputed estimates, the proportion that reported being in paid work in Sample1combW12 was compared against the remaining 349 prisoners in Sample1sepW12 that were re-interviewed and answered the question at Wave 2, as well as against the 349 imputed cases. Table 4.2 demonstrates that considerably fewer prisoners in Sample1combW12 reported being in paid work during their sentence, most likely because they were serving short sentences so were not eligible for work or did not have time to undertake work.

Table 4.2 Comparison of the proportions of SPCR prisoners working in prison from the three groups of prisoners

Have you undertaken paid work during your sentence?	Sample1combW12	Sample1sepW12 Observed Wave 2	Sample1sepW12 Imputed Wave 2
Yes	36.1%	76.2%	67.1%
No	63.9%	23.8%	32.9%
Sample size	729	349	349

The imputed results are similar (although not identical) to the 349 prisoners with observed data at Wave 2, with 67% of prisoners identified as having worked, compared to 76% of the observed cases. The difference between the imputed and observed results is consistent with the under-representation of shorter sentenced prisoners identified in the missing data models in Chapter 2 (see Table 2.4 and Table 2.5), and the reduced likelihood of working amongst prisoners serving shorter sentences. Because the imputed data is more similar to the 349 observed cases at Wave 2, when the data are combined this has the effect of 'pulling' the estimate of work in prison from the complete case analysis upwards. In this example the benefits of MI can be clearly seen, with the complete case analysis underestimating the proportion of prisoners that undertake paid work in prison.

As an initial check on the data, percentages were calculated for the imputed data and directly compared to the equivalent observed data. Here interest is in identifying any anomalies in the distributions of the imputed data which might indicate potential problems with the

imputation models. Of the 40 imputations, three, dataset numbers 1, 22 and 35, were examined. Table 4.3 demonstrates close correspondence between the three sets of imputed data. The lower estimated proportion of prisoners undertaking paid work in the imputed data reflects the under-representation of shorter sentenced prisoners in the observed data (Sample1sepW12).

Table 4.3 Selected imputations (Sample1sepW12)

Have you undertaken paid work during your sentence?	Sample1sepW12 Observed Wave 2	_1_J2Work	_22_J2Work	_35_J2Work
Yes	76.2	63.3	61.0	65.9
No	23.8	36.7	39.0	34.1
Sample size	349	349	349	349

In this example, MI was shown to have a clear impact on the estimated proportion of prisoners that reported undertaking paid work in prison. This occurs when one, or more, auxiliary variable is seen to be strongly associated with *both* the underlying values of the variable of interest and missingness (the chance of observing them). In this case serving a very short prison sentence is predictive both of whether someone is able to undertake paid work in prison (with those serving short sentences less likely to work), and being contacted or complying with the survey request at Wave 2 (those serving shorter sentences were less likely to be contacted or to comply with the survey request).

4.1.2 Sample2

The procedure for Sample2 is almost identical to Sample 1 described above, with the same list of auxiliary variables incorporated alongside J2Work. Here we drop Sample1 from the dataset before proceeding with the imputations.

Table 4.4 includes estimates of the proportion of longer term prisoners that worked based on complete cases and imputed data. It is clear that there is a high degree of consistency between the complete cases analysis and the MI results, but the latter is now based on the full sample of 2,414 respondents. Again confidence intervals were included to reflect the uncertainty associated with the imputed data.

Table 4.4 Percentage of SPCR prisoners reporting undertaking paid work during their sentence (Sample2)

Have you undertaken paid work during your sentence?	Complete cases	MI Results		
	Proportion	Proportion	95% confidence Interval	
Yes	78.8%	78.4%	76.4%	80.4%
No	21.2%	21.6%	19.6%	23.6%
Sample size	1,590	2,414		

Unlike Sample1, the lack of clear differences in this instance suggests that within the set of auxiliary variables there is negligible association with both the underlying values of the variable J2Work and the chance of values being missing (since all prisoners in Sample2 were serving sentences of at least 18 months, there is no longer a clear association between sentence length and whether a prisoner works – all longer term prisoners are eligible to work in prison). Assuming J2Work is missing at random given the other variables in the imputation model, the consequence of this is that the proportion saying they undertook paid worked is similar among those whose values are observed and those whose values are missing.

Close correspondence between observed and imputed results will happen when the simultaneous associations between the auxiliary variables and both the underlying value of the partially observed variable and the chance of a missing value occurring is small. However, it is also possible that, even when no single auxiliary variable displays such associations, a particular combination of the variables may do so, and in such circumstances the adjustment through MI may still be non-trivial, but its occurrence will be harder to predict. Hence to properly gauge the impact of MI the full procedure still needs to be performed. Note that auxiliary variables that are predictive of the outcome, but not missingness, still have a role to play, but in terms of improving the precision of the estimate rather than in correcting any bias.

4.2 Imputation for inferential analyses with incomplete independent variables

MI procedures for missing data scenarios where data are fully observed on a dependent variable of interest, but partially observed for some (or all) explanatory variables are now outlined. In this multivariate regression setting, if an explanatory variable is MAR given the response, and other covariates, then the complete case analysis can be biased (see Chapter 3). Since in such regression models the explanatory variables are usually associated with the dependent variable, and it is also often plausible that the *chance* of observing the explanatory variable is associated with the dependent variable, then the MI analysis under MAR will quite often differ from the complete case analysis.

The analyses outlined below all aim to examine whether experiences within prison (involvement in educational training and paid employment) and on release (paid employment), which are considered to be explanatory variables, are associated with the dependent variable, proven re-offending within a year of release. Proven re-offending is a binary variable derived from Police National Computer (PNC) records. Prisoners who did not receive a subsequent conviction or caution within a year of release were coded 0, and those who did were coded 1. In practice, the imputation procedures apply equally when the outcome is measured on a continuous scale (for example, rate of offending or the Copas rate), with the only difference being the use of a linear, rather than a logistic, regression model for the substantive model of interest.

4.2.1 The association between involvement in educational training programmes in prison and proven re-offending within 1 year of release (Sample2)

In the first example the association between participation in educational training programmes whilst in prison and the propensity to reoffend within a year of release, taking into account other factors which are related to reoffending are examined. This analysis uses data from Sample2 – those prisoners serving sentences between 18 months and 4 years.

An illustrative selection of explanatory variables known to be related to offending, measured at Wave 1 of SPCR (and hence fully observed), were included in the analysis (see Table 4.5 for summary details). These explanatory variables include gender, whether they were from a Black, Asian or minority ethnic (BAME) background, a centred age variable, along with information on the offence for which they were imprisoned (distinguishing those who committed burglary and theft), whether they had served any prior prison sentences, and whether they reported having a serious drug dependency on entry to prison - defined as being a daily heroin or crack cocaine user. (Further additional independent variables

measured at Wave 1 could also be included in these models if required.) A small degree of item missing data is evident across these variables (11 cases), reducing the sample to 2,403. The explanatory variable measured at Wave 2 identifies whether a prisoner was involved in an educational training programme during their sentence, but information is missing for 818 respondents who were not successfully interviewed at Wave 2.

Table 4.5 Descriptive statistics (SPCR Sample2)

Wave 1 and PNC	Frequency	%
Reconvicted within 1 year	556	23.1
Black, Asian and minority ethnic (BAME)	439	18.3
Female	397	16.5
SPCR burglary offence	316	13.2
SPCR theft offence	180	7.5
Prior prison sentence	1,248	51.9
Serious drug dependency (daily user)	444	18.5
	Mean	**S.D**
Age (centred)	0.55	10.2
Sample size	2,403	

Wave 2	Frequency	%
Participated in education course during SPCR sentence	618	39.0
Sample size	1,585	

Table 4.6 includes estimates from three models predicting reoffending. Model 1 includes estimates from the model restricted to Wave 1 data, with the sample size including all eligible respondents from Sample 2. The results from this model indicated that BAME prisoners were significantly less likely to reoffend within a year of release, conditional on other included explanatory variables. Similarly, the odds of reoffending fell as age increased, and were lower for women. In contrast the odds of proven re-offending within a year of release were significantly higher for those who served a sentence for theft or burglary, those who had served prior prison sentences, and those identified as having a drug dependency prior to custody.

Model 2 extended the analysis to incorporate the measure of participation in education training observed at Wave 2. Based on complete case analysis (the default when no adjustment is made for missing data), the sample size for the full analysis dropped to 1,585. Here information had been lost from the 818 respondents that were not interviewed at Wave 2, but who contributed to the Wave 1 only analysis. The standard errors from the model have been inflated as a result of the reduced sample size reducing the precision of results. A number of parameter estimates have also been altered a small extent in the revised model, with the effect of ethnicity no longer identified as significant (reflecting the combined effect of a weaker relationship and reduced precision). There have also been changes in the

magnitude of the odds of reoffending amongst female prisoners and those who were imprisoned for theft offences. The model shows lower odds of reoffending amongst those enrolled on an educational training course during their prison sentence (additional Wave 2 variable of interest).

Table 4.6 Predicting reoffending of SPCR prisoners using data from Wave 1 and Wave 2 (Sample2)

	Model 1: Complete cases (Wave 1)				Model 2: Complete cases (Wave 1 and Wave 2)				Model 3: MI Results			
	B	S.E	Odds Ratio	Sig	B	S.E	Odds Ratio	Sig	B	S.E	Odds Ratio	Sig
BAME	-0.31	0.15	0.74	0.04	-0.24	0.17	0.78	0.16	-0.29	0.15	0.75	0.05
Age (centred)	-0.06	0.01	0.94	<0.01	-0.07	0.01	0.93	<0.01	-0.06	0.01	0.94	<0.01
Female	-0.88	0.18	0.42	<0.01	-1.08	0.24	0.34	<0.01	-0.89	0.18	0.41	<0.01
SPCR burglary offence	0.87	0.14	2.39	<0.01	0.85	0.17	2.34	<0.01	0.87	0.14	2.39	<0.01
SPCR theft offence	0.86	0.18	2.35	<0.01	1.02	0.22	2.78	<0.01	0.84	0.18	2.32	<0.01
Prior prison sentence	0.99	0.12	2.68	<0.01	0.97	0.15	2.63	<0.01	0.97	0.12	2.65	<0.01
Serious drug dependency (daily user)	0.52	0.13	1.69	<0.01	0.53	0.15	1.70	<0.01	0.54	0.13	1.71	<0.01
Participated in education course during SPCR sentence					-0.26	0.13	0.77	0.05	-0.27	0.13	0.76	0.04
(constant)	-2.02	0.10			-1.88	0.14			-1.92	0.12		
Sample size	2,403				1,585				2,403			

The imputation model must include all variables in the substantive model of interest including the fully observed outcome, but also needs to incorporate the list of auxiliary variables linked to the process of missing data at Wave 2 (outlined in Appendix I).

