Exploring the Impact of Missing Data in Multiple Regression

Michael G Kenward

London School of Hygiene and Tropical Medicine

28th May 2015

1. Introduction

In this note we are concerned with the conduct of a conventional multiple regression analysis, in which large numbers of values are missing. Without introducing additional steps such as imputation or weighting, the analyst has few options for dealing with this problem. Perhaps the most common option is still an analysis confined to those units with complete records. This can be both very inefficient and highly biased. In the illustrative example to be used below, which is taken from a wider piece of research based on the English Longitudinal Study of Ageing (ELSA), only one quarter of the sample has complete records. To make use of the data from the remaining three quarters of the sample, Multiple Imputation (MI) will be used. However, the main aim of this note is not to introduce MI, or to show how to carry out the necessary computations. This widely used technique now has a very large literature that covers these points in great detail and at many technical levels. Here the concern is more with using and presenting results when analyses are based on MI, especially in comparison with a complete records analysis which often accompanies such an MI based analysis. Clear recommendations for carrying out and presenting analyses based on MI are given by Sterne *et al.* (2009). This note complements such guidelines by emphasising the importance of explaining, in the context of multiple regression, the differences between results obtained with MI and from a complete records analysis (should such differences occur). It will be seen that useful information can be obtained from the relationships among the variables involved and from their association with the occurrence of missing values.

2. The Illustrative Example

The multiple regression to be used as an illustrative example in this note is one of several conducted as part of a wider piece of research that is based on data from the English Longitudinal Study of Ageing (ELSA) and that is concerned with the relationships between fertility and wealth and health at older ages. ELSA has several waves, starting in 2002 (Wave 1: 2002-2003). Here we use data from Wave 1 and Wave 3 (2006–2007). The overall aim of the study from which the

Variable	Number of categories	Number missing	% missing	Wave
recall all (outcome variable)	–	190	4	3
number of children	5	0	0	1
young father	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 (quartiles)	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

Table 1: Regression covariates, with numbers and percentages missing.

current analysis is taken was to investigate relationships between numbers of children and various measures of wellbeing of the parents, adjusting for potential confounders. Here we focus on one such wellbeing measure: *Recall Score*, a measure of cognitive function. Interest is in the the regression relationship of this measure with the five category explanatory variable, *Numbers of Children* $(0, 1, 2, 3, > 3)$ for which there are no missing data. We consider here the males in the sample (total number: 5335). The variables to be used as additional covariates in the multiple regression analysis are listed in Table 1, along with the outcome (regression) variable and primary explanatory variable (number of children). Note that the covariates are all categorical, with numbers of categories ranging from 2 to 6. These include both ordinal (*e.g. wealth (quartiles)*) and nominal (*e.g. marital status*). The percentages of missing date vary greatly among the additional covariates, ranging from less than 0.1% to 46.9%. The highest level of missingness is found among the Wave 3 variables, but some of the Wave 1 variables also have high degrees of missingness. Only 25.2% have complete records.

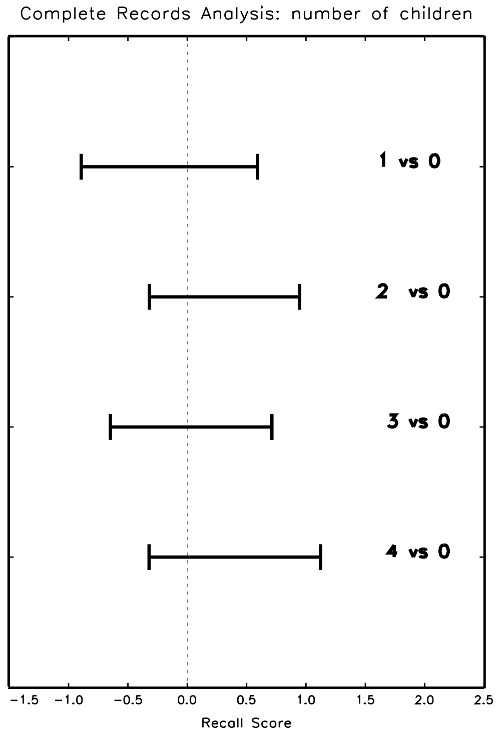

Figure 1: Complete Records Multiple Regression: 95% confidence intervals for effects of Number of Children", with 0 as reference.

