Dealing with missing data: key assumptions and methods for applied analysis

Technical Report · May 2013

CITATIONS
104

READS
2,472

1 author:

Marina Soley-Bori
King's College London
28 PUBLICATIONS **297** CITATIONS

SEE PROFILE

Technical Report No. 4 May 6, 2013

Dealing with missing data: Key assumptions and methods for applied analysis

Marina Soley-Bori
msoley@bu.edu

Boston University School of Public Health
Department of Health Policy & Management

This paper was published in fulfillment of the requirements for PM931 Directed Study in Health Policy and Management under Professor Cindy Christiansen's (cindylc@bu.edu) direction. Michal Horný, Jake Morgan, Kyung Min Lee, and Meng-Yun Lin provided helpful reviews and comments.

Contents

Executive Summary

This tech report presents the basic concepts and methods used to deal with missing data. After explaining the missing data mechanisms and the patterns of missingness, the main conventional methodologies are reviewed, including Listwise deletion, Imputation methods, Multiple Imputation, Maximum Likelihood and Bayesian methods. Advantages and limitations are specified so that the reader is able to identify the main trade-offs when using each method. The report also summarizes how to carry out Multiple Imputation and Maximum Likelihood using SAS and STATA.

Keywords: missing data, missing at random, missing completely at random, listwise deletion, imputation, multiple imputation, maximum likelihood.

Acronyms

- MCA -Missing Completely at Random
- MAR -Missing at Random
- NMAR -Not Missing at Random
- MI -Multiple Imputation
- ML -Maximum Likelihood
- MCMC- Markov Chain Monte Carlo
- FCS-Fully conditional specification
- EM-Expectation Maximization
- OCDE-Organization for Economic Cooperation and Development

1. Introduction

Missing data is a problem because nearly all standard statistical methods presume complete information for all the variables included in the analysis. A relatively few absent observations on some variables can dramatically shrink the sample size. As a result, the precision of confidence intervals is harmed, statistical power weakens and the parameter estimates may be biased. Appropriately dealing with missing can be challenging as it requires a careful examination of the data to identify the type and pattern of missingness, and also a clear understanding of how the different imputation methods work. Sooner or later all researchers carrying out empirical research will have to decide how to treat missing data. In a survey, respondents may be unwilling to reveal some private information, a question may be inapplicable or the study participant simply may have forgotten to answer it. Accordingly, the purpose of this report is to clearly present the essential concepts and methods necessary to successfully deal with missing data.

The rest of the report is organized as follows: Section 2 and 3 explain the different missing data mechanisms and the patterns of missingness. Section 4 presents the main methods for dealing with missing data. I differentiate between 'conventional methods', which include Listwise Deletion and Imputation Methods, and 'advanced methods', which cover Multiple Imputation, Maximum Likelihood, Bayesian simulation methods and Hot-Deck imputation. Finally, section 5 explains how to carry out Multiple Imputation and Maximum Likelihood using SAS and STATA. The report ends with a summary of other software available for missing data and a list of the useful references that guided this report.

Across the report, bear in mind that I will be presenting 'Second-Best' solutions to the missing data problem as none of the methods lead to a data set as rich as the truly complete one.

"The only really good solution to the missing data problem is not to have any. So in the design and execution of research projects, it is essential to put great effort into minimizing the occurrence of missing data. Statistical adjustments can never make up for sloppy research" (Paul D. Allison, 2001)

2. Missing data mechanisms

There are different assumptions about missing data mechanisms:

a) Missing completely at random (MCAR): Suppose variable Y has some missing values. We will say that these values are MCAR if the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variable in the data set. However, it does allow for the possibility that "missingness" on Y is related to the "missingness" on some other variable X. (Briggs et al., 2003) (Allison, 2001)

***Example:** We want to assess which are the main determinants of income (such as age). The MCAR assumption would be violated if people who did not report their income were, on average, younger than people who reported it. This can be tested by dividing the sample into those who did and did not report their income, and then testing a difference in mean age. If we fail to reject the null hypothesis, then we can conclude that the MCAR is mostly fulfilled (there could still be some relationship between missingness of Y and the values of Y).

b) Missing at random (MAR)-*a weaker assumption than MCAR*-: The probability of missing data on Y is unrelated to the value of Y after controlling for other variables in the analysis (say X). Formally: P(Y missing|Y,X) = P(Y missing|X) (Allison, 2001).