Model 3 (Table 4.6) includes details of the combined estimates from the imputed datasets. The sample size in the imputed datasets has returned to 2,403 confirming the inclusion in the model of all the respondents that were not re-interviewed at Wave 2. As a result of the increased sample size, the standard errors in the imputed model return to the same magnitude as Model 1. The lower odds of reoffending amongst ethnic minority prisoners is again identified as significant, and of similar magnitude to the model restricted to Wave 1 variables. Whilst the substantive interpretation of the remaining variables has remained consistent across the three models, examination of the regression coefficients (B) also reveals that the imputed analysis now more closely resembles Model 1, with the size of the effect of gender and serving a sentence for a theft offence changing most noticeably. The imputed model still demonstrates the important association between educational training and the odds of proven re-offending. While the substantive conclusions from models 2 and 3 are broadly the same in this case, the advantage of MI is clearly seen in (i) the recovery of the information on the Wave 1 variables which increases the precision of the estimates and also in particular (ii) the confirmation of the lower odds of reconviction amongst ethnic minorities when the Wave 2 variables are included.

Precision in the Wave 1 variables is primarily being recovered because information from all those with Wave 1 data observed but Wave 2 data missing were brought back into the analysis. The changes to the magnitude of the effect of gender and theft suggests that the relationship between these variables and the outcome differed between the full sample used in Model 1 and the reduced sample in Model 2. The MI model accounts for this unrepresentativeness, with the results in Model 3 returning to the magnitude observed in Model 1. The coefficient from the Wave 2 variable is similar in models 2 and 3, suggesting that given the other variables in the model the chance of this being missing is not strongly associated with the dependent variable (reoffending).

As an initial check on the data descriptive analyses were performed on the imputed data and compared to the equivalent observed data. Any noticeable differences in the distributions might indicate potential problems with the imputation models. In this example the imputations 1, 22 and 35 were examined and the results shown in Table 4.7. The three imputations closely correspond to the observed data.

Table 4.7 Selected imputations (SPCR Sample 2)

Did you participate in any educational training programmes during your sentence?	Interviewed at Wave 2	_1_EducCo	_22_EducCo	_35_EducCo
Yes	39.0%	35.9%	38.4%	40.1%
No	61.0%	64.1%	61.6%	59.9%
Sample size	1,585	818	818	818

Sensitivity analyses

Initial exploration of missing data in SPCR identified two distinct types of nonresponse: non-contact, and successful contact but refusal. The current imputation method uses the same imputation model, whatever the type of nonresponse. This could be misleading, with the potential for different relationships between variables for those who are easy to contact, those who are hard to contact, and those who refuse to respond when contact. To explore the sensitivity of the current results to the assumption that the same imputation model applies to all respondents (or that the MAR assumption is equally applicable to all groups), two related strategies were employed.

The first approach was to develop separate imputation models for non-contact and noncompliance and combine the resulting datasets. (Further details of the method and the Stata code necessary to undertake this sensitivity analysis is given in Appendix II.) Table 4.8 includes estimates from the revised imputation, shown as Model 4. The results from this model are very similar to those from the combined model which indicates that whilst it was important to distinguish between the sources of missing data when attempting to understand the mechanisms underpinning missingness, in this case it is appropriate to impute on all sources of missing data simultaneously.

Table 4.8 Sensitivity of MI to alternative assumptions (SPCR Sample2)

	Model 4: Sensitivity 1				Model 5: Sensitivity 2 (ease of contact)				Model 6: Sensitivity 2 (refusal)			
	B	S.E	Odds Ratio	Sig	B	S.E	Odds Ratio	Sig	B	S.E	Odds Ratio	Sig
BAME	-0.29	0.15	0.75	0.05	-0.29	0.15	0.75	0.05	-0.29	0.15	0.75	0.05
Age (centred)	-0.06	0.01	0.94	<0.01	-0.06	0.01	0.94	<0.01	-0.06	0.01	0.94	<0.01
Female	-0.89	0.18	0.41	<0.01	-0.89	0.18	0.41	<0.01	-0.89	0.18	0.41	<0.01
SPCR burglary offence	0.87	0.14	2.39	<0.01	0.87	0.14	2.39	<0.01	0.87	0.14	2.39	<0.01
SPCR theft offence	0.84	0.18	2.32	<0.01	0.84	0.18	2.32	<0.01	0.84	0.18	2.32	<0.01
Prior prison sentence	0.97	0.12	2.65	<0.01	0.97	0.12	2.64	<0.01	0.97	0.12	2.65	<0.01
Serious drug dependency (daily user)	0.54	0.13	1.71	<0.01	0.54	0.13	1.71	<0.01	0.53	0.13	1.71	<0.01
Participated in education course during SPCR sentence	-0.28	0.13	0.75	0.03	-0.31	0.14	0.73	0.03	-0.27	0.13	0.77	0.04
(constant)	-1.91	0.12			-1.90	0.12			-1.92	0.11		
Sample size	2,403				2,403				2,403			

The second approach was to examine whether the relationships between the auxiliary variables and missing data differed based on a respondent's 'ease of contact'. The model predicting the probability of non-contact (first introduced in Chapter 2, Table 2.4) was used to stratify respondents at Wave 1 into three groups (easy, medium, and hard to contact). Multiple imputation was then performed separately in each of these three groups, thus allowing a full interaction of the auxiliary variables with 'ease of contact' (see Appendix II). These imputed datasets were then merged, with the model of interest then estimated on each imputed dataset as before. This approach is in line with recently published guidance on the use and reporting of multiple imputation (Sterne et al, 2009).

The results are presented at Model 5 in Table 4.8. The same approach was used to explore the effect of allowing for possible differential relationships between the auxiliary variables and missing data among those who were unlikely and likely to refuse to be interviewed (first introduced in Chapter 2, Table 2.5). The results are presented as Model 6 in Table 4.8. The results from Models 5 and 6 also show close correspondence with the original imputed results, providing further confirmation that the use of a single imputation model for all cases is appropriate.

4.2.2 The association between undertaking paid employment in prison and proven re-offending within 1 year of release (Sample1)

To demonstrate MI using data from Sample1 – the representative sample of the majority of receptions to prison – the relationship between undertaking paid employment whilst in prison and reoffending within a year of release was explored. Whether a prisoner had undertaken paid work was asked of all Sample1 respondents (with sub-sample Sample1combW12 asked as part of their Wave 1 interview). As the question was asked of every member of Sample1, conceptually, the variable was amenable for imputation. The same set of Wave 1 explanatory variables, which were adopted in the previous example, were entered into the model here. Descriptive statistics are presented in Table 4.9.

Table 4.9 Descriptive statistics (SPCR Sample1)

Wave 1 and PNC	Frequency	%
Reconvicted within 1 year	708	49.7
Black, Asian and minority ethnic (BAME)	223	15.7
Female	131	9.2
SPCR burglary offence	98	6.9
SPCR theft offence	317	22.3
Prior prison sentence	1,009	70.8
Serious drug dependency (daily user)	380	26.7
	Mean	**S.D**
Age (centred)	.09	9.0
Sample size	1,425	

Wave 2	Frequency	%
Participated in paid employment during SPCR sentence	528	49.0
Sample size	1,077	

Table 4.10 includes results from three models. Model 1 is restricted to variables measured at Wave 1, with an analytic sample of 1,425 cases (the remaining 10 cases had item missing data on one or more of the included explanatory variables and were omitted). The odds of reoffending within one year are significantly lower for BAME and older prisoners. In contrast, those serving a sentence for a theft offence, those that had served a prior prison sentence, and those with a Class A drug dependency prior to their sentence had significantly higher odds of reoffending. No significant differences in the likelihood of reoffending were identified based on gender, or amongst those serving their sentence for burglary.

Model 2 includes information on whether respondents worked in prison, with the complete case analysis dropping information from the 349 cases not successfully re-interviewed at Wave 2 or part of Sample1combW12. Some changes to estimates are evident, with weaker effects of serving a prior prison sentence and drug dependency. The odds of reoffending were significantly lower amongst those who participated in employment during their prison sentence.

Table 4.10 Predicting reoffending using data from Wave 1 and Wave 2 (SPCR Sample1)

	Model 1: Complete cases (Wave 1)				Model 2: Complete cases (Wave 1 and Wave 2)				Model 3: MI Results			
	B	S.E	Odds Ratio	Sig	B	S.E	Odds Ratio	Sig	B	S.E	Odds Ratio	Sig
BAME	-0.42	0.17	0.66	0.01	-0.46	0.19	0.63	0.02	-0.43	0.17	0.65	0.01
Age (centred)	-0.03	0.01	0.97	<0.01	-0.04	0.01	0.96	<0.01	-0.03	0.01	0.97	<0.01
Female	-0.18	0.22	0.83	0.40	0.04	0.24	1.04	0.87	-0.11	0.22	0.89	0.60
SPCR burglary offence	0.17	0.23	1.18	0.47	0.24	0.26	1.28	0.34	0.22	0.23	1.24	0.34
SPCR theft offence	0.88	0.16	2.41	<0.01	0.93	0.18	2.54	<0.01	0.88	0.16	2.42	<0.01
Prior prison sentence	1.54	0.15	4.66	<0.01	1.33	0.17	3.76	<0.01	1.55	0.15	4.70	<0.01
Serious drug dependency (daily user)	0.66	0.14	1.93	<0.01	0.52	0.16	1.68	<0.01	0.65	0.15	1.91	<0.01
Participated in paid employment during prison sentence					-0.41	0.14	0.67	<0.01	-0.31	0.13	0.73	<0.01
(constant)	-1.42	0.13			-1.03	0.16			-1.27	0.15		
Sample size	1,425				1,077				1,425			

Model 3 presents the results based on 40 imputed datasets, with the sample size and standard errors now reflecting the use of all sample members. Following MI, the estimated associations between reoffending and both prior prison sentence and Class A drug dependency were of a similar magnitude to the Wave 1 only analysis. The estimated association between work in prison and reoffending reduced in magnitude in the revised model. This was consistent with the changes evident in the descriptive models (see section 4.1), with the complete case analysis giving too much weight to the data from those in 'Sample1combW12', who were shown to be significantly less likely to work in prison. The current models suggest that the association between employment in prison and subsequent reoffending was also different for this group.

Sensitivity analyses

The same series of sensitivity analyses outlined in section 4.2.1 were used to assess the assumption that the imputation model was consistent across non-contact and noncompliance. Model 4 in Table 4.11 used separate imputation models for non-contact and noncompliance. Model 5 allowed for a differential relationship between auxiliary variables and missing data based on sample members 'ease of contact'. Model 6 allowed for differential relationships amongst those that were likely and unlikely to refuse to be interviewed. Like the analyses restricted to Sample2, these alternative models revealed close correspondence across the different MI approaches, giving additional confidence that the assumptions underpinning the imputation models – that data are MAR given the included auxiliary variables – are plausible at Wave 2.

