EXPLORING THE IMPACT OF MISSING DATA IN MULTIPLE REGRESSION

Mike Kenward, LSHTM

6th ESRC Research Methods Festival, Oxford, 8–10 July, 2014

Data are from:

ELSA: English Longitudinal Study of Ageing

This has several waves, starting in 2002 (Wave 1).

The outcome here is **RECALL SCORE**: a measure of cognitive function.

The analysis will be done for MALES (total number: 5335).

We are regressing this on the **Numbers of Children:**

$[0, 1, 2, 3, > 3]$ with no missing data,

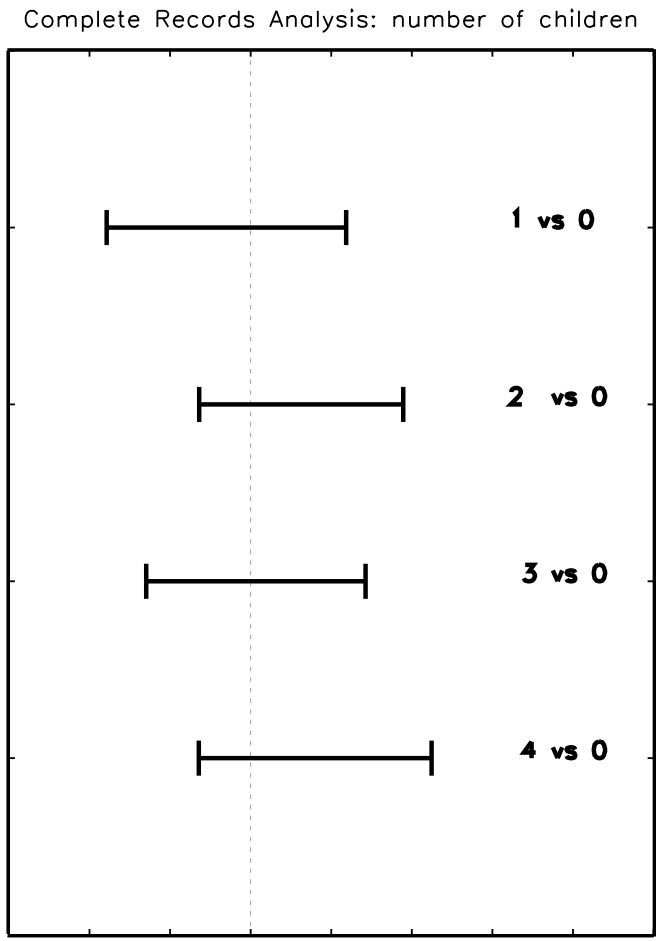
and adjusting using a set of covariates:

Variable	Number of categories	Number missing	%	Wave
young mother	2	440	8.2	1
marital status	6	2	0.03	1
decade	4	105	2.0	1
fdied70	2	2074	38.9	1
mdied70	2	1710	32.1	1
smoking status	5	92	1.7	1
father's job	6	244	4.6	1
education	4	41	0.8	1
partner's schooling	4	258	4.8	1
wealth (quart)	4	75	1.4	1
major health problems in childhood	2	2504	46.9	3
minor health problems in childhood	2	2504	46.9	3
poor health in childhood	2	2504	46.9	3

Analysis of those with complete records ($N = 1346$, 25.2%).

Overall effect of Number of Children, $F = 1.51$ on 4df, $P = 0.20$.

95% Confidence Intervals:



This analysis is valid (if potentially inefficient) under **MCAR** (Missing Completely at Random).

It will be biased when this doesn't hold:

MAR (Missing at Random) and **NMAR** (not missing at random).

Validity under MAR is often invoked for likelihood based analyses.

Reminder: under MAR the the probability that an observation is missing is conditionally independent of its value given the *observed* data.

What does MAR mean in the current setting?

The missing data pattern is non-monotone.