***Example**: The MAR assumption would be satisfied if the probability of missing data on income depended on a person's age, but within age group the probability of missing income was unrelated to income. However, this cannot be tested because we do not know the values of the missing data, thus, we cannot compare the values of those with and without missing data to see if they systematically differ on that variable.

c) Not missing at random (NMAR): Missing values do depend on unobserved values.

***Example:** The NMAR assumption would be fulfilled if people with high income are less likely to report their income.

If MAR assumption is fulfilled: The missing data mechanism is said to be ignorable, which basically means that there is no need to model the missing data mechanism as part of the estimation process. These are the method this report will cover.

If MAR assumption is not fulfilled: The missing data mechanism is said to be nonignorable and, thus, it must be modeled to get good estimates of the parameters of interest. This requires a very good understanding of the missing data process.

3. Patterns of missingness

We can distinguish between two main patterns of missingness. On the one hand, data are missing monotone if we can observe a pattern among the missing values. Note that it may be necessary to reorder variables and/or individuals. On the other hand, data are missing arbitrarily if there is not a way to order the variables to observe a clear pattern (SAS Institute, 2005).

Missing monotone				Missing arbitrarily			
v1	**v2**	**v3**	**v4**	**v1**	**v2**	**v3**	**v4**
X	X	X	X	X	X	.	X
X	X	X	X	.	X	X	.
X	X	X	.	X	.	X	.
X	X	.	.	X	X	.	.
X	.	.	.	.	X	X	X

Assumptions and patterns of missingness are used to determine which methods can be used to deal with missing data

4. Methods for handling missing data

4.1. Conventional methods

4.1.1. Listwise deletion (or complete case analysis)**:** If a case has missing data for any of the variables, then simply exclude that case from the analysis. It is usually the default in statistical packages. (Briggs et al.,2003).

Advantages: It can be used with any kind of statistical analysis and no special computational methods are required.

Limitations: It can exclude a large fraction of the original sample. For example, suppose a data set with 1,000 people and 20 variables. Each of the variables has missing data on 5% of the cases, then, you could expect to have complete data for only about 360 individuals, discarding the other 640.

It works well when the data are missing completely at random (MCAR), which rarely happens in reality (Nakai & Weiming, 2011).

4.1.2. Imputation methods**:** Substitute each missing value for a reasonable guess, and then carry out the analysis as if there were not missing values.

There are two main imputation techniques:

- *Marginal mean imputation*: Compute the mean of X using the non-missing values and use it to impute missing values of X.

  Limitations: It leads to biased estimates of variances and covariances and, generally, it should be avoided.

- *Conditional mean imputation*: Suppose we are estimating a regression model with multiple independent variables. One of them, X, has missing values. We select those cases with complete information and regress X on all the other independent variables. Then, we use the estimated equation to predict X for those cases it is missing.

  If the data are MCAR, least-squares coefficients are consistent (i.e. unbiased as the sample size increases) but they are not fully efficient (remember, efficiency is a measure of the optimality of an estimator. Essentially, a more efficient estimator, experiment or test needs fewer samples than a less efficient one to achieve a given performance). Estimating the model using weighted least squares or generalized least squares leads to better results (Graham, 2009) (Allison, 2001) and (Briggs et al., 2003).

Limitations of imputation techniques in general: They lead to an underestimation of standard errors and, thus, overestimation of test statistics. The main reason is that the imputed values are completely determined by a model applied to the observed data, in other words, they contain no error (Allison, 2001).

> Statistics has developed two main new approaches to handle missing data that offer substantial improvement over conventional methods: Multiple Imputation and Maximum Likelihood.

