A comparative evaluation of currently available software remedies to handle missing data in the context of longitudinal design and analysis.

Wiggins, R.D[1]., Ely, M[2]. & Lynch, K.[3]

[1]Department of Sociology, City University, London, U.K[+] and the Centre for Longitudinal Studies, The Institute of Education, University of London.

[2] The Centre for Research in Health and Social Care, Anglia Polytechnic University

[3] The Centre for Longitudinal Studies, The Institute of Education, University of London.

[+]Address for correspondence: Department of Sociology, City University, Northampton Square, London, EC1 OHB, U.K. Email: **R.D.Wiggins@city.ac.uk**

A paper prepared for 5th International Conference on Social Science Methodology, Cologne, October 2000.

Acknowledgements This work arises from funded activity for training and dissemination under the UK's Economic and Social Research Council's (ESRC) Analysis of Large and Complex Data (ALCD) programme. Grant No. H519255056. The authors are also extremely grateful to Professor John Bynner at The Centre for Longitudinal Studies and Dr Nicholas Longford of De Monfort University for their comments on earlier drafts of the paper.

Structure Abstract

Objectives- to provide an evaluation of a range of currently available software remedies to handle missingness in the context of longitudinal research. **Data Source**- an illustration from the 1958 British Cohort Study compares and contrasts new software remedies to handle missing data. In particular, the SPSS Missing Values Analysis module (SPSS Inc., 1997) is compared to NORM (Schafer, J., 1997) and SOLAS (Statistical Solutions, 1999). **Setting**- England and Wales. **Sample**- a ten per cent random sample (1721 men and women) from the National Child Development Survey (NCDS, otherwise referred to as the 1958 British Birth Cohort Study). The study follows 17,000 births in one week in March 1958 until the present day. **Results-** for a regression model used to predict maths attainment at age 11 years for this cohort we conclude that analyses based on fully observed data (using casewise deletion) may miss, under- or overstate substantive relationships that occur post imputation**. Conclusion-** it is not simply enough to apply missing value analysis or fill-in missing values in a single imputation without taking careful account of missing data uncertainty. It is worth making the effort to use multiple imputation and fully exploit longitudinal data to understand the nature of any potential response bias prior to analysis. Wherever, imputation procedures are used it is recommended that analysts routinely include a casewise indicator in their analysis to record the impact of any missing data adjustment.

1. Introduction

Missing data is a pervasive fact of life. Textbook discussions of missing data generally make the distinction between unit nonresponse (complete absence of any information from a sampled individual or case) and item nonresponse (an individual who cooperates but for some reason has missing values for certain items). As a convention weighting adjustments are used for the former and imputation for the latter. In longitudinal surveys, like the NCDS used in this paper, there is the additional complication of wave nonresponse: individual cohort members who respond to some but not all waves of data

collection (see, Shepherd, 1993). From a longitudinal perspective, this may be viewed as a set of item nonresponses in a longitudinal record, suggesting that imputation may be appropriate (Rovine and Delaney, 1990). From a cross-sectional perspective, it may be viewed as unit nonresponse, for which a weighting adjustment may be appropriate. Thus, weighting, imputation or a combination of both may be considered to compensate for missing data due to wave nonresponse. The main focus of this paper is an evaluation of model-based methods for item nonresponse as described in Rubin (1987) and Scahfer (1997) and, the propensity score method (Rubin, 1987; Rubin and Schenker, 1991) as applied to NCDS data. The work arose originally from a programme of investment in the development of statistical methodology by the UK's Economic and Social Research Council (ESRC). The illustrations presented in the paper use regression analysis to investigate the impact of ignoring any item nonresponse. Three software products are used; notably the SPSS Missing Values Analysis module, SOLAS for missing data analysis (Statistical Solutions Ltd, 1999) and NORM as developed by Schafer (1997). Rather than simply ignore any cases with missing data we apply the likelihood-based estimation procedures available in SPSS and NORM to average over missing data. SPSS adopts an EM (expectation-maximization) algorithm for multivariate normal data (Dempster, Laird and Rubin, (1977) and fills-in (imputes) missing values. For further comparative purpose we also explore the use of SPSS missing values analysis regression and a SOLAS model based method which use multiple linear regression for preselected predictors of missing values. Both NORM and SOLAS use multiple imputation (Rubin, 1987, Schafer, 1997) to generate several (plausible) versions of filled-in data. Each new dataset is analysed separately and the resulting parameter estimates are finally combined using Rubin's (1987) rule for scalar estimands. The next section briefly expands on these approaches for handling item nonresponse.

