MULTIPLE COMPARISONS USING MULTIPLE IMPUTATION UNDER A TWO-WAY MIXED EFFECTS INTERACTION MODEL

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By

Joseph S. Kosler, B.A., M.S.

* * * * *

The Ohio State University

2006

Dissertation Committee:  Approved by
Professor Elizabeth Stasny, Ph.D., Adviser
Professor Angela Dean, Ph.D.
Professor Jason Hsu, Ph.D.

______________________________
Adviser
Graduate Program in
Statistics
ABSTRACT

Missing data is commonplace with both surveys and experiments. For this dissertation, we consider imputation methods founded in Survey Sampling, and assess their performance with experimental data. With a two-way interaction model, missing data renders Multiple Comparisons Procedures invalid; we seek a resolution to this problem through development of a Multiple Imputation Procedure. By completing an incomplete data set, we obtain a balanced data set for which multiple comparisons of treatment effects may be performed.

We develop an imputation procedure, Repeated Measures Normal Imputation (RMNI), for use with any hierarchical linear model. The advantage of RMNI is that the procedure preserves the underlying variance-covariance matrix structure of the model. The two-way interaction model has a spherical variance-covariance matrix, and the property of sphericity is required for the existence of a valid Multiple Comparisons Procedure. With RMNI, we are assured that the imputed values do not violate assumptions regarding the structure of the variance-covariance matrix of the data. With multiple imputations, we are assured that the imputed values are not treated as real observed data. Through RMNI, we are able to demonstrate the construction of a multiply-imputed confidence interval for each treatment contrast using a standard Tukey procedure, with confidence that the width of the interval is adjusted for uncertainty due to missing data.
In

loving memory of my father

Melvin M. Kosler, Senior

and gratitude

to the love, compassion, and dedication I have known

through my inspiring mother

Evelyn Frances Kosler

and

through sterling guidance

with

Elizabeth Stasny

and

Barbara Friedman

iii
ACKNOWLEDGMENTS

Thank you Elizabeth Stasny
for sharing with me your unyielding professionalism, enthusiasm, and patience,
as well as your brilliant expertise, intuition, and insight.
I would like to express my appreciation for the committee work of
Angela Dean and Jason Hsu, as well as
Deborah Rumsey, Michael Fligner, Steven MacEachern, and Mario Peruggia.
Additionally, I would like to acknowledge the professional examples of
Michael Fligner, Douglas Wolfe, and Kimberly Kinateder,
who each did heavily influence my practice of Statistics.
My profuse thanks are extended to Bonnie Parish,
Donald Laufersweiler, Howard Fradkin, Rich Cummings, and Barbara Friedman.
Finally, I would like to acknowledge the encouragement and supportiveness of my family:
Mom, Joann, Melvin, Beth, and Amy;
as well as the enduring sincere friendship of
Christine Wong and Deborah Countess.
VITA

May 23, 1969 ........................ Born in San Antonio, Texas

1986 .................................. Diploma, Thomas Edison High School
   San Antonio, Texas

1990 ................................. B.A., Trinity University
   Adviser: Professor Diane Saphire, Ph.D.
   San Antonio, Texas

1992 ................................. M.S., The Ohio State University
   Adviser: Professor Elizabeth Stasny, Ph.D.
   Columbus, Ohio

1990 - 1996 .......................... Lecturer and Teaching Assistant
   Department of Statistics
   The Ohio State University
   Columbus, OH

Summers 1991, 1992, 1993 ... Mathematical Statistician, Research Assistant
   Mentor: Robert Tortora, Ph.D.
   Federal Bureau of the Census
   Suitland, MD

1992 - 1993 .......................... Research Assistant
   Mentor: Professor Elizabeth Stasny, Ph.D.
   The Center for Human Resource Research
   Columbus, OH

1995 ................................. Mathematical Statistician
   Lubrizol Corporation
   Wickliffe, OH
1996 - 1997 ........................... Analytical Consultant
Lincoln Electric Company
Euclid, OH

1997 - 1999 ........................... Analytical Consultant
Digital Equipment Corporation / Compaq
Cleveland, OH

2000 - 2002 ........................... Instructor, Course Coordinator, Teaching Assistant
Department of Statistics
The Ohio State University
Columbus, OH

2002 - 2004 ........................... Assistant Professor
Department of Mathematics
Indiana University of Pennsylvania
Indiana, PA

2004 - 2006 ........................... Out of State Researcher
Department of Statistics
The Ohio State University
Columbus, OH from Indiana, PA

2003 - 2005 ........................... Statistical Consultant
Airo Die Casting
Leggett & Platt Aluminum Corporation
Latrobe, PA

2006 - Present ....................... Statistician
Childhood Depression Research Project
Western Psychiatric Institute and Clinic
University of Pittsburgh Medical Center
Pittsburgh, PA

FIELDS OF STUDY

Major Field: Statistics
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Vita</td>
<td>v</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiv</td>
</tr>
<tr>
<td>Chapters:</td>
<td></td>
</tr>
<tr>
<td>1. Thesis and Basis for Research</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Thesis</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Background</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Initial Purpose and Focus of This Dissertation</td>
<td>5</td>
</tr>
<tr>
<td>1.4 Missing-Data Methods and Multiple Comparisons</td>
<td>8</td>
</tr>
<tr>
<td>1.5 Multiple Imputation</td>
<td>10</td>
</tr>
<tr>
<td>1.6 Chapter Summaries</td>
<td>12</td>
</tr>
<tr>
<td>2. Multiple Imputation</td>
<td>15</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>15</td>
</tr>
<tr>
<td>2.2 Missing-Data Environment</td>
<td>16</td>
</tr>
<tr>
<td>2.2.1 Context of Survey Sampling</td>
<td>16</td>
</tr>
<tr>
<td>2.2.2 Context of Experimental Design</td>
<td>19</td>
</tr>
<tr>
<td>2.3 Sampling Mechanism</td>
<td>22</td>
</tr>
<tr>
<td>2.3.1 Context of Survey Sampling</td>
<td>22</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Context of Experimental Design</td>
</tr>
<tr>
<td>2.4</td>
<td>Response Mechanism</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Context of Survey Sampling</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Context of Experimental Design</td>
</tr>
<tr>
<td>2.5</td>
<td>The Method of Multiple Imputation</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Complete-Data Estimator of $Q$</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Combined Estimate of $Q$</td>
</tr>
<tr>
<td>2.5.3</td>
<td>Process for Multiple Imputation</td>
</tr>
<tr>
<td>2.6</td>
<td>$m$-Asymptotic Properties of $\bar{Q}_m$</td>
</tr>
<tr>
<td>2.7</td>
<td>Confidence Regions and Significance Levels</td>
</tr>
<tr>
<td>2.8</td>
<td>Fraction of Missing Information</td>
</tr>
<tr>
<td>2.9</td>
<td>Quality of Multiply-Imputed Inference</td>
</tr>
<tr>
<td>2.9.1</td>
<td>Randomization-Valid Multiply-Imputed Inference</td>
</tr>
<tr>
<td>2.9.2</td>
<td>Proper Multiple-Imputation Procedure</td>
</tr>
<tr>
<td>2.9.3</td>
<td>Bayesianly Proper</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>3.2</td>
<td>The Two-Way Mixed-Effects Interaction Model</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Complete-Data Model Specification</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Incomplete-Data Model Specification</td>
</tr>
<tr>
<td>3.3</td>
<td>Estimation of Treatment Effects</td>
</tr>
<tr>
<td>3.3.1</td>
<td>LSE for a Fixed Treatment Effect</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Weighted Means Estimator of $\theta$</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Comparison of $\bar{\theta}$ and $\hat{\theta}$</td>
</tr>
<tr>
<td>3.4</td>
<td>Balanced-Data Estimation of Variance Components</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Balanced-Data Estimators of $\sigma^2$</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Experimentwise Subject Deletion</td>
</tr>
<tr>
<td>3.5</td>
<td>Unbalanced-Data Estimation of Variance Components</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Mean Squares for Unbalanced Data</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Expected Mean Squares</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>4.2</td>
<td>Repeated Measures Normal Imputation</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Background and Development of RMNI</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Procedural Steps for RMNI</td>
</tr>
<tr>
<td>4.3</td>
<td>Assumptions: Data Environment</td>
</tr>
<tr>
<td>4.4</td>
<td>Assumptions: Model</td>
</tr>
<tr>
<td>4.5</td>
<td>Assumptions: Missing Data</td>
</tr>
</tbody>
</table>
4.6 Assumptions: Analysis Environment
4.6.1 Complete-Data Estimates
4.6.2 Incomplete-Data Estimates
4.6.3 Multiply-Imputed Estimates
4.7 RMNI Simulation and the Two-Way Mixed Model
4.7.1 Step Zero: Complete Data Baseline and MCAR Data
4.7.2 Step Three: Box and Tiao’s Posterior Approximations
4.7.3 Step Four: Mardia, Kent, and Bibby’s \[ Y_{mis} | Y_{obs} \]
4.7.4 Step Five: Generating Random Draws from \[ Y_{mis} | Y_{obs} \]
4.8 Programming for Simulation and Analysis
4.9 Evolution of RMNI

5. Demonstration of RMNI
5.1 Basis for Comparison: Complete and Incomplete Data
5.2 Pilot Run of RMNI
5.3 RMNI Convergence Test Run with \( m = 1000 \)
5.3.1 Stable Multiply-Imputed Estimates with \( m = 400 \)
5.3.2 Box and Tiao Posterior Draws
5.3.3 Multiply-Imputed Treatment Means Using RMNI
5.3.4 Multiply-Imputed Total Variance Using RMNI
5.4 Reasonable Extension of Simulation Results to Practice
5.4.1 Proper Imputation Related with Number of Imputations
5.4.2 Heuristic Argument That RMNI is Proper for \( \hat{q} \)

6. Demonstration of MCP
6.1 Selection of the Multiple Comparisons Procedure
6.2 Simulated Assessment of Multiple Comparisons Using RMNI
6.2.1 Complete-Data Results
6.3 Confidence Intervals for Treatment Contrasts Using RMNI
6.3.1 Completed-Data Estimates of Treatment Contrasts
6.3.2 Completed-Data Estimates of Tukey Standard Error
6.4 Tukey Confidence Bounds with Imputed Values
6.5 Influence of the Imputed Value Using RMNI
6.6 Additional Simulations

7. Model-Based Multiple Imputation
7.1 Two Paradigms for Multiple Imputation
7.1.1 The Multivariate Normal Approach
7.1.2 The Model-Based Approach
7.2 Comparison of Imputed Values: RMNI and Data Augmentation . . . . 162
  7.2.1 Procedures ................................................................. 162
  7.2.2 Comparison ................................................................. 165
7.3 Conclusion ................................................................. 172

8. Prospects for Continued Research ............................................ 174
  8.1 Theoretical Statistics ...................................................... 174
  8.2 Applied Statistics ......................................................... 176
  8.3 Statistical Computing ..................................................... 177

Appendices:

A. Master Program Files ...................................................... 178
  A.1 RMNI Master File: miauto_rmnix.ksh .................................. 178
  A.2 Master File: miauto.ksh .................................................. 179

B. Programs: Utility .............................................................. 181
  B.1 Code: batchsas.ksh ....................................................... 181
  B.2 Code: fcdatain.sas ....................................................... 181
  B.3 Code: miautoclean.ksh .................................................. 182
  B.4 Code: midata_sas.ksh ..................................................... 182
  B.5 Include File: model.par .................................................. 183
  B.6 Code: parseglm.ksh ....................................................... 184
  B.7 Code: parseglm2.ksh ..................................................... 185
  B.8 Code: parseglm.mi.ksh ................................................... 186
  B.9 Code: parseglm2.mi.ksh ................................................ 187
  B.10 Code: parsevc.ksh ...................................................... 188
  B.11 Code: seq_file.ksh ...................................................... 189

C. Programs: Simulation and Subject Deletion ................................ 190
  C.1 Main Program: compdata.f .............................................. 190
  C.2 Main Program: compstat.f .............................................. 191
  C.3 Code: datgen.f .......................................................... 192
  C.4 Code: miss.f ............................................................ 194
  C.5 Main Program: missdata.f .............................................. 195
  C.6 Main Program: multimp.f .............................................. 197
  C.7 Code: refmat.f .......................................................... 199
  C.8 Code: subchk.f .......................................................... 200
D. Programs: Calculating Initial Estimates .......................... 203
   D.1 Code: batchglm_comp.sas ......................................... 203
   D.2 Code: batchglm_misdat.sas ....................................... 203
   D.3 Code: clsevar.f .................................................. 204
   D.4 Code: reduce.f .................................................. 205
   D.5 Code: varcmp.f .................................................. 206

E. Programs: Posterior Draws for Parameter Values ................. 208
   E.1 Code: sigmas.f .................................................. 208

F. Programs: Posterior Draws for $Y_{mis}$ ........................... 212
   F.1 Code: datcom.f .................................................. 212
   F.2 Code: rmni.f .................................................. 216

G. Programs: Completed-Data Analysis .................................. 224
   G.1 Code: lsmns.f .................................................. 224
   G.2 Code: mifinal.f ................................................. 225
   G.3 Code: milsevar.f ............................................... 228
   G.4 Code: milsmns.f ............................................... 229
   G.5 Code: mivarcmp.f .............................................. 230

Bibliography ........................................................................ 233
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Parameter Values for Simulation of the Two-Way Mixed Model</td>
<td>101</td>
</tr>
<tr>
<td>5.2</td>
<td>Baseline Comparison Estimates of Model Parameters</td>
<td>102</td>
</tr>
<tr>
<td>5.3</td>
<td>RMNI-Imputed Data Sets</td>
<td>104</td>
</tr>
<tr>
<td>5.4</td>
<td>RMNI-Imputed Values with Five Imputations Under Treatment One</td>
<td>105</td>
</tr>
<tr>
<td>5.5</td>
<td>RMNI-Imputed Values with Five Imputations Under Treatment Two</td>
<td>106</td>
</tr>
<tr>
<td>5.6</td>
<td>RMNI-Imputed Values with Five Imputations Under Treatment Three</td>
<td>106</td>
</tr>
<tr>
<td>5.7</td>
<td>RMNI-Imputed LSEs and Multiply-Imputed LSE for $\theta$</td>
<td>107</td>
</tr>
<tr>
<td>5.8</td>
<td>RMNI-Imputed Within Variance and Multiply-Imputed Within Variance</td>
<td>108</td>
</tr>
<tr>
<td>5.9</td>
<td>Multiply-Imputed Between Variances and Covariances</td>
<td>109</td>
</tr>
<tr>
<td>5.10</td>
<td>Rates of Missingness for Each Treatment Effect</td>
<td>130</td>
</tr>
<tr>
<td>6.1</td>
<td>PROC MIXED Statements for Performance of Pairwise Comparisons</td>
<td>140</td>
</tr>
<tr>
<td>6.2</td>
<td>PROC GLM Statements for Performance of Pairwise Comparisons</td>
<td>141</td>
</tr>
<tr>
<td>6.3</td>
<td>Complete-Data Multiple Comparisons of Treatment Effects</td>
<td>143</td>
</tr>
<tr>
<td>6.4</td>
<td>Incomplete-Data Multiple Comparisons of Treatment Effects</td>
<td>144</td>
</tr>
<tr>
<td>6.5</td>
<td>Interval Widths for Complete-Data and Incomplete-Data MCP</td>
<td>145</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>6.6</td>
<td>RMNI-Imputed Multiple Comparisons of Treatment Effects</td>
<td>158</td>
</tr>
<tr>
<td>7.1</td>
<td>NORM-Imputed Values with Five Imputations Under Treatment One</td>
<td>165</td>
</tr>
<tr>
<td>7.2</td>
<td>NORM-Imputed Values with Five Imputations Under Treatment Two</td>
<td>166</td>
</tr>
<tr>
<td>7.3</td>
<td>NORM-Imputed Values with Five Imputations Under Treatment Three</td>
<td>168</td>
</tr>
<tr>
<td>7.4</td>
<td>Summary Statistics for Twenty NORM-Imputed Values Under Treatment One</td>
<td>169</td>
</tr>
<tr>
<td>7.5</td>
<td>Performance Comparison of RMNI with NORM</td>
<td>170</td>
</tr>
<tr>
<td>7.6</td>
<td>Summary Statistics for 100 RMNI-Imputed Values Under Treatment One</td>
<td>171</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Initial Values for Subject Standard Deviation</td>
<td>111</td>
</tr>
<tr>
<td>5.2</td>
<td>Initial Values for Interaction Standard Deviation</td>
<td>112</td>
</tr>
<tr>
<td>5.3</td>
<td>Initial Values for Error Standard Deviation</td>
<td>113</td>
</tr>
<tr>
<td>5.4</td>
<td>Cumulative Means of Initial Values for Each Variance Component</td>
<td>114</td>
</tr>
<tr>
<td>5.5</td>
<td>Cumulative Standard Deviations of Initial Values for Each Variance Component</td>
<td>115</td>
</tr>
<tr>
<td>5.6</td>
<td>Completed-Data LSE for $\theta_1$ Using RMNI</td>
<td>116</td>
</tr>
<tr>
<td>5.7</td>
<td>Completed-Data LSE for $\theta_2$ Using RMNI (Above)</td>
<td>117</td>
</tr>
<tr>
<td>5.8</td>
<td>Completed-Data LSE for $\theta_3$ Using RMNI</td>
<td>118</td>
</tr>
<tr>
<td>5.9</td>
<td>Combined LSE for $\theta_1$ Using RMNI</td>
<td>119</td>
</tr>
<tr>
<td>5.10</td>
<td>Multiply-Imputed LSE for $\theta_2$ Using RMNI</td>
<td>120</td>
</tr>
<tr>
<td>5.11</td>
<td>Multiply-Imputed LSE for $\theta_3$ Using RMNI</td>
<td>121</td>
</tr>
<tr>
<td>5.12</td>
<td>RMNI-Imputed $\sigma_y$ with each of 986 Imputations</td>
<td>122</td>
</tr>
<tr>
<td>5.13</td>
<td>Multiply-Imputed Within Variance Using RMNI</td>
<td>123</td>
</tr>
<tr>
<td>5.14</td>
<td>Multiply-Imputed Within Covariance Using RMNI</td>
<td>124</td>
</tr>
<tr>
<td>5.15</td>
<td>Multiply-Imputed Between Variance Using RMNI</td>
<td>125</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>5.16</td>
<td>Multiply-Imputed Between Covariance Using RMNI</td>
<td>126</td>
</tr>
<tr>
<td>5.17</td>
<td>Multiply-Imputed Total Variance Using RMNI</td>
<td>128</td>
</tr>
<tr>
<td>5.18</td>
<td>Multiply-Imputed Total Covariance Using RMNI</td>
<td>129</td>
</tr>
<tr>
<td>6.1</td>
<td>Single Imputation Treatment Contrasts Using RMNI</td>
<td>147</td>
</tr>
<tr>
<td>6.2</td>
<td>Multiply-Imputed Treatment Contrasts Using RMNI in Series with $m$</td>
<td>148</td>
</tr>
<tr>
<td>6.3</td>
<td>Single Imputation Interaction Mean Square Using RMNI</td>
<td>149</td>
</tr>
<tr>
<td>6.4</td>
<td>Single- and Repeated Within Variance Using RMNI</td>
<td>150</td>
</tr>
<tr>
<td>6.5</td>
<td>Multiply-Imputed Between Variance for Each Treatment Contrast Using RMNI</td>
<td>152</td>
</tr>
<tr>
<td>6.6</td>
<td>Multiply-Imputed Total Variance for Each Treatment Contrast Using RMNI</td>
<td>153</td>
</tr>
<tr>
<td>6.7</td>
<td>Single Imputation Studentized Range Half Widths</td>
<td>155</td>
</tr>
<tr>
<td>6.8</td>
<td>Repeated Imputation Studentized Range Half Widths in Series with $m$</td>
<td>156</td>
</tr>
<tr>
<td>6.9</td>
<td>Multiply-Imputed Studentized Range Half Widths in Series with $m$</td>
<td>157</td>
</tr>
</tbody>
</table>
CHAPTER 1

Thesis and Basis for Research

1.1 Thesis

General Statement This dissertation demonstrates a model-based approach to imputation in the case of performing statistical analysis which is theoretically dependent upon a spherical structure of the variance-covariance matrix. For the demonstration, we have chosen the two-way mixed effects interaction model designated as Model IA by Ronald Hocking (1973). The commercially adopted approach, popularized by Joseph Schafer (1997), soundly preserves the distributional assumption of Multivariate Normal (MVN) data, but lacks the sophistication required to preserve the sensitive variance-covariance structure under this model, as demonstrated in Chapter Seven. Consequently, for the performance of multiple comparisons requiring a spherical model for the MVN data, we demonstrate that specification and incorporation of the model should be component to the imputation procedure. Donald Rubin (1987) concludes that a Multiple Imputation Procedure is proper for the intended inferential statistics only if the imputation model and the analysis model are the same model for the data. Nevertheless, specification of our imputation model is not component to the SAS MI Procedure, and cannot be accomplished with the NORM freeware algorithm for Data Augmentation offered by Joseph Schafer (1997). Therefore, we cannot expect the commercially adopted MVN approach to preserve the spherical structure
of our imputation model. We develop and demonstrate the procedure of the Repeated Measures Normal Imputation (RMNI) introduced by Daryl Yamashita (1995) under the two-way interaction model, and we develop and demonstrate a method for performing valid multiple comparisons of treatment effects under this model.

1.2 Background

Introduction This dissertation targets the missing-data problem in the context of Experimental Design. In particular, we apply Rubin’s (1987) method of multiple imputation to experimental data and assess the results of multiply-imputed inference. Little and Rubin (1987) define an imputation-based procedure as one under which “missing values are filled in and the resultant completed data are analyzed by standard methods.” Procedures for single- and multiple imputation in the presence of missing survey data are well documented in the literature, and commonly used in practice. The adaptation of survey-specific missing-data methods to the analysis of experimental data is compelling because of the common roots of theoretical development for Survey Sampling and Experimental Design. The impetus to multiply impute experimental data derives from the demonstrated validity of Multiple Comparisons Procedures performed using multiply-imputed data under a one-way fixed-effects repeated-measures model (Yamashita, 1995). We demonstrate this method of multiple imputation as extended to the two-way mixed-effects model, with repeated measures and inclusion of the interaction term. To begin, we characterize the missing-data problem and summarize the approach of multiple imputation. (We refers to the writer, Joseph Kosler, and the adviser, Elizabeth Stasny, Ph.D.)

Characterization of Missing Data Statistical analysis of data obtained through a designed experiment may be limited by missing or deficient data. A missing datum is one
that is expected to be measured given the design of the experiment, but is not actually observed. A deficient datum is one that is recorded, yet is of insufficient quality to be included in the analysis. For example, any datum that is corrupted by an incorrectible practical error is deficient. If the deficient datum is discarded by the analyst, that response value is effectively missing.

**Definition of Missingness**  Within the context of Experimental Design, we define missingness, the presence of missing data, to be analogous to survey nonresponse, which Lessler and Kalsbeek (1992) define as “the inability to obtain useful data on all questionnaire items from all members of the sample.” That is, missingness is defined to occur with the failure to obtain useful data with each and every response measurement required by the design of the experiment. Unit-missingness refers to the lack of all observations required by design for a particular experimental unit/subject; and item-missingness refers to the lack of any single observation required by the design of the experiment, as specified by the experiment’s design matrix. Specifically, for a two-way fixed-effects repeated-measures interaction model, data may be cross-tabulated, with each cell of the table holding the repeated measures of the response variable corresponding to one combination of the two main effects. Unit-missingness occurs if an entire row or column of data is missing, which typically reflects the absence of an experimental unit or subject. Item-missingness occurs if any cell in the table of observed data lacks any repeated measure which was specified by the design matrix. With regard to an experiment or survey, a fully-observed data set is commonly called complete, and a partially-observed data set is referred to as an incomplete data set. We use the term “item-missingness” to characterize the data set which is devoid of unit-missingness,
yet incomplete; analysis of such a data set avoids speculation about any level of a fixed effect for which no information is available.

**Primary Dissertation Reference**  Donald B. Rubin’s *Multiple Imputation for Nonresponse in Surveys* (1987) develops the method of multiple imputation in the context of Survey Sampling, and suggests its application in the context of Experimental Design. In one example, Rubin (1987) conducts inference regarding regression coefficients, proceeding with standard least squares analysis and using multiply-imputed survey data. Rubin (1987) claims that multiple imputation can be used to handle missingness in a nonsurvey context, but does not deeply explore any such application using experimental data. A description of Rubin’s method is given in Chapter Two of this dissertation.

**Prequel Dissertation**  Yamashita (1995) demonstrates the use of multiple imputation with incomplete experimental data. In particular, he creates confidence intervals based on Dunnett’s Multiple Comparisons procedure for the balanced one-way repeated measures model. The impact of missingness on the analyst’s ability to perform multiple comparisons of fixed treatment effects is discussed in Section 1.3. In short, missingness is likely to render Multiple Comparisons procedures invalid. Yamashita (1995) shows that the analyst can perform valid Multiple Comparisons procedures using multiply-imputed experimental data. We extend this research by applying its propositions to the two-way mixed-effects repeated-measures interaction model.
1.3 Initial Purpose and Focus of This Dissertation

**Introduction** With this study, we investigate the viability of Rubin’s (1987) method of multiple imputation in the context of Experimental Design. In particular, we assess the performance of multiply-imputed data under the study of multiple comparisons of treatment effects. Consideration is given to the robustness of procedures for multiple imputation and multiply-imputed inference with respect to the model for the data, and to the size of the experiment.

**Pilot Simulation Study** Yamashita (1995) develops a multiple-imputation procedure for use with the balanced one-way repeated measures model; this Repeated Measures Normal Imputation procedure, or RMNI, is made explicit in Chapter Four. The current research began with the notion of adapting RMNI to other linear models, and was launched with a simulation study Pilot Study results and conclusions are discussed in Chapter Three and Chapter Four, particularly Section 4.9. A complete demonstration of RMNI is presented in Chapter Five. Ultimately, we have extended the RMNI procedure for use under the two-way model, and have clarified that the RMNI procedure may be adapted for use under any hierarchical model.

