Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

An Analysis of the Missing Data Methodology for Different Types of Data

A THESIS PRESENTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE
OF

MASTER OF APPLIED STATISTICS

AT MASSEY UNIVERSITY, ALBANY
NEW ZEALAND

Judith-Anne Scheffer

2000

Abstract

Missing data is an eternal problem in data analysis. It is widely recognised that data is costly to collect, and the methods used to deal with missing data in the past relied on case deletion. There is no one overall best fix, but many different methodologies to use in different situations.

This study was motivated by the writer's time spent analysing data in the nutrition study, and realising how much data was wasted by case deletion, and subsequently how this could bias inferences formed from the results. A better method (or methods), of dealing with missing data (than case deletion) is required, to ensure valuable information is not lost.

What is being done: What is in the literature? The literature on this topic has exploded with new methods in recent times. Algorithms have been written and incorporated based on these methods into a number of statistical packages and add-on libraries.

Statistical packages are also reviewed for their practicality and application in this area. The nutrition data is then applied to different methodologies, and software packages to assess different types of imputation.

A set of questions are posed; based on type of data, type of missingness, extent of missingness, the required end use of the data, the size of the dataset, and how extensive that analysis needs to be. This can guide the investigator into using an appropriate form of imputation for the type of data at hand.

A comparison of imputation methods and results is given with the principal result that imputing missing data is a very worthwhile exercise to reduce bias in survey results, which can be achieved by any researcher analysing their own data.

Further to this, a conjecture is given for using Data Augmentation for ordinal data, particularly Likert scales. Previously this has been restricted to either person or item mean imputation, or hot deck methods. Using model based methods for imputation is far superior for other types of data. Model based methods for Likert data are achieved by means of inserting the linear by linear association model into standard missing data methodology.

Acknowledgements

I wish to offer my sincerest thanks to my supervisor, Doctor Barry W. M^{c}Donald, for all his helpful advice, comments and efforts on my behalf, and also for his encouragement and mentoring throughout the course of this degree.

My thanks also go to Doctor Howard P. Edwards for his assistance in 'Matters Bayesian', Ms Katya Ruggiero for her ability to challenge practices and ideas, Mrs Kay Rowbottom for her assistance with the production of the flowcharts, and Synthia for her encouragement.

Thanks also go to Mrs Patsy E. Watson for providing via my supervisor, the nutrition dataset; and also to Ms Janet Norton for providing her dataset, via Professor Graham R. Wood.

Lastly but not least, I would like to thank my family (the thesis orphans) for putting up with my frequent absences for long periods to do this work.

Blessed is the man who perseveres under trial,
because when he has stood the test,
he will receive the crown of life that
God has promised to those who love him.
James 1:12

Table of Contents

TABLE OF CONTENTS IV

NOTATION AND ABBREVIATIONS XIII

1 INTRODUCTION: IS IGNORANCE BLISS? 1

1.1 The thesis 1

1.1.1 An overview of the thesis 1

1.1.2 Background 1

1.1.3 The Remaining Chapters 2

2 LITERATURE REVIEW OF DATA COLLECTION METHODOLOGY 4

2.1 What is Missing Data? 4

2.1.1 Ways in which Missing Data Arise 5

2.1.2 Inference and missing data 6

2.1.3 Consequences of Missing Data 7

2.1.4 Bias 7

2.1.5 Omitting covariates 9

2.2 Forms of Nonresponse. 9

2.2.1 Unit Nonresponse. 10

2.2.2 Item Nonresponse 11

2.3 Missing Data Mechanism 12

2.3.1 Parameter distinctiveness 13

2.3.2 MCAR 13

2.3.3 MAR 15

2.3.4 NMAR 17

2.3.5 Patterns of Missing Data 17

2.4 Types of data in Surveys 19

2.4.1 Surveys 19

2.4.2 Occurrences of Nonresponse in Surveys 20
2.4.3 Inevitable missingness in Surveys 20
2.4.4 Longitudinal drop out mechanism 21
2.4.5 Quota Sampling: 22
2.4.6 Telephone Surveys 23
2.4.7 Call Backs for the Noncontactables 23
2.4.8 Sensitive questions. 24
2.4.9 Coercion 25
2.4.10 Methods of Interviewing 26
2.4.11 Incentives 27
2.4.12 Double Sampling 27

2.5 Special Types of Data 28
2.5.1 Experimental design 28
2.5.2 Case Control Studies 30

2.6 Ways to prevent Nonresponse 30

3 LITERATURE REVIEW OF METHODOLOGY FOR ANALYSING MISSING DATA 32

3.1 Cure for Missing data 32
3.1.1 Complete and Available Case Analysis 32
3.1.2 Imputation (see chapter 5, for a more detailed description of methods used) 33
3.1.3 Reweighting 34
3.1.4 Model Based Methods 35