2. Background to methodology

The dataset for our example can be considered as a rectangular matrix, Y, consisting of n= 1721 cases or rows and p= 5 columns. See figure 1. We have adopted a convention, as required in NORM, to code any item for which there is a missing response as '99' (emboldened for convenience). Full descriptions of the items will follow in section 3.

Figure1: Sample Data matrix to be imputed

99	1	9	3	5	Sample of input data. Note all missing values have been set to **99** and the data set contains no alpha characters.
7	1	19	**99**	**99**	
5	1	8	**99**	**99**	
5	2	6	3	3	
5	1	16	3	4	
5	2	5	4	4	
2	2	26	7	3	
7	2	12	3	3	
4	1	27	5	5	
99	2	6	**99**	**99**	
99	1	32	**99**	**99**	
5	1	15	**99**	**99**	

Figure 2 summarises the extent of missingness in our data according to the number of items missed per case.

Figure 2 The degree of missingness

No. of items missed	**Frequency**	**%**	**Cumulative %**
0	**825**	**47.9**	**47.9**
1	**230**	**13.4**	**61.3**
2	**331**	**19.2**	**80.3**
3	**166**	**9.6**	**90.2**
4	**169**	**9.8**	**100.0**
Total	**1721**		

Conventional textbooks in advanced statistical methods would simply assume that the matrix is fully observed and there are no missing value indicators. The simplest way to obtain a fully observed data set is, of course, to simply delete any case with at least one missing value code. This would reduce our dataset to 825 cases, a loss of 52% of our original sample. The extent of this loss is illustrated in figure 2.

Loss of information is likely to lead to inevitable bias unless the deleted cases are *missing completely at random (MCAR)*. Casewise deletion would assume that the missing values are a simple random sample of all data values. Even if this assumption were true our estimates would still be less efficient. Likelihood estimation procedures in SPSS and NORM assume that the missing data are *missing at random (MAR)*. At first, this may appear to contradict the MCAR label. If data is missing at random then surely the missing data must be a random sample of all data values? Not so, MAR is actually less restrictive than MCAR. If you like, completely means exactly that, 'completely at random' whereas 'at random' implies that missing values behave like a random sample of all values within subclasses defined by observed data. Rubin (1976) defines MAR more formally by including a distribution for response indicator variables that take the value 1 if an item is recorded and 0 otherwise (the vector R below). With the MAR assumption this distribution of missingness patterns does not depend on the missing values. The missing data mechanism is said to be *ignorable.* Schafer (1987) puts this crucial assumption made by ignorable methods 'not that the propensity to respond is completely unrelated to the missing data, but that this relationship can be explained by data that are observed'. What is observed not only includes the values included in the data matrix, but also the patterns of missingness themselves. Algebraically, we can summarise the MAR assumption as follows:

Firstly, partition Y as Y_{obs} = observed data and Y_{miss} = missing data then posit a probability model for R, $P(R|Y,\gamma)$ which depends on Y as well as some unknown parameters, γ. The MAR assumption is that this distribution does not depend on Y_{miss}.

Put another way,

$$P(R|Y, \gamma) = P(R| Y_{obs}, Y_{miss}, \gamma) = P(R| Y_{obs}, \gamma) \quad (1)$$[1]

Various approaches to filling-in incomplete data that do not assume ignorability are beyond the scope of this paper. Little and Rubin (1987) review some of these methods, whilst Schafer (1997) contains references to more recent applications.

Formally, let Y denote the 1721 x 5 (n x p) matrix of *complete* data, which is *not* fully observed, where y_i denotes the ith row of Y. By the iid assumption, the probability density function of the complete data is written

$$P(Y| \theta) = {}^{n}\Pi_{i=1} f(y_i | \theta) \quad (2)$$

where f is the probability function for a single row, and θ is a vector of unknown parameters.

Now, consider the imputation (filling-in) procedures themselves.

NORM

The implementation of EM for multivariate normal data with an arbitrary pattern of missingness is described in detail in Schafer (1997). To summarise consider a matrix of missingness patterns corresponding to Y. In general, there will be S unique patterns appearing in the data matrix (all possible ways of observing 1 missing value amongst p variables, 2 missing values amongst p and so on up to a maximum of p-1 missing values from p). This is vital information for the E(xpectation)-step. Essentially, Y_{miss} is predicted from Y_{obs} and θ . The distribution $P(y_{i\,(miss)} | y_{i\,(obs)}, \theta)$ is a multivariate normal linear regression of $y_{i(miss)}$ on $y_{i(obs)}$ for the missingness pattern corresponding to row i.

