# Statistical Analysis with Missing Data

*Roderick J. A. Little*

*Richard D. Remington Distinguished University Professor of Biostatistics, Professor of Statistics, and Research Professor, Institute for Social Rsearch, at the University of Michigan*

*Donald B. Rubin*

*Professor at Yau Mathematical Sciences Center, Tsinghua University; Murray Shusterman Senior Research Fellow, Fox School of Business, at Temple University; and Professor Emeritus, at Harvard University*

3rd Edition

WILEY

# Contents

**Preface to the Third Edition** *xi*

**Part I Overview and Basic Approaches** *1*

**1 Introduction** *3*
1.1 The Problem of Missing Data *3*
1.2 Missingness Patterns and Mechanisms *8*
1.3 Mechanisms That Lead to Missing Data *13*
1.4 A Taxonomy of Missing Data Methods *23*

**2 Missing Data in Experiments** *29*
2.1 Introduction *29*
2.2 The Exact Least Squares Solution with Complete Data *30*
2.3 The Correct Least Squares Analysis with Missing Data *32*
2.4 Filling in Least Squares Estimates *33*
2.4.1 Yates's Method *33*
2.4.2 Using a Formula for the Missing Values *34*
2.4.3 Iterating to Find the Missing Values *34*
2.4.4 ANCOVA with Missing Value Covariates *35*
2.5 Bartlett's ANCOVA Method *35*
2.5.1 Useful Properties of Bartlett's Method *35*
2.5.2 Notation *36*
2.5.3 The ANCOVA Estimates of Parameters and Missing $Y$-Values *36*
2.5.4 ANCOVA Estimates of the Residual Sums of Squares and the Covariance Matrix of $\hat{\beta}$ *37*
2.6 Least Squares Estimates of Missing Values by ANCOVA Using Only Complete-Data Methods *38*
2.7 Correct Least Squares Estimates of Standard Errors and One Degree of Freedom Sums of Squares *40*

2.8 Correct Least-Squares Sums of Squares with More Than One Degree of Freedom *42*

**3 Complete-Case and Available-Case Analysis, Including Weighting Methods** *47*
3.1 Introduction *47*
3.2 Complete-Case Analysis *47*
3.3 Weighted Complete-Case Analysis *50*
3.3.1 Weighting Adjustments *50*
3.3.2 Poststratification and Raking to Known Margins *58*
3.3.3 Inference from Weighted Data *60*
3.3.4 Summary of Weighting Methods *61*
3.4 Available-Case Analysis *61*

**4 Single Imputation Methods** *67*
4.1 Introduction *67*
4.2 Imputing Means from a Predictive Distribution *69*
4.2.1 Unconditional Mean Imputation *69*
4.2.2 Conditional Mean Imputation *70*
4.3 Imputing Draws from a Predictive Distribution *73*
4.3.1 Draws Based on Explicit Models *73*
4.3.2 Draws Based on Implicit Models – Hot Deck Methods *76*
4.4 Conclusion *81*

**5 Accounting for Uncertainty from Missing Data** *85*
5.1 Introduction *85*
5.2 Imputation Methods that Provide Valid Standard Errors from a Single Filled-in Data Set *86*
5.3 Standard Errors for Imputed Data by Resampling *90*
5.3.1 Bootstrap Standard Errors *90*
5.3.2 Jackknife Standard Errors *92*
5.4 Introduction to Multiple Imputation *95*
5.5 Comparison of Resampling Methods and Multiple Imputation *100*

**Part II Likelihood-Based Approaches to the Analysis of Data with Missing Values** *107*

**6 Theory of Inference Based on the Likelihood Function** *109*
6.1 Review of Likelihood-Based Estimation for Complete Data *109*
6.1.1 Maximum Likelihood Estimation *109*
6.1.2 Inference Based on the Likelihood *118*
6.1.3 Large Sample Maximum Likelihood and Bayes Inference *119*