Table 4.11 Sensitivity of MI to alternative assumptions (SPCR Sample1)

	Model 4: Sensitivity 1				Model 5: Sensitivity 2 (ease of contact)				Model 6: Sensitivity 2 (refusal)			
	B	S.E	Odds Ratio	Sig	B	S.E	Odds Ratio	Sig	B	S.E	Odds Ratio	Sig
BAME	-0.44	0.17	0.65	0.01	-0.44	0.17	0.64	0.01	-0.43	0.17	0.65	0.01
Age (centred)	-0.03	0.01	0.97	<0.01	-0.03	0.01	0.97	<0.01	-0.03	0.01	0.97	<0.01
Female	-0.11	0.22	0.89	0.61	-0.11	0.22	0.90	0.62	-0.13	0.22	0.88	0.56
SPCR burglary offence	0.22	0.23	1.24	0.35	0.22	0.23	1.24	0.35	0.20	0.23	1.23	0.38
SPCR theft offence	0.88	0.16	2.41	<0.01	0.88	0.16	2.41	<0.01	0.88	0.16	2.40	<0.01
Prior prison sentence	1.55	0.15	4.71	<0.01	1.55	0.15	4.69	<0.01	1.55	0.15	4.69	<0.01
Serious drug dependency (daily user)	0.65	0.15	1.91	<0.01	0.64	0.15	1.91	<0.01	0.65	0.15	1.92	<0.01
Participated in paid employment during prison sentence	-0.32	0.13	0.73	0.02	-0.32	0.15	0.73	0.04	-0.26	0.14	0.77	0.06
(constant)	-1.26	0.15			-1.27	0.15			-1.29	0.15		
Sample size	1,425				1,425				1,425			

4.2.3 The association between employment status in the two months following release from prison and proven re-offending within 1 year (Sample1)

To illustrate the use of MI to adjust for missing data in a variable measured at Wave 3, the association between undertaking paid employment at any time since release from prison and subsequent reoffending was explored using Sample1 data. An attempt was made to interview all prisoners at Wave 3, therefore it was not necessary to distinguish between the subsamples Sample1combW12 and Sample2sepW12. As a result, the procedures to adjust for missing data in a variable measured at Wave 3 using Sample2 would be identical to those outlined below (the imputation model would simply be restricted to Sample2 before undertaking MI).

An illustrative selection of explanatory variables measured at Wave 1 of SPCR that were identified as predictors of reoffending were included in the analysis (see Table 4.12 for summary details). A small number of cases had missing data on one or more of these variables (n=3), reducing the useable sample to 1,432. The explanatory variable measured at Wave 3 identifies whether a prisoner had undertaken any paid work since being released from prison, and is missing for all 622 respondents from Sample1 that were not successfully interviewed at Wave 3.

Table 4.12 Descriptive statistics (Sample 1)

Wave 1 and PNC	**Frequency**	**%**
Reconvicted within 1 year	713	49.8
Black, Asian and minority ethnic (BAME)	223	15.6
SPCR theft offence	319	22.3
SPCR breach	180	12.6
SPCR motoring offence	255	17.8
Prior prison sentence	1,015	70.9
Serious drug dependency (daily user)	383	26.8
	Mean	**S.D**
Age (centred)	0.10	9.0
Sample size	1,432	

Wave 3	**Frequency**	**%**
In paid employment since release	227	28.0
Sample size	810	

The MI procedure is identical to the examples used to recover data from Wave 2. The key difference is that the list of auxiliary variables associated with missing data at Wave 3 replaced those associated with missing data at Wave 2. The model of interest was also adjusted to reflect the new list of explanatory variables.

Table 4.13 includes estimates from the three models predicting reoffending using Sample1. Model 1 was restricted to Wave 1 data, Model 2 included employment status at Wave 3, and Model 3 included the MI estimates.

Table 4.13 Predicting reoffending using data from Wave 1 and Wave 3 (SPCR Sample1)

	Model 1: Complete cases (Wave 1)				Model 2: Complete cases (Wave 1 and Wave 3)				Model 3: MI Results			
	B	S.E	Odds Ratio	Sig	B	S.E	Odds Ratio	Sig	B	S.E	Odds Ratio	Sig
BAME	-0.40	0.17	0.67	0.02	-0.43	0.24	0.65	0.07	-0.41	0.17	0.67	0.02
Age (centred)	-0.03	0.01	0.97	<0.01	-0.04	0.01	0.96	<0.01	-0.03	0.01	0.97	<0.01
SPCR theft offence	1.04	0.16	2.83	0.00	0.87	0.22	2.39	0.00	1.03	0.16	2.81	0.00
SPCR breach	0.69	0.19	2.00	0.00	0.39	0.25	1.48	0.12	0.68	0.19	1.98	0.00
SPCR motoring offence	0.33	0.16	1.39	0.04	0.34	0.22	1.40	0.13	0.32	0.16	1.38	0.04
Prior prison sentence	1.50	0.15	4.50	0.00	1.52	0.19	4.59	0.00	1.47	0.15	4.34	0.00
Serious drug dependency (daily user)	0.63	0.14	1.88	0.00	0.58	0.20	1.78	0.00	0.57	0.15	1.76	0.00
In paid employment since release					-0.45	0.18	0.64	0.01	-0.40	0.18	0.67	0.02
(constant)	-1.58	0.14			-1.24	0.19			-1.42	0.15		
Sample size	1,432				810				1,432			

The results from Model 1 indicated that the odds of reoffending within a year of release were significantly lower for prisoners from BAME backgrounds, conditional on other included explanatory variables. Similarly, the odds of reoffending fell as age increased. In contrast, the odds of reoffending were significantly higher for those who served their SPCR sentence for theft, a breach of a former sentence, or a motoring offence, those who had served a prior prison sentence, and those identified as having a serious drug dependency prior to incarceration.

The inclusion of employment status from Wave 3 identified significantly lower odds of reoffending amongst those prisoners who were able to secure some form of paid employment in the months since release. Including this Wave 3 variable also significantly reduced the useable sample, with the results now based on the 810 respondents that were successfully interviewed in the community. This reduction in sample size led to an increase in standard errors for all explanatory variables included in the model from Wave 1, and also some changes to effect size estimates. In particular, the odds of reoffending amongst those imprisoned for a theft offence or breach were weaker than in the previous model, with the effect of a breach no longer identified as significant. The higher odds of reoffending amongst those serving a sentence for a motoring offence also failed to reach significance in the revised model.

Model 3 includes the MI results, with the sample size returning to the original sample of 1,432 respondents. Estimates and standard errors for the variables measured at Wave 1 all return to a similar magnitude as Model 1. As a result, the higher odds of reoffending amongst those sentenced for a motoring offence or breach of previous sentence were again identified as significant.

The largest changes following MI were seen when considering the effects of being imprisoned for a theft offence and breach of previous sentence. This can be explained because the sample used for Model 2 differed with respect to the relationship between these variables and the outcome (reoffending) from the original sample used for Model 1. This implied an association between missingness and these variables, which was also strongly associated with the outcome in the original sample. The MI analysis for Model 3 took account of this to correct for the unrepresentativeness of the dataset used for Model 2, and so the estimate and associated inference, for the effects of being imprisoned for theft or breach were close, under Model 3, to that seen in Model 1. The additional variable, employment since release, was not strongly associated with any of the included variables from Wave 1, and so the addition of employment status to the model did not have any additional impact on

the estimate of these effects. In more complex settings, where variables are more strongly interrelated, it can be difficult to predict the consequence of simultaneously adding variables to a model, and using MI.

To check the appropriateness of the imputations, comparisons were made between the imputed results and the observed data for imputations 1, 22 and 35. The results are presented in Table 4.14. This demonstrates a high degree of consistency, with the observed data for the three imputations displaying plausible estimates.

Table 4.14 Selected imputations (Sample1)

Have you undertaken paid work since being released?	Interviewed at Wave 3	_1_R3Job	_22_R3Job	_35_R3job
Yes	28.0%	25.9%	28.3%	26.0%
No	72.0%	74.1%	71.7%	74.0%
Sample size	810	622	622	622

Sensitivity analyses

The same sensitivity strategy as adopted previously was used here to explore the appropriateness of the missing data assumptions at Wave 3. Here three sources of missing data that reflect the data collection process must be incorporated: non-contact and noncompliance amongst the sample of respondents re-interviewed in the community, and nonresponse amongst those back in prison at the time of re-interview (recall that no details were available to distinguish between non-contact and noncompliance for this group). (Further details and the Stata code are given in Appendix II.)

Using a separate imputation model for non-contact, noncompliance, and nonresponse amongst those back in prison (presented as Model 4 in Table 4.15) shows a high degree of correspondence with the overall imputation model. The results from imputation models that allow for a full interaction with ‘ease of contact’ (Model 5), being unlikely or likely to refuse to be interviewed (Model 6), and being unlikely or likely to respond if back in prison (Model 7) are also very similar to the simple imputation model. This gives additional confidence in the plausibility of the missing data assumptions underpinning the MI models.

Table 4.15 Sensitivity of MI to alternative assumptions (Sample1)

	Model 4: Sensitivity 1			
	B	S.E	Odds Ratio	Sig
BAME	-0.41	0.17	0.67	0.02
Age (centred)	-0.03	0.01	0.97	<0.01
SPCR Theft offence	1.04	0.16	2.83	<0.01
SPCR Breach	0.68	0.19	1.98	<0.01
SPCR motoring offence	0.32	0.16	1.38	0.05
Prior prison sentence	1.47	0.15	4.36	<0.01
Serious drug dependency (daily user)	0.57	0.15	1.76	<0.01
In paid employment since release	-0.41	0.18	0.66	0.02
(constant)	-1.43	0.15		
Sample size	1,432			

Table 4.15 cont'd Sensitivity of MI to alternative assumptions (Sample1)

	Model 5: Sensitivity 2 (ease of contact)				Model 6: Sensitivity 2 (refusal)				Model 7: Sensitivity 2 (prison nonresponse)			
	B	S.E	Odds Ratio	Sig	B	S.E	Odds Ratio	Sig	B	S.E	Odds Ratio	Sig
BAME	-0.41	0.17	0.67	0.02	-0.41	0.17	0.66	0.02	-0.41	0.17	0.66	0.02
Age (centred)	-0.03	0.01	0.97	<0.01	-0.04	0.01	0.97	<0.01	-0.03	0.01	0.97	<0.01
SPCR Theft offence	1.03	0.16	2.81	<0.01	1.03	0.16	2.81	<0.01	1.03	0.16	2.81	<0.01
SPCR Breach	0.70	0.19	2.01	<0.01	0.69	0.19	1.99	<0.01	0.68	0.19	1.98	<0.01
SPCR motoring offence	0.32	0.16	1.38	0.05	0.32	0.16	1.38	0.04	0.32	0.16	1.38	0.05
Prior prison sentence	1.47	0.15	4.35	<0.01	1.47	0.15	4.36	<0.01	1.47	0.15	4.36	<0.01
Serious drug dependency (daily user)	0.57	0.15	1.77	<0.01	0.57	0.15	1.76	<0.01	0.57	0.15	1.77	<0.01
In paid employment since release	-0.41	0.19	0.66	0.03	-0.42	0.17	0.66	0.02	-0.40	0.18	0.67	0.02
(constant)	-1.42	0.15			-1.42	0.15			-1.43	0.15		
Sample size	1,432				1,432				1,432			

4.3 Imputation for inferential analyses with an incomplete dependent variable

In some instances, researchers may be interested in performing regression analyses on an incomplete response variable. The procedures for conducting this form of MI are equivalent to those above, with the imputation model making no distinction between response and explanatory variables. When data is only missing from the dependent variable, if the response is MAR given the covariates in the model, then the complete case analysis is valid but can be inefficient (Carpenter and Kenward, 2013). However, there would be a gain from an analysis using MI if auxiliary variables (not explanatory variables in the model of interest) predict both the partially observed response and the chance of this being missing.

4.3.1 Identifying the factors associated with an increased propensity to secure employment on release from prison (Sample2)

To illustrate this form of missing data adjustment, the factors associated with an increased propensity of securing employment on release from prison (as measured at the Wave 3 interview) were identified. For simplicity, it is assumed that the explanatory variables are fully observed, although the imputation procedures are the same when there is also missing data on explanatory variables. Sample2 is used for the current example – those prisoners serving sentences between 18 months and 4 years.