[1] Technically, the missing data mechanism is said to be ignorable if both MAR and *distinctness* hold. The latter condition simply implies that knowledge of θ provides little information about γ and vice-versa. Alternatively, the propensity to respond is not related to the missing values. Schafer provides a number of examples where ignorability is known not to hold.

(Schafer, 1997, pp.164-166*).* Once Y_{miss} has been predicted the M(aximistation) step is fairly straightforward. Further estimates of θ are obtained from Y_{obs} and Y_{miss} until the parameter estimates converge. The rate of convergence will depend on the fractions of missing information. As might be expected high rates of missingness will slow convergence.

Schafer argues that multiple imputation shares the same underlying philosophy as EM[2] solving an incomplete-data problem by repeatedly solving the complete-data version. It is a simulation-based approach to missing data. Our multiple imputations need to be *proper*, by which Rubin (1987) defines as iterations of Y_{miss} that are suitably separated to ensure independence. Otherwise there is a danger that successive iterates of Y_{miss} will be correlated. Iterates of Y_{miss} $\{Y_{miss}^{(t)}, Y_{miss}^{(2t)}, \ldots. Y_{miss}^{(mt)}\}$are collected using data augmentation (DA) such that t is large enough to achieve independence. DA bares strong similarity to EM where the E-and M-steps are replaced by stochastic I(mputation)- and P(osterior)-steps (Tanner and Wong, 1987). The term DA arose from applications of these algorithms to Bayesian inference with missing data. They represent a class of Markov chain Monte Carlo (MCMC) techniques for creating psuedorandom draws from probability distributions (see Gilks, Richardson and Spieglehalter (1996) for an overview). On the I-step draw a value of the missing data from Y_{miss}

$$Y_{miss}^{(t+1)} \sim P(Y_{miss} \mid Y_{obs}, \theta^{(t)}) \qquad (3a)$$

and on the P-step draw a new value of θ from the complete data-posterior $(P(\theta|Y_{obs}, Y_{miss})$.

$$\theta^{(t+1)} \sim P(\theta \mid Y_{obs}, Y_{miss}^{(t+1)}) \qquad (3b)$$

This creates a Markov chain $Y_{miss}^{(1)}, \theta^{(1)}, Y_{miss}^{(2)}, \theta^{(2)}, \ldots\ldots$where the distribution of each pair of estimates only depends on the previous draw (the Markov chain property) which

[2] and data augmenation where the E- and M-steps are replaced by stochastic I- and P-steps (Tanner and Wong, 1987).

eventually converges after repeated applications of 3a and 3b from an initial starting value for θ.[3] The chain is said to have converged (or achieved stationarity) by t iterations if $\theta^{(t)}$ is independent of $\theta^{(0)}$, $\theta^{(2t)}$ is independent of $\theta^{(t)}$, etc. This is assessed by means of time series plots (parameter estimates versus iteration number) and autocorrelation functions (ACF's) of θ (the lag-k autocorrelation versus k for parameter estimates). Step by step illustrations of a NORM imputation run are given in Wiggins et al., (1999). The data for this illustration is also available on request.

SOLAS

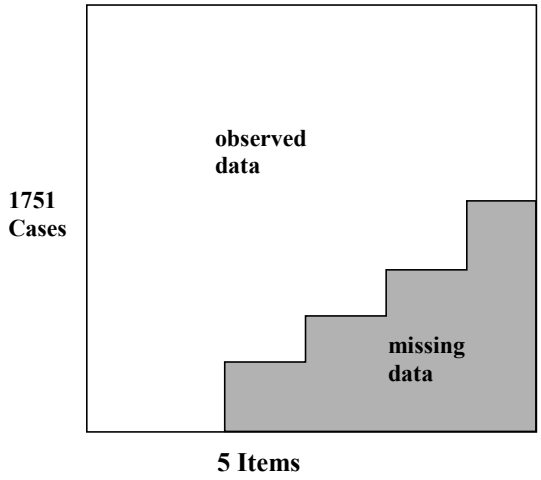
Secondly we look at methods used by SOLAS (1999). In SOLAS each variable is filled in using separate multiple regression models, unlike NORM where all missing variables with missing data are imputed simultaneously in a multivariate regression. SOLAS provides two methods for filling in data: the predictive model based method and the propensity score method. For both of these methods, the approach used by SOLAS is to first organise the missing data in a monotone pattern. A monotone pattern is one where the data matrix of observed values looks rather like a staircase as shown in figure 3. The monotone pattern is achieved in SOLAS by two steps. Firstly, data is sorted by case and variable order to get as close as possible to a monotone pattern. Secondly, any missing values that destroy the monotone pattern are multiply imputed using a series of available-case regressions. That is, for each pattern of missing data, missing values of each variable are imputed from a regression equation using all available observed and previously imputed values of the variables selected for the model. These imputations are independently generated by randomly drawing regression parameters and error terms from the posterior distribution given the observed data (Rubin, 1987).