**Approximating Posterior Distributions** This study considers the problem of approximating the posterior distribution for each each parameter of the hierarchical model, as specified in Section 3.2, as well as the posterior distribution of the missing data. We use the approximation methods of Box and Tiao (1973) to obtain posterior random draws for each variance component of the model (Equation 3.4); and we use the conditional distribution of the missing data (Section 4.2.2), given the observed data, to obtain *imputed values*
as substitutes for the missing data. The commercially adopted MVN approach applies iterative methods which require manual assessment of convergence; our approach (RMNI) requires matrix inversion (Equation 4.5).

**Incomplete-Data Estimators** The recommended Bayesian procedure for generation of imputed values under a specified model requires unbiased initial estimates of the model parameters which are computed using the observed incomplete data set. We consider various options for estimation of fixed main effects, as well as approaches to the estimation of variance components. In particular, we discuss the problem of calculating expected mean squares, and obtaining unbiased variance component estimates with unbalanced data. Results of the pilot study indicate the robustness of multiply-imputed inference to various incomplete-data estimators. Discussion and conclusions regarding initial estimates for our application of RMNI are presented in Chapter Three.

**MCAR Assumption** Rubin (1976) defines the term *missing completely at random*, or MCAR, to mean that whether a datum is missing is independent of the response variable, as well as any known covariate of the response variable. Alternatively, missingness is defined to be *missing at random*, or MAR, if it is independent of the response variable, yet dependent upon one of its known covariates. Missingness is not MCAR, for example, in the case that the interaction between a treatment and a unit or subject introduces some malfunction or toxicity which renders the response unmeasurable. For this study, we assume that missingness is MCAR unless specified otherwise.

**Procedure for Generating Missingness** We begin by simulating a complete data set. Subsequently, we generate a pattern of missingness. We use with-replacement draws from
the Discrete Uniform distribution, with parameter given by the size of the experiment, to generate a pattern of MCAR missingness. The number of draws is governed by the target rate (percentage) of missingness desired for the demonstration. We replace simulated original data with the value of 999999 at each location determined by the random draws from the Discrete Uniform. The end result is a complete data set for comparison, and an incomplete data set with MCAR data for demonstration of imputation.

Models for Missingness  Rubin (1987) explores the incorporation of a model for nonresponse into his multiple imputation procedure. He states that multiple imputation is especially valuable in the context of Survey Sampling partly because “it is rare that the missing values occur completely at random, as they might in some experimental contexts.” In the event that missingness is dependent upon the response variable, or a known covariate of it, a stochastic model for missingness is incorporated by the method of multiple imputation so that this dependency is reflected by imputed values. This dissertation does not address the problem of modeling missingness in depth. A note on further work in this area is given in Chapter Eight. For more on the occasion of MCAR data, or lack thereof, in the context of Survey Sampling, we recommend the discussion given by Scheuren (2005).

Assumption of Subject Deletion  It is inadvisable to impute values for a unit or subject that is completely unobserved. In the presence of unit-missingness, subject deletion refers to the removal of an experimental unit or subject from statistical analysis where it is nowhere observed. If experimental data are MCAR, then subject deletion will not affect the method of analysis. Consequently, if an experimental unit or subject is completely unobserved, then we delete that unit/subject from inference. If an experimental unit or subject is partially observed, but completely unobserved under a particular treatment, then
we retain that unit/subject for inference in general; however, we delete that unit/subject from inference on the treatment under which no valid observation is obtained. With the practice of subject deletion, we focus on analysis in the presence of item-missingness, as it is described in Section 1.1.

1.4 Missing-Data Methods and Multiple Comparisons

Motivation  This dissertation research derives from the problem presented by Jason Hsu that no valid Multiple Comparisons Procedure exists in the case of unbalanced data following a two way mixed interaction model (Voss and Hsu, 1998). Our approach has been to use RMNI (multiple imputation) to complete an unbalanced data set of this type, and apply a valid MCP to the multiply imputed data. Procedural development of RMNI is discussed in Chapter Four; and demonstration of RMNI under this model is presented in Chapter Five. Demonstration of the performance of multiple comparisons is presented in Chapter Six.

Assumption of Rectangular Data  Throughout this dissertation, we refer to a data set as balanced if it can be displayed with a rectangular two-way table, in the usual way; and not balanced otherwise. That is, balance refers to the inclusion of the same number of replicates for each combination of treatment and subject. Most generally, a general linear model is said to be balanced if the general least squares estimate is equal to the ordinary least squares estimate for each treatment effect (Scheffe, 1959). An experimental design is considered balanced wherever it prescribes observation of a balanced data set. Despite balance of the design, we might observe an unbalanced data set due to missingness. For more on the subject of balance and its impact upon procedures for multiple comparisons, we refer you to Wang (92).
Multiple Comparisons with Incomplete Data  Scheffe’s method for pairwise comparisons applies to unbalanced designs, but it is known to produce conservative confidence intervals for differences between treatments. Liberal alternatives for unbalanced data may or may not exist, depending upon the model, and depending upon whether the data is balanced. Hochberg and Tamhane (1987) discuss unbalanced-data alternatives only for the one-way layout, but with little enthusiasm. Voss and Hsu (1998) explicate that no unbalanced-data method under the two-way interaction model can be valid.

Multiple Imputation and Multiple Comparisons  Yamashita (1995) studies a multiple imputation procedure for the one-way repeated measures model, used with Dunnett’s Multiple Comparisons Procedure, to derive multiply-imputed confidence intervals. He shows by simulation that these multiply-imputed confidence intervals have approximately the correct coverage probabilities and widths comparable to those in the case of complete data. The results of Yamashita (1995) are limited to a specific pattern of missing data (monotonic), and are limited to the case of the One-Way Layout. With this dissertation, we extend Yamashita’s (1995) approach to demonstrate a procedure for RMNI which can be applied under any pattern of MCAR missingness. Our procedure is specifically demonstrated for the two-way mixed interaction model, but can easily be adapted to function with any hierarchical model. Under the Two-Way Layout, we refer to Hochberg and Tamhane (1983) for an exact Tukey MCP which can be conducted using our RMNI-imputed data. We present the multiply-imputed confidence intervals for treatment contrasts under the two-way interaction model in Chapter Six.
1.5 Multiple Imputation

Motivation Our goal is to complete an unbalanced data set observed under the two-way interaction model because of missingness, and not because of design. The value of completing the unbalanced data set is to be able to perform valid multiple comparisons of treatment effects. Imputation methodology offers many options, which we review briefly here.

Single- versus Multiple Imputation Single imputation has the advantage of generating a complete data set, as well as incorporating the experimenter’s knowledge into the production of imputed values. Rubin (1987), however, emphasizes that a single set of imputed values cannot dependably reflect the variability of the intended observations, especially if more than one model for nonresponse is plausible. Furthermore, single-imputation methods treat the imputed value as known, and make no correction for its inaccuracy. Consequently, variability in the data and standard error in inference are underestimated, leading to confidence intervals that are too narrow, and misleading multiple comparisons. Ideally, the analyst could impute likely values for the missing data and then conduct standard complete-data inference, while accounting for loss in accuracy due to the imputation: the method of multiple imputation accomplishes this.

Method The method of multiple imputation, which is described in detail in Chapter Two, replaces each missing datum with two or more plausible values, representing a distribution of possibilities. We specifically draw $m$ imputed values for each missing datum from its posterior distribution, according to model assumptions for the response variable. We then complete $m$ data sets using each of the $m$ sets of imputed values. Standard complete-data inference is then conducted on the $m$ completed (by imputation) data sets. The $m$ sets
of completed-data estimates are finally combined into a single set of estimates of model parameters, with associated standard errors.

**Plausible Models** With this dissertation, we focus on the two-way mixed interaction model. Nevertheless, the process given above, and the procedure of RMNI described hereafter, may be repeated for each plausible model for the data, and potentially for each plausible model for nonresponse. Rubin (1987) develops theory for combining the results of the $m$ analyses under one model into a single set of parameter estimates. Unfortunately, Rubin (1987) does not prescribe a method for combining multiply-imputed estimates under different models into a single estimate across models. This does not prevent us from comparing results of Multiple Comparisons procedures between models, where multiply-imputed inference is conducted separately under each plausible model. Furthermore, the completed data sets under each model may be used to conduct a sensitivity analysis, testing the robustness of conclusions of a Multiple Comparisons Procedure with respect to the choice of imputation model.

**Influence of the Imputation Model** In Chapter Eight, we discuss potential for further exploration of the impact of the choice of model for imputation upon the final analysis of multiply imputed data. A criticism of imputation is that the imputed values may skew inferential results in favor of the model which the imputed values do follow. We believe it is likely that the influence of the imputation model upon multiply imputed inference is a function of the fraction of missing information due to missingness. In this dissertation, for purposes of computing and simplification of discussion, we have chosen to focus on the
percentage or rate of missingness for an experiment, and we have not considered the frac-
tion of missing information directly. Additionally, we have excluded attempts to quantify
the influence of an imputation model upon inferential statistics.

1.6 Chapter Summaries

Chapter Two: Discussion of Multiple Imputation  In Chapter Two, Rubin’s (1987) me-

thod of multiple imputation is described in detail. Results relevant to the general imple-

mentation of the method, as demonstrated in the pilot simulation study of Chapters Five

and Six, are provided here. Yamashita’s (1995) RMNI procedure is also presented here.
The notions of randomization-valid and proper multiply-imputed inference are discussed
in the final section.

Chapter Three: Discussion of the Two-Way Layout  In Chapter Three, we declare the
two-way mixed model considered for this dissertation. A review of inferential statistics
under this model is provided, along with results pertinent to our work. In particular, the
controversy over Expected Mean Squares with unbalanced data is discussed here.

Chapter Four: Procedural Details of RMNI  In Chapter Four, we present a protocol for
the performance of Repeated Measures Normal Imputation. A digest of assumptions and
regularity conditions is provided here, along with a thorough discussion of each step of the
RMNI procedure. Additionally, a developmental history of RMNI is presented.

Chapter Five: Simulated Demonstration of RMNI  In Chapter Five, we present our find-
ings with a performance of RMNI using simulated data generated under our true model,
and a randomly generated pattern of missingness. A small scale pilot example of RMNI is
displayed here, along with results from a run of RMNI with 1000 imputations per missing value. We assess the stability of the RMNI procedure, and argue that RMNI is a proper Multiple Imputation Procedure for performing multiple comparisons of treatment effects.

**Chapter Six: Simulated Demonstration of Multiple Comparisons Using RMNI** In Chapter Six, we apply RMNI to the problem of performing multiple comparisons with data following the two-way mixed interaction model, in the presence of missingness. A standard complete-data Multiple Comparisons Procedure is used, and the results are promising. In particular, we stress that RMNI is intended to preserve the underlying variance-covariance structure of the designed experimental data, as requisite of the standard complete-data MCP. No valid MCP for the case of unbalanced data (incomplete data) is known (Voss and Hsu, 1998).

**Chapter Seven: Two Paradigms for Multiple Imputation** In Chapter Seven, we compare RMNI with the popular Multivariate Normal (MVN) approach to Multiple Imputation, as they are available today with standard application software. The two paradigmns for the practice of Statistics, the multivariate approach and the model-based approach, manifest separately in the area of Multiple Imputation. The multivariate approach has been advanced by Joseph Shafer (1997, NORM software application) and SAS Proc MI with SAS Proc MIAnalyze. The model-based approach is manifest with RMNI, which was introduced by Yamashita (1995), and which shares theoretically sound basis in the work of Rubin (1987). We contend that the multivariate approach may be the more effective in the context of Survey Sampling, and that the model-based approach is the more effective in the context of Experimental Design.
Chapter Eight: Prospects for Continued Research  In Chapter Eight, we suggest research and development topics for further work in the exploration of Multiple Imputation, particularly RMNI. Topics include prospects for extending RMNI beyond our assumption of ignorable missingness and our choice of model. In conclusion, we assert the potential for application software development to advance RMNI to desktop computing, which has recently become possible.
CHAPTER 2

Discussion of Multiple Imputation

2.1 Introduction

General Multiple Imputation  We present the method of multiple imputation as developed by Rubin (1987), beginning with a construction of the missing-data environment. Specifically, we include theory relevant to the simulation study discussed in Chapter Five. Nevertheless, Rubin’s (1987) notation, method, and related distributional results are presented as they apply to both the survey and the experiment. In particular, we draw distinctions between interpretations of various quantities in terms of surveys and experiments.

Valid Multiple Imputation  Application of Rubin’s (1987) method requires specification of a procedure for generating imputed values, as well as the complete-data method of inference to be employed. While the method of multiple imputation itself does not vary with imputation procedure, the validity of multiply-imputed inference depends upon properties of the imputation procedure in conjunction with the method of inference. The notion of a Bayesianly proper imputation procedure is defined within a discussion of conditions which guarantee an imputation procedure will precipitate valid inference. In particular, we ultimately establish that our imputation procedure of choice, RMNI, does yield valid multiply-imputed inference in the case of performing multiple comparisons (Chapter Six).
2.2 Missing-Data Environment

Introduction We describe the missing-data problem in the context of Experimental Design using the notation of Rubin (1987). In Section 2.2.1, basic definitions are stated in the context of Survey Sampling, as given by Rubin (1987). In Section 2.2.2, these definitions are interpreted within the context of Experimental Design. Major interpretive differences stem from the translation of the survey’s finite sampling population and sampling distribution into the language of Experimental Design.

2.2.1 Context of Survey Sampling

Y, X, R, I*, and Q Within the survey context, consider the sampling of n experimental units or subjects from a population of size N, with the intent of obtaining their n values of the p×1 response vector, y, with which to conduct inference on a k×1 population parameter Q. Let Y be the N×p matrix where each row of Y is the p-variate response of one unit or subject within the sampling population. Let X be the known N×c covariate of Y, where each row of X is the 1×c known covariate of a single population subject. Rubin (1987) defines I* to be the N×p componentwise indicator of inclusion by the survey, associated with Y; and defines R to be the N×p componentwise indicator of whether a response will be obtained for the associated component of Y if that component is included in the survey.

Subscripts for Survey Variables Let i be the index of population units or subjects with i=1,…,N; and let j be the index of scalar response components to be measured on the ith unit or subject with j=1,…,p. For each i, let y_i comprise the components Y_ij of Y, where Y_ij represents the jth scalar response component which may be measured on the ith unit or
subject. Similarly, we distinguish a component of $\mathbf{R}$ by writing $R_{ij}$, and one of $\mathbf{I}^*$ with $I_{ij}^*$. With univariate response, where $p=1$, the subscript, $j$, is dropped.

**Inclusion Indicator $\mathbf{I}^*$** The inclusion indicator variable, $I_{ij}^*$, is binary such that $I_{ij}^*=1$ if the response $Y_{ij}$ is included in the survey, and $I_{ij}^*=0$ otherwise. That is, $I_{ij}^*=1$ if an attempt is made to measure $Y_{ij}$. In the case of univariate $\mathbf{Y}$, $\mathbf{I}^*$ is reduced to an $N \times 1$ vector; and the $i$th component of $\mathbf{I}^*$, $I_{i}^*$, becomes the binary indicator of whether the $i^{th}$ unit or subject is included in the survey. It is assumed that $I_{ij}^*$ is known, given the survey design, for all response variable components and all population units or subjects.

**Response Indicator $\mathbf{R}$** The response indicator variable, $R_{ij}$, is binary such that $R_{ij}=1$ if the $i^{th}$ subject will yield a value for $Y_{ij}$ if $Y_{ij}$ is included in the survey; otherwise, $R_{ij}=0$. In the case of univariate $\mathbf{Y}$, $\mathbf{R}$ is reduced to an $N \times 1$ vector; and the $i^{th}$ component of $\mathbf{R}$, $R_{i}$, becomes the binary indicator of whether the $i^{th}$ subject will respond if included in the survey. In the case of inanimate experimental units, $R_{ij}$ indicates whether $Y_{ij}$ can be measured if $Y_{ij}$ is included in the survey. It is assumed that $R_{ij}$ is observed for all $Y_{ij}$ included in the survey, and is unknown otherwise. That is, $R_{ij}$ is known if $I_{ij}^*=1$, and $R_{ij}$ is unknown if $I_{ij}^*=0$.

$\mathbf{Y}_{inc}$ and $\mathbf{Y}_{exc}$ Rubin (1987) groups components of $\mathbf{Y}$ according to states such as being included by the survey design, and refers to each collection of $\mathbf{Y}$-components by naming the set of subscripts which defines that collection, and subscripting $\mathbf{Y}$ with this set name. For example, $inc$ is the set of subscripts identifying $\mathbf{Y}$-components that are included by the survey design, and $\mathbf{Y}_{inc}$ represents the collection of included response components. Similarly, $\mathbf{Y}_{exc}$ represents the collection of excluded response components.
Sets of Y-Component Subscripts  Rubin (1987) considers the following collections of Y-components: those included in the sample; those excluded from the sample; those observed successfully; those missing from the observed data set, yet included in the sample; and those that are not observed, whether by exclusion from the sample or nonresponse.

We define $pop$ to be the collection of all subscripts for Y-components. Sets of subscripts which define these collections are specified below:

$$\text{pop} = \{(i,j): Y_{ij} \text{ is a component of } Y\}; \quad (2.1)$$

$$\text{inc} = \{(i,j): I_{ij} = 1\}; \quad (2.2)$$

$$\text{exc} = \{(i,j): I_{ij} = 0\}; \quad (2.3)$$

$$\text{obs} = \{(i,j): I_{ij} = 1, R_{ij} = 1\}; \quad (2.4)$$

$$\text{mis} = \{(i,j): I_{ij} = 1, R_{ij} = 0\}; \quad \text{and} \quad (2.5)$$

$$\text{nob} = \{(i,j): I_{ij} = 0 \text{ or } R_{ij} = 0\}. \quad (2.6)$$

Form of the Component Collection  The form of a component collection, such as $Y_{inc}$, may be a list, set, vector, or matrix, and is to be inferred from context. A combination of component collections is denoted by a list in parentheses. For example, we may write $Y_{nob} = (Y_{mis}, Y_{exc})$ to express that $Y_{mis}$ is the complement of $Y_{exc}$ relative to $Y_{nob}$, where $Y_{nob}$, $Y_{mis}$, and $Y_{exc}$ are sets. This notation is occasionally abused to write, for instance, $Y = (Y_{inc}, Y_{exc})$, where the notation might suggest that $Y_{inc}$ has the form of a rectangular matrix. Unfortunately, item-missingness generally denies $Y_{obs}$ explicit expression as a rectangular matrix. Typically, with univariate response, we partition $Y_{inc}$ into subvectors $Y_{obs}$ and $Y_{mis}$, and write $Y_{inc} = (Y_{obs}, Y_{mis})$ to mean that $Y_{mis}$ is appended to $Y_{obs}$ to complete a permuted $Y_{inc}$.  

18
Component Collections with $X$, $R$, and $I^*$  Rubin (1987) uses (2.1) through (2.6) to reference a collection of components of $X$, $R$, or $I^*$, which is discussed with regard to its associated collection of $Y$-components. For example, $X_{inc}$ denotes the components of $X$ corresponding to $Y_{inc}$; $R_{obs}$ denotes the components of $R$ corresponding to $Y_{obs}$, each of which is observed to be one by definition; and $I^*_{exc}$ denotes the known components of $I^*$ which correspond to $Y_{exc}$, so that by definition, if used as a vector or matrix, $I^*_{exc}$ is a representation of zero.

2.2.2  Context of Experimental Design

Introduction  With this section, we define the missing-data environment in the context of Experimental Design. In general, the literature of multiple imputation is written from the perspective of Survey Sampling. We consider univariate response without specification of design and model assumptions, so that it is sufficient to discuss experimental data as a vector of scalar responses, denoted by $Y$, and inclusive of repeated measures. Otherwise, we seek to make explicit the meaning of quantities defined by Rubin (1987), as they are applied here in the context of Experimental Design.

Size of the Experiment  Let $n$ be the number of response measurements required by the design of the experiment. Suppose, for example, that the experimenter intends to observe response values with $n_r$ repetitions on each of $n_s$ units or subjects under each of $n_t$ treatments, so that each combination of unit/subject and treatment is measured $n_r$ times. For simplicity, assume that $n_t$, $n_s$, and $n_r$ are each constant with respect to a specific treatment, unit/subject, or combination thereof. The value of $n$ is then given by $n = n_t n_s n_r$. We consider $n$, as defined here, to be analogous to the survey sample size.
“Sampling” Population for Y-Components  For comparison with the finite survey population size, let $N$ be the number of scalar response variables, $Y_j$, with $j = 1, \ldots, N$, such that $Y_j$ could be included by the design of the experiment. Continuing the example of the previous paragraph, let $N_s$ be the theoretical maximum number of units or subjects, which does not vary with treatment. That is, in a study of treatments for diabetes, we might consider $N_s$ to be the global number of people with diabetes, while $n_s$ is the number of diabetic subjects within the experiment. Let $N_r$ be the theoretical maximum number of repetitions, which does not vary with combination of treatment and unit/subject. Since the number of treatments is fixed, the value of $N$, if finite, is given by $N = n_t N_s N_r$. In theory, the value of $N_r$ is generally infinite, and the value of $N_s$ may be considered essentially infinite.

Response Component Subscripts  With factorial models, a Y-component may be subscripted with indices reflecting its placement in the design of the experiment. For example, each response component subscript represented here by $j$ might include an index to specify treatment, one to specify unit or subject, and one to specify repetition. We use a single-digit subscript for generality and convenience.

Sets of Y-Component Subscripts  We define collections of Y-components by the following sets of component subscripts which are analogous to those given by (2.1) through (2.6). Let $pop$ be the population of subscripts which label the $N$ theoretically possible response components. Let $inc$ be the practical subset of $pop$ comprising subscripts which distinguish the $n$ response components included by the design of the experiment. Let $exc$ be the complement of $inc$ relative to $pop$. Let $obs$ be the subset of $inc$ comprising subscripts of Y-components which are successfully observed. Let $mis$ be the complement of $obs$ relative to $inc$, so that $mis$ is the set of subscripts corresponding to “missing” Y-component
measurements. Finally, we take *nob* to be the complement of *obs* relative to *pop*. Explicit specification of these sets depends upon the choices of experimental design and model for response.

**Y and Y\text{inc}**  Adapting Rubin’s (1987) notation, we define \(Y = Y_{\text{pop}}\) to be the \(N \times 1\) theoretical vector of response component variables, of which \(n\) will be included for measurement by the experiment’s design. We define \(Y_{\text{inc}}\) to be the \(n \times 1\) vector of scalar response variables included by experimental design. For a given experiment, \(Y_{\text{inc}}\) is equivalent to the familiar response variable in the literature of Experimental Design.

**Other Partitions of Y**  We define \(Y_{\text{obs}}\) to be the practical vector of observed \(Y_j\), where \(j \in obs\), and \(Y_{\text{mis}}\) to be the practical vector of missing \(Y_j\), where \(j \in mis\), so that \(Y_{\text{inc}} = (Y_{\text{obs}}, Y_{\text{mis}})\). Finally, we define \(Y_{\text{nob}} = (Y_{\text{exc}}, Y_{\text{mis}})\) to be the theoretical vector of response components which are not observed, either by exclusion from the experiment or by missingness.

**X, the Design Matrix, and Known Covariates**  We define \(X\) in the context of Experimental Design to be the \(N \times c\) known covariate of \(Y\). The \(n \times c\) matrix \(X_{\text{inc}}\) comprises the rows of \(X\) which are associated with components of \(Y_{\text{inc}}\), so that \(X_{\text{inc}}\) is the known covariate of \(Y_{\text{inc}}\). We define \(X_{\text{exc}}\) to comprise the rows of \(X\) which are associated with components of \(Y_{\text{exc}}\), so that \(X = (X_{\text{inc}}, X_{\text{exc}})\). For this study, we assume no known covariate of \(Y\), other than the design of the experiment, which is defined only for the components of \(Y_{\text{inc}}\). Consequently, we interpret Rubin’s (1987) \(X_{\text{inc}}\) as equivalent to the design matrix for the experiment, where \(c\) is the number of individual effects in the model for data, excluding
random error. It follows that $X_{exc}$ is a matrix of zeroes. The application of multiple imputation in the presence of known covariates of $Y_{inc}$ other than the design matrix remains a topic for further research.

**The Inclusion Indicator $I^*$** Given the design of the experiment, we define $I^*$ and $I_{inc}^*$, respectively, to be the $N \times 1$ and $n \times 1$ componentwise indicators of inclusion by the design, associated with $Y$ and $Y_{inc}$, respectively. Note that by definition, $I_{inc}^*$ is a constant vector of ones, and is fixed by the design of the experiment. More is written about $I^*$ and related distributions in section 2.3, *Sampling Mechanism*.

**The Response Indicator $R$** Given the design of the experiment, we define $R$ and $R_{inc}$, respectively, to be the $N \times 1$ and $n \times 1$ componentwise indicators of successful observation, associated with $Y$ and $Y_{inc}$. In particular, for $j \in mis$, $R_j=0$, and for $j \notin mis$, $R_j=1$. More is said about $R$ and related distributions in section 2.4, *Response Mechanism*.

### 2.3 Sampling Mechanism

#### 2.3.1 Context of Survey Sampling

**Definition of Sampling Mechanism** Let $[Y|X]$ denote the conditional distribution of $Y$ given $X$. Rubin (1987) refers to the conditional distribution $[I^*|X,Y,R]$ as the *sampling mechanism* for the survey. The sampling mechanism is the probability model used as a basis for selection of a survey sample from the target population.

**Ignorable Sampling Mechanism** The sampling mechanism is said to be *ignorable* at the observed values $(X,Y_{obs},R_{inc},I^*)$ if

$$[Y_{nob}|X, Y_{obs}, R_{inc}, I^*] = [Y_{nob}|X, Y_{obs}, R_{inc}].$$

(2.7)
That is, the sampling mechanism is ignorable if the posterior distribution of $Y_{nob}$ does not depend upon the specification of the sampling mechanism. Rubin (1987) shows that the sampling mechanism is ignorable at $(X, Y_{obs}, R_{inc}, I^*)$ if and only if

$$[I^*|X, Y, R_{inc}] = [I^*|X, Y_{obs}, R_{inc}]$$

for all possible $Y_{nob}$, that is, if and only if the mechanism depends on the response only through observed values. Otherwise, the sampling mechanism is said to be \textit{nonignorable}.