6.1.4 Bayes Inference Based on the Full Posterior Distribution *126*
6.1.5 Simulating Posterior Distributions *130*
6.2 Likelihood-Based Inference with Incomplete Data *132*
6.3 A Generally Flawed Alternative to Maximum Likelihood: Maximizing over the Parameters and the Missing Data *141*
6.3.1 The Method *141*
6.3.2 Background *142*
6.3.3 Examples *143*
6.4 Likelihood Theory for Coarsened Data *145*

**7 Factored Likelihood Methods When the Missingness Mechanism Is Ignorable** *151*
7.1 Introduction *151*
7.2 Bivariate Normal Data with One Variable Subject to Missingness: ML Estimation *153*
7.2.1 ML Estimates *153*
7.2.2 Large-Sample Covariance Matrix *157*
7.3 Bivariate Normal Monotone Data: Small-Sample Inference *158*
7.4 Monotone Missingness with More Than Two Variables *161*
7.4.1 Multivariate Data with One Normal Variable Subject to Missingness *161*
7.4.2 The Factored Likelihood for a General Monotone Pattern *162*
7.4.3 ML Computation for Monotone Normal Data via the Sweep Operator *166*
7.4.4 Bayes Computation for Monotone Normal Data via the Sweep Operator *174*
7.5 Factored Likelihoods for Special Nonmonotone Patterns *175*

**8 Maximum Likelihood for General Patterns of Missing Data: Introduction and Theory with Ignorable Nonresponse** *185*
8.1 Alternative Computational Strategies *185*
8.2 Introduction to the EM Algorithm *187*
8.3 The E Step and The M Step of EM *188*
8.4 Theory of the EM Algorithm *193*
8.4.1 Convergence Properties of EM *193*
8.4.2 EM for Exponential Families *196*
8.4.3 Rate of Convergence of EM *198*
8.5 Extensions of EM *200*
8.5.1 The ECM Algorithm *200*
8.5.2 The ECME and AECM Algorithms *205*
8.5.3 The PX-EM Algorithm *206*
8.6 Hybrid Maximization Methods *208*

**9 Large-Sample Inference Based on Maximum Likelihood Estimates** *213*
9.1 Standard Errors Based on The Information Matrix *213*
9.2 Standard Errors via Other Methods *214*
9.2.1 The Supplemented EM Algorithm *214*
9.2.2 Bootstrapping the Observed Data *219*
9.2.3 Other Large-Sample Methods *220*
9.2.4 Posterior Standard Errors from Bayesian Methods *221*

**10 Bayes and Multiple Imputation** *223*
10.1 Bayesian Iterative Simulation Methods *223*
10.1.1 Data Augmentation *223*
10.1.2 The Gibbs' Sampler *226*
10.1.3 Assessing Convergence of Iterative Simulations *230*
10.1.4 Some Other Simulation Methods *231*
10.2 Multiple Imputation *232*
10.2.1 Large-Sample Bayesian Approximations of the Posterior Mean and Variance Based on a Small Number of Draws *232*
10.2.2 Approximations Using Test Statistics or $p$-Values *235*
10.2.3 Other Methods for Creating Multiple Imputations *238*
10.2.4 Chained-Equation Multiple Imputation *241*
10.2.5 Using Different Models for Imputation and Analysis *243*

**Part III Likelihood-Based Approaches to the Analysis of Incomplete Data: Some Examples** *247*

**11 Multivariate Normal Examples, Ignoring the Missingness Mechanism** *249*
11.1 Introduction *249*
11.2 Inference for a Mean Vector and Covariance Matrix with Missing Data Under Normality *249*
11.2.1 The EM Algorithm for Incomplete Multivariate Normal Samples *250*
11.2.2 Estimated Asymptotic Covariance Matrix of $(\theta - \hat{\theta})$ *252*
11.2.3 Bayes Inference and Multiple Imputation for the Normal Model *253*
11.3 The Normal Model with a Restricted Covariance Matrix *257*
11.4 Multiple Linear Regression *264*
11.4.1 Linear Regression with Missingness Confined to the Dependent Variable *264*
11.4.2 More General Linear Regression Problems with Missing Data *266*
11.5 A General Repeated-Measures Model with Missing Data *269*