The advantage of a monotone pattern of missing data is that imputation in a multivariate data set is reduced to a series of single variable imputations. Each variable is imputed starting with the variable which has the lowest proportion of missingness and proceeding left to right throughout the monotone structure (Figure 3). The user specifies independent variables for each imputation. These predictors are either completely observed covariates

[3] For more on choice of starting values see Schafer (1987, pp.85-7)

or previously imputed variables. In our example the covariate is sex and all the variables other than the one being imputed are selected, those which are positioned to the left in the monotone structure being used. These monotone imputations are generated using either the predictive model based or the propensity score method.

Figure 3 Data with a monotone pattern of missingness



The predictive model based method uses ordinary least squares regression. The variable to be imputed is regressed on the chosen covariates and previously imputed. The m imputations are independently generated using the values predicted by the regression equation, with the parameters randomly drawn from the posterior distribution given the observed data (Rubin, 1987; Gelman et al., 1995) and a randomly drawn error term. Further details are given in SOLAS^{TM} 2.0 User Reference, Appendix C.

The propensity score method is based on a logistic regression (Everitt, 1977). An indicator variable r_j for the missing variable y_j is regressed on the chosen covariates (including the previously imputed variables). The propensity score is the conditional probability of missingness given the vector of observed covariates. The imputations for each missing value of $y_{j\,(miss)}$ are independent random draws from a subset of observed values of $y_{j(obs)}$, called a donor pool, with propensity scores close to that assigned to the

case with missing data. Donor pools can be defined in various ways (SOLAS^{TM} 2.0 User Reference, Appendix E). The method adopted here is to divide the ordered propensity scores into quintiles (Lavori et al., 1995). A random sample, equal in size to the observed responses, is drawn , with replacement, from the $y_{j(obs)}$ within the propensity quintile. This sample is called the posterior predictive distribution. For each missing case, the imputed value of y_j is a random draw (with replacement) from the posterior predictive distribution. Formally, the imputations are drawn by an Approximate Bayesian Bootstrap method (Rubin and Schenker, 1991). This process is repeated m times to produce m imputations for each missing case.

Both SOLAS and NORM are multiple imputation methods in which the Y_{miss} are replaced by a number of plausible versions (m) of Y_{miss} where m>1. Each of the m filled-in datasets are then analysed by conventional methods (multiple regression etc) and the results are combined. The variability amongst the resulting parameter estimates provides a measure of uncertainty due to missing data. This is combined with measures of ordinary variation to produce a single measure of variation (the annex contains Rubin's rule for scalar estimands). This represents a major advantage over EM which will only provide one complete dataset and no standard errors. Multiple imputation can be highly efficient , even for very few m. If the fraction of missing information about a scalar estimand is λ, the efficiency of an estimator based on m imputations is $(1+ \lambda/m)^{-1}$ on a variance scale (Rubin, 1987, p.114). When, λ =0.3, for example, an estimate based on m=5 will be 94% efficient. Put another way, the standard error will only be about 1.03 ($\sqrt{}$ (1 + 0.3/5)) times as large as the estimate with an infinite value for m.

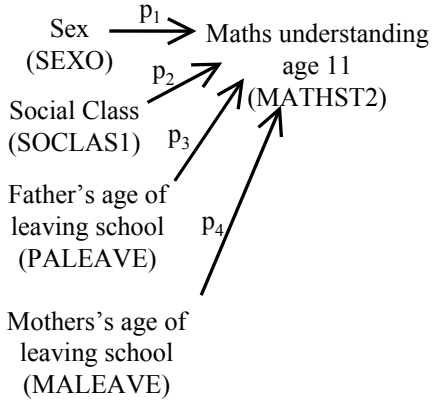
To complete our evaluation we selected two *single imputation* model based procedures available under the SPSS Missing Values Analysis (1997) option: regression and EM methods. The regression method simply uses multiple regression based on a prespecified set of independent variables to estimate missing values and adjusts the predicted values by adding a random component. Under the default option used in our example a randomly selected residual based on complete case analysis is added to the predicted

value. The EM algorithm available in SPSS is decscribed above as the first step in a NORM imputation. Each of these options fills in missing values once only.

3. Analytical context using 1958 British Cohort Study data

The decision to base our comparative evaluation on a 10 per cent sample of cases from the 1958 British Cohort Study was simply a convenience for the potential user who wishes to replicate our analyses or try alternative missing-data procedures (the data is available on request from the authors). The analytical approach is to capitalise on the temporal sequencing of the data to build a series of path diagrams in the context of exploring the relationship between a child's birth antecedents subsequent educational performance and ultimately, attitudes to politics (Bynner, Ukoumunne and Wiggins, 1996).