### 2.3.2 Context of Experimental Design

**Randomization of Design vs. Probability Sampling**  Whereas inference in the survey context is rooted in sampling according to a probability model, inference using experimental data is grounded upon randomization of the design. While the distribution of $I^*$ is central to the analysis of survey data, it is irrelevant in the analysis of experimental data. Consequently, the sampling mechanism has no familiar counterpart in the context of Experimental Design.

**$I^*$ is Ignorable with GLM**  Many results given in Rubin (1987) require the sampling mechanism to be ignorable, as defined in terms of the posterior distribution of missing response values $Y_j$, where $j \in nob$. With the assumption of a General Linear Model, or GLM, the distribution of $Y_{nob}$ is fully specified and independent of $I^*$. Consequently, the posterior distribution $[Y_{nob}|X, Y_{obs}, R_{inc}, I^*]$ is independent of $I^*$. Taking $X_{inc}$ to be the design matrix and $X_{exc}$ to be a matrix of zeroes, it follows that

$$[Y_{nob}|X, Y_{obs}, R_{inc}, I^*] = [Y_{nob}|X, Y_{obs}, R_{inc}] = [Y_{mis}|X_{inc}, Y_{obs}, R_{inc}].$$  (2.9)
Therefore, the mechanism by which the \( n \) components of \( Y_{inc} \) are included by the design of the experiment is ignorable, by definition.

### 2.4 Response Mechanism

#### 2.4.1 Context of Survey Sampling

**Definition of Response Mechanism**  
Rubin (1987) refers to the conditional distribution \([R|X,Y]\) as the *response mechanism* for the survey, which is effectively a model for survey nonresponse. If data are MCAR, a subject’s propensity to respond is uniform across all subjects. In practice, the variable \( R_{ij} \) is assigned the value 1 if \( Y_{ij} \) is actually observed; otherwise, it is assigned the value 0. The constant componentwise propensity to respond may be estimated by an average of the \( n \) observed values of \( R_{ij} \). To support the assumption of MCAR data, an average stratified by known covariates must be reasonably stable across strata. If unstable, the stratified average may suggest data are MAR, or motivate specification of a nontrivial model for nonresponse. Such a model might be based upon data from previous experiments, or advice of a field expert.

**Ignorable Response Mechanism**  
If the sampling mechanism is ignorable, Rubin (1987) defines the response mechanism to be ignorable at the observed values \((X,Y_{obs},R_{inc},I^*)\) if

\[
[Y_{nob}|X,Y_{obs},R_{inc},I^*] = [Y_{nob}|X,Y_{obs}].  \tag{2.10}
\]

Supposing the sampling mechanism to be ignorable, Rubin (1987) shows that the response mechanism is ignorable if and only if the conditional distribution of \( R \) depends upon the response, \( Y \), only through the observed data, \( Y_{obs} \), written as

\[
[R|X,Y] = [R|X,Y_{obs}].  \tag{2.11}
\]

Otherwise, the response mechanism is said to be *nonignorable.*
2.4.2 Context of Experimental Design

Ignorable Response Mechanism  Regarding the designed experiment, we define the response mechanism and the ignorable response mechanism as written for the survey context but with \( X, Y, \) and \( I^* \) as defined for an experiment. Taking \( X_{inc} \) to be the design matrix, \( X_{exc} \) to be a matrix of zeroes, and the sampling mechanism to be ignorable, we define the response mechanism in the context of Experimental Design to be ignorable at \((X, Y_{obs}, R_{inc}, I^*)\) if

\[
[Y_{nob}|X, Y_{obs}, R_{inc}, I^*] = [Y_{mis}|X_{inc}, Y_{obs}, R_{inc}]
\]

Equality (2.12) holds by the aforementioned assumptions, for any randomized designed experiment. We write this definition in stages to emphasize its specific reference to the conditioning of \( Y_{mis} \) on \( R_{inc} \) for experimental data.

MAR, MCAR, and Ignorable Response Mechanisms  In the context of Experimental Design, result (2.11) reads that the response mechanism is ignorable if and only if

\[
[R_{inc}|X_{inc}, Y_{inc}] = [R_{inc}|X_{inc}, Y_{obs}].
\]

If experimental data are MCAR, then the response mechanism is ignorable. The response mechanism is also ignorable if the data are MAR.

Implications of Ignorable Missingness  In the presence of ignorable missingness, missing data may be treated as stochastically similar to the observed data, given the covariate \( X \). Consequently, observed data provides information about the parameters of the posterior distribution of \( Y_{mis} \), which can be used to generate likely values of \( Y_{mis} \) to be used as
imputed values. With the assumption of ignorable missingness, no adjustment for missingness need be made to the standard complete-data method of inference using multiply imputed data.

2.5 The Method of Multiple Imputation

General Statement of Method  The method of multiple imputation is defined independently of any specific procedure for generating imputed values. Given an imputation procedure, such as hot-deck imputation or Yamashita’s (1995) RMNI, Rubin (1987) applies that procedure \( m \) times and uses standard complete-data methods to compute \( m \) sets of parameter estimates, one corresponding to each of the \( m \) sets of imputed data. These \( m \) sets of estimates are then combined to form a single multiply-imputed estimate of the parameter of interest, as well as an estimate of the associated standard error. Rubin’s (1987) primary goal is to correctly incorporate uncertainty regarding the value of each missing datum into that estimate of standard error. Rubin’s (1987) method is made precise in the following paragraphs.

2.5.1 Complete-Data Estimator of \( Q \)

Distribution of \( Q-\hat{Q} \)  Let \( Q=Q(X, Y) \) be a \( k \)-dimensional vector of parameters of interest and let \( \hat{Q} = \hat{Q}(X, Y_{inc}, I^*) \) be the statistic that estimates \( Q \) in the absence of missingness such that

\[
Q - \hat{Q} \sim \mathcal{N}(0, U)
\]

(2.15)

where \( \mathcal{N}(0, U) \) denotes the normal distribution with \( k \times 1 \) mean vector \( 0 \) and variance-covariance matrix \( U = U(X, Y_{inc}, I^*) \). To satisfy this distributional assumption, \( Q \) may be
a transformation of the parameter of interest. Here, \( U \) is a statistic providing the \( k \times k \) variance-covariance matrix of \( (Q - \hat{Q}) \).

**Completed- and Complete-Data Statistics** Rubin (1987) refers to a data set as *complete* if each intended observation is successfully observed, that is, if \( R_{inc} \) is observed to be a vector (or matrix) of ones. In the presence of missingness, Rubin (1987) refers to a data set as *completed* (by imputation) if values are imputed for the components of \( Y_{mis} \). Suppose that under a specified Bayesian model, \( m \) sets of repeated imputations are drawn and used to construct \( m \) completed data sets. The \( \ell^{th} \) values of \( \hat{Q} \) and \( U \) are denoted by \( \hat{Q}_{*\ell} \) and \( U_{*\ell} \), respectively, for \( \ell = 1, \ldots, m \). An asterisk is used in the subscript of a statistic to emphasize that it is calculated using completed data. The \( m \) pairs of completed-data statistics are collectively called \( S_m \), so that

\[
S_m = \{(\hat{Q}_{*\ell}, U_{*\ell}) : \ell = 1, \ldots, m\}. \tag{2.16}
\]

\( \hat{Q}, U \) with Ignorable \( R, I^* \) With ignorable response and sampling mechanisms, the statistics \( \hat{Q} \) and \( U \) can be considered complete-data statistics which provide the mean and variance of \( Q \) given \((X, Y_{inc})\). That is, with ignorable mechanisms, \( \hat{Q} \) and \( U \) can be calculated using complete-data methods. In general, however, \( \hat{Q} \) and \( U \) are completed-data statistics which provide the mean and variance of variance of \( Q \) given \((X, Y_{inc}, R_{inc}, I^*)\). The distinction lies in the conditioning upon the response and sampling mechanisms. In the latter case, the analyst must condition explicitly on \( I^* \) when analyzing each completed data set, and also consider the observed value of \( R_{inc} \) in construction of the Bayes model used to generate imputed values.
**Posterior Distributions of Q**  The completed-data posterior distribution of $Q$ is given by

$$[Q|X,Y_{inc},R_{inc},I^*], \quad (2.17)$$

whereas, the complete-data posterior distribution of $Q$ is given by

$$[Q|X,Y_{inc}]. \quad (2.18)$$

In standard scientific surveys and randomized designed experiments, the sampling mechanism is ignorable. In this case, an application of Bayes Theorem shows that (2.17) becomes

$$[Q|X,Y_{inc},R_{inc}]. \quad (2.19)$$

Therefore, the distinction between the completed-data and complete-data posteriors of $Q$ becomes the conditioning of $Q$ on $R_{inc}$. The practical distinction is incorporation of a model for missingness into the generation of imputed values, and into the calculation of multiply-imputed estimates.

**Equivalence of Completed- and Complete-Data Posteriors of Q**  If both the sampling and response mechanisms are ignorable, the completed-data posterior distribution of $Q$ reduces to the complete-data posterior distribution of $Q$. Using Bayes Theorem, Rubin (1987) writes the completed-data posterior distribution of $Q$ as a product of the complete-data posterior distribution of $Q$ and an adjustment factor which is a ratio of conditional distributions of $R_{inc}$:

$$[Q|X,Y_{inc},R_{inc}] = [Q|X,Y_{inc}][R_{inc}|X,Y_{inc},Q][R_{inc}|X,Y_{inc}]. \quad (2.20)$$

The adjustment factor ratio becomes 1.0 when the response mechanism is ignorable, so that the two posteriors are equivalent.
2.5.2 Combined Estimate of Q

Let
\[ \bar{Q}_m = \frac{1}{m} \sum_{\ell=1}^{m} \bar{Q}_{\ell} \]  \hspace{1cm} (2.21)
be the average of the \( m \) complete-data estimates of \( Q \); let
\[ \bar{U}_m = \frac{1}{m} \sum_{\ell=1}^{m} \bar{U}_{\ell} \]  \hspace{1cm} (2.22)
be the average of the \( m \) complete-data variances, which is called the *within repetition variance*; and let
\[ B_m = \frac{1}{m-1} \sum_{\ell=1}^{m} (\bar{Q}_{\ell} - \bar{Q}_m)(\bar{Q}_{\ell} - \bar{Q}_m) \]  \hspace{1cm} (2.23)
be the *variance between* (among) the \( m \) complete-data estimates. The *total variance* of (\( Q - \bar{Q}_m \)) is given by
\[ T_m = \bar{U}_m + \left( 1 + \frac{1}{m} \right) B_m. \]  \hspace{1cm} (2.24)

\( B_m \) accounts for the variability in estimates of \( Q \) due to missing data. The coefficient of \( B_m \) is an adjustment for finite \( m \) proposed by Rubin & Schenker (1986).

**Posterior Mean and Variance of \( \hat{Q} \)**
Rubin (1987) proves that if \( k > 1 \), where \( k \) is the dimension of \( Q \), and if the completed-data posterior distribution of \( Q \) has mean \( \hat{Q} = \hat{Q}(X,Y_{\text{inc}},R_{\text{inc}},I^*) \) and variance-covariance matrix \( U = U(X,Y_{\text{inc}},R_{\text{inc}},I^*) \), then the posterior distribution of \( Q \) has mean given by
\[ E(Q|X,Y_{\text{obs}},R_{\text{inc}},I^*) = E(\hat{Q}|X,Y_{\text{obs}},R_{\text{inc}},I^*) \]  \hspace{1cm} (2.25)
and variance given by
\[ V(Q|X,Y_{\text{obs}},R_{\text{inc}},I^*) = E(U|X,Y_{\text{obs}},R_{\text{inc}},I^*) + V(\hat{Q}|X,Y_{\text{obs}},R_{\text{inc}},I^*). \]  \hspace{1cm} (2.26)

With ignorable sampling and response mechanisms, these equalities hold without conditioning on \( R_{\text{inc}} \) and \( I^* \).
2.5.3 Process for Multiple Imputation

Step Zero: Method of Inference, Procedure for Imputation  Before performing a multiple imputation, assumptions regarding the inferential models for the data and observed missingness must be decided upon and supported by $Y_{obs}$. Next, it must be decided whether to proceed with inference using incomplete-data methods where available, or to impute and use complete-data methods. If imputation is appropriate, the procedure to be used must be chosen with consideration of design features so that the multiple imputation is randomization-valid (Section 2.9) for the set of complete-data statistics $\{\hat{Q}, U\}$.

Step One: Construct Completed Data Sets  Generate $m$ vector-values from the posterior distribution of the missing data,
\[
[Y_{mis}|X, Y_{obs}, R_{inc}, I^*].
\] (2.27)
In the case of ignorable sampling and response mechanisms, the posterior distribution of missing data becomes
\[
[Y_{mis}|X, Y_{obs}].
\] (2.28)
Combining each synthesized observation from (2.27) with $Y_{obs}$ yields $m$ distinct completed data sets. In general, an approximation method must be used to simulate draws from (2.27), such as by Gibbs Sampler or by Data Augmentation (Schafer, 1997).

Step Two: Compute Multiply-Imputed Estimates  In the case of ignorable sampling and response mechanisms, the completed-data and complete-data posterior distributions of $Q$ are equivalent, so that it is sufficient to use the standard complete-data estimators for
inference on \( Q \). Calculate the completed-data estimate, \( \hat{Q}_{s, \ell} \), and its associated variance-covariance, \( U_{s, \ell} \), for each \( \ell = 1, \ldots, m \). Combine these \( m \) distinct estimates into a single multiply-imputed estimate of \( Q \), and its associated variance-covariance matrix.

### 2.6 \( m \)-Asymptotic Properties of \( \bar{Q}_m \) and \( T_m \)

**\( \bar{Q}_\infty, \bar{U}_\infty, \text{and } B_\infty \)** With \( m \) independent draws from \([Y_{mis}|X, Y_{obs}, R_{inc}, I^*]\), where \( m \) is arbitrarily large, \( m \) completed data sets are created, along with \( m \) values of the statistics \( \hat{Q}_{s, \ell} \) and \( U_{s, \ell} \). Taking \( m \) to be essentially infinite, the \( m \) values of \( \hat{Q}_{s, \ell} \) and \( U_{s, \ell} \), simulate aspects of \([\hat{Q}, U]|X, Y_{obs}, R_{inc}, I^*\]. In particular, the average of the \( m \) calculated values of \( \hat{Q}_{s, \ell} \) approaches the posterior mean of \( \bar{Q} \):

\[
\bar{Q}_\infty \equiv \lim_{m \to \infty} \frac{1}{m} \sum_{\ell=1}^{m} \hat{Q}_{s, \ell} = E(\hat{Q}|X, Y_{obs}, R_{inc}, I^*). \tag{2.29}
\]

The average of the \( m \) calculated values of \( U_{s, \ell} \), approaches the posterior mean of \( \bar{U} \):

\[
\bar{U}_\infty \equiv \lim_{m \to \infty} \frac{1}{m} \sum_{\ell=1}^{m} U_{s, \ell} = E(U|X, Y_{obs}, R_{inc}, I^*). \tag{2.30}
\]

The variance among the \( \hat{Q}_{s, \ell} \), approaches the posterior variance of \( \bar{Q} \):

\[
\bar{B}_\infty \equiv \lim_{m \to \infty} \frac{1}{m-1} \sum_{\ell=1}^{m} (\hat{Q}_{s, \ell} - \bar{Q}_\infty)'(\hat{Q}_{s, \ell} - \bar{Q}_\infty) = V(\hat{Q}|X, Y_{obs}, R_{inc}, I^*). \tag{2.31}
\]

**Posterior Mean and Variance of \( Q \)** Using the posterior mean and variance of \( Q = Q(X, Y) \) given by (2.25) and (2.26), Rubin (1987) proves that with sufficiently large \( m \), the statistics \( \bar{Q}_m \) and \( T_m \) are asymptotically unbiased for the posterior mean and variance covariance matrix of \( Q \), respectively. Formally,

\[
E(Q|X, Y_{obs}, R_{inc}, I^*) = \bar{Q}_\infty \tag{2.32}
\]

and

\[
V(Q|X, Y_{obs}, R_{inc}, I^*) = T_\infty \tag{2.33}
\]
where
\[ T_\infty = \bar{U}_\infty + B_\infty. \] (2.34)

**Practical Significance** In the context of Experimental Design, the size of the experimental data set, \( n \), is typically small. These asymptotic results suggest that with a sufficiently large number of imputations, \( m \), we may obtain a stable estimate of \( Q \), as well as a stable calculation of its associated standard error.

### 2.7 Confidence Regions and Significance Levels

**Confidence Interval for Scalar \( Q \)** In the case of scalar estimand \( Q \), Rubin (1987) suggests an estimate of the significance level of a confidence interval based on the student \( t \) reference distribution with \( \nu \) degrees of freedom, where
\[ \nu = (m - 1)(1 + \frac{1}{r_m})^2 \] (2.35)
and
\[ r_m = (1 + \frac{1}{m}) \frac{B_m}{U_m} \] (2.36)
is defined to be the relative increase in variance due to the uncertainty of imputation for missingness. It follows that an estimate of the 100(1-\( \alpha \))% confidence interval for \( Q \) is given by
\[ \bar{Q}_m \pm t_{\nu}^{\alpha/2} T_m^{1/2}, \] (2.37)
where \( t_{\nu}^{\alpha/2} \) is the upper \( \frac{\alpha}{2} \)th percentage point of the student \( t \) distribution on \( \nu \) degrees of freedom. With a normal posterior distribution, an estimate of the 100(1-\( \alpha \))% confidence interval for \( Q \) is given by
\[ \bar{Q}_m \pm z_{\alpha/2} T_m^{1/2}, \] (2.38)
where \( z_{\alpha/2} \) is the upper \( \frac{\alpha}{2} \)th percentage point of the standard normal distribution.
**Significance Levels for Scalar Q**  
Corresponding to (2.37), the significance level associated with the null value of Q, Q₀, is given by

\[
P(\text{F}_{1,v} > \frac{(Q₀ - \bar{Q}_m)^2}{T_m})
\]  
(2.39)

where \(\text{F}_{1,v}\) is a random variable following Snedecor’s F distribution on 1 and \(v\) degrees of freedom. With a normal posterior distribution, the significance level associated with the null value of Q, Q₀, is given by

\[
p\text{-value}(Q|X,Y_{obs},R_{inc},I^*) = P(\chi_k^2 > kD_\infty)
\]  
(2.40)

where \(\chi_k^2\) is a random variable following the chi-squared distribution on \(k\) degrees of freedom, and \(D_\infty\) is given by

\[
D_\infty = \frac{1}{k}(Q₀ - \bar{Q}_\infty)T_{m}^{-1}(Q₀ - \bar{Q}_\infty)'.
\]  
(2.41)

**F-Based Confidence Region for k-Dimensional Q**  
For \(k\)-dimensional Q, where \(k>1\), Rubin (1987) and Rubin (1991) propose estimators of the significance level of the null value, Q₀, of Q. Rubin (1987) gives that if the number of completed data sets, \(m\), is large relative to \(k\), \((m > 5k)\), this significance level is found by calculating

\[
D_m = \frac{1}{k}(Q₀ - \bar{Q}_m)T_{m}^{-1}(Q₀ - \bar{Q}_m)'
\]  
(2.42)

and letting the significance level be \(P(\text{F}_{k,v} > D_m)\) where \(v\) is given by (2.35), and \(r_m\) is generalized to be the average relative increase in variance due to missingness,

\[
r_m = (1 + \frac{1}{m})\frac{1}{k} \text{Trace}(B_m \bar{U}_m^{-1}).
\]  
(2.43)

For \(m<p\), Rubin (1987) recommends replacing \(D_m\) with \(\bar{D}_m\) where

\[
\bar{D}_m = \frac{1}{k(1+r_m)}(Q₀ - \bar{Q}_m)\bar{U}_m^{-1}(Q₀ - \bar{Q}_m)'
\]  
(2.44)
and letting the significance level be

$$P(F_{k, \frac{k+1}{2}n} > \hat{D}_m)$$  \hfill (2.45)$$

If $k=1$, then (2.45) reduces to (2.39), the significance level for scalar $Q$ given above.

**$\chi^2$-Based Confidence Region for $k$-Dimensional $Q$**

As an alternative to $D_m$ and $\tilde{D}_m$ which depend upon $k \times k$ matrices, Rubin (1987) proposes a test statistic, $\hat{D}_m$, which depends upon only scalar statistics and is asymptotically equivalent to $\tilde{D}_m$. Let the significance level of the null value, $Q_o$, of $Q$ for the $\ell^{th}$ completed data set, $\ell=1, \ldots, m$, be found as $P(\chi^2_k > d_{\ell})$ where $d_{\ell}$ is the $\ell^{th}$ repeated value of the complete-data $\chi^2$ statistic on $k$ degrees of freedom associated with the null, $Q_o$. Then

$$\hat{D}_m = \frac{\bar{d}_m - m^{-1} \frac{m-1}{m+1} r_m}{1 + r_m}$$  \hfill (2.46)$$

where

$$\bar{d}_m = \frac{1}{m} \sum_{\ell=1}^m d_{\ell}$$  \hfill (2.47)$$

is the average repeated $\chi^2$ statistic, and is referred to Snedecor’s F distribution on $k$ and $\frac{k+1}{2}n$ degrees of freedom and $r_m$ is as given in 2.43.

**P-Value based on $d_{\ell}$**

Li, Meng, Raghunathan, and Rubin (1991) describe a procedure for computing a p-value when only the completed-data test statistics, $d_{\ell}$, are available. They propose the test statistic, $\hat{D}_d$, given by

$$\hat{D}_d = \frac{\bar{d}_m k^{-1} - m^{-1} \frac{m+1}{m-1} \hat{r}_d}{1 + \hat{r}_d}$$  \hfill (2.48)$$

where

$$\hat{r}_d = (1 + \frac{1}{m}) \left[ \frac{1}{m-1} \sum_{\ell=1}^m (\sqrt{d_{\ell}} - \sqrt{\bar{d}})^2 \right]$$  \hfill (2.49)$$

where $\sqrt{\bar{d}}$ is the mean of the $m$ values of $d_{\ell}$.  

34
Significance Levels  \( \hat{D}_d \) is similar in form to the statistic, \( \hat{D}_m \) given in (1.28), with adjustments made to the calculation of \( r_m \). The performance of many such procedures for obtaining the significance level of confidence intervals calculated from multiply-imputed data is tested by simulation study (and some theoretical calculation) in Li (1985), Rubin (1987), Raghunathan (1987), and Li, Raghunathan, and Rubin (1990). Li, Meng, Raghunathan, and Rubin (1991) comment that the fraction of missing information is the most important factor influencing the performance of their procedure.

2.8 Fraction of Missing Information

Significance of the FMI  The fraction of missing information is an alternative measure of missingness, and may in some cases be preferred to the rate of missingness. We do believe that further investigation of the FMI would be interesting in the characterization of the influence of the imputed model, as well as characterization of the impact of pattern of missingness. The FMI is discussed in Chapter Eight as component to further research.

Expression for the FMI  The variance-covariance matrices, \( T_\infty \), \( B_\infty \), and \( \hat{U}_\infty \), are used to characterize the information about \( Q \) available in the experimental data, and the fraction of information lost due to missingness. Define the expected total information if \( Y_{mis} \) were also observed to be \( [E(U|X,Y_{obs},R_{inc},I^*)]^{-1} = \hat{U}_\infty^{-1} \). Uncertainty in the estimation of \( Q \) due to missingness is quantified by \( B_\infty \), so that the information actually observed is defined to be \( (\hat{U}_\infty + \hat{B}_\infty)^{-1} = \hat{T}_\infty^{-1} \). We then have the missing information quantified by \( \hat{U}_\infty^{-1} - \hat{T}_\infty^{-1} \), and the fraction of information that is missing, \( \gamma_\infty \), is written as \( \gamma_\infty = \hat{U}_\infty^{-1/2} (\hat{U}_\infty^{-1} - \hat{T}_\infty^{-1}) \hat{U}_\infty^{-1/2} \) using the symmetric form of the matrix generalization of dividing the missing information by the expected total information. Further simplification gives \( \gamma_\infty = \hat{U}_\infty^{1/2} B_\infty \hat{U}_\infty^{1/2} \).
2.9 Quality of Multiply-Imputed Inference

Introduction  Rubin (1987) defines conditions for a multiple-imputation procedure which guarantee the desired results of unbiased estimation of $Q$ along with an associated standard error that is “correct” in that it fully incorporates the analyst’s uncertainty regarding the missing data. Rubin (1987) argues requirements for valid inference from the frequentist or randomization-based repeated-imputation perspective, and relates these requirements to the perspective of inference derived from a Bayesian multiple-imputation procedure. Schafer (1997) discusses Rubin’s (1987) conditions for “correct” multiply-imputed inference in the context of a hierarchical Bayes model for the data, where imputed values are drawn from the posterior distribution of $Y_{mis}$. Rubin’s (1987) conditions for proper and randomization-valid multiply-imputed inference are generally arduous to verify. Schafer (1997) cites Binder and Sun (1996) as one of the most elaborate examples of such verification. Luckily, Schafer’s (1997) extension of Rubin’s (1987) notion of a proper procedure simplifies the task of arguing the randomization-validity of Bayesian multiply-imputed inference in the presence of ignorable sampling and response mechanisms. We evaluate multiple-imputation procedures with respect to Schafer’s (1997) concept of their being Bayesianly proper, relying upon Rubin’s (1987) imprecise, yet practical, conclusion that Bayesian procedures tend to be randomization-valid for randomization-valid complete-data estimators, $(\hat{Q}, U)$.