3. Complete Records Analysis

Regression procedures in all major statistical packages will, by default, drop any unit that has missing data. The resultant analysis can be termed a *completer records analysis.* Here such an analysis includes only 1346 (25.2%) of those individuals in the original sample. There are two obvious concerns with such an approach. First, much relevant information has been discarded, *i.e.* the analysis is potentially *inefficient.* Second, if the relationships among the variables of those excluded is not the same as those included there is the potential for *bias*.

We begin by carrying out the complete records multiple regression. The resultant estimated relationships of primary interest, *i.e.* those between the recall score and numbers of children are summarised graphically in Figure 1 as 95% confidence intervals for the four comparisons with the the reference category, no children. Note that the overall term is not satistically significant at the 5% level.

The impact of the other covariates is summarised in Table 2 in terms of the overall F tests for covariate. Although based only on the complete records analysis, it is instructive to see which covariates appear to be making a non-trivial contribution to this regression. The three covariates

	Num	Den		
Effect	DF	DF	F Value	Prob.
number of children	4	1307	1.51	0.20
young father	1	1307	0.42	0.5188
marital status	5	1307	1.00	0.4182
decade	**3**	**1307**	**47.50**	< 0.001
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**	< 0.001
partner's schooling	3	1307	2.55	0.0540
wealth (quartile)	**3**	**1307**	**5.65**	< 0.001
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

Table 2: Covariates in the complete records regression: overall F tests

with strong relationships in this analysis, *Decade*, *Education* and *Wealth* are highlighted in the table. We return to the relevance of the variables in the handling of missing data below.

4. Underlying Assumptions concerning the Missing Data Mechanisms

Two issues were raised earlier about the complete records analysis. One concerned bias (and hence validity). Under what circumstances will the above analysis provide a valid (if inefficient) analysis? A thorough discussion of this question is given in Chapter 1 of Carpenter and Kenward (2013). A simple and sufficient requirement for validity is that the missing data be Missing Completely at Random (MCAR). This means, in this setting, that the probability of an observation being missing is independent of any of the data values, whether observed or not. This implies that the 1346 members of the sample with complete records is a random sample of the original 5335. This is a very strong assumption and unlikely to be wholly true.

A less restrictive assumption is so-called covariate dependent Missing at Random (MAR). In the current setting this can be expressed as the requirement that the probability of an observation being missing may depend on the observed values of other variables, and, if a covariate, on its own value, provided that, conditional on these, it is not associated with the outcome variable. This includes situations in which a covariate value itself may be deemed to be missing not at random (MNAR). The key point here is that we are estimating multiple regression coefficients in which the analysis is wholly conditioned on the covariates. When dealing with true multivariate analyses, such as with structural equation models, all variables on which we do not condition would need to be treated in the above definition as outcome variables, making the MAR requirement potentially much more complex.

The assumption of MAR is used in practice as a justification for many analyses with missing data, especially those using likelihood, multiple imputation or inverse probability weighting. See Molenberghs *et al.* (2015) for a recent and thorough treatment of these approaches. However, as Robins and Gill (1997) and others have pointed out, the justification of MAR when missing data are *non-monotone* is problematic. To clarify this, a monotone missing data pattern is one in which the variables can be ordered in such a way that a missing value in one variable implies all variables of higher order are also missing. Attrition (or dropout) in longitudinal settings is the obvious and most common realisation of monotone missingness. Settings like the one considered here, and in most cross-sectional data, missingness is not monotone. Robins and Gill (1997) argue that MAR in non-monotone settings implies a practically implausible missing data mechanism. One such justification requires different units (here individuals) to have different missing value mechanisms and these must exactly match the particular pattern of observed values for that unit. Hence we must be careful in using MAR to justify analyses such as the one that follows, in which multiple imputation is used. More realistically, the adjustments implied by analyses that invoke MAR, such as multiple imputation models that use what is available for each unit, can be regarded as ways of improving on the bias and efficiency of complete records analyses rather completely eliminating bias. The latter is anyway unrealistic in nearly all missing data settings as its confirmation requires information not available from the data (Carpenter and Kenward, 2013, Chapter 1). Another view of the same issue is that in practice MAR can rarely be justified from the data under analysis (*e.g.* Molenberghs *et al.* 2008). Our justification of the use of multiple imputation in what follows will therefore follow this rather looser argument.