Figure 4:Path Diagram for evaluation



Data from the NCDS have been collected for people born in the first week of March 1958 at ages 7, 11, 16, 23 and 33. The original sample was collected as part of a perinatal mortality survey (PMS). The illustration that follows simply focuses on the relationship between birth antecendents and maths understanding age 11 years. The path diagram is presented in figure 4.

Table 1 lists the 5 variables used in our first regression model based on fully observed data. This analysis is subsequently used to illustrate the impact of ignoring incomplete cases in section 6. *Every* case analysed here has a code for gender. For NCDS wherever there was a missing value for gender (1.8% of all cases) there was no other information for the variables selected in this illustration. Otherwise the level of item non-response varied from around 22% (social class of father at birth) to 39% (father's age of leaving

school). Viewed from a perspective of unit non-response the presence of a gender code for every case ensures complete recovery of lost information under subsequent imputation.

Table 1: Variables used in regression modelling

SPSS label	**NORM label***	**Scoring** (% item non-response in parentheses)
SOCLAS1	VAR_1	1=high to 7=low (21.8)
SEX	VAR_2	1=girl, 0=boy (0.0)
MATHST2	VAR_3	Standardised maths score (22.7)
PALEAVE3	VAR_4	Father's age of leaving school (38.8)
MALEAVE3	VAR_5	Mother's age of leaving school (37.8)

*Used in section 6.

4. Analytical strategy for comparative evaluation

Having a strategy to handle missing data is necessary for exploiting all of the available information. Schafer (1998) argues that imputation modelling should not only include variables that are crucial to the analysis but also predictive of them and/or missingness. In a birth cohort study we have rich baseline data that can help us understand the impact of any panel attrition (Shepherd, 1993). In particular we are able to gain insight about the characteristics of those cases with partial (item) nonresponse by comparing their characteristics at birth with those of the cases who have fully observed data. We have extended our imputation procedures to include certain items recorded during the PMS wave (0). Four variables are included for illustrative purposes. These are listed below in table 2:

Table 2: Additional variables included from the Perinatal Mortality Survey used in imputation procedures

NORM label	Description
VAR_6	Persons per room
VAR_7	Birthweight
VAR_8	Mother's smoking during pregnancy
VAR_9	Age of mother at first birth

Each of these variables predicts the degree of missingness (as indicated by NM in the analysis: a simple count of the number of missing values per case). Low birthweight cohort members appear to be underrepresented amongst those with fully observed data (NM=0) as do those whose mothers had their first child as a teenager, smoked heavily during the pregnancy (leading to the cohort member's birth) and those who were living in accommodation with more than 2 persons per room. Full details are given in Wiggins et al., (1999).

In addition to using the PMS variables to enhance our multiple imputations we ran all regressions with and without NM present in the modelling. If cases with greater degrees of missingness are more similar to those without any information at all than those who fully cooperate then one might expect this indicator to act as a proxy to detect any differences between cases that might remain after filling-in missing values. In this way, NM may flag the need to review the ignorability assumption.

In sum, the comparative analysis that follows first contrasts regression parameter estimates based on the fully observed (casewise deletion) data alone with that based on NORM, SOLAS predictive model, SOLAS propensity score, SPSS regression and SPSS EM for missing values analysis. All analyses are then repeated in the presence of the NM indicator. Finally, all missing data procedures are replicated to include the four PMS

variables as predictors of missing values for the explanatory variables included in the regression.

It was decided average estimates for multiple imputation over 5 imputations. Based on a complete matrix of 1721x5 (=8605) cells of information the fraction of missingness was 24 per cent.[4] Under these conditions (Schafer, 1998) multiple imputation estimates with m=5 are expected to be around 95 per cent efficient.

5. Results

All analyses first present the regression results obtained under casewise deletion (n=825 fully observed cases) with single imputation and then with the multiple imputation models. Tables are therefore divided as a) and b). For convenience the results for analyses of fully observed data is reproduced in both sets of tables. Wherever estimated regression coefficients are more than double their standard error they have been italiscised and emboldened. Wherever a standard error is *greater than* the equivalent value for the fully observed analysis it has been underscored.

[4] This can be derived directly from figure 2.

5.1 Regression analyses using model variables to fill-in missing values

Table 4 Comparison of regression analyses for fully observed data (n=825) with imputation analyses (n=1721).