2.9.1 Randomization-Valid Multiply-Imputed Inference

Randomization-Valid  Define the notation $A \sim (B, \ll C)$ to mean that the distribution of $A$ tends to be centered at $B$ with each component having variability substantially less than each (positive) component of $C$. Rubin (1987) proposes the following notion of a valid
repeated-imputation inference: Suppose inferential analysis is conducted on an infinite number of completed data sets under one model for the data involving standard complete-data statistics, \( \hat{Q} = \hat{Q}(X, Y_{inc}, I^*) \) and \( U = U(X, Y_{inc}, I^*) \). In addition, suppose inference is based upon the normal reference distribution such that

\[
(Q - \bar{Q}_\infty | X, Y_{obs}, R_{inc}, I^*) \sim \mathcal{N}(0, T_\infty).
\]  

The resulting multiply-imputed inference will be valid from the perspective of having a randomization-based experiment under the posited (not necessarily actual) response mechanism if

\[
(\bar{Q}_\infty | X, Y) \sim \mathcal{N}(Q, T_0)
\]

and

\[
(\bar{T}_\infty | X, Y) \sim (T_0, << T_0)
\]

where \( Q = Q(X, Y) \) and \( T_0 = T_0(X, Y) \) are fixed by the true values of \( (X, Y) \); \( T_\infty = U_\infty + B_\infty \); and \( \bar{Q}_\infty, \bar{U}_\infty, \) and \( B_\infty \) are functions of \( (X, Y_{inc}, R_{inc}, I^*) \). Note here that the random variable in (2.51) and (2.52) is \( (R, I^*) \). Rubin (1987) defines multiply-imputed inference in the case where (2.51) and (2.52) hold to be randomization-valid.

Implications of Random-Validity

If the multiply-imputed inference on scalar \( Q \) is randomization-valid, then by (2.38)

\[
\bar{Q}_\infty \pm 1.96 T_\infty^{1/2}
\]

is a 95% confidence interval for \( Q \). In addition, the p-value given by (2.40) is uniformly distributed on \((0, 1)\). To be randomization-valid, the multiple-imputation procedure must generate imputed values in such a way that the completed-data estimate of total variability, \( T_m \), given by (2.24), is not systematically underestimated. Note that randomization-validity
is dependent upon the choice of completed-data estimators. A multiple-imputation procedure is typically not randomization-valid if it is the basic repetition of an existing frequentist single-imputation procedure. As an example of a procedure that is not randomization-valid, Rubin (1987) cites modification of hot-deck imputation where

$$E(B_{\infty}|X, Y, I^*) = \frac{n_{\text{obs}}}{N}B$$

(2.54)

where \(n_{\text{obs}}\) is the number of successfully observed univariate response values. In this case, between-imputation variability is insufficient to reflect the uncertainty in an imputed value, because each repetition imputes draws from the same population subgroup of observed responses.

**Models for Missingness (Nonresponse)** Note that the given conditions for randomization-valid inference allow any posited model for missingness, not necessarily a correct one. The central theme of multiple imputation, as Rubin explains, is the ability to produce a valid inference based upon a myriad of suggested models for nonresponse. The quality of frequentist inference with multiple imputations is exploited in the case of missingness of experimental data for which the precise model is unknown. We have that if (2.51) and (2.52) hold, our inference under any suggested response mechanism will be randomization-valid, and need not be adjusted to facilitate fair comparison of results across plausible models for missingness.

**Conditions that Guarantee Randomization-Validity** Rubin (1987) verifies more general conditions for randomization-validity. The following two conditions guarantee that (2.51) and (2.52) hold for a particular multiple-imputation procedure and complete-data
statistic $Q$. First, the complete-data inference must be randomization-valid.

$$\hat{Q}|X,Y \sim \mathcal{N}(Q,U_0) \quad (2.55)$$

and

$$U|X,Y \sim (U, \ll U_0) \quad (2.56)$$

where $Q = Q(X,Y)$ and $U_0 = U_0(X,Y)$ are fixed by the true values of $X$ and $Y$, and the underlying variable in (2.55) and (2.56) is $I^\ast$. Second, the multiple-imputation procedure must be proper for the set of complete-data statistics $\{Q, U\}$, as defined in Section 2.9.2.

### 2.9.2 Proper Multiple-Imputation Procedure

**Definition of Proper** Rubin (1987) defines a multiple-imputation procedure to be proper for the set of complete-data statistics $\{\hat{Q}, U\}$ if three conditions are satisfied:

1. Treating $(X, Y, I^\ast)$ as fixed, under the posited response mechanism, the $m=\infty$ multiple-imputation procedure provides randomization-valid inferences for the complete-data statistic $\hat{Q} = \hat{Q}(X, Y_{inc}, I^\ast)$ based on the statistics $\bar{Q}_\infty$ and $B_\infty$:

$$\bar{Q}_\infty|X,Y,I^\ast \sim \mathcal{N}(\hat{Q}, B) \quad (2.57)$$

$$B_\infty|X,Y,I^\ast \sim (B, \ll B) \quad (2.58)$$

where $B = B(X, Y_{inc}, I^\ast)$ is defined by

$$B = V(\bar{Q}_\infty|X,Y,I^\ast). \quad (2.59)$$

2. Treating $(X, Y, I^\ast)$ as fixed, under the posited response mechanism, the $m=\infty$ imputation estimate of the complete-data statistic $U = U(X, Y_{inc}, I^\ast)$, that is, $\bar{U}_\infty$, is centered at $U$ with variability of a lower order than that of $\bar{Q}_\infty$:

$$(\bar{U}_\infty|X,Y,I^\ast) \sim (U, \ll B). \quad (2.60)$$
3. Treating \((X, Y)\) as fixed, over repeated samples the variability of \(B\) is of lower order than that of \(\hat{Q}\):

\[
(B|X,Y) \sim (B_0, \ll U_0),
\]

(2.61)

where \(B_0 = B_0(X,Y)\) is defined by

\[
B_0 = E(B|X,Y),
\]

(2.62)

and \(U_0 = U_0(X, Y)\) is defined by (2.55).

The underlying variable in (2.57)-(2.60) is \(R\) with distribution specified by the response mechanism, \([R|X,Y]\), whereas the underlying random variable in (2.61) and (2.62) is \(I^*\) with distribution \([I^*|X]\).

**Bayesian Procedures Tend to be Proper** Asserting that exceptions can be found, Rubin (1987) draws the following conclusion “as a guide to practice:” If imputations are drawn to approximate repetitions from a Bayesian posterior distribution of \(Y_{mis}\) under the posited response mechanism and an appropriate model for the data, then in large samples the imputation method is proper. An “appropriate model for the data” is one such that \((\bar{Q}_\infty, \bar{U}_\infty)\), the posterior mean of \((\hat{Q}, U)\), is approximately unbiased for \((\hat{Q}, U)\) under the posited response mechanism. That is,

\[
E(\bar{Q}_\infty|X,Y,I^*) \approx \hat{Q}
\]

(2.63)

\[
E(\bar{U}_\infty|X,Y,I^*) \approx U.
\]

(2.64)

Based on this conclusion, Rubin (1987) suggests three guidelines:

1. Draw imputed values following the Bayesian paradigm as repetitions from a Bayesian posterior distribution of the missing values under the chosen models for nonresponse
and data, or an approximation to this posterior distribution that incorporates appropriate between-imputation variability.

2. Choose models of nonresponse appropriate for the posited response mechanism.

3. Choose models for the data that are appropriate for the complete-data statistics likely to be used; if the model for the data is correct, then the model is appropriate for all complete-data statistics.

**Simplified Restatement of Proper Conditions**  
Schafer (1997) restates the conditions necessary for multiply-imputed inference to be proper for the complete-data statistics \((\hat{Q}, U)\) as follows:

1. As the number of imputations becomes large, \((\hat{Q}_m - \hat{Q}) / \sqrt{B_m}\) should become approximately standard normal over the distribution of the response indicators \(R\) with \(Y\) held fixed.

2. As the number of imputations becomes large, \(\bar{U}_m\) should be a consistent estimate of \(U\), with \(R\) regarded as random and \(Y\) regarded as fixed.

3. The true between-imputation variance (i.e. the variance of \(\hat{Q}_m\) over an infinite number of multiple imputations) should be stable over repeated samples of the complete data \(Y\), with variability of a lower order than that of \(\hat{Q}\).
2.9.3 Bayesianly Proper

Bayesian vs. Frequentist Multiply-Imputed Inference  Schafer (1997) argues that it is not necessary to thoroughly understand the meaning of Rubin’s (1987) proper multiple-imputation procedure if we use a procedure derived from the principles of Bayesian inference. Schafer (1997) makes two assumptions which allow simplification of conditions for valid inference: First, Schafer (1997) assumes that valid inferential statements can be obtained through summaries of a likelihood function or posterior distribution arising from a parametric model. Second, Schafer (1997) assumes the response mechanism is ignorable, so that its specification is not necessary for data analysis.

Bayesianly Proper  Schafer (1997) defines multiple imputations to be Bayesianly proper if they are independent realizations of \( Y_{mis|obs} \), the posterior predictive distribution of the missing data under some complete-data model, parameterized by some \( \theta \), and some prior distribution for \( \theta \). Writing this posterior distribution as

\[
[Y_{mis|obs}] = \int P(Y_{mis|obs}, \theta)P(\theta|Y_{obs}) \, d\theta,
\]

Schafer asserts that Bayesianly proper multiple imputations reflect uncertainty about \( Y_{mis} \) given the parameters of the complete-data model, as well as uncertainty about the unknown model parameters. Again, it is assumed that the sampling and response mechanisms are ignorable in this definition. Schafer’s concept of proper is applicable throughout this dissertation.
CHAPTER 3

Discussion of the Two-Way Layout

3.1 Introduction

Strategy To demonstrate the method of multiple imputation within the context of Experimental Design, we present a simulation study wherein we tested Yamashita’s (1995) RMNI procedure on a two-way mixed-effects repeated-measures interaction model, in the absence of known covariates of the response. We follow the multiple-imputation process of Section 2.5.3 to obtain multiply-imputed estimates of fixed effects and associated standard errors. With a need for initial parameter estimates per Step One of the process, we present various incomplete-data estimators of both fixed effects and variance components and assess them in terms of their relative contributions to the standard error of multiply-imputed estimates. In particular, we discuss estimation of variance components for the said model using unbalanced data. Ultimately, we assess the quality of multiply-imputed estimates as gauged through their comparison with complete-data estimates, to be discussed in Chapter Four and demonstrated in Chapter Five.

Chapter Contents Within this chapter, we provide details of the two-way interaction model construct we chose for the demonstration of RMNI. For alternative constructs of
the model, see Hocking (1973). Additionally, we explore options for handling of unit-
missingness, calculation of incomplete data estimators for treatment effects and variance 
components, and calculation and interpretation of expected mean squares for our construct 
of the model. The simulated results included in this chapter summarize properties of our 
choices for estimators of treatment effects and variance components requisite of the RMNI 
procedure. The simulated results for the RMNI procedure itself are presented separately in 
Chapter Five.

**Notation** Throughout this chapter, we refer to the response variable as \( Y \) or \( y \), and to 
the design matrix as \( X \). This is consistent with the literature of Experimental Design, and 
preferred in the absence of direct reference to Rubin (1987). For interpretation of Rubin’s 
(1987) results in the context of Experimental Design, we defer to the notation of Section 
2.2.2. We denote the transpose of a vector (matrix) by inserting a prime as a superscript 
after the name or bracketed specification of that vector (matrix).

### 3.2 The Two-Way Mixed-Effects Interaction Model

**Introduction** In Section 3.2.1, we outline a generic designed experiment with which the 
two-way mixed-effects repeated-measures interaction model corresponds by construction; 
and hereafter, we refer to this design as Design 3.2.1. Our intention is to study multiple im-
putation in the context of data that truly follow the said model, which is made explicit with 
Equation (3.4). We assume the observed incomplete data to be MCAR with a pattern of 
missingness that does not prohibit point estimation of any parameter of the said model us-
ing incomplete data. In particular, we force the pattern of missingness to allow calculation 
of an initial estimate of the interaction variance. Our particular focus is the predicament 
of requiring a balanced data set for inference, yet observing an unbalanced data set due to
missingness. Consequently, we present the unbalanced two-way mixed-effects repeated-measures interaction model in relation to the reference balanced model which is assumed to be the accurate complete-data model associated with Design 3.2.1.

3.2.1 Complete-Data Model Specification

**Design Perspective**  We consider data which is collected per specification of a randomized experimental design, where each datum is the measurement of a scalar response variable, y. Our perspective is the experiment which entails the administration of each of $n_t$ treatments to each of $n_s$ subjects with $n_r$ repetitions of each treatment/subject combination. Let $n_s$ and $n_r$ be integers that are constant with respect to treatment and treatment/subject combination, respectively. The size of the experiment is then given by $n$, where $n = n_t n_s n_r$.

For simulation of incomplete MCAR data under this design, we assume that each treatment $t$ is at least partially observed, and that any subject nowhere observed is dropped per practice of subject deletion (Section 1.2). Therefore, to characterize the incomplete data set observed under this design, we modify the complete-data dimensions $n_s$ and $n_r$ only, as described below. Hereafter, we refer to this perspective as Design 3.2.1.

**Number of Subjects per Treatment**  With respect to a data set observed under Design 3.2.1, we reserve $n_s$ to represent the constant complete-data upper bound on the subject index $s$. Let $n_s(t)$ denote the number of subjects with whom an individual treatment effect $\theta_t$ is observed. With incomplete data, $n_s(t)$ may vary with $t$, and $n_s(t) \leq n_s$ for each $t \in \{1, \ldots, n_t\}$. We assume data is MCAR and treat $n_s(t)$ as the incomplete-data upper bound on the subject index $s$. In the case of unit missingness, we denote the number of subjects somewhere observed by $n_s^*$, where $n_s^* < n_s$ and $\max_{t \in \{1, \ldots, n_t\}} \{n_s(t)\} \leq n_s^*$.
Number of Replicates per Combination  With respect to a data set observed under Design 3.2.1, we reserve $n_r$ to represent the constant complete-data upper bound on the replicate index $r$. To distinguish cell sizes of an unbalanced incomplete data set, we define $n_r(t,s)$ to represent the observed number of replicates under the combination of treatment $t$ with subject $s$; for each combination $(t,s)$, where $t \in \{1, \ldots, n_t\}$ and $s \in \{1, \ldots, n_s(t)\}$, we have that $n_r(t,s) \leq n_r$. With MCAR data, we treat $n_r(t,s)$ as the incomplete-data upper bound of the replicate index $r$ for the combination $(t,s)$.

Specification of Effects  The effects included with the two-way mixed effects repeated-measures interaction model are defined for treatments, subjects, and interactions. Let $\theta$ denote the $n_t \times 1$ vector of fixed treatment effects $\theta_t$, where $t = 1, \ldots, n_t$. Let $u$ be the $n_s \times 1$ vector of random subject effects $u_s$, where $s = 1, \ldots, n_s$. Let $c$ be the $n_t n_s \times 1$ vector of random interaction effects $c_{ts}$, where $t$ identifies treatment and $s$ identifies subject for all pairs $(t,s)$. Finally, let $\varepsilon$ be the $n \times 1$ vector of random measurement errors $\varepsilon_{tsr}$, where $r = 1, \ldots, n_r$.

Distributions of Random Effects and the Error Term  We assume that each random effect and random error component is distributed as a Normal random variate with mean zero and unknown variance. In addition, we assume random effects to be independent of each other, as well as of each component of $\varepsilon$. We denote the variance of each subject effect $u_s$ by $\sigma_u^2$; we denote the variance of each interaction effect $c_{ts}$ by $\sigma_c^2$; and we denote the variance of each random error component $\varepsilon_{tsr}$ by $\sigma_{\varepsilon}^2$. The distributions of these random effects and error components are summarized by

$$u_s \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2),$$  \hspace{1cm} (3.1)
\begin{align*}
    c_{ts} & \overset{iid}{\sim} \mathcal{N}(0, \sigma_c^2), \text{ and} \\
    \varepsilon_{tsr} & \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2),
\end{align*}

which hold for all combinations of \( t, s, \) and \( r \).

**Specification of Our Model**  Let \( y_{tsr} \) represent the response of the \( s^{th} \) subject to the \( t^{th} \) treatment on the \( r^{th} \) repetition of their combination. The two-way mixed-effects repeated-measures interaction model for \( y_{tsr} \) is the linear combination of fixed and random effects with random error given by

\[
y_{tsr} = \theta_t + u_s + c_{ts} + \varepsilon_{tsr}
\]

for each triplet \((t,s,r)\). Where model (3.4) is referenced as the *complete-data model*, we mean that \((t,s,r) \in \{1,\ldots,n_t\} \times \{1,\ldots,n_s\} \times \{1,\ldots,n_r\} \). Where model (3.4) is referenced as the *incomplete-data model*, we mean that \( t \in \{1,\ldots,n_t\} \); \( s \in \{1,\ldots,n_s(t)\} \) for each treatment \( t \); and \( r \in \{1,\ldots,n_r(t,s)\} \) for each combination \((t,s)\).

**Distribution of the Response Variable**  The distribution of an individual response measurement under model (3.4) is given by

\[
y_{tsr} \sim \mathcal{N}(\theta_t, \sigma_y^2)
\]

for each triplet \((t,s,r)\), where

\[
    \sigma_y^2 = \sigma_u^2 + \sigma_c^2 + \sigma_e^2.
\]

The summands of (3.6) are collectively termed *variance components*; and we use \( \sigma^2 \) to denote the vector of individual variance components, so that \( \sigma^2 = [\sigma_u^2, \sigma_c^2, \sigma_e^2]' \).
3.2.2 Incomplete-Data Model Specification

Specification of the Incomplete-Data Model  In general, we treat an incomplete data set as if it is fully observed under an alternative incomplete-data design, and as following a corresponding incomplete-data model. Here, the incomplete-data model is a modification of the complete-data model such that features of the incomplete data set, especially observed cell sizes, are accommodated. The incomplete-data model may be a reduced form of the complete-data model if, for instance, a parameter of the complete-data model is rendered inestimable by missingness. For example, under Model (3.4), the incomplete-data estimator of $\sigma_c^2$ (Section 3.5) may be unstable or incalculable due to several observed cell sizes of one. In this case, the interaction effect for those cells may be excluded from the incomplete-data model, while included in the complete-data model.

Our Incomplete-Data Model  We treat model (3.4) as the reference model in construction of an incomplete-data model. To maintain focus on model (3.4), we assume all fixed effects and variance components to be estimable using observed data despite missingness. Consequently, we do not consider reduced forms of model (3.4) as incomplete-data models in this simulation study. Our incomplete-data model differs from (3.4) only in the specifications for subscripts. Specifically, the incomplete-data model may or may not be unbalanced, whereas model (3.4) is balanced.

The Balanced Incomplete Data Set  We emphasize that balanced and complete are distinct characterizations of an observed data set. An incomplete data set is balanced in the
case of strict unit-missingness (Section 1.1), where observed data is nonetheless rectangular. In the presence of item-missingness (Section 1.1), incomplete data is generally unbalanced, but may be balanced. Under Design 3.2.1 with MCAR data, for example, if any but only one replicate is unobserved with each treatment/subject combination, the incomplete data set is effectively balanced. Despite anecdotal counter-examples, we generally assume an incomplete data set to be unbalanced.

**Inference with Incomplete Data** Hypothetically, incomplete data may suffice in terms of the necessary contribution of inference to meeting the goals of an experiment. For example, if the goal is to produce summary statistics, and if conservative confidence intervals for the treatment effects will suffice, then imputation may not be necessary. Incomplete data methods do exist for these purposes, though not for every model. However, if meeting the goal requires multiple comparisons of fixed treatment effects, (multiple) imputation may be necessary to facilitate the prescribed Multiple Comparisons procedure. For example, no valid MCP exists for unbalanced incomplete data following the two-way mixed interaction model (Voss and Hsu, 1998). With unbalanced incomplete data, multiple imputation is adopted to facilitate valid inference. We still require incomplete-data point estimators. However, we do not base inference on these, but use them to compute initial estimates of the parameters of the inferential model per Step One (Section 2.5.3) of RMNI. Once incomplete-data estimators are employed in the process of generating a complete data set, we use only balanced-data estimators to compute multiply-imputed estimates. For Model (3.4), we consider incomplete-data point estimators of \( \theta \) in Section 3.3; and we discuss methods for estimating \( \sigma^2 \) in Sections 3.4 and 3.5.
**Specification of Observed Subjects and Replicates**  With respect to model (3.4), given an incomplete data set and a nonignorable response mechanism (Section 2.4.2), the distributional properties of an incomplete-data estimator of \( \{\theta, \sigma^2\} \) may depend upon the observed pattern of missingness. In this case, the calculating formula for the estimator must be robust to the indices of the observed responses over which it is calculated. Therefore, we define \( n_t \) sets \( \mathcal{S}_t \) such that for each \( t \), \( \mathcal{S}_t \) is the set of subjects somewhere observed under treatment \( t \); and we define \( \mathcal{S} = \bigcup_{t=1}^{n_t} \mathcal{S}_t \) to be the set of all observed subjects. In addition, we define the sets \( \mathcal{C}, \mathcal{C}_t \) for each \( t \), and \( \mathcal{C}_s \) for each \( s \in \mathcal{S} \), as follows:

\[
\mathcal{C} = \{ (t,s) : \text{at least one measurement is observed} \},
\]

\[
\mathcal{C}_t = \{ (t,s) : \text{at least one measurement is observed with } \theta \text{ fixed} \},
\]

\[
\mathcal{C}_s = \{ (t,s) : \text{at least one measurement is observed with } s \text{ fixed} \}.
\]

For each \( t \), \( \mathcal{C}_t \) is the set of all cells holding at least one observation under treatment \( t \); for each \( s \in \mathcal{S} \), \( \mathcal{C}_s \) is the set of all cells holding at least one observation under subject \( s \); and \( \mathcal{C} = \bigcup_{t=1}^{n_t} \mathcal{C}_t = \bigcup_{s=1}^{n_s} \mathcal{C}_s \) is the set of all nonempty cells. If the response mechanism is ignorable (Section 2.4.2), this construction is unnecessary for the presentation of incomplete-data estimators and their properties.

### 3.3 Estimation of Treatment Effects

**Synopsis**  In the context of Model (3.4) under Design 3.2.1 with MCAR data, we present the Least Squares and Weighted Means estimators of the fixed treatment effect \( \theta \) in Sections 3.3.1 and 3.3.2, respectively. Both estimators are unbiased for \( \theta \). In Section 3.3.3, we compare these estimators in terms of their variance-covariance matrices. Note that we do not consider methods of maximum likelihood estimation because we intend to develop a method which is free of manual assessment of convergence.
Notation for Sums and Averages  To denote an average over a specific variable subscript, we substitute a dot for that subscript. To denote a sum over a specific variable subscript, we substitute a “+” for that subscript. In the case that the average (or sum) is taken over an index given as an argument, such as the argument \( t \) of \( n_s(t) \), we substitute a dot (or “+”) for that argument. Throughout this chapter, data are assumed to be MCAR so that estimators and their properties are presented without explicit incorporation of the observed pattern of missingness, as discussed in the previous section (3.2.2). Consequently, summands may be indexed by a sequence of Natural numbers to simplify representation of calculating formulas.

3.3.1 LSE for a Fixed Treatment Effect

Specification of \( \hat{\theta} \)  In the context of Model (3.4) under Design 3.2.1 with MCAR data, we present the Least Squares Estimator (LSE) \( \hat{\theta}_t \) of an individual treatment effect \( \theta_t \), for \( t = 1, \ldots, n_r \); and we denote the vector of \( n_r \) treatment effects by \( \hat{\theta} = [\hat{\theta}_1, \ldots, \hat{\theta}_{n_r}]' \). The general form of \( \hat{\theta}_t \) is given first by (3.11), and is simplified for the case of balanced data as given by (3.15). The number of observations obtained under treatment \( t \) is given by

\[
n_r(t+) = \sum_{s=1}^{n_s(t)} n_r(t,s)
\]

(3.10)

for each \( t \). We assume the practice of subject deletion (Section 1.2), so that each summand of (3.10) is nonzero. The LSE for an individual treatment effect, \( \theta_t \), is given by

\[
\hat{\theta}_t = \frac{\sum_{s=1}^{n_s(t)} \sum_{r=1}^{n_r(t,s)} y_{tsr}}{n_r(t+)}
\]

(3.11)

for each treatment \( t \).

Distribution of \( \hat{\theta} \)  Standard distribution theory verifies that

\[
\hat{\theta} \sim \mathcal{N}(\theta, \text{V}(\hat{\theta}))
\]

(3.12)
where $V(\hat{\theta})$ is the $n_t \times n_t$ variance-covariance matrix of $\hat{\theta}$. The $t^\text{th}$ diagonal element of $V(\hat{\theta})$ is the variance of $\hat{\theta}_t$, which is given for each $t$ by

$$V(\hat{\theta}_t) = \frac{\sum_{s=1}^{n_r(t)} n_s^2(t,s)}{n_r^2(t+)} (\sigma_u^2 + \sigma_e^2) + \frac{1}{n_r(t+)} \sigma_e^2; \quad (3.13)$$

the off-diagonal element of $V(\hat{\theta})$ for row $t$ and column $t'$ is given by

$$\text{COV}(\hat{\theta}_t, \hat{\theta}_{t'}) = \frac{\sum_{s=1}^{n_r(t)} n_r(t,s)n_r(t',s)}{n_r(t+)n_r(t'+)} \sigma_u^2 \quad (3.14)$$

for $t \neq t'$.

**\hat{\theta} with Balanced Data**  Assume $n_r(t,s) = n_r^*$ is the constant number of observed replicates under each treatment/subject combination, and $n_s(t) = n_s^*$ is the constant number of subjects observed under each treatment $t$. We assume that $n_r^*$ and $n_s^*$ are nonzero. In this case, the LSE $\hat{\theta}_t$ reduces from the form given by (3.11) to

$$\hat{\theta}_t = \frac{\sum_{s=1}^{n_r^*} \sum_{r=1}^{n_r^*} \hat{Y}_{t,sr}}{n_r^* n_s^*} \quad (3.15)$$

for $t = 1, \ldots, n_t$. In conjunction with (3.15), the variance-covariance matrix of $\hat{\theta}$, $V(\hat{\theta})$, reduces from the form given by (3.13) and (3.14) to

$$V(\hat{\theta}) = \frac{1}{n_r^* n_s^*} [(n_r^* \sigma_e^2 + \sigma_e^2)I + n_r^* \sigma_u^2 J]. \quad (3.16)$$

Within (3.16), $I$ represents the $n_t \times n_t$ identity matrix, and $J$ represents the $n_t \times n_t$ matrix of ones. Formulas (3.15) and (3.16) apply in the case of an observed balanced data set, whether complete or incomplete.