Table 4.16 includes summary details of the dependent and explanatory variables included in the analysis. An illustrative selection of explanatory variables measured at Wave 1 is included; ethnicity, a centred age variable, along with information on whether the prisoner was a regular user of crack cocaine or heroin prior to their sentence, or identified themselves as having a long term limiting illness. Information is also included on whether they reported working in the four weeks prior to their sentence. Amongst these variables there is a small degree of item missing data (n=19).

Table 4.16 Descriptive statistics (Sample2)

Wave 1	**Frequency**	**%**
Black, Asian or minority ethnic (BAME)	438	18.3
Serious drug dependency (daily user)	441	18.4
Worked in 4 weeks prior to sentence	855	35.7
Long term limiting illness	713	29.8
	Mean	**S.D**
Age (centred)	0.55	10.2
Sample size	2,395	

Wave 3	**Frequency**	**%**
In paid employment since release	442	30.4
Sample size	1,454	

Table 4.17 includes estimates from two models predicting whether a respondent reported being in paid employment at any time since release from prison. Model 1 includes estimates from a model restricted to complete cases, with data lost on 941 respondents from Wave 1 that were not successfully re-interviewed at Wave 3. This identifies significantly lower odds of being in employment after release amongst BAME respondents, those who reported that they were a daily user of crack cocaine or heroin prior to their sentence, and those who identified themselves as having a long-term limiting illness. In contrast, those who were in work in the 4 weeks prior to their prison sentence had significantly higher odds of being in employment again on release

Table 4.17 Predicting re-employment on release from prison (Sample2)

	Model 1: Complete cases (Wave 1 and Wave 3)				Model 2: MI Results			
	B	S.E	Odds Ratio	Sig	B	S.E	Odds Ratio	Sig
BAME	-0.36	0.17	0.70	0.04	-0.36	0.17	0.70	0.04
Age	-0.03	0.01	0.97	<0.01	-0.03	0.01	0.97	<0.01
Serious drug dependency (daily user)	-0.47	0.18	0.63	0.01	-0.49	0.18	0.61	0.01
Worked in 4 weeks prior to sentence	1.36	0.13	3.89	<0.01	1.32	0.13	3.75	<0.01
Long term limiting illness	-0.58	0.15	0.56	<0.01	-0.59	0.14	0.56	<0.01
(constant)	-1.12	0.10			-1.10	0.10		
Sample size	1,454				2,395			

Model 2 includes the MI results. This model identifies a high degree of correspondence with the non-imputed results, all being of a similar magnitude. The very small difference between the results from the completers and MI analyses, both in terms of estimates and standard errors, indicated that there were no auxiliary variables or combinations of auxiliary variables that were strongly associated with both the unobserved values on the dependent variable and the chances of being missing. However, although both analyses lead to similar results, this is only known through doing both analyses, so the MI analysis retains its value even when showing a negligible impact on the results of the completers analysis.

5. Conclusions

When faced with missing data, whatever the research context, the first objective is to understand the nature and complexity of the problem and above all the reasons for its occurrence. Exploration of the problem entails analysing the dataset to identify gaps and in the case of SPCR this revealed that the main concern was unit nonresponse – there was little item nonresponse. But to fully diagnose, and hence be in a position to effectively address the problem, it will be necessary to appraise all aspects of the processes that generated the data; the organisation of the fieldwork, the time and place of interviews, the definitions used etc. This information will reveal complexities but also hopefully patterns and explanations. It was only through discussions with Ipsos MORI that the structural issues in obtaining interviews became clear. With this background knowledge it was possible to identify factors that were associated with, and predictive of, missingness.

Variables predictive of missingness, it was shown, can be used as auxiliary variables to impute missing data. Here it was assumed that the data was missing at random, an assumption plausible for SPCR both because there was comparatively little bias when examining the distribution of Wave 1 variables across the observed data at each Wave, and because missingness could be directly related to variables reflecting the process of data collection, with few respondents actively refusing to be interviewed.

Unfortunately, however, it is not possible to distinguish between missing at random and missing not at random based on the observed data and we emphasise the need to perform sensitivity analyses to examine the effects of different assumptions. Approaches to sensitivity analysis are presented in the report. It is important to be clear which dataset will form the basis of any analysis. In SPCR the organisation of the fieldwork meant that some prisoners could not refuse to opt out at Wave 2 as they were asked a selection of Wave 2 questions at Wave 1. Thus it makes no conceptual sense to include this group when imputing Wave 2 missing data.

The next decision is the choice of imputation method and of the alternatives multiple imputation (MI) was judged to be the most attractive for analysing the SPCR data, because it is straightforward to use and is readily available in statistical software. Furthermore, once the missing data have been imputed, the imputed data sets can be analysed as would have been done if no data were missing.

MI is a procedure that yields appropriate standard errors for parameter estimates relatively simply in a wide range of settings, encompassing all those we anticipate in analysis of the SPCR data. If anything, given the MAR assumption, inferences from MI are conservative. Once the imputation model is chosen, MI proceeds automatically. The key is thus appropriate specification of the imputation model. This should (i) be consistent (also known as congenial) with the research model of interest and (ii) appropriately incorporate relevant auxiliary variables.

Multiple imputation should not simply be undertaken routinely but its adoption should proceed carefully and appropriately. In a recent paper, Sterne and his colleagues (Sterne *et al*, 2009 - see also Carpenter and Kenward, 2013) set out some of the pitfalls that may occur and thus which analysts should seek to avoid. The authors point out that most problems arise because of inappropriate choice of imputation model. In particular, the imputation model needs to contain all the variables in the substantive model of interest, including the dependent variable. Too often the analyst only explores the association between explanatory variables and the dependent variable. But often, however, one or more of the explanatory variables themselves contain missing values and the dependent variable may contain information about the missing data on the explanatory variable. In order to benefit from this relationship, the dependent variable should be incorporated in the imputation model.

It is important to note that missing at random is an assumption and ‘not a property of the data’. For this assumption to be reasonable the imputation model needs to include all the variables that are predictive of missingness. If any of the important auxiliary variables are omitted this will make the MAR assumption less tenable. This report points out that it is better to err on the side of caution and include all relevant auxiliary variables (as well as all variables in the substantive model).

The above two pitfalls are concerned with the specification of the imputation model. There is however, another pitfall that arises with handling non-normally distributed variables, such as binary or categorical data. As many of the variables in SPCR are not normally distributed this is a potential problem when conducting multiple imputations for this dataset. It is to overcome these problems that the ICE routine in Stata was recommended and used, as ICE automatically recognises the variable type and conducts the analysis appropriately. If other routines in Stata or other statistical software are employed to perform MI, then this may be a potential problem that will need to be addressed.

References

Boorman, R., and Hopkins, K. (2012) Prisoners' criminal backgrounds and proven re-offending after release: results from the Surveying Prisoner Crime Reduction (SPCR) survey. Ministry of Justice Research Summary 8/12.

Brunton-Smith, I., Carpenter, J. R., Kenward, M. G. and Tarling, R. (2014) Multiple Imputation for handling missing data in social research. *Social Research Update*, 65.

Brunton-Smith, I. and Hopkins, K. (2013) The factors associated with proven re-offending following release from prison: findings from Waves 1 to 3 of SPCR. Results from the Surveying Prisoner Crime Reduction (SPCR) longitudinal cohort study of prisoners. Ministry of Justice Analytical Series 2013

Brunton-Smith, I. and Hopkins, K. (2014) The impact of experience in prison on the employment status of prisoners after release: findings from the first 3 waves of Surveying Prisoner Crime Reduction (SPCR). Ministry of Justice Analytical Series 2014

Carpenter, J. R. and Kenward M. G. (2013) *Multiple Imputation and Its Application*. Chichester: Wiley.

Carpenter J. R., Kenward, M. G. and Vandsteelandt, S. (2006) A comparison of multiple imputation and doubly robust estimation for analysis with missing data. *Journal of the Royal Statistical Society*, Series A, 169, 571-584.

Carpenter, J. R. and Plewis, I. F. (2010) Analysing longitudinal studies with nonresponse: issues and statistical methods. In: M. Williams and P. Vogt (Eds.) *The Sage Handbook of Methodological Innovation*. London: Sage.

Cleary, A., Ames, A., Kostadintcheva, K., and Muller, H. (2012a) Surveying Prisoner Crime Reduction (SPCR): Wave 1 (Reception) Samples 1 and 2 Technical Report. Ministry of Justice Research Series 5/12.

Cleary, A., Ames, A., Kostadintcheva, K., and Muller, H. (2012b) Surveying Prisoner Crime Reduction (SPCR): Wave 2 (Pre-Release) Samples 1 and 2 Technical Report. Ministry of Justice Research Series 6/12.

Cleary, A., Ames, A., Kostadintcheva, K., and Muller, H. (2014) Surveying Prisoner Crime Reduction (SPCR): Waves 3 and 4 (Post-release) Samples 1 and 2 Technical Report. Ministry of Justice Analytical Series 2014

Cunniffe, C., Van de Kerckhove, R., Williams, K., and Hopkins, K. (2012) Estimating the prevalence of disability amongst prisoners: Results from the Surveying Prisoner Crime Reduction (SPCR) survey. Ministry of Justice Research Summary 4/12.

Diggle, P. J. and Kenward, M. G. (1994) Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, 43, 49-94.

Enders, C. K. (2010) *Applied Missing Data Analysis*. London: The Guilford Press.

Goldstein, H. (2003) *Multilevel Statistical Models (Third Edition)*. London: Arnold.

Hopkins, K. (2012) The pre-custody employment, training and education status of newly sentenced prisoners: Results from the Surveying Prisoner Crime Reduction (SPCR) longitudinal cohort study of prisoners. Ministry of Justice Research Series 3/12.

Hopkins, K., and Brunton-Smith, I. (2014) Prisoners experience of prison and outcomes on release: results from Surveying Prisoner Crime Reduction (SPCR) survey. Ministry of Justice Analytical Series 2014.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Kenward, M. G. and Carpenter, J.R (2007) Multiple Imputation: Current Perspectives. *Statistical Methods in Medical Research*, 16(3): 199-218.

Light, M., Grant, E., and Hopkins, K. (2013) Gender differences in substance misuse and mental health amongst prisoners: Results from the Surveying Prisoner Crime Reduction (SPCR) longitudinal cohort study of prisoners. Ministry of Justice Analytical Series X/13.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data (Second Edition)*. Chichester: Wiley.

Ministry of Justice (2010) Compendium of reoffending statistics and analysis. Ministry of Justice Statistics Bulletin

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*. Chichester: Wiley.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, 63, 581-592.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data.* London: Chapman and Hall.

Scharfstein, D. O., Rotnizky, A. and Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*, 94, 1096-1146.

Sterne, J. A. C., White I. R., Carlin J. B., Spratt M, Royston, P. Kenward, M. G., Wood A. M. and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*, 339, 157-160.