4.2. Advanced Methods

4.2.1. Multiple Imputation

The imputed values are draws from a distribution, so they inherently contain some variation. Thus, multiple imputation (MI) solves the limitations of single imputation by introducing an additional form of error based on variation in the parameter estimates across the imputation, which is called “between imputation error”. It replaces each missing item with two or more acceptable values, representing a distribution of possibilities (Allison, 2001).

MI is a simulation-based procedure. Its purpose is not to re-create the individual missing values as close as possible to the true ones, but to handle missing data to achieve valid statistical inference (Schafer, 1997).

It involves 3 steps:

a) Running an imputation model defined by the chosen variables to create imputed data sets. In other words, the missing values are filled in *m* times to generate *m* complete data sets. *m*=20 is considered good enough. Correct model choices require considering:
   - Firstly, we should identify which are the variables with missing values.
   - Secondly, we should compute the proportion of missing values for each variable.
   - Thirdly, we should assess whether different missing value patterns exist in the data (**SAS** helps us doing this), and try to understand the nature of the missing values. Some key questions are:

     -Are there a lot of missing values for certain variables? (E.g. Sensitive question, data entry errors?)

     -Are there groups of subjects with very little information available? (E.g. Do they have something in common?)

     -Which is the pattern of missingness? Monotone or arbitrary?
b) The *m* complete data sets are analyzed by using standard procedures
c) The parameter estimates from each imputed data set are combined to get a final set of parameter estimates.

Advantages: It has the same optimal properties as ML, and it removes some of its limitations. Multiple imputation can be used with any kind of data and model with conventional software. When the data is MAR, multiple imputation can lead to consistent, asymptotically efficient, and asymptotically normal estimates.

Limitations: It is a bit challenging to successfully use it. It produces different estimates (hopefully, only slightly different) every time you use it, which can lead to situations where different researchers get different numbers from the same data using the same method (Nakai & Weiming, 2011), (Allison, 2001).

4.2.2. Maximum Likelihood

We can use this method to get the variance-covariance matrix for the variables in the model based on all the available data points, and then use the obtained variance- covariance matrix to estimate our regression model (Schafer, 1997).

Compared to MI, MI requires many more decisions than ML (whether to use Markov Chain Monte Carlo (MCMC) method or the Fully Conditional Specification (FCS), how many data sets to produce, how many iterations between data sets, what prior distribution to use-the default is Jeffreys-, etc.). On the other hand, ML is simpler as you only need to specify your model of interest and indicate that you want to use ML (SAS Institute, 2005).

There are two main ML methods:

a) Direct Maximum Likelihood: It implies the direct maximization of the multivariate normal likelihood function for the assumed linear model. **Advantage**: It gives efficient estimates with correct standard errors. **Limitations:** It requires specialized software (it may be challenging and time consuming).

b) The Expectation-maximization (EM) algorithm: It provides estimates of the means and covariance matrix, which can be used to get consistent estimates of the parameters of interest. It is based on an expectation step and a maximization step, which are repeated several times until maximum likelihood estimates are obtained. It requires a large sample size and that the data are missing at random (MAR). **Advantage:** We can use **SAS**, since this is the default algorithm it employs for dealing with missing data with Maximum Likelihood. **Limitations:** Only can be used for linear and log-linear models (there is neither theory nor software developed beyond them). (Allison, 2001) (Graham, 2009) (Enders & Bandalos, 2001) and (Allison, 2003).

4.3. Other advanced methods

4.3.1. Bayesian simulation methods

There are two main methods:

a) Schafer algorithms: It uses Bayesian iterative simulation methods to impute data sets assuming MAR. Precisely, it splits the multivariate missing problem into a series of univariate problems based on the assumed distribution of the multivariate missing variables (e.g. multivariate normal for continuous variables, multinomial loglinear for categorical variables). In other words, it uses an iterative algorithm that draws samples from a sequence of univariate regressions.

b) Van Buuren algorithm: It is a semi-parametric approach. The parametric part implies that each variable has a separate imputation model with a set of predictors that explain the missingness. The non-parametric part implies the specification of an appropriate form (e.g. linear), which depends on the

kind of variables (Briggs et al., 2003) (Kong et al., 1994).