In fact, only one variable (number of children) is fully observed.

To hold exactly in such a setting, MAR requires different individuals to have different missing value mechanisms:

and these must exactly match the particular pattern of observed values for each individual.

This is implausible.

In such settings, when we say our likelihood analysis is valid under MAR, we really mean valid when the missing data mechanism is *ignorable*.

So in practice, such analyses that assume MAR (*i.e.* that make appropriate adjustments in some way) are not assumed to be strictly valid, but in most (but not all situations) will reduce the bias due to non-MCAR missing data mechanisms.

Methods for the incorporation of incomplete records typically rest on such assumptions.

When is an impact of this likely to be seen in terms of bias (rather than precision).

From a missing data perspective, the key incomplete variables are those that are

1. related to the outcome, and
2. for which the probability of being missing is also related to the outcome.

(1) Other covariates in the complete records regression: overall F tests

Effect	Num DF	Den DF	F Value	Pr > F	
young mother	1	1307	0.42	0.5188	
marital status	5	1307	1.00	0.4182	
decade	3	1307	47.50	<.0001	<===
fdied70	1	1307	1.77	0.1836	
mdied70	1	1307	0.10	0.7520	
smoking status	4	1307	0.91	0.4602	
father's job	5	1307	1.17	0.3221	
education	3	1307	16.32	<.0001	<===
partner's schooling	3	1307	2.55	0.0540	
wealth (quart)	3	1307	5.65	0.0008	<===
major health problems	1	1307	0.51	0.4763	
minor health problems	1	1307	0.17	0.6846	
phealth	1	1307	2.34	0.1261	

[n.b. these results could well be influenced by the missing data mechanism]

(2) Logistic regressions of observed/missing (1/0) on outcome (recall).

Effect	[n miss]	Estimate	Standard Error	P	
young mother	[440]	-0.0909	0.0149	<.0001	<===
marital status	[2]	-0.7294	0.3981	0.0669	
decade	[105]	-0.1338	0.0307	<.0001	<===
fdied70	[2074]	0.0536	0.0082	<.0001	<===
mdied70	[1710]	0.0807	0.0086	<.0001	<===
smoking status	[244]	0.3389	0.3076	0.2705	
father's job	[258]	0.0543	0.0190	0.0042	<===
education	[41]	0.4047	0.0688	<.0001	<===
partner's schooling	[258]	-0.0559	0.0189	0.0031	<===
wealth (quart)	[75]	-0.0637	0.0369	0.0849	
major health problems	[2504]	0.1180	0.0084	<.0001	<===
minor health problems	[2504]	0.1180	0.0084	<.0001	<===
phealth	[2504]	0.1180	0.0084	<.0001	<===

Incorporating the incomplete records

The complete records analysis is inefficient and potentially biased.

To use the incomplete records it is necessary to introduce a joint distribution for partially observed covariates.

This can be done explicitly or implicitly (*e.g.* weighting methods).

This can be done directly (a formal joint model) or indirectly (an implied joint model based on conditional models).

We have used **Multiple Imputation (MI)**.

[Carpenter and Kenward (2013) *Multiple Imputation and its Application*, Wiley]

Two possibilities for generating the imputations:

1. direct: impute from proper joint models (*e.g.* REALCOM),
2. indirect: fully conditional specification (FCS) which uses a Gibbs type sampling scheme based on univariate conditional models for each partially observed variable, given all the others including the outcome.