5. Incorporating the Incomplete Records

The potential for the complete records given analysis above to be both inefficient and potentially biased has been discussed above. One route to tackling both issues is to bring the incomplete records into the analysis. A commonly used method in the past for this has been to introduce *missing value indicators*. When covariates are categorical, as here, an additional category can be added to each partially incomplete variable, to which this variable is assigned when it is missing. In this way values are assigned to all units for all variables and so can be incorporated in the multiple regression. Unfortunately this method is seriously flawed and can introduce addition large biases, depending on the actual missing data mechanism and model of interest. See for example Greenland and Finkle (1995) and Carpenter and Kenward (2015).

To use the incomplete records in a more statistically principled way it is necessary to introduce a joint distribution for the partially observed covariates. Note that inverse probability weighting methods can be viewed (with some oversimplification) as an example of this in which a fully non-parametric distribution is used. In contrast to this, a parametric joint distribution can be used as the basis for a full likelihood analysis, or arguably more conveniently, through Multiple Imputation (MI; Rubin 1986, Carpenter and Kenward 2013). The imputation model uses the joint distribution of all the partially observed variables conditional on those that are fully observed. This can be done directly from a properly defined joint multivariate distribution, such as that implemented in REALCOM (Carpenter *et al.* 2011), or indirectly through a series of univariate conditional models, the so-called *Fully Conditional Specificiation (FCS)* (*e.g.* van Buuren, 2015). The former has great advantages when data are hierarchical, while the latter, although lacking a rigorous justification, is very convenient to implement and use in cross-sectional settings, and has been shown to work well in a range of settings. As we have no hierarchical structure in the present setup, FCS will be used, through the Stata MI procedure (Statacorp, 2013).

Briefly, in an MI analysis, the missing data are “filled in” using Bayesian predictive draws from an appropriate imputation model. The completed dataset is then analysed in the way originally intended. Here this would be using our full multiple imputation model. This is repeated several times (M say) and the results from each of the M analyses are combined using Rubin’s formulae to give a single set of estimates and appropriate measure of precision. These formulae ensure, among other things, that a proper account is taken of the process of imputation, that is, the information in the MI analysis reflects that in the observed data and modelling assumptions, and no more. In the application here, we use $M = 5$ imputations. In some circumstances much larger values of M are advisable (Carpenter and Kenward, 2013, Section 2.6).

The MI based multiple regression analysis leads to a borderline statistically significant adjusted effect of ‘number of children’, $F = 2.32$ on 4df, $P = 0.054$. Recall that from the complete records we obtained $P = 0.20$. The estimated comparisons from the Number of Children are compared between the completers and MI analyses in Table 2 again in terms of 95% confidence intervals. It can be seen that the intervals from MI are considerably shorter than those from the complete records analysis (as expected) and are consistently shifted towards the right. That is the measures of cognitive function (Recall Score) tend to be higher for those fathers with children compared to those without. There is no indication of a trend however among the numbers of children beyond the zero versus greater than zero difference. If different conclusions are drawn from an MI analysis compared with one based on complete records it is important to establish what is leading to this

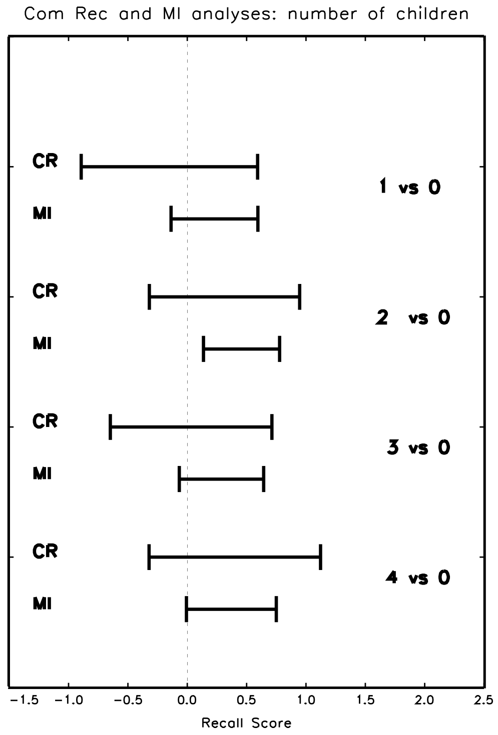

Figure 2: Complete Records and MI based Multiple Regression: 95% confidence intervals for effects of Number of Children", with 0 as reference.

difference. This can be regarded as part of the MI analysis. In the next section a closer investigation is made of the impact of MI on the regression analysis.