### 3.3.2 Weighted Means Estimator of $\theta$

**Specification of $\tilde{\theta}$**  In the context of model (3.4) with MCAR data, we present the Weighted Means Estimator of an individual treatment effect $\theta_t$, denoted $\tilde{\theta}_t$ for $t = 1, \ldots, n_t$, as an
alternative to the LSE $\hat{\theta}_t$ given by (3.11). We denote the vector of $n_t$ weighted means estimators by $\tilde{\theta}$, where $\tilde{\theta} = [\tilde{\theta}_1, \ldots, \tilde{\theta}_{n_t}]'$. The estimator $\tilde{\theta}_t$ equally weights each combination of a subject with treatment $t$, and is well-described as the average of nonempty-cell averages under treatment $t$. For each observed combination $(t,s)$, where $t \in \{1, \ldots, n_t\}$ and $s \in \mathcal{S}_t$, the observed cell average is given by

$$y_{ts} = \frac{\sum_{r=1}^{n_r(t,s)} y_{tsr}}{n_r(t,s)}.$$  \hfill (3.17)

The weighted means estimator is given by

$$\tilde{\theta}_t = \frac{\sum_{s=1}^{n_s(t)} y_{ts}}{n_s(t)},$$  \hfill (3.18)

for each treatment $t$. The practice of subject deletion (Section 1.2) guarantees that $n_r(t,s)$ is nonzero for each combination $(t,s)$ included in the calculation of $\tilde{\theta}$. The removal of subjects unobserved under treatment $t$ is reflected in the count $n_s(t)$, which we assume to be nonzero for each $t$.

**Distribution of $\tilde{\theta}$** Standard distribution theory verifies

$$\tilde{\theta} \sim \mathcal{N}(\theta, V(\tilde{\theta})),$$  \hfill (3.19)

where $V(\tilde{\theta})$ is the $n_t \times n_t$ variance-covariance matrix of $\tilde{\theta}$. For each treatment effect $\theta_t$, the variance of the estimator $\tilde{\theta}_t$ is denoted $V(\tilde{\theta}_t)$ and is given by

$$V(\tilde{\theta}_t) = \frac{1}{n_s(t)} (\sigma_u^2 + \sigma_e^2) + \frac{\sigma_e^2}{n_s(t)} \sum_{s=1}^{n_s(t)} \frac{1}{n_s(t)}.$$  \hfill (3.20)

For each pair of treatments $(t, t')$, where $t \neq t'$, the covariance of $\theta_t$ and $\theta_{t'}$ is denoted $\text{COV}(\tilde{\theta}_t, \tilde{\theta}_{t'})$ and is given by

$$\text{COV}(\tilde{\theta}_t, \tilde{\theta}_{t'}) = \frac{1}{n_s(t)} \sigma_u^2.$$  \hfill (3.21)
The \( t^{th} \) diagonal element of \( V(\tilde{\theta}) \) is specified by (3.20), and the off-diagonal element of \( V(\tilde{\theta}) \) for row \( t \) and column \( t', t \neq t' \), is specified by (3.21).

**\( \hat{\theta} \) with Equal Cell Sizes** Assume \( n_r(t,s) = n_r^* \) is the constant number of observed replicates under each treatment/subject combination, and \( n_s(t) = n_s^* \) is the constant number of subjects observed under each treatment \( t \). The form of the weighted means estimator, \( \tilde{\theta}_t \), reduces from formula (3.18) to

\[
\tilde{\theta}_t = \frac{\sum_{s=1}^{n_s^*} y_{ts}}{n_s^*},
\]  
(3.22)

for each treatment \( t \). In conjunction with (3.22), the variance-covariance matrix of \( \tilde{\theta} \), \( V(\tilde{\theta}) \), reduces in form from the representation given by (3.20) and (3.21) to

\[
V(\tilde{\theta}) = \frac{1}{n_s^* n_r^*} [(n_r^* \sigma_c^2 + \sigma_c^2) I + n_r^* \sigma_u^2 J],
\]  
(3.23)

where matrices \( I \) and \( J \) are as defined for (3.16).

### 3.3.3 Comparison of \( \tilde{\theta} \) and \( \hat{\theta} \)

**Basis for Comparison** Both the Least Squares (3.11) and Weighted Means (3.18) estimators are unbiased for \( \theta \). Given two unbiased estimators, we’d like to establish the conditions under which one estimator has the lower variance. Therefore, we compare the LSE \( \hat{\theta} \) with the alternative \( \tilde{\theta} \) in terms of their variance-covariance matrices. We calculate incomplete-data estimates of \( \theta \) as the posterior mean for missing data \( Y_{mis} \). To minimize the standard error of multiply-imputed estimates, we choose the estimator of \( \theta \) with the lowest variability. In the case of balanced data, \( V(\hat{\theta}) = V(\tilde{\theta}) \), as given by (3.16) and (3.21). Consequently, we focus our comparison on the case of unbalanced incomplete data. To compare \( V(\hat{\theta}) \) with \( V(\tilde{\theta}) \), we consider the minimum of their traces, as well as the minimum of the maximums of their diagonal components, as described below.
Results of Comparison  We are unable to prove a general result for this comparison; such a result necessarily depends upon the observed pattern of missing data. Both $V(\hat{\theta})$ and $V(\tilde{\theta})$ are functions of $\sigma^2$ and the observed cell sizes. Determination of the minimum variance incomplete-data estimator of $\theta$ is therefore dependent upon the observed pattern of missingness. Given incomplete data, we may calculate $V(\hat{\theta})$ and $V(\tilde{\theta})$, up to $\sigma^2$, and select an estimator of $\theta$ based upon the preferred criterion for minimum variance.

Expression for Minimax Estimator $\theta^*$  We denote the $t^{th}$ diagonal component of $V(\hat{\theta})$ by $\hat{v}_{tt}$, and the $t^{th}$ diagonal component of $V(\tilde{\theta})$ by $\tilde{v}_{tt}$, for each $t$. Let $\hat{v}_{\text{max}}$ denote the maximum diagonal value of $V(\hat{\theta})$, where

$$\hat{v}_{\text{max}} = \max(\hat{v}_{11}, \ldots, \hat{v}_{nn}).$$

(3.24)

Similarly, let $\tilde{v}_{\text{max}}$ denote the maximum diagonal value of $V(\tilde{\theta})$, where

$$\tilde{v}_{\text{max}} = \max_t(\tilde{v}_{11}, \ldots, \tilde{v}_{nn}).$$

(3.25)

The minimax estimator of $\theta$, which we denote $\theta^*$, is given by

$$\theta^* = \begin{cases} \hat{\theta} & : \hat{v}_{\text{max}} \leq \tilde{v}_{\text{max}} \\ \tilde{\theta} & : \hat{v}_{\text{max}} > \tilde{v}_{\text{max}} \end{cases}$$

(3.26)

where we arbitrarily favor the LSE in the case that $\hat{v}_{\text{max}} = \tilde{v}_{\text{max}}$.

Evaluating the Minimax Estimator  The functional forms of $\hat{v}_{\text{max}}$ and $\tilde{v}_{\text{max}}$ are given by (3.13) and (3.20), respectively. Comparison of $\hat{v}_{\text{max}}$ with $\tilde{v}_{\text{max}}$ is algebraically similar to the simplest-case comparison of $\hat{v}_{tt}$ with $\tilde{v}_{tt}$, for an arbitrary treatment $t$. Assume $\hat{v}_{\text{max}} = \tilde{v}_{\text{tt}}$ and that $\tilde{v}_{\text{max}} = \tilde{v}_{tt}$ and consider the comparison of the difference $(\hat{v}_{\text{max}} - \tilde{v}_{\text{max}})$ with zero. The difference is given by

$$\left[ \sum_{s=1}^{n_{(t)}} \frac{n_s^2(t,s)}{n_r^2(t+)} - \frac{1}{n_r(t)} \right] (\sigma_u^2 + \sigma_c^2) + \left[ \frac{1}{n_r(t)} - \frac{1}{n_r^2(t)} \sum_{s=1}^{n_r(t)} \frac{1}{n_s(t)} \right] \sigma_e^2;$$

(3.27)
for a fixed $t$. The coefficient of $\sigma^2_v$ is negative given unbalanced data, and with an application of Hölder’s inequality, we verify that the coefficient of $(\sigma^2_u + \sigma^2_v)$ is strictly positive. Therefore, comparison of $(\hat{v}_{max} - \tilde{v}_{max})$ with zero is inconclusive for unknown $\sigma^2$. In this way, we are prevented from any generalization of a minimax estimator such as $\theta^*$ across patterns of missingness. Given data, however, the calculation of $\theta^*$ is straightforward, using an incomplete-data estimate of $\sigma^2$.

**Expression for the Minimal Trace Estimator of $\theta$** In addition to $\theta^*$, we consider $\theta^*_{trace}$, which we define to be the estimator determined by finding the minimum trace with respect to the variance-covariance matrices of $\hat{\theta}$ and $\tilde{\theta}$. Let $\hat{t}$ be the trace of $V(\hat{\theta})$, and let $\tilde{t}$ be the trace of $V(\tilde{\theta})$. We define $\theta^*_{trace}$ by

$$\theta^*_{trace} = \begin{cases} 
\hat{\theta} : \hat{t} \leq \tilde{t} \\
\tilde{\theta} : \hat{t} > \tilde{t}
\end{cases}$$

(3.28)

where we arbitrarily favor the LSE in the case that $\hat{t} = \tilde{t}$.

**Evaluating the Minimal Trace Estimator** To determine $\theta^*_{trace}$, we recommend direct calculation of $\hat{t}$ and $\tilde{t}$ using observed data and an incomplete-data estimator of $\sigma^2$. As with $\theta^*$ given by (3.26), we derive no general result for the determination of $\theta^*_{trace}$ with unknown $\sigma^2$ across pattern of missingness. A general result is prohibited by algebraic constraints similar to those illuminated by (3.27).

**Rule of Thumb** For an unbiased estimator of $\theta$, we consider formulation of a statistic which might suggest a rule of thumb for preference of $\hat{\theta}$ or $\tilde{\theta}$. For each treatment $t$, let $\overline{\sigma}(t)$ be the average observed cell size excluding zeros, and let $\sigma^2_{\overline{\sigma}(ts)}$ be the sample variance of observed cell sizes under treatment $t$. Finally, let $g(\sigma^2)$ be a function of the variance
components given by
\[ g(\sigma^2) = \frac{\sigma_u^2 + \sigma_c^2}{\sigma_e^2}. \] (3.29)

We found that \( \hat{v}_{tt} < \tilde{v}_{tt} \) wherever
\[
g(\sigma^2) \geq \frac{n_r(t)}{(n_s(t) - 1) \sigma_n^2} \left[ n_r(t) \sum_{s=1}^{n_r(t)} \frac{1}{n_r(t,s)} - n_s \right]
\] (3.30)
where the right hand side of (3.30) is strictly positive. The right hand side of (3.30) may be calculated for any pattern of missingness prior to observing the data. Unfortunately, calculation of an estimate of \( g(\sigma^2) \), or any nonlinear function of \( \sigma^2 \), is prohibitive with unbalanced data. While an unbiased estimate \( \hat{\sigma}^2 \) of \( \sigma^2 \) may be obtained by the Method of Fitting Constants (Section 3.5), the components of \( \hat{\sigma}^2 \) are not independent, and \( \hat{\sigma}^2 \) is not an MLE for \( \sigma^2 \). The ratio \( g(\sigma^2) \) may be estimated by \( \hat{g}(\sigma^2) \) given by
\[ g(\sigma^2) \equiv \frac{\hat{\sigma}_u^2 + \hat{\sigma}_c^2}{\hat{\sigma}_e^2}. \] (3.31)

The properties of this estimator remain unclear and are perhaps indeterminable. Consequently, the usefulness of (3.30) in practice remains dubious.

**Conclusion of Comparison** We choose to demonstrate RMNI using the Least Squares Estimator. The exploration of comparison of the LSE with the Weighted Means Estimator proved to be an interesting window onto apparent connections between the variance-covariance matrix of the treatment estimate and the pattern of missingness. We support stochastic investigation of the Fraction of Missing Information and its relationship to pattern of missing data as well as rate of missingness. Exploration of the FMI is discussed in Chapter Eight as a topic for further research.
3.4 Balanced-Data Estimation of Variance Components

Synopsis An incomplete-data estimate of $\sigma^2$ is necessary for Step One of a multiple imputation (Section 2.5.3). In the context of Model (3.4) under Design 3.2.1 with MCAR data, we present methods for the estimation of $\sigma^2$. Procedures which generate unbiased estimates of $\sigma^2$ exist in the literature for the cases of both balanced- and unbalanced data. With balanced data, unbiased estimation of $\sigma^2$ is straightforward. For unbalanced mixed models in general, $\sigma^2$-estimation poses theoretical complications which vary with model assumptions. With Model (3.4), we adopt the Fitting Constants Method, or Henderson’s Method 3, as recommended by Searle (1971). Since the Fitting Constants Method is computing-intensive, we present the method of Experimentwise Subject Deletion as an alternative. Section 3.4 proposes options for estimating $\sigma^2$ using observed data, complete and incomplete.

Notation The variance components are defined in Section 3.2.1 with (3.1)-(3.3) and (3.6), and correspond to random subject effect ($\sigma^2_u$), random interaction effect ($\sigma^2_c$), and random error component ($\sigma^2_e$). We refer to the variance components collectively as the vector $\sigma^2$ defined with (3.6).

Assumptions Given the design and model assumptions of Section 3.2.1, with the intent of conducting inference on parameters of Model (3.4), we assume that data is observed to be MCAR. We also assume that the observed pattern of missingness does not prohibit estimation of any variance component, where the estimability of $\sigma^2_c$ is of particular concern. We also note that the model assumptions of Section 3.2.1 match the model construction used by Box and Tiao (1973) to derive closed form expressions for approximations of the
posterior distributions of the variance components, and ultimately the missing data, from the two-way mixed interaction model.

**Prior and Posterior Distributions of \( \sigma^2 \)** By assumption, \( \sigma^2 \) is unknown yet estimable with the observed data set. We assume the analyst has no information regarding \( \sigma^2 \) other than the observed data set. Given the observed data set, we utilize the prior distribution as discussed by Box and Tiao (1973), which is constructed by model assumptions (as with Design 3.2.1) for the response. The posterior distribution of each variance component is constructed by Box and Tiao (1973). These posterior distributions are functions of the estimates of variance components which are calculated using observed data. In sum, given data we have incomplete-data estimates of each variance component, and a fully prescribed posterior distribution for each variance component. (Incidentally, the prior and posterior distributions for the treatment effect are obtained from the same reference point; the incomplete-data estimate of the treatment effect and its posterior are used jointly with the same for the variance components to arrive at a posterior distribution for the missing data.) For distributional properties of these posteriors, we refer you to Section 5.3 of Box and Tiao (1973).

### 3.4.1 Balanced-Data Estimators of \( \sigma^2 \)

**Introduction** We present the Analysis of Variance method (Scheffe, 1959) for estimating \( \sigma^2 \) under Model (3.4) with balanced data; and we explore the method’s applicability in the case of incomplete data. In particular, we use this method to compute a baseline complete-data estimate of \( \sigma^2 \) for comparison with incomplete-data estimates which vary with simulated pattern of missingness. Balance is required for this computation method, and so we must consider when this method may be useful in the handling of incomplete
data, which is generally unbalanced. Briefly, we note here that balanced-data estimation of variance component estimates enables closed-form expression and clear interpretation of the Expected Mean Squares; unbalanced-data estimation does not.

**Instances of Balance with RMNI and Missing Data** Under Design 3.2.1, a complete data set is necessarily balanced. Therefore, we use the balanced-data estimators to compute complete-data estimates of variance components as a baseline comparison. With data that is MCAR, an incomplete data set is typically unbalanced in the presence of item-missingness (Section 1.1), but may be balanced, as explained in Section 3.2.2. If the observed data is MCAR and effectively balanced, our incomplete-data model is Model (3.4) with diminished size $n^*$, where $n^* = n_s n_r$ with $n_s$ and $n_r$ as defined for estimator (3.15). Incomplete data may be MCAR and balanced by either luck or strategic re-arrangement of data so that the pattern of missingness allows balanced-data estimation of variance components. (Re-arrangement of data to establish a desired pattern of missingness is discussed by Little and Rubin (1987), though not for the case of experimental data under Design 3.2.1.) If the observed data is MCAR and unbalanced, we consider the method of Experimentwise Subject Deletion (Section 3.4.2) which enables use of a balanced-data estimator for $\sigma^2$ with the unbalanced data set. In short, the data set with which an initial estimate of $\sigma^2$ is calculated may be balanced through one of three phenomena: observance of a complete data set, luck of the incomplete-data analyst, or intervention of the analyst. In each case, assuming data are MCAR and follow Model (3.4), we obtain an unbiased point estimator of $\sigma^2$, which we generically denote $\hat{\sigma}^2$.

**Complete-Data Mean Squares Estimators** We define the vector $\textbf{MS}$, where $\textbf{MS} = [\text{MS}_1, \text{MS}_2, \text{MS}_3]'$, to be the vector of mean squares estimators for balanced data following
Model (3.4), as provided by Scheffe (1959):

\[
MS_1 = \frac{n_t n_r \sum_{x=1}^{n_x} (y_{s..} - y_{...})^2}{n_t - 1},
\]

\[MS_2 = \frac{n_r \sum_{j=1}^{n_r} \sum_{x=1}^{n_x} (y_{t..} - y_{..s} - y_{s..} + y_{..})^2}{(n_t - 1)(n_r - 1)},\]

\[
MS_3 = \frac{\sum_{i=1}^{n_i} \sum_{r=1}^{n_r} \sum_{s=1}^{n_s} (y_{t.s..} - y_{t..})^2}{n_t n_r (n_r - 1)}.
\]

**Complete-Data Expected Mean Squares**  We define the vector given by \( E(\mathbf{MS}) = [E(MS_1), E(MS_2), E(MS_3)]' \) to be the vector of *expected mean squares* associated with \( \mathbf{MS} \), where \( E(\mathbf{MS}) \) is the expected value of \( \mathbf{MS} \) with respect to Model (3.4). Scheffe (1959) also provides expressions for the expected mean squares:

\[
E(MS_1) = \sigma_e^2 + n_r \sigma_c^2 + n_t n_r \sigma_s^2,
\]

\[
E(MS_2) = \sigma_e^2 + n_r \sigma_c^2, \text{ and}
\]

\[
E(MS_3) = \sigma_e^2.
\]

We assume that \( n_t \geq 2, n_s \geq 2, \text{ and } n_r \geq 2 \).

**Analysis of Variance Method**  Given the balanced data set, the observed value of \( \mathbf{MS} \) is calculated using (3.24)-(3.26). Equating the observed value of \( MS_k \) with \( E(\mathbf{MS}_k) \), for \( k=1,2,3 \), establishes a system of three linear equations in three variables, the variance components. We denote the solution to this system by the vector \( \hat{\sigma}^2 = [\hat{\sigma}_u^2, \hat{\sigma}_c^2, \hat{\sigma}_e^2]' \). By construction of \( \hat{\sigma}^2 \), it is evident that \( \hat{\sigma}^2 \) is componentwise an unbiased estimator of \( \sigma^2 \). That is, \( E(\hat{\sigma}^2) = \sigma^2 \). For a complete discussion of this method, we refer you to Scheffe (1959).

**Negative Variance Component Estimates**  While a unique solution is guaranteed for the system \( \{MS_k = E(\mathbf{MS}_k): k=1,2,3\} \), components of its solution are not constrained to
be nonnegative. We require the initial estimate of \( \sigma^2 \) to apply Box and Tiao’s (1973) method for approximating draws from the posterior distribution of \( \sigma^2 \), where \( \hat{\sigma}^2 \) is key to formulation of the prior on \( \sigma^2 \). Negative variance component values cannot be used as substitutes. For this study, we replace each negative component of \( \hat{\sigma}^2 \) with zero. Note that this adjustment renders \( \hat{\sigma}^2 \) biased for \( \sigma^2 \). Furthermore, the case of a zero variance component renders the model assumptions questionable. In our simulation, the occurrence of negative estimate of the variance component is sufficiently rare that we do not include such cases in the presentation of results. These comments are included for the sake of thoroughness. If a variance component is deemed to be no different from zero, then it is advisable to proceed with RMNI using an approximated posterior for a hierarchical model with fewer terms, as available in Box and Tiao (1973) and demonstrated by Yamashita (1995).

3.4.2 Experimentwise Subject Deletion

The Method In the context of Model (3.4) and Design 3.2.1 with MCAR data, we present a two-step method for obtaining an incomplete-data estimate of \( \sigma^2 \) with a balanced-data estimator where observed data is unbalanced; and we refer to the method as Experimentwise Subject Deletion, or ESD. The value of this method is that it enables us to move forward with computing an approximate posterior distribution for each variance component, and ultimately a posterior distribution for the missing data given the observed data. Given enough completely observed subjects, this approach is advisable for the sake of clear calculation and interpretation of the expected mean squares.

Step One The first step of ESD is to construct an artificially balanced incomplete data set, \( A \), by removing each partially-observed subject from the observed incomplete data
set. Let \( n_s^{(A)} \) represent the number of fully-observed subjects, with \( n_s^{(A)} \leq n_s \). The data set \( A \) is consequently balanced and follows Model (3.4) with diminished size \( n_s^{(A)} \), where 
\[
    n_s^{(A)} = n_t n_s^{(A)} n_r,
\]
with \( n_t \) and \( n_r \) fixed by Design 3.2.1. Note that the value of \( n_s^{(A)} \) is dependent upon both pattern and rate of missingness, and is potentially zero or one.

**Step Two** The second step of ESD is to compute an estimate of \( \sigma^2 \) using the balanced data set \( A \) and the estimator \( \hat{\sigma}^2 \) described in Section 3.4.1. We refer to the estimator defined by Steps One and Two above as \( \hat{\sigma}^2_{ESD} \). With MCAR data, \( \hat{\sigma}^2 \) is not biased by the construction of \( A \), so that \( E(\hat{\sigma}^2_{ESD}) = \sigma^2 \).

**Application** The method of experimentwise subject deletion may be appropriate if missingness is sparse relative to the size of the experiment, \( n \), and missingness affects few subjects relative to \( n_s \). However, under Model (3.4), with \( n = n_t n_s n_r \), if the observed rate of missingness exceeds \( (n_t n_r)^{-1} \times 100\% \), it is possible that no subject is completely observed, in which case the data set reduced by this method is empty. If the observed data are MCAR, the method of experimentwise subject deletion is prone to exclude sufficient information from the estimation of \( \sigma^2 \) that the method’s simplicity does not merit the cost in accuracy of the estimate. Having said this, we recommend that Experimentwise Subject Deletion be considered only in the case of sparsely missing data which eliminates only a few subjects from calculation of initial estimates of the variance components. Given a substantial rate of missingness on very few subjects, we would question our assumption that the data are MCAR.

**Simulation Study** We conducted a simulation study to investigate the performance of the complete data mean squares estimators with the reduced data set. For the simulation, \( n_t \) and
nr are set at 3, with ns = 50 subjects. We considered the number of subjects and percentage of observed information used to calculate estimates for the expected mean squares.

3.5 Unbalanced-Data Estimation of Variance Components

3.5.1 Mean Squares for Unbalanced Data

Henderson’s Mean Squares Estimators  Henderson’s Method I sums of squares, or an adaptation of it, may be used to calculate mean squares estimates for random effects models. To do so, treat the incomplete data set as a completely observed data set with an unbalanced design. The calculating formulas for the mean squares are given by

\[ MS_1^H = \frac{n_0 \sum_{s=1}^{n_s} n_{+s} (y_{s} - \bar{y})^2}{n_s - 1} \]  \hspace{1cm} (3.38)

\[ MS_2^H = \frac{\sum_{t=1}^{n_t} \sum_{s=1}^{n_s} n_{ts} (y_{ts} - \bar{y}_{ts} - \bar{y}_{s} + \bar{y})^2}{(n_t - 1)(n_s - 1)} \]  \hspace{1cm} (3.39)

\[ MS_3^H = \frac{\sum_{t=1}^{n_t} \sum_{s=1}^{n_s} \sum_{r=1}^{n_r} (y_{tsr} - \bar{y}_{ts})^2}{\sum_{t=1}^{n_t} \sum_{s=1}^{n_s} n_{ts} - n_t n_s} \]  \hspace{1cm} (3.40)

The expected values of each of these mean squares can be computed using the method called synthesis which is detailed by Milliken and Johnson (1984). Their exact expectations will vary with observed pattern of missingness. Once obtained, however, an unbiased estimate of each variance component of a random effects model can be found as the solution to a system of equations as done in the case of complete and balanced data. Henderson’s Method I is not appropriate under a mixed effects model; we consider a correction proposed by Milliken and Johnson (1984) for the case of mixed effects below.

Adjusting Mixed Effects Models for Fixed Effects  Milliken and Johnson (1984) propose adjusting data observed under a mixed effects model by subtracting the LSE for the
appropriate fixed effect from each datum. After subtraction of this LSE, the data is distributionally equivalent to data observed under a random effects model. Ultimately, we establish that this approach is theoretically questionable for the two-way mixed interaction model. Searle (1971, Section 10.2) discusses this approach and the use of synthesis (Hartley, 1967) in the case of model (3.4). In sum, Searle asserts that for the mixed model, the estimates of variance components computed by this process will be biased. Ultimately, we see that adjusting our model for fixed effects and employing synthesis is worth considering, but do not incorporate it into our RMNI procedure.

