

Nicholas T. Longford

Missing Data and Small-Area Estimation

Modern Analytical Equipment
for the Survey Statistician

With 45 Figures

|y Springer

Contents

PrefaceVII

Part I Missing data

1 Prologue 3

1.1 Terminology. Some basics 3

1.1.1 Efficiency 8

1.1.2 Classes and types of estimators 10

1.2 Populations and variables 11

1.3 Missing data 13

1.4 Suggested reading 15

1.5 Exercises 16

2 Describing incompleteness 19

2.1 The problem of incompleteness 19

2.2 The extent of missing data and the response pattern 22

2.2.1 Monotone response patterns 26

2.3 Sampling and nonresponse processes 28

2.3.1 The nature of the nonresponse process 30

2.3.2 The importance of MAR 33

2.4 Exercises 34

3 Single imputation and related methods 37

3.1 Data reduction 39

3.2 Data completion 40

3.2.1 Mean imputation 40

3.2.2 Imputation from another variable 41

3.2.3 Nearest-neighbour imputation 42

3.2.4 Hot deck 43

3.2.5 Weight adjustment 44

3.2.6	Regression imputation.	45
3.2.7	Using experts' judgements.	47
3.2.8	Data editing.	48
3.2.9	Single imputation. Summary.	48
3.3	Models for imputation.	49
3.3.1	Operating with uncertainty.	50
3.3.2	Models for the nonresponse process.	52
3.4	EM algorithm.	53
3.5	Suggested reading.	56
3.6	Exercises.	57
4	Multiple imputation.	59
4.1	The consequences of imperfect imputation.	60
4.2	The method.	61
4.2.1	Fitting a model for missing values.	61
4.2.2	Generating plausible values.	62
4.2.3	Analysis of each completed dataset.	64
4.2.4	The MI estimator.	64
4.2.5	The lost information.	64
4.2.6	Assumptions and properties.	66
4.3	Conditional distributions.	66
4.3.1	Normally distributed data.	66
4.3.2	Categorical variables.	67
4.3.3	Categorical and continuous variables.	68
4.3.4	Multivariate and multi-stage imputation.	69
4.3.5	Imputation with monotone response patterns.	70
4.3.6	The method of chained equations.	71
4.3.7	From MAR to NMAR models.	72
4.4	From theory to practice.	72
4.4.1	Organising MI.	72
4.4.2	Validity of the assumptions.	73
4.4.3	MI adaptation of LOCF.	74
4.4.4	Mi-proper hot deck.	75
4.4.5	Propensity scoring.	77
4.5	NMAR and sensitivity analysis.	78
4.6	Other applications of MI.	79
4.6.1	Measurement error.	80
4.6.2	Misclassification.	83
4.6.3	Coarse data and rounding.	84
4.6.4	Summary.	92
4.7	Suggested reading.	93
4.8	Exercises.	93

Case studies	97
5.1 The UK Labour Force Survey	97
5.1.1 From LOCF to hot deck	101
5.1.2 Results and discussion	104
5.1.3 Imputation for absentees	108
5.2 The National Survey of Health and Development	110
5.2.1 Eliciting information about alcohol consumption	112
5.2.2 Excessive alcohol consumption	115
5.2.3 Sensitivity analysis	117
5.3 The International Social Survey Programme	119
5.3.1 Imputation for 'national identity' items	121
5.3.2 Attitudes to immigration	124
5.3.3 Sensitivity analysis	127
5.4 The Scottish House Condition Survey	130
5.4.1 Missing information	133
5.4.2 Estimating the misclassification probabilities	135
5.4.3 Generating plausible scores	137
5.5 Suggested reading. Software	138