3.2 Older Methods used an 'ad hoc approach': Early Literature on Missing Observations 37
3.2.1 Performance of Different Methods: 38

3.3 More Modern Methods 40
3.3.1 Imputation using Box-Cox Transformations 40
3.3.2 More on Regression Imputation 42
3.3.3 Imputation using Coarsening, or Discretising Data 43
3.3.4 Multiple Imputation 44
3.3.5 Uncongenial sources of input. 48

3.3.6 EM Based, MCMC Based Methods 51

3.4 Little's test for MCAR 53

3.4.1 Σ known. 54

3.4.2 Σ unknown. 54

3.4.3 Monotone missing 54

3.4.4 Monotone data patterns 55

3.5 Ignorable Nonresponse 57

3.5.1 EM algorithm: what is it applied to Missing data 60

3.5.2 MLE for multivariate normal 61

3.5.3 Contingency Tables (Categorical) 62

3.5.4 MLE for Multinomial Model 66

3.5.5 MLE for Loglinear Model 66

3.5.6 Longitudinal 67

3.5.7 Repeated Binary outcomes 67

3.5.8 Mixed models 68

3.5.9 Likert-type scales 69

3.6 Non-Ignorable Missing. 72

3.6.1 Non-Random Missingness. 73

3.7 Data Models 74

3.7.1 Multivariate Normal 74

3.7.2 Multinomial (Saturated) 74

3.7.3 Loglinear 75

3.7.4 General Location Model 76

3.8 Likelihood theory 77

3.8.1 Coarsening 77

3.8.2 Sensitivity to Normality 77

3.8.3 Categorical 78

3.8.4 Bayesian Approach 78

3.9 Analysis of missing data 79

3.9.1 Rubin's Rules for Recombining Estimates 79

3.9.2 Rules for Analysis: % missing categorical, mixed, and continuous. 80

3.9.3 Longitudinal data 80
3.9.4 Bayesian Methods (Multiple Imputation): as applied to Frequentist Ideas. 81
3.9.5 Parameter Expansion for Data Augmentation 82
3.9.6 Nonparametric Method 82
3.9.7 MCMC Algorithm. 82

4 MOTIVATION AND DATA DESCRIPTION 83

4.1 The problem: 83

4.2 Motivation for this study: 83

4.3 The two data sets used here. 84
4.3.1 Nutrition Data set. 84
4.3.2 Genetics Foods Data Set. 87

5 IMPUTATION 94

5.1 What is Imputation, and why Impute? 94

5.2 Complete Case Methods Overview 96
5.2.1 Case Deletion 97
5.2.2 Available case 98
5.2.3 Logical substitution and Look-up tables 98

5.3 Mean Based Methods Overview 99
5.3.1 Mean Substitution 99
5.3.2 Mode Substitution (categorical) 99
5.3.3 Median Substitution (robust) 100
5.3.4 Discriminant Analysis 100
5.3.5 Stochastic Mean Substitution. 101
5.3.6 Mean within category substitution (conditional)- class mean. 101

5.4 Data Substitution Methods Overview 102
5.4.1 Colddeck 102
5.4.2 Hotdeck- random 103
5.4.3 Hotdeck- next available case. 103