**Incomplete-Data Design Matrix**  For a complete discussion of adjusting mixed models for fixed effects, we refer you to Milliken and Johnson (1984). It is helpful to our discussion to define the design matrix as expressed and annotated under an unbalanced two-way mixed model. First, let $W^*$ be the design matrix obtained by adjusting the complete-data design matrix, $W$, for missing data. That is, $W^*$ is effectively the design matrix for an unbalanced design corresponding to our incomplete observed data. Partition the design matrix, $W^*$, into submatrices denoted $W^*_t$, $W^*_s$, and $W^*_c$, where $W^*_t$ comprises the first $n_t$ columns, which correspond to treatment allocation, $W^*_s$ the next $n_s$ columns, which correspond to subject allocation, and $W^*_c$ the last $n_t n_s$ columns.

**Expressions for $Y_{obs}$ and $Y_{adj}$**  Given expressions for partitions of the unbalanced design matrix $W^*$, the model for the incomplete observed response data may be expressed by

$$Y_{obs} = W^*_t \tau + \varepsilon$$

(3.41)

where

$$V(\varepsilon) = \sigma^2_e W^*_t W^*_t' + \sigma^2_c W^*_c W^*_c' + \sigma^2 I.$$  

(3.42)
The LSE described by Equation (3.11) is written \( \hat{\theta} = (W_t^*W_t^*)^{-1}W_t^*Y_{obs} \), and the adjusted data set is then given by

\[
Y_{adj} = Y_{obs} - W_t^*\hat{\theta} = (I - W_t^*W_t^{*\prime})Y_{obs},
\]

as shown by Milliken and Johnson (1984) in their discussion of synthesis.

**Mean Squares for** \( Y_{adj} \)  The adjusted data effectively follows a two-way random effects model with interaction term, and so Henderson’s Method I mean squares given by (3.37) through (3.39), are the appropriate calculation method. With unbalanced data, a closed-form expression for the expected mean squares does not exist. The expected mean squares can be computed using Hartley’s (1967) method of synthesis. The method of synthesis does apply with any pattern of missingness and any unbalanced data set. Hartley (1967) asserts that synthesis gives the correct expected mean squares under “any unbalanced mixed or random effects model and any analysis of variance.”

**Reservations for Use of Synthesis**  We were inspired by the convictions of Milliken and Johnson (1984) and Hartley (1967) regarding adjustment for fixed effects and synthesis, but disappointed by discovery of Searle’s (1971) comments regarding these approaches under Model (3.4). The approach of synthesis using Henderson Method 1 Mean Squares is undisputed for completely random models, and may be valuable for RMNI under such a model. For the sake of reliability and interpretation of results under a mixed interaction model, we resolved to employ methods made available by SAS (Statistical Application Software), as described in Section 3.5.2 below. In so doing, we avoid the controversy and confusion from lack of consensus amongst experts of Experimental Design regarding computation of
variance component estimates with unbalanced data. In further discussion, we acknowledge that controversy and lack of consensus remains under our model, as discussed by Searle (1987, *Linear Models for Unbalanced Data*). In the end, we opt for a commonplace method for computation of expected mean squares with the mixed interaction model which is free of convergence issues, as described in Section 3.5.2.

### 3.5.2 Expected Mean Squares

**Controversy**  As discussed in the previous section, we must acknowledge that no consensus exists for an exact, correct, or appropriate method for computation of expected mean squares under the two-way mixed interaction model with unbalanced data. Complications ensue from having a mixed model as well from inclusion of the interaction term. Our goal is to demonstrate RMNI which requires initial estimates of variance components under Model (3.4) with unbalanced incomplete data. For the initial estimates, we have chosen to employ a standard commonplace approach rather than the controversial alternatives exhibited in the previous section.

**General Form for Mean Squares** A valid mean squares estimator $Q(y)$ of $\sigma^2 = \sigma_s^2 + \sigma_c^2 + \sigma_e^2$ is one that can be written as a quadratic form in the response vector $y$ as

$$Q(y) = y' Ay,$$  \hspace{1cm} (3.44)

where $A$ is a symmetric matrix of constants. Graybill (1976) shows that any sum of squares can be written in this form. The Mean Squares formulations given in Section 3.3 satisfy this criterion. For complete data, we use (1.31) through (1.33). For incomplete data, we use (1.37) through (1.39).
General Form for Expected Mean Squares  The difficulty in applying any particular method for mean squares calculations in the case of an unbalanced design is calculating the expectation of the mean squares. Typically, a general closed-form expression for the expected mean squares is theoretically intractable without balanced data; it does not exist. For the purpose of creating a general method for handling missingness in Experimental Design, we need a method of calculating the expected mean squares given any pattern of missingness, or any unbalanced data set. For complete data, we use (1.34) through (1.36), provided in closed form by Scheffe (1959). For incomplete data, we did consider the methods of Experimentwise Subject Deletion (3.4.2) and synthesis with adjusted data (3.4.3) in situations that are not general. Limitations on these approaches and their general acceptance led us to adopt a standard procedure available in SAS.

Preferred Method for Calculating Means Squares and Expectations  We use SAS PROC GLM Sums of Squares III, and the Expected Mean Squares computed therewith by SAS. SAS Sums of Squares III is the Fitting Constants Method (also known as Henderson’s Method III) described and demonstrated for Model (3.4) by Searle (1987). With no general consensus in the field regarding a correct choice of procedure under our model, we concluded to be expeditious by using a method clearly demonstrated and advocated by Searle in the literature, and involving the least complexity. Alternatives rely on adjusting the data for fixed effects, as discussed in Section 3.5.1, and applying the methods of synthesis, Method of Moments, Maximum Likelihood Estimation, and MINQUE. For a given data set, any of these methods may be a viable option; for a random model, these methods may be preferred.
Synthesis and Random Hierarchical Models  The exceptional advantage of synthesis is that the matrix $A$ which characterizes a particular sums of squares calculation does not need to be known explicitly. The expected mean squares are calculated exactly by supplying columns of the design matrix, $W^*$, as data to the computer algorithm which computes the sums of squares. With random models, an unbiased estimate of each variance component is found by setting each mean squares value, $MS_p$, equal to its expectation, $E( MS_p )$, and then solving the resultant system of linear equations given by

$$E( MS_p ) = MS_p \quad p = 1,2,3 \quad (3.45)$$

for $\hat{\sigma}^2 = (\hat{\sigma}_x^2, \hat{\sigma}_e^2, \hat{\sigma}_e^2)$. This estimate of $\sigma^2$ is unbiased with our assumption that the data are MCAR. Another advantage of synthesis is that it is relatively simple to program an algorithm, relieving dependence upon SAS. If the response mechanism is nonignorable, further consideration must be made to obtain an unbiased estimate of $\sigma^2$, if it is possible to do so.

Negative Estimate of Variance Component  Finally, the unbiased estimate of any one variance component may be negative. In that event, we use zero as a practical estimate of that variance component. Further discussion is provided at the end of Section 3.4.1 above.
CHAPTER 4

Procedural Details of RMNI

4.1 Introduction

Contents With this chapter, we present a procedural schematic for Repeated Measures Normal Imputation (RMNI). The final section summarizes the evolution of RMNI from its beginnings with Rubin (1987) to today under the research guidance of Elizabeth Stasny through two dissertations (Yamashita, 1995, and this dissertation). Throughout this chapter, background for RMNI is presented incrementally, as relevant to the discussion. In particular, we discuss details of the RMNI procedure which tailors the imputation to Model (3.4). That is, the imputation model is Model (3.4), and the imputation procedure generates imputed values which preserve the underlying spherical variance-covariance matrix. A summary digest of assumptions made for the simulated demonstration of RMNI is also provided.

RMNI Procedure Summary The following steps for the procedure of RMNI are intended to serve as a general guide, and are not model-specific. Note that no consideration is given to the incorporation of a response mechanism for non-ignorable missingness. Also note that Rubin (1987) makes a focus of obtaining multiply-imputed means and variances; multiply-imputed estimates of other forms should be handled with care.
**Step Zero:** Observe MCAR data following a hierarchical model with repeated measures and ignorable nonresponse/missingness; remove unobserved subjects from the study.

**Step One:** Compute initial estimates of parameters using unbalanced-data methods, or incomplete-data methods.

**Step Two:** Obtain the posterior distribution \([(\theta, \sigma^2)|Y_{obs}]\), where \(\theta\) is a fixed treatment effect.

**Step Three:** Generate \(m\) random draws from \([(\theta, \sigma^2)|Y_{obs}]\), where \(m\) is the intended number of imputations.

**Step Four:** Obtain the posterior distribution for the missing data, \([Y_{mis}|Y_{obs}]\).

**Step Five:** Generate \(m\) sets of \(n_{mis}\) random draws from \([Y_{mis}|Y_{obs}]\); use these to construct \(m\) completed data sets.

**Step Six:** Use standard complete-data estimators to calculate completed-data estimates of \((\hat{\theta}, \text{Var}(\hat{\theta}))_\ell\), for \(\ell = 1, \ldots, m\). Additionally, one may compute any other statistics desired using each of the completed data sets.

**Step Seven:** Combine \(m\) estimates into one multiply-imputed estimate as described in Section 2.5.

**Step Eight:** Perform statistical inference using the combined multiply-imputed estimates.

**Example** A simple application of the steps for RMNI is the following process guideline for construction of a confidence region for \(\theta\), the \(n_r\)-variate fixed treatment effect: Step Zero intends a validation of the assumptions for RMNI to be performed with the natural or
simulated data set. Steps One through Five constitute the core components of constructing the imputed values for a multiple imputation. In Step Six, the analyst is intended to calculate the LSE (OLS, Equation 3.11) for $\theta$, and the variance-covariance matrix for $\hat{\theta}$; RMNI enables use of standard complete-data methods with each of the $m$ completed data sets. In Step Seven, these $m$ estimates of $\theta$ and $\text{Var}(\hat{\theta})$ are combined according to the rules provided by Rubin (1987). In Step Eight, the results from Step Six, and/or Step Seven, may be used to construct the confidence region for $\theta$.

**Note** In Step Six, a complete ANOVA, performance of multiple comparisons, or inferential test may be performed on each of the $m$ data sets. Rubin (1987) does not specifically detail how to combine the various estimates that may be generated in this case. However, Rubin’s (1987) construct for means and variances is useful as a guideline to proper combination of any set of $m$ statistics which may be calculated in this case. It is crucial to ascertain whether and how the additional component of variance due to missingness is incorporated into the combined standard error for any estimator used in this case. See Section 2.5.2 for details.

### 4.2 Repeated Measures Normal Imputation

#### 4.2.1 Background and Development of RMNI

**Donald Rubin** The origin of the Repeated Measures Normal Imputation (RMNI) procedure lies in Rubin (1987) with Lemma 2.1 on page 41 of his text; the Fully Normal Bayesian Repeated Imputation procedure developed with Example 2.1 (page 40), Example 3.2 (page 83), and Example 4.2 (page 123); and Conclusion 4.1 (page 125) which relies on Result 4.1 (page 119).
Rubin’s Lemma 2.1 In summary, Rubin’s (1987) Lemma 2.1 supposes that given a scalar mean $\mu$ and scalar variance $\sigma^2$, and given a population of $N$ observable responses $Y_i$, $i = 1, \ldots, N$, which are iid $\mathcal{N}(\mu, \sigma^2)$, and given that the prior distribution of $(\mu, \sigma^2)$ has density proportional to $\sigma^{-2}$, then two results follow:

1. The distribution of $\mu$ given $(\sigma^2, Y_1, \ldots, Y_n)$ is $\mathcal{N}(\bar{y}, \sigma^2/n)$ where $\bar{y}$ is the usual sample mean.

2. The distribution of $\sigma^2$ given $(Y_1, \ldots, Y_n)$ is $(n-1)s^2\chi_{n-1}^{-2}$ where $s^2$ is the usual sample standard deviation, and $\chi_{n-1}^{-2}$ is the inverted chi-squared distribution with $n-1$ degrees of freedom.

Fully Normal Imputation Fully Normal Imputation is discussed, explored, and named variously by Rubin and Schenker (1986), Rubin (1987), and Yamashita (1995). This procedure is designed for an imputation of scalar sample survey response variates in the absence of known covariates. As with Lemma 2.1 of Rubin (1987), the population of $N$ observable responses is assumed to be iid normal with constant mean $\mu$, and with constant variance $\sigma^2$. Given Lemma 2.1, the following steps are postulated for an imputation of $n_{\text{mis}}$ observations given $n_{\text{obs}}$ valid responses, where the sample has size $n = n_{\text{mis}} + n_{\text{obs}}$:

1. Generate an $\chi_{n_{\text{obs}}-1}^{-2}$ random variate, $x$, and let

   $$\sigma_*^2 = s_{n_{\text{obs}}}^2 (m - 1)/x. \quad (4.1)$$

2. Generate a standard normal random variate, $z_o$, and let

   $$\mu_* = \bar{y}_{n_{\text{obs}}} + \sigma_* z_o / n_{\text{obs}}. \quad (4.2)$$
3. Generate $n_{mis}$ independent $\mathcal{N}(0,1)$ random variates, $z_i, i \in mis$, and impute each component of $Y_{mis}$ using

$$Y_{i*} = \mu + \sigma z_i.$$  

(4.3)

4. Repeat this imputation process $m$ times to accomplish a multiple imputation. Summary statistics may be computed using each data set, and then a combined multiply-imputed estimate is calculated as described in Section 2.5 of this dissertation. Note that $\bar{y}_{n_{obs}}$ and $s^2_{n_{obs}}$ are the usual sample mean and sample variance computed with observed data, where $n_{obs}$ is substituted for the intended sample size $n$.

Rubin’s Conclusion 4.1  
Rubin’s (1987) Conclusion 4.1 is discussed in Section 2.9.2 of this dissertation. We offer this paraphrased restatement of Rubin’s said conclusion: A multiple imputation procedure is reasonably guaranteed to be proper if it extends a distribution-based single imputation procedure by means of a Bayesian procedure. For example, Rubin (1987) demonstrates that a multiple imputation procedure is valid if it extends the modeling approach with Gaussian error, but not valid if it extends the Hot Deck procedure to repeated Hot Deck imputations (Rubin, 1993).

Yamashita Dissertation  
Yamashita (1995) adapts the Fully Normal procedure as provided above to accommodate the One-Way Layout with repeated measures, where observable responses are, by assumption and by design, normal variates with prescribed means and variances. For our purposes, the term RMNI distinguishes the imputation procedure as a model-based imputation procedure, standing in contrast with the Multivariate Normal imputation methods, as popularized by Schafer (1997). The RMNI procedure for the One-Way Layout with repeated measures is described in Section 5.2 of Yamashita (1995).
This Dissertation  As stated in Chapter One, we extend the work of Yamashita (1995) with this dissertation to adapt the RMNI procedure for use with the Two-Way Layout, with repeated measures and a random interaction term. For Design 3.2.1, we did not need to alter the steps of RMNI as developed by Daryl Yamashita (1995); however, we did develop a methodology for accomplishing each step which is specific to our model, and which can be extended to function for the entire family of hierarchical models, as described with Step Two below.

4.2.2 Procedural Steps for RMNI

Step Zero  Obtain data according to Experimental Design and observe missingness; eliminate any subject who is completely unobserved. Verify the appropriate model for the design, and determine the intended analysis of the data in advance. It is a tenet of Multiple Imputation that the imputation model should match the analysis model.

Step One  Use incomplete data to obtain an initial estimate of each parameter of the chosen model. For example, with our model, we compute initial estimates for each fixed treatment effect and each variance component using unbalanced-data methods. If an unbalanced-data method is intractable, consider utilizing Experimentwise Subject Deletion to obtain the initial estimates with balanced complete-data methods (Section 3.4.2).

Step Two  Obtain the posterior distribution \([((\theta, \sigma^2)|Y_{obs}, R, I])\), which is equivalent to \([(\theta, \sigma^2)|Y_{obs}]\), given the assumption of an ignorable response mechanism, \(R\). To accomplish this, we employ the approximation to the joint posterior distribution of \((\theta, \sigma^2)\) as provided by Box and Tiao (1973) for our model. Box and Tiao provide similar results for
fixed, random, and mixed hierarchical models with one-way, two-way, and three-way classifications. With these, RMNI might be extended to the complete family of hierarchical models.

**Step Three** Generate $m$ random draws from $[(\theta, \sigma^2)|Y_{obs}]$, where $m$ is the requisite number of imputations. For our model, we obtain $m$ posterior values of $[(\theta, \sigma^2_u, \sigma^2_c, \sigma^2_e)]$. This step involves simulating random draws from the $\chi^2$ distribution with adjusted degrees of freedom, as prescribed by Box and Tiao (1973), and as determined in Step Two. See Section 4.7.2 for details regarding the algorithm for producing these posterior values with the two-way mixed interaction model.

**Step Four** Obtain the posterior distribution for the missing data, $[Y_{mis}|Y_{obs}, R, I]$, which is equivalent to $[Y_{mis}|Y_{obs}]$ given the assumption of ignorable $R$. To accomplish this, we utilize the conditional distribution provided by Mardia, Kent, and Bibby (1979, Theorem 3.2.4, page 63 of text). This conditional distribution is obtained by partitioning a data vector $Y$ into two subvectors $Y_1$ and $Y_2$, which effectively correspond to our $Y_{obs}$ and $Y_{mis}$:

1. Let $Y$ be a $p$-dimensional multivariate normal random variable which can be partitioned as

$$
Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_p \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right),
$$

where $Y_1$ is a $p_1$-dimensional vector.

2. Then

$$
Y_2|Y_1 \sim N_{p-p_1} \left( \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(Y_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \right).
$$

This result provides us with a conditional posterior for $Y_{mis}$ given $Y_{obs}$, which is fully evaluated with the random draws for parameter values obtained in Step Three. Note that
this posterior distribution for $Y_{mis}$ is dependent upon the assumption that the response mechanism $R$ is ignorable. See Section 4.7.3 for details regarding the application of this result under the two-way mixed interaction model.

**Step Five** Generate $m$ sets of $n_{mis}$ random draws from $[Y_{mis}|Y_{obs}]$. See Section 4.7.4 for details regarding the algorithm for producing these posterior values.

**Step Six** Use standard complete-data estimators to calculate completed-data estimates of $(\hat{\theta}, \text{Var}(\hat{\theta}))_\ell$, for $\ell = 1, \ldots, m$. The procedure for combining the completed-data estimates into a single multiply-imputed estimate is described in Section 2.5 and Rubin (1987). Distributional properties for these estimators are provided in Section 3.3.1 under our model.

**Step Seven** Combine the $m$ RMNI-imputed estimates (completed-data estimates) into one multiply-imputed estimate as described in Section 2.5. In general, one obtains multiply-imputed treatment means, with associated estimated error, at this stage. One may choose to compute other summary statistics at this time with each completed data set; however, care must be taken in the method by which these sundried statistics are combined to yield a single multiply-imputed result. See comments at the end of Section 4.1 for further explanation.

**Step Eight** Perform the intended analysis of the data. At this stage, one has estimates of each treatment effect, and the estimated variance-covariance matrix as a gauge for error. For our demonstration, we perform a standard complete-data Multiple Comparisons Procedure at this step, as presented in Chapter Six.
4.3 Assumptions: Data Environment

**Experiment**  We assume the collection of data by designed experiment according to the Two-Way Layout with mixed effects and repeated measures. It is intended that the interaction effect be valid for inclusion in the model. Details of the design are given in Section 3.2.1. Our specification of the model is given with Equation 3.4, which corresponds to Model Ia of Hocking (1973) and Hochberg and Tamhane (1983). As example of our intended experiment, consider a test for the effects of three blood thinning agents on each of 50 patients with high blood pressure, where each patient yields a response of blood pressure measured three times under each medication being studied.

**Replicates**  We assume that each combination of treatment and subject is measured three times. We perform the imputation in the case where any one or all of the replicates are missing for a specific treatment/subject combination. Note that two replicates are required to enable estimation of an interaction effect; we choose \( n_r = 3 \) for purposes of our simulation and demonstration.

**Fixed Effect**  We assume that the treatment effect is fixed. Any fixed treatment effect that is completely unobserved is eliminated from the study before imputation. For our simulation study, we demonstrate RMNI in the case of three treatments.

**Random Effect**  We assume that the subject effect is random. Any subject who is completely unobserved is eliminated from the study before imputation. For our simulation study, we demonstrate RMNI in the case of 50 subjects.
**Interaction Effect**  We assume that the interaction effect is random and necessary to the model. An interaction effect is eliminated from the study only in the case that the corresponding subject is eliminated from the study. For our simulation study, we demonstrate RMNI with three treatments and fifty subjects, yielding 150 interaction effects in the complete-data model.

**Size of Experiment**  We fix the size of the experiment at $n=450$ (3 by 50 by 3), as described above. We do not explore the impact of the size of the experiment on multiply-imputed estimates. We contend that increasing the number of multiple imputations will compensate for a reduced sample size of the experiment, as made evident with the simulation study of Chapter Five.

### 4.4 Assumptions: Model

**Treatment Effect**  Each treatment effect is treated as a fixed effect. We provide further specification of each effect, and the construction of our model, in Section 3.2.

**Subject Effect**  Each subject effect is assumed to follow the normal distribution, with a mean of zero and a constant variance. We assume the subject effects to be mutually independent, and both individually and collectively independent of each interaction effect and each error term.

**Interaction Effect**  Each interaction effect is assumed to follow the normal distribution, with a mean of zero and a constant variance. We assume the interaction effects to be mutually independent, and both collectively and individually independent of each error term.
Variance Components  The subject variance component is constant from one subject to the next. The interaction variance component is constant from one measurement to the next. The error variance component is constant from one measurement to the next. We provide further specification of each variance component in Section 3.6.

Estimability of the Interaction Effect  We do not need to estimate any interaction effect directly; we need only estimate the interaction variance component using unbalanced data. With completed data, standard complete-data methods suffice by design for estimation and inference regarding the interaction effect, given two replicates at minimum. Note that RMNI does completely impute empty cells of the Two-Way Layout.

4.5 Assumptions: Missing Data

Percent Missing or Rate of Missingness  With this dissertation, we have explored model-based multiple imputation for our model in the case of one, five, and ten percent missingness. At the rate of ten percent, we begin to see some computational difficulty with the computation of initial estimates using Experimentwise Subject Deletion (Section 3.4.2). For the simulation study presented in Chapter Five, we choose an instructive missingness rate of ten percent.

Unit Missingness and Subject Deletion  We delete subjects which are nowhere observed. We do consider Experimentwise Subject Deletion (Section 3.4.2) for the calculation of initial estimates, and recognize it is most useful if very few subjects need be deleted, as discussed in Section 3.4.2.
**MCAR**  We assume that the data are MCAR: Missing Completely at Random, as defined by Little and Rubin (1987). Effectively, this means that whether a datum is missing is completely independent of its value.

**MAR**  Missing at Random (MAR), as defined by Little and Rubin (1987), refers to the case in which missingness is dependent upon a known covariate. For example, nonresponse may be more likely to occur within higher tax brackets than middle-income tax brackets, while nonresponse occurs completely at random within each tax bracket. We do not consider MAR data in this dissertation.

**Response Mechanism**  We assume that the response mechanism is ignorable (Section 2.4). We make no consideration for non-ignorable response mechanisms in our simulation study. This may be a worthy topic for further research.

**Pattern of Missingness**  We consider any randomly generated pattern of missingness where the data are MCAR; we do not make a targeted assessment of RMNI involving any otherwise classifiable pattern of missingness. RMNI with monotonic missing data is discussed by Yamashita (1995).

**Fraction of Missing Information**  We do not explore the rate of missingness in the form of the fraction of missing information. It is apparent that the strategic re-ordering of observed data may yield benefit with certain incomplete data sets vis-a-vis re-structuring of the variance-covariance matrix of the incomplete data. That is, it is evident that there is not
a one-to-one correspondence between rate of missingness and fraction of missing information. This is not explored here, but is discussed with topics for further research in Chapter Eight.

4.6 Assumptions: Analysis Environment

4.6.1 Complete-Data Estimates

Usage Complete-data estimators are used in two places for the simulation study presented in Chapter 5. First, the complete-data estimators are used with the simulated complete data set to compute complete-data treatment means, with associated error. Second, the complete-data estimators are used with the RMNI-imputed data sets to compute completed-data treatment means, with associated errors.

Treatment Effect We consider both the Least Squares Estimator and a Weighted Means Estimator for each treatment effect (Section 3.3). With complete and incomplete data, we prefer the LSE for presentation of our simulation study. See Equation (3.15).

Variance Components With complete data, we use the standard Henderson’s Method I Sums of Squares. The associated Mean Squares Estimators are provided as Equations (3.32) - (3.34).

Expected Mean Squares With complete data, the Expected Mean Squares exist in closed form, and are easily computed using Scheffe’s Analysis of Variance method. Further specification of this process is provided with Equations (3.35) - (3.37).
4.6.2 Incomplete-Data Estimates

**Usage**  Incomplete-data estimators are used to calculate initial estimates of the treatment effects and variance components in Step One of the RMNI procedure. The initial estimates are supplied to an algorithm which generates simulated random draws from the joint posterior distribution of the effects and variance components. See Section 4.2.

**Treatment Effect**  With incomplete data, we use the standard Least Squares Estimator to compute an unbiased initial estimate for each treatment effect. The calculating formula, with distributional properties, are provided with Equation (3.11).

**Variance Components**  With incomplete data, we use the Type III Sums of Squares, and the associated expected mean squares, of the SAS GLM Procedure, to obtain an unbiased initial estimate of each variance component. This is also known as the Method of Fitting Constants and Henderson’s Method III. Further discussion of this process is provided in Section 3.5.

4.6.3 Multiply-Imputed Estimates

**Treatment Effect**  The multiply-imputed estimate of the treatment effect is calculated as described with Equation (2.21). In sum, it is computed as the mean of the $m$ completed-data treatment means calculated using each of $m$ RMNI-imputed data sets.

**Variance-Covariance Matrix of the Estimated Treatment Effect**  The multiply-imputed estimate of $V(\hat{\theta})$ is given by Equation (2.24), and termed the Total Variance. The evaluated variance-covariance matrix is dominated by the value of the Within Variance, $\bar{U}_m$, which is computed as the mean of the $m$ completed-data values of $V(\hat{\theta}_\ell)$, where $\ell = 1, 2, \ldots, m$. 