Table 4 (a) Analysis for fully observed data compared to single imputation models

	Fully Observed		**SPSS Reg**		**SPSS EM**	
	B	**StErr**	**B**	**StErr**	**B**	**StErr**
Constant	19.868	2.221	17.362	1.496	17.748	1.560
Sex	-.916	.637	***-.970***	.446	***-.849***	.387
Soclas1	***-1.719***	.229	***-1.315***	.155	***-1.669***	.161
Maleave3	***.652***	.309	***.810***	.193	***.592***	.230
Paleave3	***1.203***	.263	***1.082***	.162	***1.558***	.198

Table 4 (b) Analysis for fully observed data compared to multiple imputation models

	Fully Observed		**SOLAS Propensity**		**SOLAS Predictive**		**Norm**	
	B	**StErr**	**B**	**StErr**	**B**	**StErr**	**B**	**StErr**
Constant	19.868	2.221	18.052	1.838	17.524	2.321	19.166	1.933
Sex	-.916	.637	-0.661	0.487	-0.985	0.542	-0.982	0.495
Soclas1	***-1.719***	.229	***-1.516***	0.189	***-1.447***	0.246	***-1.596***	0.168
Maleave3	***.652***	.309	***0.71***	0.195	***0.832***	0.252	0.6708	0.387
Paleave3	***1.203***	.263	***1.149***	0.265	***1.161***	0.183	***1.0931***	0.249

An analysis based on the fully observed data alone would find little evidence that gender has little effect on maths attainment aged 11 years whereas being born to a father with

low social class and having parents who have left school early will significantly lower a cohort member's attainment score. SPSS regression and EM analyses would only strictly affect the substantive finding for gender. In both analyses, girls are significantly disadvantaged in their maths performance at age 11 years. However, under SPSS EM the impact of father's age of leaving school looks to be relatively stronger than under SPSS regression. Averaging over both SOLAS and NORM imputations leaves the gender non-significant. Mother's age of leaving school is now non-significant under NORM whereas the effect of both father's and mother's age of leaving school remain strong under SOLAS models. It would appear that the effect for social class is most consistent across these analyses. What about if we now include the missing indicator, NM, following the imputation procedures?

The results are presented in table 5 below:

Table 5: Comparison of regression analyses for fully observed data (n=825) with post-imputation analyses (n=1721) that include a degree of missingness indicator (NM).

Table 5 (a) Analysis for fully observed data compared to single imputation models

	Fully Observed		**SPSS Reg**		**SPSS EM**	
	B	**StErr**	**B**	**StErr**	**B**	**StErr**
Constant	19.868	2.221	17.532	1.496	18.150	1.560
Sex	-.916	.637	***-.980***	.446	***-.877***	.386
Soclas1	***-1.719***	.229	***-1.316***	.155	***-1.662***	.161
Maleave3	***.652***	.309	***.804***	.193	***.599***	.230
Paleave3	***1.203***	.263	***1.085***	.162	***1.576***	.198
NM			-.133	.156	***-.385***	.135

Table 5 (b) Analysis for fully observed data compared to multiple imputation models

	Fully Observed		**SOLAS Propensity**		**SOLAS Predictive**		**Norm**	
	B	**StErr**	**B**	**StErr**	**B**	**StErr**	**B**	**StErr**
Constant	19.868	2.221	18.209	1.858	17.88	2.313	19.600	1.837
Sex	-.916	.637	-0.667	0.486	-1.000	0.543	-0.996	0.494
Soclas1	***-1.719***	.229	***-1.514***	0.190	***-1.442***	0.245	***-1.600***	0.168
Maleave3	***.652***	.309	***0.709***	0.195	***0.828***	0.257	0.667	0.382
Paleave3	***1.203***	.263	***1.151***	0.265	***1.177***	0.185	***1.0941***	0.243
NM			-0.133	0.198	-0.335	0.245	-0.3198	0.227

Compared to the previous contrasts our substantive conclusions remain broadly intact. However, what we glean from the SPSS regression and EM modelling is that the impact of the degree of missingness reported for a case is somewhat contradictory. Under both SPSS regression and SPSS EM we witness a negative effect but statistical significance is only reported under SPSS EM. The effect remains negative under both SOLAS and NORM but not significant. Our substantive conclusions also concur with those obtained under the previous multiple imputation analyses in table 4(b). Social class and the age father's left school remain consistently significant in these analyses. Gender remains non-significant throughout whereas the age mother's left school is only significant under the SOLAS analyses. Under all analyses that include a degree of missingness indicator we would conclude that those individuals with higher degrees of missingness are much more likely to have lower maths scores. However, the seriousness of the relative impact of the indicator varies. But what would happen, if we now take greater account of birth characteristics (from the PMS) that may be predictive of missingness in our imputation procedures? The results are presented in table 6 below.