Part II Small-area estimation

6 Introduction	143
6.1 Preliminaries	146
6.2 Choosing the estimator	149
6.2.1 Uniform choice	149
6.2.2 Tailored choice	150
6.3 Composition	151
6.3.1 Combining the district-level means	157
6.3.2 Suboptimal composition	159
6.4 Estimating the district-level variance	160
6.4.1 The sampling variance of \bar{y}_d	161
6.4.2 The impact of uncertainty about \bar{y}_d	164
6.5 Spatial similarity	167
6.6 Suggested reading	169
6.7 Exercises	170
7 Models for small areas	173
7.1 Analysis of variance	173
7.2 Auxiliary information	176
7.2.1 Several covariates	178
7.2.2 Two-level models and small-area estimation	181
7.3 Computational procedures	182
7.3.1 Restricted maximum likelihood	186
7.3.2 Implementing ML and REML	188

7.3.3	Computational issues.	189
7.4	Model selection issues.	192
7.4.1	Residuals and model diagnostics.	195
7.5	District-level models.	197
7.6	Generalised linear models.	200
7.6.1	Two-level GLMs.	202
	Appendix. The REML adjustment of the Hessian.	203
7.7	Suggested reading.	204
7.8	Exercises.	205
8	Using auxiliary information.	207
8.1	From models to small-area estimates.	208
8.1.1	Synthetic estimation.	208
8.2	Composite estimation.	212
8.2.1	Shrinkage and borrowing strength.	214
8.3	Multivariate composition.	215
8.3.1	How to choose x ?.	217
8.3.2	Estimating E_B	219
8.4	Applications.	220
8.4.1	Related variables in a survey.	220
8.4.2	Estimation for several subpopulations.	221
8.4.3	Estimating compositions.	223
8.4.4	Survey and register.	224
8.4.5	Historical data as auxiliary information.	226
8.4.6	Summary. Using all the relevant information.	228
8.5	Planning and design for small-area estimation.	229
8.5.1	Optimal design for the composite estimator.	231
8.5.2	Variable subsample sizes and several divisions.	234
8.6	Suggested reading.	235
8.7	Exercises.	236
9	Using small-area estimators.	239
9.1	Non-linear transformations of the estimates.	239
9.1.1	How important is bias?.	241
9.2	Ranking and ordering.	241
9.2.1	Inference about selected districts.	244
9.3	Estimating many variances and precisions.	246
9.3.1	Estimated or guessed variance ratio.	248
9.3.2	Estimating precisions.	252
9.4	Suggested reading.	254
9.5	Exercises.	254

10 Case studies	257
10.1 The UK Labour Force Survey	257
10.1.1 Multivariate shrinkage	262
10.1.2 Distribution of district-level rates	268
10.1.3 Estimation for age-by-sex subpopulations	271
10.1.4 Pooling information across time	274
10.2 Samples of Anonymised Records	276
10.3 Norwegian municipalities	282
10.3.1 Composition of the labour force by industrial sectors	289
10.4 The Scottish House Condition Survey	291
10.4.1 Estimation for subpopulations	296
10.5 Suggested reading	297

Part III Combining estimators

11 Model selection	303
11.1 The problem	303
11.1.1 EM algorithm	307
11.1.2 Example	308
11.2 Why model selection fails	310
11.2.1 Limitations of model selection	311
11.3 Synthetic estimation	313
11.3.1 One submodel	314
11.4 Analysis of variance	316
11.4.1 Minimax estimation	318
11.4.2 Estimating $a\%j$	319
11.4.3 Estimated coefficient 6^*	321
11.4.4 Simulations	322
11.4.5 ANOVA with random effects	323
11.5 Ordinary regression	325
11.5.1 Estimating a^2	326
11.5.2 Several covariates	327
11.6 Discussion	329
11.7 Other applications of synthesis	331
11.7.1 Meta-analysis	331
11.7.2 Multiple sources and prior information	332
11.7.3 Secondary outcomes and auxiliary information	333
11.8 Suggested reading	334
11.9 Exercises	334

References	337
-------------------	-----

Index	353
--------------	-----