83
Note that the additional component of variability due to the presence of missing data is incorporated into this calculation as the Between Variance, $B_m$. The total variance is given by $T_m = \bar{U}_m + (1 + 1/m)B_m$

### 4.7 RMNI Simulation and the Two-Way Mixed Model

**Tailoring RMNI to Your Model**  
The procedural steps for RMNI outlined in Section 4.2 are general and apply to any hierarchical model. To accomplish specific tasks, particularly obtaining draws from the posterior distributions of the variance components and the posterior distribution of $Y_{mis}$ given $Y_{obs}$, it is necessary to develop an algorithm specific to the model being studied. These algorithms have been developed and examined for the One-Way Layout by Yamashita (1995). We attempt to demonstrate analogous algorithms for the Two-Way Layout. Our approach is tailored to the two-way mixed interaction model, but would accommodate a random model with or without interaction with minor adjustments.

### 4.7.1 Step Zero: Complete Data Baseline and MCAR Data

**Complete Data Simulation**  
For the demonstration of RMNI, we begin with the simulation of a complete data set. We use these data to compute baseline statistics for comparison with multiply-imputed estimates. The advantage of the simulated data over a natural data set is that we are assured of the distributional properties of the simulated data. By using a SEED = 32397, we are also able to reproduce results easily. Also, we are able to fix the model parameter values so that we are assured of distinct fixed effects and sensible variance component values. Finally, we are confident in the distributional properties of the complete data estimates as well as the derived MCAR data. With natural data, many of our assumptions could be called into question. Statistical assessment of the simulations of complete data and MCAR data is presented in Section 5.1.
MCAR Data Simulation  Given a complete data set, $Y_{inc}$, simulated to follow the two-way mixed interaction model, we obtain MCAR data, $Y_{mis}$, by generating an MCAR pattern of missingness and removing select data from $Y_{inc}$. The deleted values constitute $Y_{mis}$. The indices for values of $Y_{inc}$ to be removed are obtained as draws from the Discrete Uniform distribution, as described in Section 1.3. Finally, any completely unobserved subject (or treatment) is removed from the data set.

4.7.2  Step Three: Box and Tiao’s Posterior Approximations

Two-Way Layout  For the Two-Way Layout, Box and Tiao (1973) derive an approximation for the posterior distribution of each variance component under the random model, assuming no fixed effects. See Table 5.3.3 on page 292 of their text. Box and Tiao (1973) discuss our model (mixed, including interaction) in Section 6.4.2 on page 364, where they determine that inferences regarding the variance components can be accomplished using the same distributional results provided for the random model (Table 5.3.3).

Box and Tiao Notation  For the Two-Way Layout, Box and Tiao (1973) refer to the error variance as $\sigma_1^2$, the interaction variance as $\sigma_2^2$, and the subject variance as $\sigma_3^2$. That is, for our purposes, $[\sigma_1^2, \sigma_2^2, \sigma_3^2]' = [\sigma_{err}^2, \sigma_{sub}^2, \sigma_{int}^2]'$. For purposes of reference, we restate the approximation formulas here for the posterior distributions of the three variance components. These approximations involved several constants for which the specification is too lengthy to reproduce here. These constants include $\nu_1', \nu_2', \text{ and } \nu_3$, which are variously adjusted degrees of freedom; as well as $m_1, m_1', m_2, m_2', \text{ and } m_3$, which are variously adjusted mean squares (Type III). The formulations for these constants are given by Box and Tiao (1973) in Section 5.3 with Equations 5.3.24, 5.3.27, and 5.3.33 of the text; a summary example is given with Table 5.3.3 on page 292 of their text. It should be noted
that these constants rely upon computation of the Incomplete Beta function denoted \( I(\ast, \ast) \) in their text. Finally, recall that \( n_t \) is the number of treatments, \( n_s \) is the number of subjects, and \( n_r \) is the number of replications, so that Box and Tiao’s \([I, J, K]' = [n_t, n_s, n_r]'\).

**Approximation: Error Variance Component**  
Box and Tiao (1973) provide the following result:

\[
\frac{\sigma^2 \epsilon}{\nu_1' m_1'} \sim \chi_{\nu_1'}^{-2}.
\]  
(4.6)

To approximate a random draw from the posterior distribution of \( \sigma^2 \epsilon \) given \( Y_{obs} \), we calculate

\[
\sigma^2 \epsilon = \nu_1' m_1' x^2,
\]  
(4.7)

where \( x \) is a random variate generated using the IMSL subroutine RNCHI, and transformed to follow the \( \chi^{-2} \) distribution on \( \nu_1' \) degrees of freedom. For further discussion, see Box and Tiao (1973).

**Approximation: Subject Variance Component**  
Box and Tiao (1973) provide the following result:

\[
\frac{m_1 + n_r \sigma^2 u}{\nu_2' m_2'} \sim \chi_{\nu_2'}^{-2}/
\]  
(4.8)

To approximate a random draw from the posterior distribution of \( \sigma^2 u \) given \( Y_{obs} \), we calculate

\[
\sigma^2 u = \frac{\nu_2' m_2' x^2 - m_1}{n_r},
\]  
(4.9)

where \( x \) is a random variate generated from the \( \chi^{-2} \) distribution on \( \nu_2' \) degrees of freedom. For further discussion, see Box and Tiao (1973).
Approximation: Interaction Variance Component  Box and Tiao (1973) provide the following result:

\[
\frac{m_2'' + n_s n_r \sigma_c^2}{v_3 m_3} \sim \chi_{v_3}^{-2}.
\] (4.10)

To approximate a random draw from the posterior distribution of \( \sigma_c^2 \) given \( Y_{obs} \), we calculate

\[
\sigma_c^2 = \frac{v_3 m_3 x^2 - m_2''}{n_s n_r},
\] (4.11)

where \( x \) is a random variate generated from the \( \chi^{-2} \) distribution on \( v_3 \) degrees of freedom.

For further discussion, see Box and Tiao (1973).

Negative Estimate of a Variance Component  It should be noted that this approximation algorithm can yield a negative estimate for any variance component. For this simulation, we substitute any such negative estimate with a value of 0.0. Should this occur with a natural incomplete data set, we would consider an alternative model for the data. In simulation, however, we do find that a negative subject or interaction variance component estimate is occasionally obtained, even though the data are constructed to follow the two-way mixed interaction model. From this we may conclude that it is not correct to assume that a negative variance component is an indicator of improper model selection.

4.7.3  Step Four: Mardia, Kent, and Bibby’s \([Y_{mis} | Y_{obs}]\)

Usage of \([Y_{mis} | Y_{obs}]\)  RMNI is distinguished from the Fully Normal Imputation Procedure of Rubin (1987) by use of the conditional distribution \([Y_{mis} | Y_{obs}]\) derived and presented by Mardia, Kent, and Bibby (1979). This conditional distribution is specified in Section 4.2.2 with Equation (4.5). Usage of this result requires the partitioning of the data.
into two subvectors, and the partitioning of the variance-covariance matrix into four submatrices. We need to emphasize that our usage of the result does not strictly partition the data as it occurs naturally. Therefore, we include here a few comments regarding the partitions of the data to which we apply the result of Mardia, Kent, and Bibby (1979). In particular, the following examples are provided in demonstration of the algorithm we use to maintain software programs which are robust to observed pattern of missingness.

**Partitions of \( Y_{inc} \)** Note that writing \( Y_{inc} = (Y_{obs}, Y_{mis})' \) partitions the data set into distinct groups, tantamount to a permutation of the data. Since missing data does not conveniently occur for our first \( n_{mis} \) observations of the data vector \( Y_{inc} \), we must must sort the data into two subvectors, \( Y_{mis} \) and \( Y_{obs} \). The grouping \( Y_{mis} \) is MCAR, and comprises observations of the treatment/subject combinations which vary with each natural observation of \( Y_{mis} \) under our model. For example, \( Y_{mis} \) may include one missing observation per treatment/subject combination, and \( Y_{mis} \) may include two missing observations on each of only three possible treatment/subject combinations.

**Partitions of \( \text{Var}(Y_{inc}) \)** In the notation of Mardia, Kent, and Bibby (1979), the variance-covariance matrix of \( Y_{mis} \) is represented by \( \Sigma_{22} \), and the variance-covariance matrix of \( Y_{obs} \) is represented by \( \Sigma_{11} \), where \( \Sigma_{11} \) and \( \Sigma_{22} \) are partitions of \( \Sigma_{inc} \), the variance-covariance matrix of \( Y_{inc} \). For reference we have

\[
\Sigma_{inc} = \begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix}.
\] (4.12)
Sample $\Sigma_{inc}$  For discussion, we include here a sample variance-covariance matrix to emphasize the complexity of the structure we intend to preserve with RMNI. Let $n_I=2$ (treatments), $n_u=2$ (subjects), and $n_c=3$ (repetitions); let the data $Y_{inc}$ be written as

$$Y_{inc} = \begin{pmatrix} y_{111} \\ y_{112} \\ y_{113} \\ y_{121} \\ y_{122} \\ y_{123} \\ y_{211} \\ y_{212} \\ y_{213} \\ y_{221} \\ y_{222} \\ y_{223} \end{pmatrix}$$

(4.13)

It follows under model (3.4) that $\text{Var}(Y_{inc})$ is given by

$$\Sigma_{inc} = \begin{pmatrix} A & B & B & 0 & 0 & 0 & C & C & C & 0 & 0 & 0 \\ B & A & B & 0 & 0 & 0 & C & C & C & 0 & 0 & 0 \\ B & B & A & 0 & 0 & 0 & C & C & C & 0 & 0 & 0 \\ 0 & 0 & 0 & A & B & B & 0 & 0 & 0 & C & C & C \\ 0 & 0 & 0 & B & A & B & 0 & 0 & 0 & C & C & C \\ 0 & 0 & 0 & B & B & A & 0 & 0 & 0 & C & C & C \\ C & C & C & 0 & 0 & 0 & A & B & B & 0 & 0 & 0 \\ C & C & C & 0 & 0 & 0 & B & A & B & 0 & 0 & 0 \\ C & C & C & 0 & 0 & 0 & B & B & A & 0 & 0 & 0 \\ 0 & 0 & 0 & C & C & C & 0 & 0 & 0 & A & B & B \\ 0 & 0 & 0 & C & C & C & 0 & 0 & 0 & B & A & B \\ 0 & 0 & 0 & C & C & C & 0 & 0 & 0 & B & B & A \end{pmatrix}$$

(4.14)

where $A = \sigma_u^2 + \sigma_c^2 + \sigma_e^2$, $B = \sigma_u^2 + \sigma_c^2$, and $C = \sigma_u^2$.

Sample Permutation of $Y_{inc}$  In (4.13) and (4.14), we have arranged the data by order of treatment/subject combination so that $Y_{inc}$ is a series of subvectors $Y_{(ij)}$, where $i = 1, 2, \ldots, n_I$ and $j = 1, 2, \ldots, n_s$. Yamashita (1995) considers $Y_{inc}$ to be a series of subvectors $Y_{(j)}$ where $j = 1, 2, \ldots, n_s$. In contrast to the permutation explored in Yamashita (1995), we
consider $Y_{inc}$ and $\Sigma_{inc} = \text{Var}(Y_{inc})$ given by

$$Y_{inc} = \begin{pmatrix} y_{111} \\ y_{112} \\ y_{113} \\ y_{121} \\ y_{122} \\ y_{123} \\ y_{211} \\ y_{212} \\ y_{213} \\ y_{221} \\ y_{222} \\ y_{223} \end{pmatrix}$$

\hspace{1cm} (4.15)

and

$$\Sigma_{inc} = \begin{pmatrix} A & B & B & C & C & C & 0 & 0 & 0 & 0 & 0 & 0 \\ B & A & B & C & C & C & 0 & 0 & 0 & 0 & 0 & 0 \\ B & B & A & C & C & C & 0 & 0 & 0 & 0 & 0 & 0 \\ C & C & A & B & B & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ C & C & C & B & A & B & 0 & 0 & 0 & 0 & 0 & 0 \\ C & C & C & B & B & A & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & A & B & B & C & C & C \\ 0 & 0 & 0 & 0 & 0 & 0 & B & A & B & C & C & C \\ 0 & 0 & 0 & 0 & 0 & 0 & B & B & A & C & C & C \\ 0 & 0 & 0 & 0 & 0 & 0 & C & C & A & B & B & B \\ 0 & 0 & 0 & 0 & 0 & 0 & C & C & C & B & A & B \\ 0 & 0 & 0 & 0 & 0 & 0 & C & C & C & B & B & A \end{pmatrix}$$

\hspace{1cm} (4.16)

where $A = \sigma_u^2 + \sigma_c^2 + \sigma_e^2$, $B = \sigma_u^2 + \sigma_c^2$, and $C = \sigma_u^2$.

**Two-Way No-Interaction Model and $\Sigma_{inc}$** It is visibly apparent with Matrices (4.14) and (4.16) that a permutation of observations in $Y_{inc}$ impacts the structure of $\text{Var}(Y_{inc})$. We believe it is instructive also to display the impact of the interaction effect on the structure of $\text{Var}(Y_{inc})$. Under Model (3.4) with $\sigma_c^2$ set to zero, we obtain the following
result for $Y_{inc}$ and $\text{Var}(Y_{inc})$, where $Y_{inc}$ is constructed as in Expression (4.15) above:

\[
\Sigma_{inc} = \begin{pmatrix}
A & C & C & C & C & 0 & 0 & 0 & 0 & 0 & 0 \\
C & A & C & C & C & 0 & 0 & 0 & 0 & 0 & 0 \\
C & C & A & C & C & 0 & 0 & 0 & 0 & 0 & 0 \\
C & C & C & A & C & 0 & 0 & 0 & 0 & 0 & 0 \\
C & C & C & C & A & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & A & C & C & C & C & C \\
0 & 0 & 0 & 0 & 0 & C & A & C & C & C & C \\
0 & 0 & 0 & 0 & 0 & C & C & A & C & C & C \\
0 & 0 & 0 & 0 & 0 & C & C & C & A & C & C \\
0 & 0 & 0 & 0 & 0 & C & C & C & C & A & C \\
0 & 0 & 0 & 0 & 0 & C & C & C & C & C & A
\end{pmatrix}
\] (4.17)

where $A = \sigma_u^2 + \sigma_e^2$; $B = C = \sigma_u^2$.

$Y_{mis}$ Permutation and $\Sigma_{inc}$ Given the sample structures for $\Sigma_{inc}$ above, we point out that permuting the observations of $Y_{inc}$ to accommodate a partition $Y_{mis}$ has variable impact upon the structure of partition $\Sigma_{11}$ of $\Sigma_{inc}$. That is, in the construction of our imputed values, we need an algorithm for imputation which is robust to the observed pattern of missing data, in terms of its impact upon the observed structure of the variance-covariance matrix of $Y_{obs}$. In Yamashita (1995), there is the convenience that any permutation of observations within $Y_{(j)}$ will yield the same structure of covariance for $\text{Var}(Y_{(j)})$, as evident with Matrix (4.17). In the presence of an interaction effect with repeated measures, this convenience is absent, as evident with Matrices (4.14) and (4.16). We do have, however, that any permutation of observations within $Y_{(ij)}$ will yield the same structure of covariance for $\text{Var}(Y_{(ij)})$. Ultimately, we are interested in the structure of $\text{Var}(Y_{mis})$, as it is component to the formulation of Mardia, Kent, and Bibby’s $[Y_{mis} | Y_{obs}]$ (1979).
Closed-Form Calculating Expression for $[Y_{mis} | Y_{obs}]$  Given the general expression of $[Y_{mis} | Y_{obs}]$ provided by Mardia, Kent, and Bibby (1979), it is fair to consider that an expression for a calculating formula of $[Y_{mis} | Y_{obs}]$ might be obtainable given a particular set of model conditions. Yamashita (1995) accomplishes this for the One-Way Layout with repeated measures, and equivalently for the two-way no-interaction model, by focusing on $Y(j)$ and $\text{Var}(Y(j))$ for each subject $j$. This approach fails in the presence of an interaction term because the structure of $\text{Var}(Y(j))$ is dependent upon the ordering of observations within $Y(j)$, as evident with Matrices (4.14) and (4.16). We do consider the Model (3.4) analog of the process presented by Yamashita (1995), which is to consider $Y(ij)$ and $\text{Var}(Y(ij))$: the structure of $\text{Var}(Y(ij))$ is unchanged by permutations of repetitions within $Y(ij)$. Using standard distribution theory and Equation (4.5), we obtain the following result:

$$
\text{Var}(Y(ij),mis | Y(ij),obs) = \frac{\sigma_e^2(\sigma_u^2 + \sigma_c^2)}{\sigma_e^2 + n(i,j),obs(\sigma_u^2 + \sigma_c^2)} J(n(i,j),mis)x(n(i,j),mis) + \frac{\sigma_c^2 I(n(i,j),mis)x(n(i,j),mis)}{(n(i,j),mis)} \tag{4.18}
$$

where $n(i,j),obs$ is the number of repetitions observed on treatment/subject combination $(i,j); i=1, \ldots, n_t$; and $j=1, \ldots, n_s$, and $n(i,j),mis$ is the number of repetitions missing.

Note on Calculating Formula (4.18) Regrettably, Expression (4.18) is not particularly useful. On its own, it would yield an imputed value or two for missing replicates of a given combination, based only upon observed replicates under that one combination. We did hope to generalize this calculating form to $\text{Var}(Y_{mis})$, but such derivation is intractible, as it depends upon the pattern of the missing data, which varies from one natural incomplete data set to the next.
Reduction of $Y_{inc}$ and $\Sigma_{inc}$ To illustrate the task of the automated algorithm for Step Five (Section 4.2.2) of RMNI under our model, we provide the following example: if $Y_{inc}$ is given by

$$Y_{inc} = \begin{pmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \end{pmatrix}$$  \quad (4.19)$$

with $\text{Var}(Y_{inc})$ given by

$$\Sigma_{inc} = \begin{pmatrix} A & B & C & C & 0 & 0 & 0 & 0 \\ B & A & C & C & 0 & 0 & 0 & 0 \\ C & C & A & B & 0 & 0 & 0 & 0 \\ C & C & B & A & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & A & B & C & C \\ 0 & 0 & 0 & 0 & B & A & C & C \\ 0 & 0 & 0 & 0 & C & C & A & B \\ 0 & 0 & 0 & 0 & C & C & B & A \end{pmatrix}$$  \quad (4.20)$$

where $A = \sigma_u^2 + \sigma_c^2 + \sigma_e^2$; $B = \sigma_u^2 + \sigma_c^2$; and $C = \sigma_u^2$; we see that the variance-covariance matrix associated with the $Y_{inc}$ of (4.19) is easily obtained from the variance-covariance matrix for $Y_{inc}$ of (4.15), by deleting the corresponding rows and columns of Matrix (4.16) to obtain Matrix (4.20). That is, every third observation of $Y_{inc}$ (4.15) and every third row and column of $\text{Var}(Y_{inc})$ (4.16) are eliminated to obtain (4.19) and (4.20). Note that the removal of the third replicate from the model, as displayed with Expressions (4.15) and (4.19), is analogous to the observance of missing replicates with incomplete data. We proceed to make this analogy more precise.

Sample $Y_{mis}$ and $\text{Var}(Y_{mis})$ Assume $Y_{inc}$ and $\text{Var}(Y_{inc})$ are as provided with Expressions (4.19) and (4.20). Also assume we observe missingness with two intended observations, say $y_{212}$ and $y_{121}$, so that $Y_{mis} = [y_{212}, y_{121}]'$. The locations of $y_{212}$ and $y_{121}$ are as the third
and fourth elements of $Y_{inc}$. By removal of the corresponding (third and fourth) rows and columns of $\text{Var}(Y_{inc})$, we obtain $\Sigma_{11}$, the variance-covariance matrix of $Y_{obs}$. Since $y_{212}$ and $y_{121}$ share no commons subject effect, it is directly observable that

$$
\Sigma_{22} = \begin{pmatrix}
A & 0 \\
0 & A
\end{pmatrix}
$$

(4.21)

where $A = \sigma^2_u + \sigma^2_c + \sigma^2_\epsilon$. Permuting the observations of $Y_{inc}$ allows us to write

$$
Y_{inc*} = \begin{pmatrix}
Y_{obs} \\
Y_{mis}
\end{pmatrix}
$$

(4.22)

By direct derivation, we obtain $\text{Var}(Y_{inc*})$ given by

$$
\Sigma_{inc*} = \begin{pmatrix}
A & B & C & 0 & 0 & 0 & C & 0 \\
A & C & 0 & 0 & 0 & C & 0 & 0 \\
A & 0 & 0 & 0 & B & 0 & 0 & 0 \\
A & C & C & 0 & B & 0 & 0 & 0 \\
A & B & 0 & C & 0 & 0 & 0 & 0 \\
A & 0 & C & C & 0 & B & 0 & 0 \\
A & 0 & C & C & 0 & B & 0 & 0 \\
\end{pmatrix}
$$

(4.23)

where $A = \sigma^2_u + \sigma^2_c + \sigma^2_\epsilon$; $B = \sigma^2_u + \sigma^2_\epsilon$; and $C = \sigma^2_u$. If we observe $Y_{mis} = [Y_{121}, Y_{222}]'$, we have

$$
\Sigma_{22*} = \begin{pmatrix}
A & C \\
C & A
\end{pmatrix}
$$

(4.24)

where $A = \sigma^2_u + \sigma^2_c + \sigma^2_\epsilon$ and $C = \sigma^2_u$. Finally, if we observe $Y_{mis} = [Y_{121}, Y_{122}]'$, we have

$$
\Sigma_{22*} = \begin{pmatrix}
A & B \\
B & A
\end{pmatrix}
$$

(4.25)

where $A = \sigma^2_u + \sigma^2_c + \sigma^2_\epsilon$ and $B = \sigma^2_u + \sigma^2_\epsilon$. Note that we adopt the notation $\Sigma_{22*}$ to denote $\text{Var}(Y_{mis})$ and distinguish the partition $\Sigma_{22*}$ from the generic $\Sigma_{22}$ of Equations (4.5) and (4.12).
Summary of Sample $\Sigma_{22}$ It is notable that the structure of $\Sigma_{22}$ will vary with both pattern of missingness and number of missing values. Most notably, coding an algorithm that is robust to pattern of missingness for application of Mardia, Kent, and Bibby’s $[Y_{mis}|Y_{obs}]$ (Equation 4.5) needs to estimate $\text{Var}(Y_{inc*})$; this algorithm cannot be based reliably upon a permutation rows and columns of $\text{Var}(Y_{inc})$. Given this characterization of the complexity of $\text{Var}(Y_{inc*})$, we present the following algorithm for obtaining random variates from $[Y_{mis}|Y_{obs}]$, which are to be used as imputed values.

4.7.4 Step Five: Generating Random Draws from $[Y_{mis}|Y_{obs}]$

Obtain $Y_{inc*}$ and $\Sigma_{inc*}$ First we must obtain $Y_{inc}$ as a vector of intended observations for the experiment partitioned into two subvectors $Y_{mis}$ and $Y_{obs}$: $Y_{inc*} = (Y_{obs}; Y_{mis})'$. Once $Y_{inc}$ is permuted in this way, we must obtain $\text{Var}(Y_{inc*})$. The structure of $\text{Var}(Y_{inc*})$ is dependent upon the pattern of missing data, and so no closed-form expression for $\text{Var}(Y_{inc*})$ can be provided. Alternatively, we propose an iterative algorithm for generating $\text{Var}(Y_{inc*})$ once $Y_{inc*}$ has been clearly determined.

Requirements To evaluate $\Sigma_{inc*} = \text{Var}(Y_{inc*})$ for a given $Y_{inc*}$, we need the specification of model assumptions, and a random draw from the posterior distribution of each variance component, as obtained in Step Three using Box and Tiao’s (1973) approximations. Each imputation will require an independent set of draws from Step Three; multiple imputation repeats this step $m$ times to obtain $m$ completed data sets. Let $\sigma^2 = (\sigma^2_{u\ell}, \sigma^2_{c\ell}, \sigma^2_{e\ell}), \ell = 1, 2, \ldots, m$, denote the $\ell^{th}$ requisite random posterior draw. Using the model assumptions, we can construct $\text{Cov}(y_{ijk}, y'_{i'j'k'})$ for each cell of the desired variance-covariance matrix, $\Sigma_{inc*}$. 

95
Algorithm for Constructing $\Sigma_{inc*}$  

We choose to code a FORTRAN subroutine to iteratively populate each cell of $\Sigma_{inc*}$ by observing the indices on the pair $(y_{ijk}, y'_{i'j'k'})$ yielding a covariance for that cell.

- First Pass: If $i=i'$ and $j=j'$ and $k=k'$, then populate the cell with $\hat{\sigma}^2_{\ell\ell}$; otherwise, populate the cell with a zero.
- Second Pass: If $i=i'$ and $j=j'$, then add $\hat{\sigma}^2_{\ell\ell}$ to the cell value.
- Third Pass: If $j=j'$, then add $\hat{\sigma}^2_{u\ell}$ to the cell value.

Note that this process is repeated $m$ times, as indexed by $\ell$, generating a value for $\Sigma_{inc*}$ for each of the $m$ imputations to be performed; we will refer to each obtained value as $\Sigma_{inc*\ell}$, where $\ell = 1, \ldots, m$.

Partitioning $\Sigma_{inc*}$  

Once $\Sigma_{inc*\ell}$ is evaluated using the $\ell^{th}$ set of variance component posterior draws, we partition $\Sigma_{inc*\ell}$ easily into the four submatrices required by Mardia, Kent, and Bibby’s $[Y_{mis}|Y_{obs}]$ (Equation 4.5). We use IMSL subroutines and standard mathematics to obtain $\mathbb{E}[Y_{mis}|Y_{obs}]$ and $\text{Var}(Y_{mis}|Y_{obs})$; we use IMSL subroutines to obtain random draws from the Multivariate Normal Distribution with mean $\mathbb{E}[Y_{mis}|Y_{obs}]$ and variance $\text{Var}(Y_{mis}|Y_{obs})$. An incomplete-data initial estimate of $\theta$ will be required to obtain the final result (Equation 4.5).