5.4.4 Last value carried forward (Hot deck) 104

5.5 Time Series Models Overview 104

5.5.1 ARIMA models 105

5.5.2 Kalman Filter models 105

5.5.3 Period on Period Movements Ratio. 106

5.5.4 Within Case Year on Year Movements Ratio. 106

5.6 Regression Imputation Overview 107

5.6.1 Predictive Regression Imputation 107

5.6.2 Predictive Mean Matching 107

5.6.3 Random (Stochastic) Regression Imputation 108

5.6.4 Logistic Regression Imputation 109

5.7 Other single imputation methods Overview 109

5.7.1 Nearest Neighbour Imputation 109

5.7.2 Neural Networks 110

5.8 Model Based Imputation Methods Overview 112

5.8.1 EM Based Single Imputation. 113

5.8.2 Multiple Imputation - Bayesian 114

5.8.3 Multiple Imputation MCMC based - Bayesian 114

5.8.4 Multiple Imputation - Conditional 115

5.8.5 Multiple imputation for GEE (Generalised Estimating Equations) 118

5.8.6 MI for Case Control Studies 118

6 SOFTWARE FOR MISSING DATA 120

6.1 Overview of Software Available 120

6.2 Commercial Packages 121

6.2.1 Minitab 121

6.2.2 SAS 122

6.2.3 S-PLUS 125

6.2.4 Base SPSS (Data step) 126

6.2.5 SPSS MVA 126

6.2.6 Statistica 127

6.2.7 Systat 128

6.2.8 Matlab 128

6.3 Commercial Packages which are lesser known 129

6.3.1 BMDP: 129

6.3.2 Dalsolution 129

6.3.3 Solas 130

6.4 Specialist Freeware Missing Data Packages 132

6.4.1 Amelia 132

6.4.2 Cat 132

6.4.3 IVEWARE 133

6.4.4 MDM 133

6.4.5 MICE 134

6.4.6 MIX 134

6.4.7 NORM 135

6.4.8 OSWALD 135

6.4.9 PAN 137

6.4.10 TRANSCAN 137

6.5 Other Packages which may be Useful 137

6.5.1 MULTIMIX 137

6.5.2 SNOB 138

7 RULES FOR IMPUTATION 141

7.1 Imputation Strategies 141

7.2 Type of Missingness: Is the missingness MCAR, MAR, NMAR? 142

7.2.1 Continuous Data, MCAR. 142

7.2.2 Continuous Data MAR 143

7.2.3 Continuous data NMAR 143

7.3 Categorical data. 144

7.3.1 Ordinal data, MCAR. 144

7.3.2 Ordinal data, MAR 145

7.3.3 Ordinal data NMAR. 145

7.3.4 Binary, Nominal data MCAR 146
7.3.5 Binary, Nominal MAR data 146
7.3.6 Binary Nominal NMAR 146

7.4 Mixed data 147
7.4.1 Mixed data MCAR. 147
7.4.2 Mixed data MAR 147
7.4.3 Mixed data NMAR 148

7.5 Time series data 148
7.5.1 Time Series MCAR 148
7.5.2 Time Series MAR 148
7.5.3 Time series NMAR 149

7.6 Other longitudinal studies (Repeated measures) 149
7.6.1 Repeated measures MCAR 149
7.6.2 Repeated Measures MAR 149
7.6.3 Repeated measures NMAR 149

7.7 Panel data, and Clustered data 150

7.8 Case control studies. 150

8 SOME APPROACHES TO ORDINAL CATEGORICAL DATA IMPUTATION: LIKERT DATA IN PARTICULAR (A CONJECTURE) 151

9 ANALYSIS AND IMPUTATION OF DATA 157

9.1 Preparation of the data. 157
9.1.1 SPSS MVA Imputation 159
9.1.2 Solas 161
9.1.3 S-Plus 162

9.2 Analysis of data using Minitab 165
9.2.1 Results 165
9.2.2 Validity of Imputations, and results. 167

9.3 Further Analysis 169

10 CONCLUSION 170

10.1 The Ethics of Imputation 170

10.2 Conclusion 172

APPENDIX 175

BIBLIOGRAPHY 184

List of Tables and Figures

Table 3.1. Construction of a look-up table: 65

Figure 5.1. Efficiency of Imputation Table 113

Table 9.1. Estimates of coefficients under different Imputation schemes 165

Table 9.2. Standard deviations under different Imputation schemes. 166

Figure 9.1. Normal probability plot of the residuals 167

Figure 9.2. Histogram of the residuals 168

Figure 9.3. Plot of residuals versus fitted values 168

Notation and Abbreviations

BLR	Binary Logistic Regression
CD	Case Deletion
EM	Expectation Maximisation (algorithm)
EM Imp	Imputation via the EM algorithm
GLM Imp	General Location Model Imputation
HD	Hotdeck (Imputation)
iid	Independent identically distributed
LUM	Look up methods
LVCF	Last Value Carried Forwards
MCAR	Missing Completely at Random
MAR	Missing at Random
Mean Imp	Mean family of Imputation
MI	Multiple Imputation
MI BB	Multiple Imputation Bayesian Bootstrap
MICE	Multiple Imputation by Chained Equations
MI DA	Multiple Imputation via Data Augmentation
MI EM	Multiple Imputation via the EM algorithm
N.Neighbour	Nearest Neighbour
N Nets	Neural Networks
NLR	Nominal Logistic Regression
NMAR	Not Missing at Random (Informatively Missing)
OLR	Ordinal Logistic Regression
PMM	Predictive Mean matching
Reg Imp	Regression Imputation
SHHD	Sequential and/or Hierarchical Hotdeck
SI	Single Imputation
St Reg	Stochastic regression Imputation

W	Indicator for Missingness
X	Co-variate in model
Y	Variable of interest
$\hat{\alpha}$	Gamma Parameter (Ch 8)
$\hat{\beta}$	Gamma Parameter (Ch 8)
$\hat{\beta}$	Regression Coefficient Estimate (Ch 9)
θ	Distribution Parameter
$\hat{\theta}$	Maximum Likelihood Estimate of the Parameter
ψ	Missingness Parameter in Model