Williams, K., Poyser, J., and Hopkins, K. (2012a) Accommodation, homelessness and reoffending of prisoners: Results from the Surveying Prisoner Crime Reduction (SPCR) survey. Ministry of Justice Research Summary 3/12

Williams, K., Papadopoulou, V., and Booth, N. (2012b) Prisoners' childhood and family backgrounds: Results from the Surveying Prisoner Crime Reduction (SPCR) longitudinal cohort study of prisoners. Ministry of Justice Research Series 4/12

Appendix I:
List of auxiliary variables

Variable name	Description
WAVE 2	
aux_age18to20	Aged 18 to 20 years
aux_SentLE6month	Serving a sentence of less than or equal to 6 months
aux_Sent6month1year	Serving a sentence greater than 6 months, less than or equal to 1 year
aux_Sent1year18month	Serving a sentence greater than 1 year, less than or equal to 18 months
aux_Sent3year4year	Serving a sentence greater than 3 years, less than or equal to 4 years
aux_SPCRtheft	SPCR sentence for theft offence
aux_SPCRdrug	SPCR sentence for drug offence
aux_earlyrelease	Released in the first 12 months of data collection
aux_prevprison	Served a prior prison sentence
aux_priorburglary	Previously convicted of burglary
aux_PrisonNoncontact	In a prison with significantly higher non-contact level
aux_PrisonRefusal	In a prison with significantly higher refusal level
WAVE 3	
aux_age18to20	Aged 18 to 20 years
aux_black	Black ethnic origin
aux_edqual	Has some form of educational qualification
aux_SPCRdrug	SPCR sentence for drug offence
aux_SPCRbreach	SPCR sentence for breach of conditions
aux_SPCRburglary	SPCR sentence for burglary
aux_neverworked	Never worked prior to SPCR
aux_EFL	English a foreign language
aux_livfam	Lived with family prior to SPCR
aux_nogp	not registered with a GP
aux_dailycrack	Had a daily crack cocaine habit prior to SPCR
aux_priorburglary	Previously convicted of burglary
aux_priorrobbery	Previously convicted of robbery
aux_ConsentAddr	Did not consent to address matching
aux_ConsentMatch	Did not consent to DWP/HRC matching
aux_refusalW2	Refused to be interviewed at Wave 2
aux_noncontactW2	Was not successfully contacted at Wave 2

Appendix II:
Stata code for examples in Chapter 4

This section provides details of the full Stata code used to perform all analyses in Chapter 4. Where possible, all variables names from the raw dataset have been retained. All auxiliary variables can be identified with the aux_prefix. The variable Stype identifies which sample is being used (Sample1, Stype=1; Sample2, Stype=2). Qtype is used to identify the subset of Sample 1 prisoners that were only interviewed once in prison ('Sample1combW12', Qtype=2). The variables MissingW2 and MissingW3 identify those individuals that were not successfully re-interviewed at Wave 2 and Wave 3 respectively.

Imputation for descriptive statistics: Sample1

The initial code reduces the dataset to contain only variables used in the imputation models. This includes the variable with missing data (J2Work) which identifies whether a prisoner undertook paid work during their sentence, as well as the series of auxiliary variables associated with missing data at Wave 2 (aux_). Descriptive information is required for Sample1 – representative of all receptions to prison – therefore Sample2 is removed before proceeding.

```
preserve

keep J2Work Stype Qtype MissingW2 aux_SentLE6month aux_Sent6month1year
aux_Sent1year18month aux_Sent3year4year aux_age18to20 aux_SPCRtheft aux_SPCRdrug
aux_earlyrelease aux_prevprison aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal

drop if Stype==2
```

ICE automatically detects binary data, however it needs to be coded 0, 1. The following code generates a new variable (J2Workr) in this format, and removes all item missing. Descriptive data is also produced.

```
gen J2Workr=1 if J2Work==1

replace J2Workr=0 if J2Work==2

drop if MissingW2==0 & J2Work==.

tab J2Work Qtype, m

proportion J2Workr
```

The imputation model includes the variable with missing data and the list of auxiliary variables associated with missing data at Wave 2. Importantly, the imputation model is only estimated for those in Sample1 that were also eligible for interview at Wave 2 (not those in 'Sample1combW12', identified by if Qtype==1). ICE automatically detects that J2Workr is binary and uses the appropriate logistic imputation model (ICE also automatically detects continuous variables). For ordinal variables the prefix o. should be attached to the variable for imputation. For categorical variables, the prefix m. should be used. A total of 40 imputations are requested, and the data from each are combined using Rubin's rules. The results produced are now model based estimates of the proportion of prisoners that reported working during their sentence, therefore it is advisable to retain the confidence intervals to demonstrate uncertainty in the estimates.

```
mi set wide

mi ice J2Workr aux_SentLE6month aux_Sent6month1year aux_Sent1year18month
aux_Sent3year4year aux_age18to20 aux_SPCRtheft aux_SPCRdrug aux_earlyrelease aux_prevprison
aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal if Qtype==1, add(40)

mi estimate: proportion J2Workr
```

As an initial check of the imputation models, the observed data are compared to the imputed data. _mi_miss is automatically produced by ICE and identifies those cases where data has been imputed.

```
tab _1_J2Workr _mi_miss if Qtype==1, column

tab _22_J2Workr _mi_miss if Qtype==1, column

tab _35_J2Workr _mi_miss if Qtype==1, column

restore
```

Imputation for descriptive statistics: Sample2

The procedure when interest is in data from Sample2 – prisoners serving sentences between 18 months and 4 years – is identical, except that data from Sample1 is removed before proceeding.

```
preserve

keep J2Work Stype Qtype MissingW2 aux_SentLE6month aux_Sent6month1year
aux_Sent1year18month aux_Sent3year4year aux_age18to20 aux_SPCRtheft aux_SPCRdrug
aux_earlyrelease aux_prevprison aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal
```

```
drop if Stype==1

gen J2Workr=1 if J2Work==1

replace J2Workr=0 if J2Work==2

drop if MissingW2==0 & J2Work==.

tab J2Work, m

proportion J2Workr
```

The same list of auxiliary variables are used to recover data from Sample2. No prisoners in Sample2 served sentences of less than 18 months, therefore the auxiliary variables identifying prisoners on sentences of less than or equal to 6 months, and more than 6 months but less than or equal to 1 year are not included (more than 1 year but less than or equal to 18 months is retained to capture those prisoners serving exactly 18 months). 40 imputations are produced, and the proportion involved in prison based work programmes is calculated for each 'completed' dataset.

```
mi set wide

mi ice J2Workr aux_SentLE6month aux_Sent6month1year aux_Sent1year18month
aux_Sent3year4year aux_age18to20 aux_SPCRtheft aux_SPCRdrug aux_earlyrelease aux_prevprison
aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal, add(40)

mi estimate: proportion J2Workr
```

The same procedure is used to examine the distribution of the imputed data. Any number of the imputed datasets can be examined.

```
tab _1_J2Workr _mi_miss if Qtype==1, column

tab _22_J2Workr _mi_miss if Qtype==1, column

tab _35_J2Workr _mi_miss if Qtype==1, column

restore
```

The association between involvement in educational training programmes in prison and proven re-offending within 1 year of release (Sample2)

The following code imputes missing data for an explanatory variable measured at Wave 2 that is included in a logistic regression model. The procedure is identical for linear, ordinal and multinomial regression, with only the regression command changing. Data is retained

from the explanatory variable with missing data (Mclass), the remaining explanatory variables with no missing data (non-white, centred_age, female, SPCRburglary, SPCRtheft, prevprison, and daily_drug), the outcome of interest (reoffend_1yr) which has been coded with no missing data, and the auxiliary variables associated with missing data at Wave 2 (aux_). Before proceding with MI, those respondents in Sample1 are removed.

```
preserve

keep Mclass Stype Qtype MissingW2 reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year aux_age18to20 aux_SPCRtheft aux_SPCRdrug aux_earlyrelease aux_prevprison aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal

drop if Stype==1
```

First, any item missing data is removed from the reduced dataset to ensure that ICE does not try to impute these observations (in practice, there is very little item missing data in the SPCR). A variable identifying whether a prisoner was involved in educational training as part of their sentence is constructed (MEducationCourse), with data missing for those prisoners from Sample2 not successfully re-interviewed at Wave 2.

```
gen nonwhite_m=1 if nonwhite==.

gen prevprison_m=1 if prevprison==.

gen daily_drug_m=1 if daily_drug ==.

gen MEducationCourse = 0

replace MEducationCourse =1 if Mclass ==1 | Mclass ==2 | Mclass==3

replace MEducationCourse =. if MissingW2==1

gen MEducationCourse_m=1 if MissingW2==0 & Mclass==.

drop if nonwhite_m==1 | prevprison_m==1 | daily_drug_m==1 | MEducationCourse_m==1
```

Two logistic regression models are produced (with logits and odds ratios estimated), with the first restricted to those explanatory variables measured at Wave 1 (and hence observed for all prisoners). The second includes the Wave 2 variable with missing data.

```
logistic reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug

logit reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug
```

logistic reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug MEducationCourse

logit reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug MEducationCourse

The imputation model includes the variable with missing data, all other variables included in the model of interest, and the full list of auxiliary variables. aux_SPCRtheft and aux_prevprison are not included because these variables were also included in the model of interest. A total of 40 imputations are produced, and then the same logistic regression models are estimated on each 'complete' dataset and combined using Rubin's rules. To request odds ratios, the command , or is included along with mi estimate.

mi set wide

mi ice MEducationCourse reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_Sent1year18month aux_Sent3year4year aux_age18to20 [aux_SPCRtheft] aux_SPCRdrug aux_earlyrelease [aux_prevprison] aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal, add(40)

mi estimate, or: logit reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug MEducationCourse

mi estimate: logit reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug MEducationCourse

A selection of imputed datasets are examined to check the distribution of the imputed values.

tab _1_MEducationCourse _mi_miss, column

tab _22_MEducationCourse _mi_miss, column

tab _35_MEducationCourse _mi_miss, column

restore

Sensitivity analyses

Non-contact and refusal separately

The first sensitivity analysis assesses the extent that the MI results change when data missing as a result of non-contact is imputed separately from data missing as a result of noncompliance. The list of auxiliary variables in each imputation model is restricted to the set of variables identified as associated with the specific form of missing data. Interest is in the extent that involvement in educational training programmes is associated with a different

propensity to reoffend within a year of release from prison, restricted to those prisoners serving sentences of between 18 months and 4 years.