4.3.2. Hot deck imputation methods

It is used by the US Census Bureau. This method completes a missing observation by selecting at random, with replacement, a value from those individuals who have matching observed values for other variables. In other words, a missing value is imputed based on an observed value that is closer in terms of distance.SAS macro developed by Lawrence Altmayer, of the U.S. Census Bureau. Can be found in Ahmed Kazi et al; 2009. (Briggs et al., 2003).

5. Dealing with missing data using SAS

For illustrative purposes, I will use a data set constructed to analyze the socioeconomic determinants of health in the OCDE (32 countries) from 1980-2010 (in 5 year intervals). The dependent variable of interest is the Health Index, which is part of the United Nations Human Development Index. The explanatory variables are the growth rate of the GDP per capita, the unemployment rate, the inequality in the distribution of wealth (measured by the Gini coefficient), and 3 dummy variables capturing the level of social and health expenditure and the existence of a National Health System. There are 88 missing values in the dataset, which reduces the sample size from 224 to 136 observations.

COUNTRY	YEAR	Health Index	GDP	U	GINI	SE	NHS	HE
Australia	1980	0.857	3.40	6.04	.	0	1	0
Australia	1985	0.876	4.51	8.18	.	0	1	0
Australia	1990	0.895	-0.61	6.62	.	0	1	0
Australia	1995	0.917	4.18	8.16	.	0	1	0
Australia	2000	0.939	1.98	6.24	.	0	1	0
Australia	2005	0.961	1.23	5.02	.	0	1	0
Australia	2010	0.974	3.26	5.21	30.11	0	1	0
Austria	1980	0.829	1.78	1.86	.	1	0	1
Austria	1985	0.851	2.46	3.60	23.6	1	0	1
Austria	1990	0.875	4.17	3.24	.	1	0	1
Austria	1995	0.894	2.54	3.74	23.8	1	0	1
Austria	2000	0.916	3.65	3.50	25.20	1	0	1
Austria	2005	0.94	-3.89	5.15	26.5	1	0	1
Austria	2010	0.957	1.96	4.49	26.53	1	0	1
Belgium	1980	0.84	4.48	6.69	.	1	0	1
Belgium	1985	0.864	1.65	10.12	27.4	1	0	1
Belgium	1990	0.884	3.14	6.55	.	1	0	1
Belgium	1995	0.898	2.38	9.69	28.7	1	0	1
Belgium	2000	0.912	3.68	6.86	28.90	1	0	1
Belgium	2005	0.931	-2.75	8.50	27.1	1	0	1
Belgium	2010	0.946	2.12	8.56	27.10	1	0	1
Canada	1980	0.868	2.16	7.52	.	0	1	1
Canada	1985	0.889	4.78	10.64	28.7	0	1	1

The variables with missing values are GINI coefficient, GDP and Unemployment (U).

5.1. Multiple Imputation (MI)

There are several imputation methods in PROC MI. The method of choice depends on the pattern of missingness in the data and the type of the imputed variable, as the table below summarizes:

Pattern of missingness	Type of imputed variable	Available Methods
Monotone	Continuous	•Parametric method that assumes multivariate normality: 1)Monotone regression 2)Monotone predicted mean matching
		•Nonparametric method that assumes propensity scores: Monotone propensity score.
	Ordinal	•Monotone logistic regression
	Nominal	•Monotone discriminant function
Arbitrary	Continuous	•Markov Chain Monte Carlo (MCMC) full-data imputation
		•MCMC monotone-data imputation
		•Fully conditional specification (FCS), which assumes the existence of a joint distribution for all variables. FCS regression
		•FCS predicted mean matching
	Ordinal	•FCS logistic regression
	Nominal	•FCS discriminant function

Source: Table modified from "The MI Procedure Imputation Methods, SAS 9.3, The SAS Institute" http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_mi_sect019.htm