5.2 Regression analyses using model variables and selected birth characteristics to fill-in missing values

Table 6: Comparison of regression analyses for fully observed data (n=825) with imputation procedures (n=1721) that include PMS characteristics.

Table 6 (a) Analysis for fully observed data compared to single imputation models

	Fully Observed		**SPSS Reg**		**SPSS EM**	
	B	**StErr**	**B**	**StErr**	**B**	**StErr**
Constant	19.868	2.221	19.577	1.578	18.488	1.548
Sex	-.916	.637	-.710	.455	***-.888***	.388
Soclas1	***-1.719***	.229	***-1.566***	.163	***-1.732***	.159
Maleave3	***.652***	.309	***.633***	.191	***.749***	.228
Paleave3	***1.203***	.263	***.955***	.168	***1.434***	.199
NM			***-.368***	.159	***-.433***	.136

Table 6 (b) Analysis for fully observed data compared to multiple imputation models

	Fully Observed		**SOLAS Propensity**		**SOLAS Predictive**		**Norm**	
	B	**StErr**	**B**	**StErr**	**B**	**StErr**	**B**	**StErr**
Constant	19.868	2.221	21.583	1.657	19.265	1.853	19.850	1.671
Sex	-.916	.637	-0.742	0.497	-0.704	0.480	-0.704	0.622
Soclas1	***-1.719***	.229	***-1.735***	0.170	***-1.674***	0.206	***-1.739***	0.246
Maleave3	***.652***	.309	***0.579***	0.253	***0.797***	0.234	***0.768***	0.262
Paleave3	***1.203***	.263	***0.742***	0.191	***1.056***	0.192	***1.103***	0.262
NM			***-0.404***	0.191	***-0.48***	0.184	-0.335	0.268

Again, the substantive findings are very similar to those above. Although, if one were dependant on SPSS EM alone a significant gender effect would be reported. Now, it would also appear that NORM has produced a significant finding for the impact of mother's age of leaving school. Ironically, the substantive conclusions under multiple imputation analyses concur with those obtained under casewise deletion.

Under single imputation methodologies estimates may appear to be more precise than they really are. This may well account for the significant findings for the effect of gender under these analyses. The estimated standard errors for multiple imputation methods reflect the uncertainty in the filling-in procedure. They are larger than those obtained under single imputation. Generally, however the standard errors under multiple imputation are smaller than those obtained under the analysis of fully observed data (there are some exceptions that confirm the cautious averaging under Rubin's Rule). This is comforting as the sample size has doubled.

What we able to argue is that the multiple imputation results in table 6 when birth (PMS) characteristics are used in the imputation procedure are high quality. The imputation models in table 6 are more informative about incompleteness. However, the conclusion for the impact of the degree of missingness indicator (NM) is less clear. Results under NORM and SOLAS diverge and, therefore, suggest the need to investigate the ignorability assumption.

6. Discussion

This paper has confirmed the importance of carrying out analyses in the presence of missing data. Whilst it may be convenient and easy to jettison partial information it is a risky action which carries with it the rigid assumption of MCAR. With longitudinal birth cohort data we are in an advantageous position. It is possible to carefully select predictors of missingness and routinely include them in multivariate analyses. One barometer of the efficacy of so doing is to include an indicator of the level of imputation present for any case. The illustration has demonstrated that even with easy application remedies in SPSS substantive conclusions might be awry and the propensity of those with partial information to respond poorly adjusted.

NORM is theoretically sound. The procedure assumes multivariate normality. For categorical variables such as gender this can be problematic. At present special provision has to be made to truncate imputed missing values for these variables. Whilst (Schafer, 1997) presents the necessary theory to handle non-normal data the software is not yet readily available. NORM also depends on the iterative MCMC process converging.

A clear advantage under the SOLAS propensity score method is that imputed values are drawn from observed values. Thus ensuring that only plausible values are presented without any underlying distributional assumptions. However the authors do not generally recommend the SOLAS propensity score method for inferences about associations as opposed to marginal distributions. The relationships between variables are not well preserved under this approach (Ely, 2001).

Both NORM and SOLAS require a number of iterates or replicates as essential means to adjust for missing data uncertainty. This may be a deterrent to the data analyst who will have to store resulting filled-in data sets (typically around 5) and average parameter estimates and compute standard errors for subsequent analyses.

In longitudinal research where the pattern of attrition falls naturally into a monotone structure data is managed equally well under both SOLAS and NORM. Whilst SOLAS does not require data to be monotone, the method used to impute data that does not fit a monotone pattern is less principled than under NORM (Rubin, 2000).