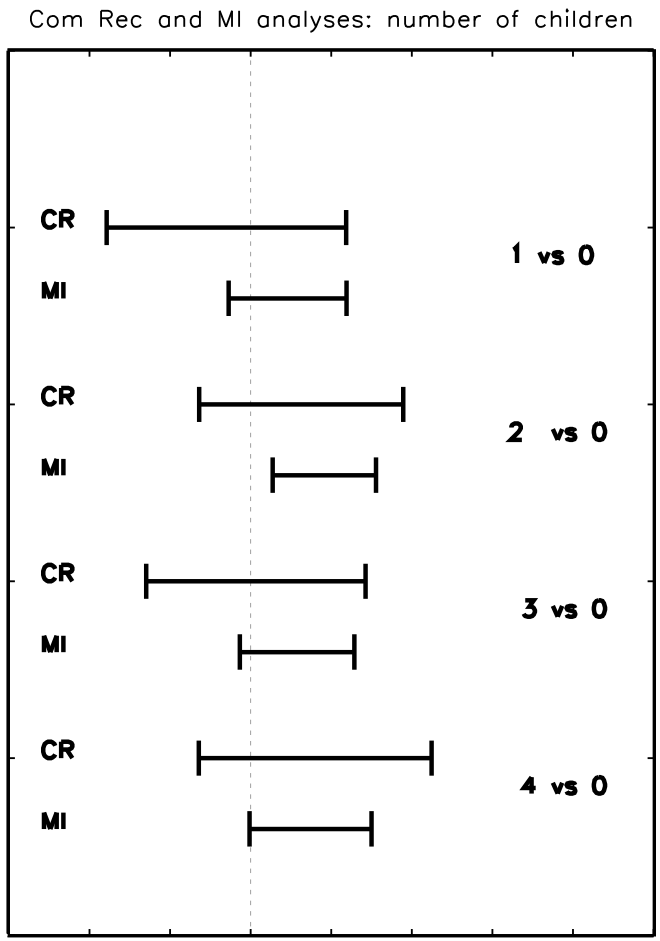
   This can be done conveniently in Stata (mi procedures), and in the latest release of SAS PROC MI.

Here we use FCS in Stata (with 5 imputations).

MI analysis (5 imputations).

Overall effect of ‘number of children’, $F = 2.32$ on 4df, $P = 0.054$ (complete records: $P = 0.20$).

95% Confidence Intervals:



What is the MI analysis implying for the missing data?

Comparison of observed and (average) imputed percentages.

young mother

(Completers (MI) regression: $P = 0.52(0.98)$, MV logistic regression: $P < 0.001$)

	N	0	1
observed:	4895	93.1	6.9
imputed:	440	92.2	7.8

major health problems

(Completers (MI) regression: $P = 0.68(0.59)$, MV logistic regression: $P < 0.001$)

	N	0	1
observed:	2831	97.9	2.1
imputed:	2504	97.1	2.9

wealth (quartiles)

(Completers (MI) regression: $P < 0.001 (< 0.001)$, MV logistic regression: $P = 0.08$)

	N	1	2	3	4
observed:	5260	23.50	23.86	25.61	27.03
imputed:	75	17.87	22.67	24.80	34.67

decade

(Completers (MI) regression: $P < 0.001 (< 0.001)$, MV logistic regression: $P < 0.001$)

	N	50	60	70	80
observed:	5230	37.3	31.0	22.5	9.3
imputed:	105	61.0	27.2	9.9	1.9

education

(Completers (MI) regression: $P < 0.001 (< 0.001)$, MV logistic regression: $P < 0.001$)

	N	1	2	3	4
observed:	5294	49.8	23.9	20.2	6.2
imputed:	41	60.5	22.0	12.7	4.8

Some conclusions

1. Incorporation of the incomplete records has
   - greatly improved efficiency, and
   - had a non-negligible, but small, impact on the main effects of interest.
2. The key variables in this (small) impact are probably ‘decade’ and ‘education’.
3. But these have only small proportions missing (2%, 0.8%).

   Suppose that these had been of the same order of the wave three variables: about 50%?

If the incorporation of the incomplete records leads to substantial changes in the results, it is important to understand what is driving the changes in the analysis.

We might also consider appropriate sensitivity analyses, perhaps targetting the key incomplete variables.

See for example Chapters 10 and 12 of Carpenter and Kenward (2013).