11.6 Time Series Models *273*
11.6.1 Introduction *273*
11.6.2 Autoregressive Models for Univariate Time Series with Missing Values *273*
11.6.3 Kalman Filter Models *276*
11.7 Measurement Error Formulated as Missing Data *277*

**12 Models for Robust Estimation** *285*
12.1 Introduction *285*
12.2 Reducing the Influence of Outliers by Replacing the Normal Distribution by a Longer-Tailed Distribution *286*
12.2.1 Estimation for a Univariate Sample *286*
12.2.2 Robust Estimation of the Mean and Covariance Matrix with Complete Data *288*
12.2.3 Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values *290*
12.2.4 Adaptive Robust Multivariate Estimation *291*
12.2.5 Bayes Inference for the *t* Model *292*
12.2.6 Further Extensions of the *t* Model *294*
12.3 Penalized Spline of Propensity Prediction *298*

**13 Models for Partially Classified Contingency Tables, Ignoring the Missingness Mechanism** *301*
13.1 Introduction *301*
13.2 Factored Likelihoods for Monotone Multinomial Data *302*
13.2.1 Introduction *302*
13.2.2 ML and Bayes for Monotone Patterns *303*
13.2.3 Precision of Estimation *312*
13.3 ML and Bayes Estimation for Multinomial Samples with General Patterns of Missingness *313*
13.4 Loglinear Models for Partially Classified Contingency Tables *317*
13.4.1 The Complete-Data Case *317*
13.4.2 Loglinear Models for Partially Classified Tables *320*
13.4.3 Goodness-of-Fit Tests for Partially Classified Data *326*

**14 Mixed Normal and Nonnormal Data with Missing Values, Ignoring the Missingness Mechanism** *329*
14.1 Introduction *329*
14.2 The General Location Model *329*
14.2.1 The Complete-Data Model and Parameter Estimates *329*
14.2.2 ML Estimation with Missing Values *331*
14.2.3 Details of the E Step Calculations *334*

14.2.4 Bayes' Computation for the Unrestricted General Location Model *335*
14.3 The General Location Model with Parameter Constraints *337*
14.3.1 Introduction *337*
14.3.2 Restricted Models for the Cell Means *340*
14.3.3 Loglinear Models for the Cell Probabilities *340*
14.3.4 Modifications to the Algorithms of Previous Sections to Accommodate Parameter Restrictions *340*
14.3.5 Simplifications When Categorical Variables are More Observed than Continuous Variables *343*
14.4 Regression Problems Involving Mixtures of Continuous and Categorical Variables *344*
14.4.1 Normal Linear Regression with Missing Continuous or Categorical Covariates *344*
14.4.2 Logistic Regression with Missing Continuous or Categorical Covariates *346*
14.5 Further Extensions of the General Location Model *347*

**15 Missing Not at Random Models** *351*
15.1 Introduction *351*
15.2 Models with Known MNAR Missingness Mechanisms: Grouped and Rounded Data *355*
15.3 Normal Models for MNAR Missing Data *362*
15.3.1 Normal Selection and Pattern-Mixture Models for Univariate Missingness *362*
15.3.2 Following up a Subsample of Nonrespondents *364*
15.3.3 The Bayesian Approach *366*
15.3.4 Imposing Restrictions on Model Parameters *369*
15.3.5 Sensitivity Analysis *376*
15.3.6 Subsample Ignorable Likelihood for Regression with Missing Data *379*
15.4 Other Models and Methods for MNAR Missing Data *382*
15.4.1 MNAR Models for Repeated-Measures Data *382*
15.4.2 MNAR Models for Categorical Data *385*
15.4.3 Sensitivity Analyses for Chained-Equation Multiple Imputations *391*
15.4.4 Sensitivity Analyses in Pharmaceutical Applications *396*

**References** *405*
**Author Index** *429*
**Subject Index** *437*