Whatever reservations the analyst might have about using NORM or SOLAS they hold inferential superiority over easy to implement single imputation methods, such as those currently available in SPSS (Heijan and Rubin, 1990; Raghunathan and Paulin, 1998). This is because single imputation methods fill-in data but leave uncertainty improperly estimated.

There is no quick or easy solution to the problem of how to handle missing data. Analyses based on imputations will only be ‘approximately correct’ (Schafer, 1998). This paper demonstrates of how the gap between theory and practice can be bridged by making our assumptions about the response process (or missing data mechanism) transparent. This can advance our understanding of the complexities and richness of data

sources and fully exploit the information they contain. To conclude with a suitable reminder from Little and Rubin (1987):

'knowledge, or absence of knowledge, of the mechanisms that led to certain values being missing is a key element in choosing an appropriate analysis and in interpreting the results'

References

Bynner, J., Ukoumunne, O. and Wiggins, R.D. (1996) Modelling childhood antecedents of political cynicism using structural equation modelling. NCDS Working Paper, No.51, Centre for Longitudinal studies, Institute of Education, University of London, U.K.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, 39, 1-38.

Ely, M. (2001) Practical ways of dealing with missing data on alcohol consumption using diet diaries in a birth cohort study. PhD thesis Department of Epidemiology and Public Health, University College London. Forthcoming.

Everitt, B. (1977). The Analysis of contingency tables. Chapman and Hall, London.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. B. (1995) Bayesian Data Analysis. New York: Chapman and Hall.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (eds.) (1996) *Markov-Chain Monte Carlo in Practice.* Chapman & Hall, London.

Heijan, D. F. and Rubin, D. B. (1990) Inferences from course data via multiple imputation with application to age heaping. Journal of the American Statistical Association, 85 410 304-314.

Lavori, P.W., Dawson, R. and Shera, D. (1995) A multiple imputation strategy for clinical trials with truncation of patient data Statistics in Medicine 14, 1913-1925.

Raghunathan, T. E. and Paulin, G. D. (1998) Multiple Imputation in the Consumer Expenditure Survey: evaluation of statistical inference. Proceedings of the Business and Economics Section of the American Statistical Association, 1-10.

Rovine, M.J. and Delaney, M.(1990) Missing data estimation in developmental research, in Von Eye, A. (ed.) *Statistical Methods in Longitudinal Research, Vol.1., Principles and Structuring Change,* New York: Academic Press.

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

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

Rubin, D. (2000). Software for multiple imputation. In conference proceedings on 'Challenging statistical issues in clinical trials', Cambridge., sponsored by Statistical Solutions Ltd.

Rubin, D. B. and Schenker, N. (1991) Multiple Imputation in health-care databases: an overview and some applications. Statistics in Medicine 10, 585-598.

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

Schafer, J.L. (1998) *The practice of multiple imputation*. Lecture notes prepared for Centre for Applied Social Surveys, University of Southampton, U.K.

Sheherd, P. (1993) The analysis of response bias., Appendix in *Life at 33*, edited by Ferri, E.

SPSS Inc. (1997). *SPSS Missing Value Analysis 7.5*.

Statistical Solutions Ltd (1999) SOLAS™ 2.0 For Missing data Analysis User Reference Manual.

Tanner, M.A. and Wong, W.H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association,* 82, 528-550.

Wiggins, R.D., Lynch, K., Gleave, S. and Bynner, J. (1999) Teaching applied multivariate analysis in the context of missing data: a comparative evaluation of current software remedies. Paper presented to ICSN, Portland, Oregon, USA.

Annex : Rubin's Rule for obtaining scalar estimands and corresponding estimates of variation

After obtaining m imputation s of Y_{mis} analyze the m completed datasets and combine the result

Rubin' s (1987) rule for scalar estimands :

$\hat{Q} = \text{complete data point estimate}$

$U = \text{complete - data variance estimate}$

$$\overline{Q} = m^{-1} \sum_{t=1}^{m} \hat{Q}^{(t)}$$

$$B = (m-1)^{-1} \sum_{t=1}^{m} (\hat{Q}^{(t)} - \overline{Q})^2$$

$$\overline{U} = m^{-1} \sum_{t=1}^{m} U^{(t)}$$

$$T = \overline{U} + (1 + m^{-1}) B$$

Interval estimate is $\overline{\mathrm{Q}} \pm \mathrm{t}_{\mathrm{v}} \sqrt{T}$ where

$$v = (m-1)\left[1 + \frac{\overline{U}}{(1+m^{-1})B}\right]^2$$