In our case, the pattern of missingness is arbitrary and all variables with missing values are continuous. We choose MCMC full-data imputation, which uses a single chain to create 5 imputations. The posterior mode, the highest observed-data posterior density, with a noninformative prior, is computed from the expectation-maximization (EM) algorithm and is used as the starting value for the chain.

```
proc mi data=Final seed=501213 out=outmi;
mcmc;
var GDP GINI U;
```

Description of the output:

Model Information

Data Set	WORK.FINAL
Method	MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	5
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	501213

200 burn-in iterations were used before the first imputation and 100 iterations between imputation, which are used to eliminate the series of dependence between the two imputations

Missing Data Patterns

						Group Means		
Group	GDP	GINI	U	Freq	Percent	GDP	GINI	U
1	X	X	X	142	63.39	1.953521	31.311408	7.534225
2	X	X	.	4	1.79	2.797500	32.465000	.
3	X	.	X	57	25.45	2.946842	.	5.212456
4	X	.	.	10	4.46	3.945000	.	.
5	.	X	X	1	0.45	.	33.600000	8.210000
6	.	X	.	3	1.34	.	23.853333	.
7	.	.	X	2	0.89	.	.	5.300000

It lists different missing data patterns with the corresponding frequency and percentage they represent. ‘X’ implies that the variable is observed, while ‘.’ that it is missing. Group means are also displayed.

Multiple Imputation Variance Information

	Variance			
Variable	Between	Within	Total	DF
GDP	0.000931	0.055795	0.056912	207.56
GINI	0.031848	0.211965	0.250182	88.516
U	0.008019	0.060918	0.070541	99.873

It presents the between-imputation variance, within-imputation variance, and total variance for combining complete-data inferences for each variable, along with the degrees of freedom for the total variance.

Multiple Imputation Parameter Estimates

Variable	Mean	Std Error	95% Confidence Limits		DF
GDP	2.299948	0.238562	1.82963	2.77026	207.56
GINI	31.070102	0.500182	30.07618	32.06403	88.516
U	6.833374	0.265595	6.30643	7.36032	99.873

Multiple Imputation Parameter Estimates

Variable	Minimum	Maximum	Mu0	t for H0: Mean=Mu0	Pr > |t|
GDP	2.262426	2.340693	0	9.64	<.0001
GINI	30.800967	31.285536	0	62.12	<.0001
U	6.718310	6.944839	0	25.73	<.0001

It summarizes basic descriptive statistics for the imputed values by variable.

The imputed data sets are stored in the outmi data set, with the index variable _Imputation_indicating the imputation numbers. The data set can now be analyzed using standard statistical procedures with _Imputation_ as a BY variable (Yuan, 2011) and (http://www.ats.ucla.edu/stat/sas/seminars/missing_data/part1.htm).

5.2. Maximum Likelihood (ML)

We consider two options to deal with missing values on the independent variables:

a) The EM algorithm in PROC MI

PROC MI uses the default algorithm (EM) to do maximum likelihood of the means and the covariance matrix and, then, it considers these estimates as starting values for multiple imputation algorithms. We can use the estimates obtained on the first step.

```
proc mi data=Final nimpute=0;
var HI GDP U GINI SE NHS HE;
em outem=Finalem;
```

This option suppresses multiple imputation

It requests EM (expectation-maximization) estimates and writes the means and covariance matrix into a SAS data set called 'Finalem'

Next, we use the output data set from PROC MI (the EM covariance matrix) as input to PROC REG to estimate our model:

```
proc reg data=Finalem;
model HI=GDP U GINI SE NHS HE;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.86248	0.00441	195.45	<.0001
GDP	1	-0.00166	0.00016602	-9.99	<.0001
U	1	0.00132	0.00016477	8.01	<.0001
GINI	1	-0.00046947	0.00011576	-4.06	<.0001
SE	1	0.00230	0.00146	1.58	0.1145
NHS	1	0.00660	0.00119	5.56	<.0001
HE	1	0.04950	0.00130	37.99	<.0001

The parameter estimates are the true maximum likelihood estimates of the regression coefficients, so we have accomplished our goal.

However, p-values are useless (we get a warning from SAS saying so). Solution: estimate the standard errors and p-values by bootstrapping[1].

b) The "full-information" maximum likelihood method in PROC CALIS

Advantage: Single procedure available in SAS 9.2.

It is maximum likelihood estimation adjusting for the cases with missing data, which is called 'Full Information Maximum Likelihood' (FIML) or 'Direct Maximum Likelihood'.

PROC CALIS was originally designed to estimate Structural Equation Models and its default estimation method is maximum likelihood under the assumption of multivariate normality.