Preserve

keep Mclass Stype Qtype MissingW2 reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year aux_age18to20 aux_SPCRtheft aux_SPCRdrug aux_earlyrelease aux_prevprison aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal contact2 compliance2

drop if Stype==1

All item missing data is first removed from the data, and a binary indicator is created that indicates whether a prisoner was involved in an educational training programme during their sentence.

gen nonwhite_m=1 if nonwhite==.

gen prevprison_m=1 if prevprison==.

gen daily_drug_m=1 if daily_drug ==.

gen MEducationCourse = 0

replace MEducationCourse =1 if Mclass ==1 | Mclass ==2 | Mclass==3

replace MEducationCourse =. if MissingW2==1

gen MEducationCourse_m=1 if MissingW2==0 & Mclass==.

drop if nonwhite_m==1 | prevprison_m==1 | daily_drug_m==1 | MEducationCourse_m==1

Two binary indicators are constructed that identify whether a prisoner was missing because of non-contact (STnoncontact2), or refusal (STrefusal2). In each case, those prisoners with observed data at Wave 2 are also identified and incorporated in the imputation model.

gen STnoncontact2=0

replace STnoncontact2=1 if contact2==0

replace STnoncontact2=1 if (contact2==1 & compliance2==1)

gen STrefusal2=0

replace STrefusal2=1 if compliance2==0

replace STrefusal2=1 if (contact2==1 & compliance2==1)

The same strategy is used to impute the missing data, with the addition of a conditional statement that restricts the imputation of missing data to non-contact or noncompliance respectively. MI is performed twice, with the list of auxiliary variables reflecting the source of missing data. In each case, a total of 40 imputations are requested.

```
mi set wide

mi ice MEducationCourse reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_Sent1year18month aux_Sent3year4year [aux_SPCRtheft] aux_SPCRdrug aux_earlyrelease [aux_prevprison] aux_priorburglary aux_PrisonNoncontact if STnoncontact2==1, add(40)

mi ice MEducationCourse reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_Sent1year18month aux_Sent3year4year aux_age18to20 aux_PrisonRefusal if STrefusal2==1, add(40)
```

The imputation procedure generates 80 imputed datasets (40 each for non-contact and refusal). The first 40 imputed datasets have ‘completed’ data for all respondents that could not be contacted, but data is still identified as missing for those that were successfully contacted, but refused to be interviewed. Imputations 41 to 80 have ‘completed’ data for those respondents that refused to be interviewed, but data is identified as missing for those that could not be contacted. As a result it is necessary to merge the imputed datasets, with the following stata code repeated for all imputed datasets.

```
replace _1_MEducationCourse= _41_MEducationCourse if _1_MEducationCourse==.

…

replace _40_MEducationCourse= _80_MEducationCourse if _40_MEducationCourse==.
```

Finally, the model of interest is estimated as usual, with the results from each ‘complete’ dataset combined using Rubin’s rules. The addition of the command nimputations(40) ensures that the imputed datasets that are used are those with merged data.

```
mi estimate, nimputations(40) or: logit reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug MEducationCourse

mi estimate, nimputations(40): logit reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug MEducationCourse

restore
```

Ease of contact

To examine whether the general imputation model is appropriate across all levels of 'ease of contact', our model predicting contact at Wave 2 introduced in Chapter 2 (Table 2.4) is used to stratify all respondents into three groups based on their probability of being contacted. This is achieved by generating the linear predicted probabilities from the contact model, and grouping respondents into terciles based on their predicted scores (identified with the variable diff_contact). The imputation model can then be estimated separately for each category of diff_contact, with the imputed data from each model combined and analysed as usual.

```
preserve

keep Mclass Stype Qtype MissingW2 reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year aux_age18to20 aux_SPCRtheft aux_SPCRdrug aux_earlyrelease aux_prevprison aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal contact2 compliance2 FinalPrison

drop if Stype==1

xtmelogit contact2 aux_Sent1year18month aux_Sent3year4year aux_SPCRtheft aux_SPCRdrug aux_earlyrelease aux_prevprison aux_priorburglary aux_PrisonNoncontact || FinalPrison:, variance or

predict contact2_lp, xb

pctile pct= contact2_lp, nq(3)

xtile diff_contact = contact2_lp, cutpoints(pct)

tab diff_contact MissingW2
```

All item missing data is first removed from the data, and a binary indicator is created that indicates whether a prisoner was involved in an educational training programme during their sentence.

```
gen nonwhite_m=1 if nonwhite==.

gen prevprison_m=1 if prevprison==.

gen daily_drug_m=1 if daily_drug ==.

gen MEducationCourse = 0

replace MEducationCourse =1 if Mclass ==1 | Mclass ==2 | Mclass==3

replace MEducationCourse =. if MissingW2==1

gen MEducationCourse_m=1 if MissingW2==0 & Mclass==.
```

```
drop if nonwhite_m==1 | prevprison_m==1 | daily_drug_m==1 | MEducationCourse_m==1
```

MI is then conducted separately for each category of diff_contact. A total of 40 imputed datasets are generated for each category, with the three sets of data combined before the model of interest is fitted to each imputation and combined.

```
mi set wide

mi ice MEducationCourse reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft
prevprison daily_drug aux_Sent1year18month aux_Sent3year4year aux_age18to20 [aux_SPCRtheft]
aux_SPCRdrug aux_earlyrelease [aux_prevprison] aux_priorburglary aux_PrisonNoncontact
aux_PrisonRefusal if diff_contact==1, add(40)

mi ice MEducationCourse reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft
prevprison daily_drug aux_Sent1year18month aux_Sent3year4year aux_age18to20 [aux_SPCRtheft]
aux_SPCRdrug aux_earlyrelease [aux_prevprison] aux_priorburglary aux_PrisonNoncontact
aux_PrisonRefusal if diff_contact==2, add(40)

mi ice MEducationCourse reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft
prevprison daily_drug aux_Sent1year18month aux_Sent3year4year aux_age18to20 [aux_SPCRtheft]
aux_SPCRdrug aux_earlyrelease [aux_prevprison] aux_priorburglary aux_PrisonNoncontact
aux_PrisonRefusal if diff_contact==3, add(40)

replace _1_MEducationCourse= _41_MEducationCourse if _1_MEducationCourse==.

…

replace _40_MEducationCourse= _80_MEducationCourse if _40_MEducationCourse==.

replace _1_MEducationCourse= _81_MEducationCourse if _1_MEducationCourse==.

…

replace _40_MEducationCourse= _120_MEducationCourse if _40_MEducationCourse==.


mi estimate, nimputations(40) or: logit reoffend_1yr nonwhite centred_age female SPCRburglary
SPCRtheft prevprison daily_drug MEducationCourse

mi estimate, nimputations(40): logit reoffend_1yr nonwhite centred_age female SPCRburglary
SPCRtheft prevprison daily_drug MEducationCourse

restore
```

Likely or unlikely to refuse

A similar strategy is used to examine difficulty in obtaining compliance.

```
preserve
```

```
keep Mclass Stype Qtype MissingW2 reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year aux_age18to20 aux_SPCRtheft aux_SPCRdrug aux_earlyrelease aux_prevprison aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal contact2 compliance2 FinalPrison

drop if Stype==1
```

Here we use the model predicting compliance given contact to estimate the linear predicted probabilities (Chapter 2, Table 2.5). Given the comparatively low number of refusals, we only distinguish two separate categories: low likelihood of refusal, and high likelihood of refusal.

```
xtmelogit compliance2 aux_age18to20 aux_Sent1year18month aux_Sent3year4year aux_PrisonRefusal if contact2==1 || FinalPrison:, variance or

predict compliance2_lp, xb

pctile pct2= compliance2_lp, nq(2)

xtile diff_refusal = compliance2_lp, cutpoints(pct2)

tab diff_refusal MissingW2
```

All item missing data were removed, and an indicator of involvement in an educational training programme generated.

```
gen nonwhite_m=1 if nonwhite==.

gen prevprison_m=1 if prevprison==.

gen daily_drug_m=1 if daily_drug ==.

gen MEducationCourse = 0

replace MEducationCourse =1 if Mclass ==1 | Mclass ==2 | Mclass==3

replace MEducationCourse =. if MissingW2==1

gen MEducationCourse_m=1 if MissingW2==0 & Mclass==.

drop if nonwhite_m==1 | prevprison_m==1 | daily_drug_m==1 | MEducationCourse_m==1
```

As usual, 40 imputations were requested, with data from each combined.

```
mi set wide

mi ice MEducationCourse reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_Sent1year18month aux_Sent3year4year aux_age18to20 [aux_SPCRtheft]
```

aux_SPCRdrug aux_earlyrelease [aux_prevprison] aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal if diff_refusal==1, add(40)

mi ice MEducationCourse reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_Sent1year18month aux_Sent3year4year aux_age18to20 [aux_SPCRtheft] aux_SPCRdrug aux_earlyrelease [aux_prevprison] aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal if diff_refusal==2, add(40)

replace _1_MEducationCourse= _41_MEducationCourse if _1_MEducationCourse==.

…

replace _40_MEducationCourse= _80_MEducationCourse if _40_MEducationCourse==.

mi estimate, nimputations(40) or: logit reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug MEducationCourse

mi estimate, nimputations(40): logit reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug MEducationCourse

restore

The association between undertaking paid employment in prison and proven re-offending within 1 year of release (Sample1)

The imputation procedure is the same when using Sample1, with the exception that the imputation model is restricted to those prisoners that were eligible for re-interview at Wave 2 (not those in Sample1combW12'). Sample2 are removed before proceeding with the MI.

Preserve

keep J2Work Stype Qtype MissingW2 reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year aux_age18to20 aux_SPCRtheft aux_SPCRdrug aux_earlyrelease aux_prevprison aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal

drop if Stype==2

All item missing data is removed, and a new variable (J2Workr) is constructed that is coded 0, 1. This identifies whether a prisoner undertook paid employment during their sentence.

gen nonwhite_m=1 if nonwhite==.

gen prevprison_m=1 if prevprison==.

gen daily_drug_m=1 if daily_drug ==.

gen J2Workr=1 if J2Work==1

```
replace J2Workr=0 if J2Work==2

gen J2Workr_m=1 if MissingW2==0 & J2Work==.

drop if nonwhite_m==1 | prevprison_m==1 | daily_drug_m==1 | J2Workr_m==1
```

Before running the MI, logistic regression models predicting reoffending using data from Wave 1 only, and Wave 1 and Wave 2 are produced.

```
logistic reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug

logit reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug

logistic reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug J2Workr

logit reoffend_1yr nonwhite Cage female IIS2burglary IIS2theft prevprison dailyHdrug J2Workr
```

The imputation model includes the variable(s) with missing data, all other variables in the model of interest (including the dependent variable), and all auxiliary variables associated with missing data at Wave 2. Crucially, the imputation model is only estimated for those prisoners eligible for re-interview at Wave 2 (not Sample1combW12, identified with if Qtype==1).

```
mi set wide

mi ice J2Workr reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year aux_age18to20 [aux_SPCRtheft] aux_SPCRdrug aux_earlyrelease [aux_prevprison] aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal if Qtype==1, add(40)

mi estimate, or: logit reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug J2Workr

mi estimate: logit reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug J2Workr
```

Descriptive data from specific imputations can also produced to assess the appropriateness of the imputed data, although we do not report these in the example in Chapter 4.

```
tab _1_J2Workr _mi_miss, column

tab _22_J2Workr _mi_miss, column

tab _35_J2Workr _mi_miss, column
```

restore

Sensitivity analyses

Non-contact and refusal separately

The Stata code to assess the sensitivity of results to imputing for non-contact and refusal separately is similar when using Sample1. Here it is also necessary to identify those sample members included in Sample1combW12 and exclude them from the imputation model.