6. The Impact of Using Multiple Imputation.

If data were missing completely at random we would expect MI to produce some increase in precision, but not alter estimates, beyond random variation. So we might ask when we should expect to see a greater impact of MI on the results, at least in terms of location of the regression estimates? Or, turning this around, what leads to such changes when these are in fact seen?

Key covariates that influence the impact of missing data, and that should be in any imputation model, are typically ones that

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

Effect	no. missing	Estimate	Standard Error	P
young father	440	-0.0909	0.0149	< 0.001
marital status	2	-0.7294	0.3981	0.0669
decade	105	-0.1338	0.0307	< 0.001
fdied70	2074	0.0536	0.0082	< 0.001
mdied70	1710	0.0807	0.0086	< 0.001
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	< 0.001
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	< 0.001
minor health problems	2504	0.1180	0.0084	< 0.001
phealth	2504	0.1180	0.0084	< 0.001

Table 3: Single variable logistic regressions of observed/missing covariate (1/0) on the outcome (Recall Score).

If such variables have different distributions among those with complete data and among those without, then these have the capability to produce to non-trivial changes in estimates obtained with and without the partially observed data. The exact impact depends on the actual inter-relationships among the variables in both the model of interest and the missing data mechanism and on the nature of the missing data mechanism itself. We cannot answer these questions exactly but can learn much from the identification of potential candidate variables with these properties. This process can only be approximate because it must be done using only the observed data. First, the role of the variables in the complete records analysis (Table 2) are examined. Three have been identified as strongly linked to the outcome (conditionally on the other covariates), at least among the complete records. Second, the association between missingness in each covariate in turn and the outcome (Recall All) is assessed using logistic regression. These analyses are necessarily confined to those individuals with the outcome observed, fortunately the proportion these is high (96%). The estimates from these logistic regressions and associated statistics are presented in Table 3. The majority show very strong associations. Only missingness in Smoking Status, Marital Status and Wealth shows little evidence of a direct relationship with the outcome.

From theory we know that, even if missingness of a variable is very strongly related with outcome, its distribution in the original, and imputed data sets, will be similar provided that they are not themselves predictive of outcome. This can be explored in the example by comparing the proportions falling in each category of a covariate among the observed data, with the corresponding proportions observed in the imputed data. We do this for a selection of covariates, covering different combinations of relationship/no relationship with outcome (from Table 2) and missingness having relationship/no relationship with outcome (from Table 3). These proportions are presented in Table 4 together with the associated P-values from Tables 2 and 3. The first covariate in this table is Young Father (8.2% missing), the mnissingness of which is strongly associated with outcome, but not itself directly associated with outcome. The proportions are very similar indeed in the observed and imputed data. Hence the use of MI for this variable would be expected to have a negligible impact on the estimated regression coefficients in the final model of interest. The second covariate, Major Health Problems (46.9% missing) has a similar patterns of association with outcome, and again the observed and imputed proprtions are very similar. A covariate with such a large proportion missing does have the potential to be influential when assessing the impact of methods, such as MI, for incorporating the partially observed records. This impact depends mainly on the predicted behaviour of the unobserved values. Here, their predicted behaviour is the same as those observed and the impact is expected to be negligible. The third covariate, Wealth (1.4% missing), is itself strongly associated with outcome, but its missingess only very weakly with outcome. Again, no difference among observed and imputed proportions is expected, and only small differences are seen. Note that the very small proportion of imputed values is reflected in the variability of the imputed proportions. The final two covariates in the table, Education (0.8% missing) and Decade (2.0%) missing are the only two for which both associations very high both with outcome. For these it is much more likely that observed and imputed proportions will differ substantially, and this is indeed the case. The differences are particularly extreme in Decade, with for example, category 50 having 37.3% of the observed values but 61.0% of the imputed. Such differences can potentially affect the resultant regression results in a non-trivial way. However, in this case the proportion missing is very small and the impact is not great. It is interesting to note that in this example, those covariates with high proportions of missingness are either not directly related to outcome or whose missingness is not directly outcome, or both, and the overall impact of incorporating the incomplete records through MI has not had a great impact on the estimate of the principle relationship of interest.