```
proc calis data=Final method=fiml;
path HI <- GDP U GINI SE NHS HE;
```

Specifying METHOD=FIML the missing values are also considered.

Notice that in this method we are making the same assumptions as with PROC MI (missing at random and multivariate normality). Also, it can handle almost any linear model.

Disadvantage: It cannot deal with logistic regression (dependent dichotomous variable) or negative binomial regression (count data). In these cases, we have to use MI. (Yuan, 2011) http://www.ats.ucla.edu/stat/sas/seminars/missing_data/part1.htm)

[1] It involves: 1) From the original sample size N, draw many samples of size N with replacements, 2) Obtain the EM estimates of the means/covariance matrix for each sample, 3) Estime the regression model from each covariance matrix, 4) Compute the standard deviation of each regression coefficient across samples.

6. Dealing with missing data using STATA

6.1. Multiple imputation

Basic steps:

1) Declare the data to be 'mi' data:

```
use dataset
mi set mlong
*On the last statement, we choose the data in marginal
long style (mlong)-a memory efficient style*
```

2) Register all variables relevant for the analysis as imputed, passive or regular

```
mi register imputed v1
mi register regular v2 v3 v4
```

3) Imputate the missing values

```
mi impute regress v1 v2 v3 v4, add(20) rseed(2232)
*To lessen the simulation (Monte Carlo) error, we
arbitrarily choose to create 20 imputations
```

4) Check that the imputation was done correctly: Compute basic descriptive statistics of v1 for some imputations (e.g. m=0 -the one with missing values-, m=1 and m=20)

```
mi xeq 0 1 20: summarize v1
```

5) Run the analysis using the mi estimate prefix command