Preserve

keep J2Work Stype Qtype MissingW2 reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year aux_age18to20 aux_SPCRtheft aux_SPCRdrug aux_earlyrelease aux_prevprison aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal

drop if Stype==2

gen nonwhite_m=1 if nonwhite==.

gen prevprison_m=1 if prevprison==.

gen daily_drug_m=1 if daily_drug ==.

gen J2Workr=1 if J2Work==1

replace J2Workr=0 if J2Work==2

gen J2Workr_m=1 if MissingW2==0 & J2Work==.

drop if nonwhite_m==1 | prevprison_m==1 | daily_drug_m==1 | J2Workr_m==1

gen STnoncontact2=0

replace STnoncontact2=1 if contact2==0

replace STnoncontact2=1 if (contact2==1 & compliance2==1)

gen STrefusal2=0

replace STrefusal2=1 if compliance2==0

replace STrefusal2=1 if (contact2==1 & compliance2==1)

mi set wide

mi ice J2Workr reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year aux_age18to20 [aux_SPCRtheft] aux_SPCRdrug aux_earlyrelease [aux_prevprison] aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal if STnoncontact2==1 & Qtype==1, add(40)

```
mi ice J2Workr reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison
daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year
aux_age18to20 [aux_SPCRtheft] aux_SPCRdrug aux_earlyrelease [aux_prevprison]
aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal if STrefusal2==1 & Qtype==1, add(40)

replace _1_ J2Workr = _41_J2Workr if _1_J2Workr==.

…

replace _40_ J2Workr = _80_ J2Workr if _40_ J2Workr ==.

mi estimate, nimputations(40) or: logit reoffend_1yr nonwhite centred_age female SPCRburglary
SPCRtheft prevprison daily_drug J2Workr

mi estimate, nimputations(40): logit reoffend_1yr nonwhite centred_age female SPCRburglary
SPCRtheft prevprison daily_drug J2Workr
```

Ease of contact

To examine whether the general imputation model is appropriate across all levels of 'ease of contact', the same procedure used with Sample2 is adopted (based on the model predicting contact at Wave 2 first introduced in Chapter 2, Table 2.4).

```
Preserve

keep J2Work Stype Qtype MissingW2 reoffend_1yr nonwhite centred_age female SPCRburglary
SPCRtheft prevprison daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month
aux_Sent3year4year aux_age18to20 aux_SPCRtheft aux_SPCRdrug aux_earlyrelease aux_prevprison
aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal

drop if Stype==2

xtmelogit contact2 aux_Sent1year18month aux_Sent3year4year aux_SPCRtheft aux_SPCRdrug
aux_earlyrelease aux_prevprison aux_priorburglary aux_PrisonNoncontact || FinalPrison:, variance or

predict contact2_lp, xb

pctile pct= contact2_lp, nq(3)

xtile diff_contact = contact2_lp, cutpoints(pct)

tab diff_contact MissingW2

gen nonwhite_m=1 if nonwhite==.

gen prevprison_m=1 if prevprison==.

gen daily_drug_m=1 if daily_drug ==.

gen J2Workr=1 if J2Work==1

replace J2Workr=0 if J2Work==2
```

gen J2Workr_m=1 if MissingW2==0 & J2Work==.

drop if nonwhite_m==1 | prevprison_m==1 | daily_drug_m==1 | J2Workr_m==1

mi ice J2Workr reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year aux_age18to20 [aux_SPCRtheft] aux_SPCRdrug aux_earlyrelease [aux_prevprison] aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal if diff_contact==1 & Qtype==1, add(40)

mi ice J2Workr reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year aux_age18to20 [aux_SPCRtheft] aux_SPCRdrug aux_earlyrelease [aux_prevprison] aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal if diff_contact==2 & Qtype==1, add(40)

mi ice J2Workr reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year aux_age18to20 [aux_SPCRtheft] aux_SPCRdrug aux_earlyrelease [aux_prevprison] aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal if diff_contact==3 & Qtype==1, add(40)

replace _1_ J2Workr = _41_J2Workr if _1_J2Workr==.

…

replace _40_ J2Workr = _80_ J2Workr if _40_ J2Workr ==.

replace _1_ J2Workr = _81_J2Workr if _1_J2Workr==.

…

replace _40_ J2Workr = _120_ J2Workr if _40_ J2Workr ==.

mi estimate, nimputations(40) or: logit reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug J2Workr

mi estimate, nimputations(40): logit reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug J2Workr

Likely or unlikely to refuse

The Stata code to assess the sensitivity of results to a full interaction with whether a respondent is likely or unlikely to refuse is as follows (using the model predicting compliance at Wave 2 first introduced in Chapter 2, Table 2.4)..

Preserve

keep J2Work Stype Qtype MissingW2 reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year aux_age18to20 aux_SPCRtheft aux_SPCRdrug aux_earlyrelease aux_prevprison aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal

drop if Stype==2

```
xtmelogit compliance2 aux_age18to20 aux_Sent1year18month aux_Sent3year4year
aux_PrisonRefusal if contact2==1 || FinalPrison:, variance or

predict compliance2_lp, xb

pctile pct2= compliance2_lp, nq(2)

xtile diff_refusal = compliance2_lp, cutpoints(pct2)

tab diff_refusal MissingW2

gen nonwhite_m=1 if nonwhite==.

gen prevprison_m=1 if prevprison==.

gen daily_drug_m=1 if daily_drug ==.

gen J2Workr=1 if J2Work==1

replace J2Workr=0 if J2Work==2

gen J2Workr_m=1 if MissingW2==0 & J2Work==.

drop if nonwhite_m==1 | prevprison_m==1 | daily_drug_m==1 | J2Workr_m==1

mi set wide

mi ice J2Workr reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison
daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year
aux_age18to20 [aux_SPCRtheft] aux_SPCRdrug aux_earlyrelease [aux_prevprison]
aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal if diff_refusal==1 & Qtype==1, add(40)

mi ice J2Workr reoffend_1yr nonwhite centred_age female SPCRburglary SPCRtheft prevprison
daily_drug aux_SentLE6month aux_Sent6month1year aux_Sent1year18month aux_Sent3year4year
aux_age18to20 [aux_SPCRtheft] aux_SPCRdrug aux_earlyrelease [aux_prevprison]
aux_priorburglary aux_PrisonNoncontact aux_PrisonRefusal if diff_refusal==2 & Qtype==1, add(40)

replace _1_ J2Workr = _41_J2Workr if _1_J2Workr==.

…

replace _40_ J2Workr = _80_ J2Workr if _40_ J2Workr ==.

mi estimate, nimputations(40) or: logit reoffend_1yr nonwhite centred_age female SPCRburglary
SPCRtheft prevprison daily_drug J2Workr

mi estimate, nimputations(40): logit reoffend_1yr nonwhite centred_age female SPCRburglary
SPCRtheft prevprison daily_drug J2Workr
```

The association between employment status in two months following release from prison and proven re-offending within 1 year (Sample 1)

To impute data for variables with missing data at Wave 3, the same procedures are used, but the list of auxiliary variables is those that were associated with missing data at Wave 3. All

prisoners in Sample1 were eligible for re-interview at Wave 3, therefore the imputation model can be fitted to all cases. R3Job identifies whether a prisoner had a job on release from prison, and is missing for those individuals not successfully re-interviewed.

```
preserve

keep R3Job Stype Qtype MissingW3 reoffend_1yr nonwhite centred_age female SPCRtheft
SPCRbreach SPCRmotor prevprison daily_drug aux_age18to20 aux_black aux_edqual
aux_SPCRdrug aux_SPCRbreach aux_SPCRburglary aux_neverworked aux_EFL aux_livfam
aux_nogp aux_dailycrack aux_priorburglary aux_priorrobbery aux_ConsentAddr aux_ConsentMatch
aux_refusalW2 aux_noncontactW2

drop if Stype==2
```

All item missing data are removed, and a new variable (R3Jobr) is constructed that is coded 0, 1. This identifies those sample members that reported having been in paid employment since release from prison.

```
gen nonwhite_m=1 if nonwhite==.

gen prevprison_m=1 if prevprison==.

gen daily_drug_m=1 if daily_drug ==.

gen R3Jobr=0 if R3Job==2

replace R3Jobr=1 if R3Job==1

gen R3Jobr_m=1 if MissingW3==0 & R3Job==.

drop if nonwhite_m==1 | prevprison_m==1 | daily_drug_m==1 |  R3Jobr_m==1
```

Logistic regression models with and without the Wave 3 variable are estimated.

```
logistic reoffend_1yr nonwhite centred_age female SPCRtheft SPCRbreach SPCRmotor prevprison
daily_drug

logit reoffend_1yr nonwhite centred_age female SPCRtheft SPCRbreach SPCRmotor prevprison
daily_drug

logistic reoffend_1yr nonwhite centred_age female SPCRtheft SPCRbreach SPCRmotor prevprison
daily_drug R3Jobr

logit reoffend_1yr nonwhite centred_age female SPCRtheft SPCRbreach SPCRmotor prevprison
daily_drug R3Jobr
```

The imputation model includes the variable with missing data, all variables from the model of interest, and all auxiliary variables associated with missing data at Wave 3. 40 imputed datasets are produced, with the same logistic regression model fitted to each 'complete' dataset.

```
mi set wide

mi ice R3Jobr reoffend_1yr nonwhite centred_age female SPCRtheft SPCRbreach SPCRmotor
prevprison daily_drug aux_age18to20 aux_black aux_edqual aux_SPCRdrug [aux_SPCRbreach]
aux_SPCRburglary aux_neverworked aux_EFL aux_livfam aux_nogp aux_dailycrack
aux_priorburglary aux_priorrobbery aux_ConsentAddr aux_ConsentMatch aux_refusalW2
aux_noncontactW2, add(40)

mi estimate, or: logit reoffend_1yr nonwhite centred_age female SPCRtheft SPCRbreach SPCRmotor
prevprison daily_drug R3Jobr

mi estimate: logit reoffend_1yr nonwhite centred_age female SPCRtheft SPCRbreach SPCRmotor
prevprison daily_drug R3Jobr
```

As with the previous examples, initial exploration of the imputed datasets can be undertaken to assess the distribution of observations.

```
tab _1_R3Jobr _mi_miss, column

tab _22_R3Jobr _mi_miss, column

tab _35_R3Jobr _mi_miss, column

restore
```

Sensitivity analyses

Non-contact, refusal, and nonresponse amongst prison returnees separately

To assess whether the general imputation model is sensitive to differences in the type of missing data at Wave 3, separate imputation models are specified to impute data from those that could not be contacted in the community, those that refused to be interviewed if successfully contacted in the community, and those that were back in prison at the time of the Wave 3 interview but could not be contacted. The dataset is restricted to those prisoners in Sample1 – designed to be representative of all receptions to prison.

```
preserve

keep R3Job Stype Qtype MissingW3 reoffend_1yr nonwhite centred_age female SPCRtheft
SPCRbreach SPCRmotor prevprison daily_drug aux_age18to20 aux_black aux_edqual
aux_SPCRdrug aux_SPCRbreach aux_SPCRburglary aux_neverworked aux_EFL aux_livfam
```

aux_nogp aux_dailycrack aux_priorburglary aux_priorrobbery aux_ConsentAddr aux_ConsentMatch aux_refusalW2 aux_noncontactW2

drop if Stype==2

Three binary indicator variables are constructed that identify whether data was missing as a result of noncontact (STnoncontact3), refusal given contact (STrefusal3), or lack of contact amongst the sample that were back in prison (STprison3). In each case, the indicator also identifies all cases with observed data at Wave 3 (excluding those back in prison when considering STnoncontact3 and STrefusal3, and restricted to those back in prison when considering STprison3).