It should be noted that in this development the subtleties that arise due to the interplay between the covariates in both outcome and missingness models have been ignored. However, one would not expect these typically to have a major impact on the overall picture as described.

6. Discussion

In the example studied above, the incorporation of the incomplete records has through Multiple Imputation has greatly improved efficiency, and had a non-negligible, but small, impact on the

							Regression P values		
			Outcome Percentage				Compl.†	MI*	MV‡
Young Father		N	0	1			0.51	0.98	< 0.001
	Observed	4895	93.1	6.9					
	Imputed	440	92.2	7.8					
Major Health Problems		N	0	1			0.68	0.59	< 0.001
	Observed	2831	97.9	2.1					
	Imputed	2504	97.1	2.9					
Wealth (quartiles)			1	2	3	4	< 0.001	< 0.001	0.08
	Observed	5260	23.5	23.9	25.6	27.0			
	Imputed	75	17.9	22.7	24.8	34.7			
Education			1	2	3	4	< 0.001	< 0.001	< 0.001
	Observed	5294	49.8	23.9	20.2	6.2			
	Imputed	41	60.5	22.0	12.7	4.8			
Decade			50	60	70	80	< 0.001	< 0.001	< 0.001
	Observed	5230	37.3	31.0	22.5	9.3			
	Imputed	105	61.0	27.2	9.9	1.9			

† Complete records multiple regression, * Multiple Imputation based multiple regression, ‡ logistic regression on missing value indicator.

Table 4: Observed and imputed proportions among a selection of covariates

main effects of interest. The key covariates behind this (small) impact are probably Decade and Education, the two variables that both strongly associated with outcome and whose missingness is also associated with outcome. However, these variables have only small proportions missing (2% and 0.8% respectively), and so are not expected to greatly affect the final regression estimates. We might consider what impact may have arisen had their proportions missing been of the same order of the wave three variables: *i.e.* about 50%.

The message behind this note is that, if the incorporation of the incomplete records in a statistical analysis leads to substantial changes in the results when compared with the complete records analysis, it is important to explore what is driving the changes in the analysis. Some simple statistics calculated from model fits and from the imputations have have been presented that can help with this, but these should not be regarded as exhaustive. It is argued here that the presentation of the results from such a process of exploration should be regarded as an essential part of the handling of missing data in any setting, *i.e.*, it is important not to treat methods of handling missing data as "black box" procedures that "solve" the problem of missing data.

Any non-trivial statistical analysis in which data are missing relies on assumptions that cannot be assessed from the data under analysis. It is now widely agreed that appropriate sensitivity analyses, also play an important rôle, see for example Part V of Molenberghs and Kenward (2007) and Chapters 10 and 12 of Carpenter and Kenward (2013).

REFERENCES

Carpenter JR and Kenward MG (2013) *Multiple Imputation and its Application.* Chichester: Wiley.

Carpenter JR and Kenward MG (2015) Developments of methods and critique of ad hoc methods. Chapter 2 in G Molenberghs *et al.* eds (2015) *Handbook of Missing Data Methodology.* Boca Raton: Champam&Hall/CRC, pp. 23–46.

Greenland S and Finkle WD (1995) A critical look at methods for handling missing covariates in epidemiological regression analyses. *American Journal of Epidemiology,* **142**, 1255–1264.

Molenberghs G, Beunckens C, Sotto C and Kenward MG (2008) Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of the Royal Statistical Society, Series B*, **70**, 371–388.

Molenberghs G, Fitzmaurice G, Kenward MG, Tsiatis A and Verbeke G, eds (2015) *Handbook of Missing Data Methodology.* Boca Raton: Champam&Hal/CRC.

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

Robins JM and Gill R (1997) Non-response models for analysis of non-monotone ignorable missing data. *Statistics in Medicine,*, **16**, 39–56.

Rubin DB (1986) *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

StataCorp (2013) Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.

Sterne JAC, White IR, Carlin, JB, Spratt, M, Kenward MG, Wood AM, and Carpenter JR (2009) Multiple imputation for missing datain epidemiological and clinical research: potential and pitfalls. *British Medical Journal*, **339**, 157–160.

van Buuren S (2015) Fully conditional specification. Chapter 13 in G Molenberghs *et al.* , eds (2015) *Handbook of Missing Data Methodology.* Boca Raton: Champam&Hal/CRC, pp. 267–294.