```
mi estimate, dots: logit v1 v2 v3 v4
```

6.2. Other imputation methods available in STATA

- Hot deck imputation: Sg116
- Weighted logistic regression for data with missing values using the mean score method: Sg156

- Imputed values by best sub-set regression: ‘ Impute’ procedure

  It involves 3 steps: 1) Identify *all* of the possible regression models derived from all of the possible combinations of the candidate predictors, 2) from the possible models identified in the first step, determine the one-predictor models that do the "best" at meeting some well-defined criteria, the two-predictor models that do the "best, the three-predictor models that do the "best," and so on. 3) Further evaluate and refine the handful of models identified in the last step. More information:

  https://onlinecourses.science.psu.edu/stat501/node/89 (StataCorp, 2009)

Final question: What if our data is missing but not at random? We must specify a model for the probability of missing data, which can be pretty challenging as it requires a good understanding of the data generating process. The Sample Selection Bias Model, by James Heckman, is a widely used method that you can apply in SAS using PROC QLIM (Heckman et al., 1998). The motivating example for this approach is a regression that predicts women’s wages, where wages data are missing for women who are not in the labor market force. Those women who know that their wages will be very low based on their education and previous job experience may be less likely to enter the job market. Thus, the data are not missing at random (Allison, 2002).

7. Other software

- SOLAS http://www.solasmissingdata.com/software ,
- SPSS http://www-142.ibm.com/software/products/us/en/spss-missing-values/
- S-PLUS http://www.msi.co.jp/splus/support/download/missing.pdf

8. Sources and useful resources

8.1. Overview

Allison, P., 2001. Missing data — Quantitative applications in the social sciences. Thousand Oaks, CA: Sage. Vol. 136. *>> A very useful book to understand both the theoretical and practical implications of the different methods to deal with missing data.*

Briggs, A., Clark, T., Wolstenholme, J., Clarke, P., 2003. Missing.... presumed at random: cost-analysis of incomplete data. Health Economics 12, 377–392. *>> A great article to get an overview of the different methods (including advanced methods such as Bayesian simulation methods) and assumptions. It also includes a useful example about missing data imputation in cost datasets.*

Graham, J.W., 2009. Missing data analysis: making it work in the real world. Annu Rev Psychol 60, 549–576. *>> It presents an excellent summary of the main missing data literature. Solutions are given for missing data challenges such as handling longitudinal, categorical, and clustered data*

Nakai M and Weiming Ke., 2011. Review of Methods for Handling Missing Data in Longitudinal Data Analysis. Int. Journal of Math. Analysis. Vol. 5, no.1, 1-13. *>> It reviews and discusses general approaches for handling missing data in longitudinal studies*

SAS Institute, 2005. Multiple Imputation for Missing Data: Concepts and New Approaches. *>> A useful overview of the different methods to deal with missing data using SAS.*

Schafer, J. L. ,1997. Analysis of Incomplete Multivariate Data, New York: Chapman and Hall *>>Excellent book aimed at bridging the gap between theory and practice. It presents a unified, Bayesian approach to the analysis of incomplete multivariate data.*

STATA 11, 2009, Multiple Imputation.StataCorp. *>> Comprehensive manual for dealing with missing data using STATA. It provides many useful examples and detailed explanations.*

Yuan Yang C., 2011. Multiple imputation for Missing Data: Concepts and New Development (SAS Version 9.0). SAS Institute Inc., Rockville, MA). *>> A useful overview of the different methods to deal with missing data using SAS.*

8.2. Advanced

Ahmed K, et al., 2009. Applying Missing Data Imputation Methods to HOS Household Income Data. Prepared by the National Committee for Quality Assurance (NCQA) for the Centers for Medicare and Medicaid Services

Allison P., 2000. Multiple imputation for missing data. A caution tale. Sociological Methods and Research, Vol. 28, No.3.

Allison P., 2003. Handling Missing Data by Maximum Likelihood. SAS Global Forum 2012. Development (Version 9.0)

Heckman, J., Ichimura, H., Smith, J., Todd, P., 1998. Characterizing Selection Bias Using Experimental Data. Econometrica 66, 1017–1098.

Enders, C.K., Bandalos, D.L., 2001. The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. Structural Equation Modeling: A Multidisciplinary Journal 8, 430–457.

Kong A., Liu KJ and Hung Wong W., 1994. Sequential Imputations and Bayesian Missing Data Problems. Journal of the American Statistical Association 89, 425, 278-288

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592.

STATA. Multiple –Imputation Reference Manual. 2009.

Twisk, J., De Vente, W., 2002. Attrition in longitudinal studies. How to deal with missing data. Journal of Clinical Epidemiology 55, 329–337.

8.3. Applications

Bernhard, J., Cella, D.F., Coates, A.S., Fallowfield, L., Ganz, P.A., Moinpour, C.M., Mosconi, P., Osoba, D., Simes, J., Hürny, C., 1998. Missing quality of life data in cancer clinical trials: serious problems and challenges. Statistics in Medicine 17, 517–532.

Burton, A., Billingham, L.J., Bryan, S., 2007. Cost-effectiveness in clinical trials: using multiple imputation to deal with incomplete cost data. Clin Trials 4, 154–161.

Little, R.J.A., 1988. Missing-Data Adjustments in Large Surveys. Journal of Business & Economic Statistics 6, 287–296.

Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M., Carpenter, J.R., 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 338, b2393–b2393.

8.4. Useful links

- Data sets with missing values that can be downloaded in different formats including SAS, STATA, SPSS and Splus http://www.ats.ucla.edu/stat/examples/md/default.htm
- Introduction to missing data with useful examples in SAS http://www.ats.ucla.edu/stat/sas/modules/missing.htm
- Multiple imputation in SAS. Comprehensive explanations. http://www.ats.ucla.edu/stat/sas/seminars/missing_data/part1.htm