gen STnoncontact3=0

replace STnoncontact3=1 if contact3==0 & Prisoner==0

replace STnoncontact3=1 if (contact3==1 & compliance3==1 & Prisoner==0)

gen STrefusal3=0

replace STrefusal3=1 if compliance3==0

replace STrefusal3=1 if (contact3==1 & compliance3==1 & Prisoner==0)

gen STprison3=0

replace STprison3=1 if Prisoner==1

All item missing data is removed, and a binary indicator identifying whether a prisoner was able to secure paid employment at any time since their initial release from prison was constructed.

gen nonwhite_m=1 if nonwhite==.

gen prevprison_m=1 if prevprison==.

gen daily_drug_m=1 if daily_drug ==.

gen R3Jobr=0 if R3Job==2

replace R3Jobr=1 if R3Job==1

gen R3Jobr_m=1 if MissingW3==0 & R3Job==.

drop if nonwhite_m==1 | prevprison_m==1 | daily_drug_m==1 | R3Jobr_m==1

Three separate imputation models are then run, with 40 imputations requested at each time. The first imputation model is restricted to imputing cases that were not successfully contacted at Wave 3, with the auxiliary variables limited to those variables associated with non-contact.

```
mi set wide

mi ice R3Jobr reoffend_1yr nonwhite centred_age female SPCRtheft SPCRbreach SPCRmotor prevprison daily_drug aux_age18to20 aux_black aux_edqual aux_SPCRdrug [aux_SPCRbreach] aux_neverworked aux_EFL aux_livfam aux_nogp aux_dailycrack aux_ConsentMatch aux_refusalW2 if STnoncontact3==1, add(40)
```

The second model is restricted to imputing cases that refused to be re-interviewed at Wave 3, despite being successfully contacted. In this instance, the auxiliary variables are those associated with refusal (given contact).

```
mi ice R3Jobr reoffend_1yr nonwhite centred_age female SPCRtheft SPCRbreach SPCRmotor prevprison daily_drug aux_priorburglary aux_ConsentAddr aux_refusalW2 aux_noncontactW2 if STrefusal3==1, add(40)
```

The final model is restricted to the sample of prisoners that were back in prison at the time of the interview. The auxiliary variables are those that were significantly associated with nonresponse.

```
mi ice R3Jobr reoffend_1yr nonwhite centred_age female SPCRtheft SPCRbreach SPCRmotor prevprison daily_drug aux_age18to20 aux_SPCRburglary aux_priorrobbery if STprison3==1, add(40)
```

40 imputations were requested using each imputation model, resulting in a total of 120 partially complete datasets. The three groups of 40 imputations were then merged to produce a single set of 40 ‘complete’ datasets.

```
replace _1_R3Jobr= _41_R3Jobr if _1_R3Jobr==.

…

replace _40_R3Jobr= _80_R3Jobr if _40_R3Jobr==.

replace _1_R3Jobr= _81_R3Jobr if _1_R3Jobr==.

…

replace _40_R3Jobr= _120_R3Jobr if _40_R3Jobr==.
```

The model of interest was fitted to each of the ‘complete’ datasets, with the 40 sets of results combined.

```
mi estimate, nimputations(40) or: logit reoffend_1yr nonwhite centred_age female SPCRtheft SPCRbreach SPCRmotor prevprison daily_drug R3Jobr

mi estimate, nimputations(40): logit reoffend_1yr nonwhite centred_age female SPCRtheft SPCRbreach SPCRmotor prevprison daily_drug R3Jobr

restore
```

Ease of contact, likely or unlikely to refuse, likely or unlikely to respond in prison

To assess whether the imputation model is sensitive to ‘ease of contact’ or ‘likelihood of refusal’, the strategy outlined for Wave 2 is extended to account for the Wave 3 data structure. Three types of missing data were identified at Wave 3, therefore the imputation model is assessed across three different forms of missingness. For brevity, we only outline the differences in the strategy for the three different forms of missingness below.

```
preserve

keep R3Job Stype Qtype MissingW3 reoffend_1yr nonwhite centred_age female SPCRtheft SPCRbreach SPCRmotor prevprison daily_drug aux_age18to20 aux_black aux_edqual aux_SPCRdrug aux_SPCRbreach aux_SPCRburglary aux_neverworked aux_EFL aux_livfam aux_nogp aux_dailycrack aux_priorburglary aux_priorrobbery aux_ConsentAddr aux_ConsentMatch aux_refusalW2 aux_noncontactW2 contact3 compliance3 Prisoner ContactPrison prisonerdetails

drop if Stype==2
```

If interest is in assessing whether the imputation model applies across differing levels of ‘ease of contact’ in the community sample, the model predicting contact amongst the community sample (Chapter 2, Table 2.6) is used to generate the linear predicted probabilities of contact. This allows us to split the sample into terciles (diff_contact).

```
logistic contact3 aux_age18to20 aux_black aux_edqual aux_SPCRdrug aux_SPCRbreach aux_ConsentMatch aux_neverworked aux_EFL aux_livfam aux_nogp aux_dailycrack aux_refusalW2 if Prisoner==0

predict contact3_lp, xb

pctile pct= contact3_lp, nq(3)

xtile diff_contact = contact3_lp, cutpoints(pct)
```

To assess whether the imputation model applies across levels of 'likelihood of refusal', the model predicting compliance at Wave 3 amongst the community sample is used (Chapter 2, Table 2.7). Again we request the linear predicted probabilities to split the observed sample into two groups (diff_refusal).

```
logistic compliance3 aux_priorburglary aux_ConsentAddr aux_refusalW2 aux_noncontactW2 if contact3==1 & Prisoner==0

predict compliance3_lp, xb

pctile pct2= compliance3_lp, nq(2)

xtile diff_refusal = compliance3_lp, cutpoints(pct2)
```

Finally, to assess whether the imputation model applies across levels of 'ease of contact' amongst those back in prison, the model predicting contact at Wave 3 amongst those back in prison is used (Chapter 2, Table 2.8). Again we request the linear predicted probabilities (diff_prison).

```
xtmelogit ContactPrison aux_age18to20 aux_SPCRburglary aux_priorrobbery if Prisoner ==1 || prisonerdetails:, variance or

predict prison3_lp, xb

pctile pct3= prison3_lp, nq(2)

xtile diff_prison = prison3_lp, cutpoints(pct3)
```

All item missing data is removed, and an indicator of whether an individual was able to secure paid work at any time since their initial release from prison is constructed.

```
gen nonwhite_m=1 if nonwhite==.

gen prevprison_m=1 if prevprison==.

gen daily_drug_m=1 if daily_drug ==.

gen R3Jobr=0 if R3Job==2

replace R3Jobr=1 if R3Job==1

gen R3Jobr_m=1 if MissingW3==0 & R3Job==.

drop if nonwhite_m==1 | prevprison_m==1 | daily_drug_m==1 |  R3Jobr_m==1
```

The general imputation model is specified, but is imputed separately for each category of ‘ease to contact’ (diff_contact), ‘unlikely or likely to refuse’ (diff_refusal), ‘unlikely or likely to respond in prison’ (diff_prison) respectively. 40 imputations were requested.

```
mi set wide

mi ice R3Jobr reoffend_1yr nonwhite centred_age female SPCRtheft SPCRbreach SPCRmotor
prevprison daily_drug aux_age18to20 aux_black aux_edqual aux_SPCRdrug [aux_SPCRbreach]
aux_SPCRburglary aux_neverworked aux_EFL aux_livfam aux_nogp aux_dailycrack
aux_priorburglary aux_priorrobbery aux_ConsentAddr aux_ConsentMatch aux_refusalW2
aux_noncontactW2 if diff_contact==1/2/3 [if diff_refusal==1/2] [if

diff_prison==1/2], add(40)
```

The imputations are then combined and the model of interest fitted to each.

```
replace _1_R3Jobr= _41_R3Jobr if _1_R3Jobr==.

…

replace _40_R3Jobr= _80_R3Jobr if _40_R3Jobr==.

replace _1_R3Jobr= _81_R3Jobr if _1_R3Jobr==.

…

replace _40_R3Jobr= _120_R3Jobr if _40_R3Jobr==.

mi estimate, nimputations(40) or: logit reoffend_1yr nonwhite centred_age female SPCRtheft
SPCRbreach SPCRmotor prevprison daily_drug R3Jobr

mi estimate, nimputations(40): logit reoffend_1yr nonwhite centred_age female SPCRtheft
SPCRbreach SPCRmotor prevprison daily_drug R3Jobr

restore
```

Identifying the factors associated with an increased propensity to secure employment on release from prison (sample2)

The final example outlines the procedures used when data is missing from the dependent variable – in this case whether a prisoner had undertaken any paid work since release from prison (R3Job). In practice, the MI proceeds exactly as before. The model is restricted to prisoners in Sample2.

```
Preserve
```

```
keep R3Job Stype Qtype MissingW3 nonwhite centred_age daily_drug fourwk_emp ltermill
aux_age18to20 aux_black aux_edqual aux_SPCRdrug aux_SPCRbreach aux_SPCRburglary
aux_neverworked aux_EFL aux_livfam aux_nogp aux_dailycrack aux_priorburglary
aux_priorrobbery aux_ConsentAddr aux_ConsentMatch aux_refusalW2 aux_noncontactW2

drop if Stype==1
```

All item missing data is first removed from the data, and a binary indicator is created that indicates whether a prisoner was in paid work (R3Jobr).

```
gen nonwhite_m=1 if nonwhite==.

gen daily_drug_m=1 if daily_drug ==.

gen fourwkemp_m=1 if fourwkemp==.

gen ltermill_m=1 if ltermill==.

gen R3Jobr=0 if R3Job==2

replace R3Jobr=1 if R3Job==1

gen R3Jobr_m=1 if MissingW3==0 & R3Job==.

drop if nonwhite_m==1 | prevprison_m==1 | daily_drug_m==1 | fourwkemp_m==1 | ltermill_m==1 |
R3Jobr_m==1
```

A single logistic regression model is produced, predicting employment on release from prison. This analysis is restricted to those respondents successfully re-interviewed at Wave 3.

```
logistic R3Jobr nonwhite centred_age daily_drug fourwk_emp ltermill

logit  R3Jobr nonwhite centred_age daily_drug fourwk_emp ltermill
```

R3Jobr is included alongside the explanatory variables in the model of interest, and the full list of auxiliary variables associated with missing data at Wave 3. 40 imputations are produced and the logistic regression models estimated for each imputation. The results are combined to produce the final estimates.

```
mi set wide

mi ice R3Jobr nonwhite centred_age daily_drug fourwk_emp ltermill aux_age18to20 aux_black
aux_edqual aux_SPCRdrug aux_SPCRbreach aux_SPCRburglary aux_neverworked aux_EFL
aux_livfam aux_nogp aux_dailycrack aux_priorburglary aux_priorrobbery aux_ConsentAddr
aux_ConsentMatch aux_refusalW2 aux_noncontactW2, add(40)
```

```
mi estimate, or: logit R3Jobr nonwhite centred_age daily_drug fourwk_emp ltermill
```

```
mi estimate: logit R3Jobr nonwhite centred_age daily_drug fourwk_emp ltermill
```

As before, it is advisable to check the imputations for potential anomalies.

```
tab _1_R3Jobr _mi_miss, column
```

```
tab _1_R3Jobr _mi_miss, column
```

```
tab _1_R3Jobr _mi_miss, column
```

```
restore
```