Note on Imputed Values  

The $n_{mis}$-variate draw from

$$Y_{mis} \sim \mathcal{N}(\mathbb{E}[Y_{mis}|Y_{obs}], \text{Var}(Y_{mis}|Y_{obs}))$$  \hspace{1cm} (4.26)

is supplied to $Y_{inc*}$ as an analyzable substitute for $Y_{mis}$. $Y_{inc*}$ is then counter-permuted to its original ordering $Y_{inc}$, from which the usual Two-Way Layout table may be obtained.
This is accomplished for each of \( m \) multiple imputations, and yields \( m \) completed data sets. In Steps Six, Seven, and Eight of RMNI, these \( m \) data sets may be analyzed using standard complete-data methods, with results being combined and interpreted per guidelines set by Rubin (1987).

### 4.8 Programming for Simulation and Analysis

**Appendix of Programs** Appendices A through H include every significant program, subroutine, and shell script used to perform RMNI, conduct the simulation, and compute analysis and display graphics of the multiply-imputed data. Appendix A contains the Master Program Files used for RMNI and an alternative method of multiple imputation based upon model assumptions, which was explored in the General Exam for doctoral candidacy. Intermediary appendices attempt to sort the programs by function, where each appendix corresponds to a single step of RMNI, as outlined in Section 4.2.2 and detailed for our model in Section 4.7 of this dissertation. For quick reference, it is best to use the Table of Contents to locate a program by name in the listing. Also, program descriptions are provided with each program, as well as within coded comments.

**Resource Applications** This simulation was conducted using a UNIX environment on SUN servers with the following utilities: FORTRAN, IMSL, KORN Shell Script, C, SAS, and SPlus. Programming would be greatly simplified by an intended use for one multiple imputation with a natural data set. The programs can be extended to function for the family of hierarchical models with random effects. Further comments are provided in Section 8.3.
4.9 Evolution of RMNI

Fully Normal Imputation The foundations for RMNI are described in Section 4.2.1, and focus squarely on the Fully Normal Multiple Imputation method explored by Rubin (1987). It is this method which has been enhanced by results of Box and Tiao (1973) and Mardia, Kent, and Bibby (1979) to schematize the method of RMNI. RMNI was first discussed by Yamashita (1995) as \textit{RMNI}, and the name refers to a process of model-based imputing, as intended to preserve the underlying variance-covariance structure imposed by repeated measures in an experimental design.

Yamashita’s RMNI Yamashita (1995) lays the groundwork and explores the potential for RMNI within the One-Way Layout. Within this context, Dr. Yamashita explores the performance of Dunnett’s Procedure for multiple comparisons using multiply-imputed data with success. At this stage, RMNI is established for use in the context of hierarchical mixed models with repeated measures. Yamashita (1995) does not fully explore the impact of imputed values upon the underlying variance covariance structure of the model. It is evident that his usage of RMNI does preserve this underlying structure, and so RMNI was considered thereafter to be a desirable option for multiply imputing two-way mixed data in the presence of an interaction term. Particularly, RMNI is desirable for preserving the covariance structure requisite of Multiple Comparisons Procedures with mixed model data.

General Exam Method Our first attempt at RMNI with two-way mixed interaction data did utilize Box and Tiao’s approximation methods to obtain imputed values. However, at first we did not utilize Mardia, Kent, and Bibby’s result for $[Y_{mis}|Y_{obs}]$. Alternatively, we used draws obtained in Step Three of RMNI to construct imputed values directly using
the model assumptions for the missing value. Given an initial estimate for $\theta$ and posterior draws $\hat{\sigma}_u$, $\hat{\sigma}_c$, and $\hat{\sigma}_e$, we constructed an imputed value as follow:

1. Obtain a random draw, $s \sim \mathcal{N}(0, \hat{\sigma}_u^2)$. 
2. Obtain a random draw, $c \sim \mathcal{N}(0, \hat{\sigma}_c^2)$. 
3. Obtain a random draw, $e \sim \mathcal{N}(0, \hat{\sigma}_e^2)$. 
4. Obtain an imputed value as $y_i = \hat{\theta}_i + s + c + e$. 
5. Repeat.

While this algorithm was much easier to code than Result (4.5), it did not fully utilize the observed data except to calculate initial estimates for the treatment effects and variance components. Programs used specifically to demonstrate this procedure are provided in Appendices A and F.

**This Dissertation** The generation of RMNI discussed and presented in this dissertation has the advantages of fully incorporating observed data into the imputation process. The primary benefit of this is to carefully preserve the delicate covariance structure of the two-way mixed interaction model, as displayed in Section 4.7.3. The importance of preserving this underlying structure is discussed further in Chapter Six. In Chapter Seven, we compare our RMNI approach to the popularized Multivariate Normal Methods approach available in commercial software today.
CHAPTER 5

Simulated Demonstration of RMNI

Synopsis We simulate a complete data set following the two-way mixed interaction model, and we apply a simulated MCAR pattern of missingness to obtain an incomplete data set. The missing values are replaced with imputed values generated with the procedure of Repeated Measures Normal Imputation, as described in Chapter Four. Using $m=1000$ sets of imputed values, we create $m=1000$ completed data sets, and assess the stability of the multiply-imputed estimate of $\theta$. We find that with a missingness rate of ten percent, the number of imputations necessary for a stable RMNI-imputed estimate of $\theta$ is roughly $m_{RMNI}=400$. Furthermore, we find that it is evident that the RMNI imputation procedure is proper for performing multiple comparisons of treatment effects.

5.1 Basis for Comparison: Complete and Incomplete Data

Complete Data We begin with the simulation of a complete data set following the Two-Way Layout as specified with Model (3.4). For the simulation study presented in this chapter, the model parameters are set as listed in Table 5.1. We choose the treatment effects and variance components for clarity of interpretation of output: the standard deviation for an observation $\sigma_y$ is approximately 17.32, and our treatment effect parameter values are approximately six standard deviations of an observation apart. In this way, we are able
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Simulated Parameter Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect: Treatment One</td>
<td>$\theta_1$</td>
</tr>
<tr>
<td>Effect: Treatment Two</td>
<td>$\theta_2$</td>
</tr>
<tr>
<td>Effect: Treatment Three</td>
<td>$\theta_3$</td>
</tr>
<tr>
<td>Subject Variance Component</td>
<td>$\sigma^2_u$</td>
</tr>
<tr>
<td>Interaction Variance Component</td>
<td>$\sigma^2_c$</td>
</tr>
<tr>
<td>Error Variance Component</td>
<td>$\sigma^2_e$</td>
</tr>
<tr>
<td>Observation Variance</td>
<td>$\sigma^2_y$</td>
</tr>
<tr>
<td>Observation Standard Deviation</td>
<td>$\sigma_y$</td>
</tr>
<tr>
<td>Number of Treatments</td>
<td>$n_t$</td>
</tr>
<tr>
<td>Number of Subjects</td>
<td>$n_s$</td>
</tr>
<tr>
<td>Number of Replicates</td>
<td>$n_r$</td>
</tr>
</tbody>
</table>

Table 5.1: Parameter Values for Simulation of the Two-Way Mixed Model

to clearly distinguish whether an imputed value is strategically placed with a treatment observation of comparable magnitude. For clarification, see the RMNI-imputed values provided in Tables 5.4, 5.5, and 5.6.

Incomplete Data Once the complete data set is available to us, we simulate a pattern of missingness which is classified as MCAR. Using the incomplete data, we use unbalanced-data methods to compute estimates of the treatment effects, as well as initial estimates of the variance components. These initial estimates serve two purposes: first, they are used as seeds for the Box and Tiao (1973) approximation algorithms provided in Section 4.7.2; second, they serve as a basis for comparison with the multiply-imputed estimates of treatment effects and associated error, which are computed with complete-data estimators using RMNI-imputed data. The initial values computed with incomplete data are provided
Table 5.2: Baseline Comparison Estimates of Model Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Computed Using $Y_{inc}$</th>
<th>Computed Using $Y_{obs}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>95.73547</td>
<td>95.66685</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>198.4444</td>
<td>197.6296</td>
</tr>
<tr>
<td>$\theta_3$</td>
<td>298.8291</td>
<td>299.1122</td>
</tr>
<tr>
<td>$\sigma_u^2$</td>
<td>85.60878 (9.25)</td>
<td>86.53203 (9.30)</td>
</tr>
<tr>
<td>$\sigma_e^2$</td>
<td>87.73600 (9.37)</td>
<td>88.50552 (9.41)</td>
</tr>
<tr>
<td>$\sigma_\eta^2$</td>
<td>101.5000 (10.07)</td>
<td>102.0600 (10.1)</td>
</tr>
<tr>
<td>$\sigma_{y^2}$</td>
<td>$\approx$ 274.85</td>
<td>$\approx$ 277.10</td>
</tr>
<tr>
<td>$\sigma_y$</td>
<td>$\approx$ 16.58</td>
<td>$\approx$ 16.65</td>
</tr>
<tr>
<td></td>
<td>Complete-Data Estimate</td>
<td>Incomplete-Data Estimate</td>
</tr>
</tbody>
</table>

in Table 5.2, along with their complete-data counterparts. The simulated complete data set is referred to as $Y_{inc}$, and the simulated incomplete data set is referred to as $Y_{obs}$.

5.2 Pilot Run of RMNI

**Scope** We begin exploration of RMNI-imputed data with a pilot run of $m=5$ imputations. With this pilot data, we are able to observe clearly the functioning of our RMNI programs, as well as the nature of the RMNI imputed value. With this manageable number of imputations, we display the results of performing the imputation, computing interim completed-data estimates of parameters, and combining the RMNI-imputed parameter estimates into multiply-imputed estimates of the treatment effects and associated error.

**Completed Data** For the pilot run of RMNI, we choose to simulate a missingness rate of ten percent, yielding (by chance) 45 missing values. No subjects were deleted with this simulation because the pattern of missing data did not present any completely unobserved...
subject. Therefore, we have 45 missing values with 450 intended observations. The missing
data are referred to as $Y_{mis}$, and the observed data are referred to as $Y_{obs}$. A sample of the
completed data sets is provided in Table 5.3.

Table 5.3  With Table 5.3, we display select portions of the RMNI-imputed data sets from
a pilot run with $m=5$. The table has three sections corresponding to each of three treat-
ment effects. Notice that the RMNI-imputed values vary from one imputation to the next.
Notice, in particular, that the RMNI-imputed value is tailored to the treatment/subject com-
bination being imputed: for example, look carefully at the imputed values for observations
$y_{143}$ and $y_{151}$. For each missing repetition, imputed values are variable, yet on the order of
magnitude of the other repetitions for the same treatment/subject combination. Not only
do the RMNI-imputed values appear reasonable for the magnitude of the treatment ef-
fect, but they genuinely reflect variation between and within subjects. The imputed values
for Subject One under Treatment Two demonstrate the attempt of RMNI to capture ob-
served variability within an individual subject’s response values, and reflect that distinctive
variability with imputed values. In this way, the RMNI procedure preserves the delicate
structure of the underlying variance-covariance matrix of $Y_{inc}$.

Table 5.4  With Table 5.4, we display the RMNI-imputed values for each missing datum
under Treatment One. The missing data pattern is provided in detail by the columns which
identify treatment, subject, and replication, for each missing value. Notice that the imputed
values under Treatment One are tailored to the simulated parameter value of $\theta_1=100$ by the
RMNI procedure: this is due to an initial estimate of $\theta_1$ computed using $Y_{obs}$, which is
used as a seed for the generation of imputed values. The variability of the RMNI-imputed
<table>
<thead>
<tr>
<th>Status</th>
<th>RMNI-1</th>
<th>RMNI-2</th>
<th>RMNI-3</th>
<th>RMNI-4</th>
<th>RMNI-5</th>
<th>Trt</th>
<th>Sub</th>
<th>Rep</th>
</tr>
</thead>
<tbody>
<tr>
<td>95.86371</td>
<td>95.86371</td>
<td>95.86371</td>
<td>95.86371</td>
<td>95.86371</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>79.34015</td>
<td>79.34015</td>
<td>79.34015</td>
<td>79.34015</td>
<td>79.34015</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>83.49093</td>
<td>83.49093</td>
<td>83.49093</td>
<td>83.49093</td>
<td>83.49093</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>74.73473</td>
<td>74.73473</td>
<td>74.73473</td>
<td>74.73473</td>
<td>74.73473</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>86.0107</td>
<td>86.0107</td>
<td>86.0107</td>
<td>86.0107</td>
<td>86.0107</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>103.8634</td>
<td>103.8634</td>
<td>103.8634</td>
<td>103.8634</td>
<td>103.8634</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>107.7222</td>
<td>107.7222</td>
<td>107.7222</td>
<td>107.7222</td>
<td>107.7222</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>76.00217</td>
<td>76.00217</td>
<td>76.00217</td>
<td>76.00217</td>
<td>76.00217</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>103.1599</td>
<td>103.1599</td>
<td>103.1599</td>
<td>103.1599</td>
<td>103.1599</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>96.29846</td>
<td>96.29846</td>
<td>96.29846</td>
<td>96.29846</td>
<td>96.29846</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>114.3153</td>
<td>114.3153</td>
<td>114.3153</td>
<td>114.3153</td>
<td>114.3153</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>106.426</td>
<td>106.426</td>
<td>106.426</td>
<td>106.426</td>
<td>106.426</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>77.26614</td>
<td>77.26614</td>
<td>77.26614</td>
<td>77.26614</td>
<td>77.26614</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>75.538</td>
<td>75.538</td>
<td>75.538</td>
<td>75.538</td>
<td>75.538</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>73.0075</td>
<td>73.0075</td>
<td>73.0075</td>
<td>73.0075</td>
<td>73.0075</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>192.6409</td>
<td>192.6409</td>
<td>192.6409</td>
<td>192.6409</td>
<td>192.6409</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>181.9324</td>
<td>181.9324</td>
<td>181.9324</td>
<td>181.9324</td>
<td>181.9324</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>166.3579</td>
<td>166.3579</td>
<td>166.3579</td>
<td>166.3579</td>
<td>166.3579</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>199.8045</td>
<td>199.8045</td>
<td>199.8045</td>
<td>199.8045</td>
<td>199.8045</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>192.4673</td>
<td>192.4673</td>
<td>192.4673</td>
<td>192.4673</td>
<td>192.4673</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>181.4937</td>
<td>181.4937</td>
<td>181.4937</td>
<td>181.4937</td>
<td>181.4937</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>195.9169</td>
<td>195.9169</td>
<td>195.9169</td>
<td>195.9169</td>
<td>195.9169</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>209.2067</td>
<td>209.2067</td>
<td>209.2067</td>
<td>209.2067</td>
<td>209.2067</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>213.0209</td>
<td>213.0209</td>
<td>213.0209</td>
<td>213.0209</td>
<td>213.0209</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>203.9311</td>
<td>203.9311</td>
<td>203.9311</td>
<td>203.9311</td>
<td>203.9311</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>209.2531</td>
<td>209.2531</td>
<td>209.2531</td>
<td>209.2531</td>
<td>209.2531</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>308.8419</td>
<td>308.8419</td>
<td>308.8419</td>
<td>308.8419</td>
<td>308.8419</td>
<td>3</td>
<td>16</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>309.9046</td>
<td>309.9046</td>
<td>309.9046</td>
<td>309.9046</td>
<td>309.9046</td>
<td>3</td>
<td>16</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>314.2614</td>
<td>314.2614</td>
<td>314.2614</td>
<td>314.2614</td>
<td>314.2614</td>
<td>3</td>
<td>16</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.3: RMNI-Imputed Data Sets
values reflects the structure of the variance-covariance matrix as specified by the model (Model 3.4).

Tables 5.5 and 5.6   With Table 5.5 and Table 5.6 on the following page, we further display RMNI-imputed values which are sensible for the simulated parameter value of each treatment effect, and which reflect the underlying variance-covariance matrix as estimated with \( Y_{obs} \) and structured by model design. Notice that with the simulated pattern of missingness, the number of missing observations varies with subject as well as with treatment. With this

<table>
<thead>
<tr>
<th>RMNI-1</th>
<th>RMNI-2</th>
<th>RMNI-3</th>
<th>RMNI-4</th>
<th>RMNI-5</th>
<th>Trt</th>
<th>Sub</th>
<th>Rep</th>
</tr>
</thead>
<tbody>
<tr>
<td>106.426</td>
<td>106.9688</td>
<td>105.3416</td>
<td>105.4184</td>
<td>100.1232</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>77.26614</td>
<td>81.61195</td>
<td>73.5518</td>
<td>70.65361</td>
<td>78.87846</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>89.74035</td>
<td>89.2618</td>
<td>95.01228</td>
<td>86.86656</td>
<td>84.73296</td>
<td>1</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>94.94321</td>
<td>102.7066</td>
<td>105.1422</td>
<td>102.606</td>
<td>104.3782</td>
<td>1</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>101.1116</td>
<td>94.25659</td>
<td>99.69329</td>
<td>104.0334</td>
<td>100.8252</td>
<td>1</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>99.2011</td>
<td>89.97668</td>
<td>97.14843</td>
<td>90.40361</td>
<td>89.87254</td>
<td>1</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>81.91866</td>
<td>82.77619</td>
<td>73.99704</td>
<td>86.99535</td>
<td>86.28274</td>
<td>1</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>119.5053</td>
<td>110.5831</td>
<td>110.2951</td>
<td>116.6231</td>
<td>108.4515</td>
<td>1</td>
<td>26</td>
<td>3</td>
</tr>
<tr>
<td>82.36634</td>
<td>86.14687</td>
<td>91.59434</td>
<td>85.70605</td>
<td>77.29899</td>
<td>1</td>
<td>28</td>
<td>2</td>
</tr>
<tr>
<td>96.24065</td>
<td>95.7733</td>
<td>90.80009</td>
<td>97.67955</td>
<td>98.94714</td>
<td>1</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>98.74939</td>
<td>100.7692</td>
<td>95.20663</td>
<td>97.58227</td>
<td>100.7943</td>
<td>1</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>89.13589</td>
<td>90.82841</td>
<td>89.62483</td>
<td>95.76648</td>
<td>86.51546</td>
<td>1</td>
<td>33</td>
<td>2</td>
</tr>
<tr>
<td>114.1841</td>
<td>109.2202</td>
<td>105.0113</td>
<td>102.6182</td>
<td>108.8136</td>
<td>1</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>93.219</td>
<td>96.22518</td>
<td>95.19788</td>
<td>98.47768</td>
<td>99.4915</td>
<td>1</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>99.63472</td>
<td>99.88519</td>
<td>104.0189</td>
<td>102.331</td>
<td>101.497</td>
<td>1</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>112.8776</td>
<td>108.452</td>
<td>114.0871</td>
<td>108.6188</td>
<td>113.873</td>
<td>1</td>
<td>42</td>
<td>2</td>
</tr>
<tr>
<td>104.414</td>
<td>98.97247</td>
<td>102.7164</td>
<td>99.63391</td>
<td>102.3715</td>
<td>1</td>
<td>46</td>
<td>1</td>
</tr>
<tr>
<td>101.1663</td>
<td>101.1772</td>
<td>93.21192</td>
<td>90.97829</td>
<td>106.7117</td>
<td>1</td>
<td>50</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 5.4: RMNI-Imputed Values with Five Imputations Under Treatment One
<table>
<thead>
<tr>
<th>RMNI-1</th>
<th>RMNI-2</th>
<th>RMNI-3</th>
<th>RMNI-4</th>
<th>RMNI-5</th>
<th>Trt</th>
<th>Sub</th>
<th>Rep</th>
</tr>
</thead>
<tbody>
<tr>
<td>181.9324</td>
<td>183.2208</td>
<td>180.4642</td>
<td>187.2801</td>
<td>182.5555</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>209.2531</td>
<td>204.8001</td>
<td>205.9166</td>
<td>204.2655</td>
<td>205.7271</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>221.085</td>
<td>218.2635</td>
<td>214.606</td>
<td>205.9802</td>
<td>217.8657</td>
<td>2</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>186.4608</td>
<td>185.5432</td>
<td>189.5944</td>
<td>191.0422</td>
<td>187.5957</td>
<td>2</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>218.0578</td>
<td>218.5351</td>
<td>208.5677</td>
<td>218.2704</td>
<td>213.0967</td>
<td>2</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>200.5385</td>
<td>200.4931</td>
<td>202.9422</td>
<td>206.014</td>
<td>203.0194</td>
<td>2</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>214.7753</td>
<td>219.3298</td>
<td>226.026</td>
<td>214.1673</td>
<td>222.1615</td>
<td>2</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>214.6705</td>
<td>219.8162</td>
<td>216.0623</td>
<td>218.2765</td>
<td>215.663</td>
<td>2</td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td>203.0911</td>
<td>206.4412</td>
<td>198.9543</td>
<td>200.7764</td>
<td>202.0696</td>
<td>2</td>
<td>38</td>
<td>2</td>
</tr>
<tr>
<td>187.9848</td>
<td>186.3464</td>
<td>191.8767</td>
<td>191.7425</td>
<td>190.3478</td>
<td>2</td>
<td>39</td>
<td>2</td>
</tr>
<tr>
<td>196.1024</td>
<td>199.6071</td>
<td>193.2928</td>
<td>190.2773</td>
<td>193.0884</td>
<td>2</td>
<td>43</td>
<td>2</td>
</tr>
<tr>
<td>201.2008</td>
<td>202.7258</td>
<td>201.4579</td>
<td>203.2582</td>
<td>199.633</td>
<td>2</td>
<td>44</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5.5: RMNI-Imputed Values with Five Imputations Under Treatment Two

<table>
<thead>
<tr>
<th>RMNI-1</th>
<th>RMNI-2</th>
<th>RMNI-3</th>
<th>RMNI-4</th>
<th>RMNI-5</th>
<th>Trt</th>
<th>Sub</th>
<th>Rep</th>
</tr>
</thead>
<tbody>
<tr>
<td>286.2051</td>
<td>297.3583</td>
<td>290.1886</td>
<td>291.4765</td>
<td>289.4419</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>284.4205</td>
<td>285.7208</td>
<td>275.8711</td>
<td>282.3199</td>
<td>281.8906</td>
<td>3</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>289.6137</td>
<td>292.8304</td>
<td>294.0612</td>
<td>295.2653</td>
<td>297.7617</td>
<td>3</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>282.2361</td>
<td>280.5605</td>
<td>278.3385</td>
<td>279.9991</td>
<td>281.6863</td>
<td>3</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>287.2569</td>
<td>291.346</td>
<td>292.1042</td>
<td>284.0814</td>
<td>291.2207</td>
<td>3</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>308.8419</td>
<td>320.7031</td>
<td>311.642</td>
<td>304.8372</td>
<td>306.0373</td>
<td>3</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>309.9046</td>
<td>304.9449</td>
<td>306.8303</td>
<td>306.5943</td>
<td>299.4432</td>
<td>3</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>282.6641</td>
<td>277.7794</td>
<td>275.9927</td>
<td>279.1695</td>
<td>269.4939</td>
<td>3</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>304.1287</td>
<td>309.918</td>
<td>304.5729</td>
<td>303.357</td>
<td>312.7245</td>
<td>3</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>292.2508</td>
<td>293.8741</td>
<td>295.838</td>
<td>295.4502</td>
<td>297.359</td>
<td>3</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>297.6064</td>
<td>295.0442</td>
<td>290.468</td>
<td>289.8542</td>
<td>292.8084</td>
<td>3</td>
<td>31</td>
<td>2</td>
</tr>
<tr>
<td>292.7441</td>
<td>297.5819</td>
<td>304.867</td>
<td>294.3876</td>
<td>297.6552</td>
<td>3</td>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>307.8444</td>
<td>300.9872</td>
<td>298.2807</td>
<td>309.1489</td>
<td>308.459</td>
<td>3</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td>295.9792</td>
<td>296.3075</td>
<td>300.2728</td>
<td>296.5422</td>
<td>289.509</td>
<td>3</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>290.8327</td>
<td>290.1568</td>
<td>291.2361</td>
<td>292.1383</td>
<td>291.4614</td>
<td>3</td>
<td>41</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5.6: RMNI-Imputed Values with Five Imputations Under Treatment Three
### Table 5.7: RMNI-Imputed LSEs and Multiply-Imputed LSE for $\theta$

<table>
<thead>
<tr>
<th></th>
<th>Treatment One</th>
<th>Treatment Two</th>
<th>Treatment Three</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMNI-1</td>
<td>95.93417</td>
<td>198.0536</td>
<td>298.6178</td>
</tr>
<tr>
<td>RMNI-2</td>
<td>95.82411</td>
<td>198.1201</td>
<td>298.7684</td>
</tr>
<tr>
<td>RMNI-3</td>
<td>95.79785</td>
<td>198.0176</td>
<td>298.6047</td>
</tr>
<tr>
<td>RMNI-4</td>
<td>95.80679</td>
<td>198.0283</td>
<td>298.5651</td>
</tr>
<tr>
<td>RMNI-5</td>
<td>95.85942</td>
<td>198.0380</td>
<td>298.5806</td>
</tr>
<tr>
<td>Combined</td>
<td>95.84310</td>
<td>198.0515</td>
<td>298.6273</td>
</tr>
</tbody>
</table>

Simulation, there is no instance of a completely unobserved treatment/subject combination; however, the RMNI procedure is capable of generating imputed values should this occur.

**Multiply-Imputed Estimate of $\theta$**  The multiply-imputed estimate of the fixed treatment effect is obtained with a two-step process: first, the LSE of $\theta$ is computed using each RMNI-imputed data set; second, the mean of these RMNI-imputed estimates is computed, and treated as the combined estimate described in Section 2.5.2. The combined mean of the $m=5$ RMNI-imputed estimates of $\theta$ is termed the *multiply-imputed estimate* for $\theta$. The RMNI-imputed LSEs for each treatment effect are provided in Table 5.7, along with the combined mean.

**Multiply-Imputed Variance-Covariance of $\hat{\theta}$**  The multiply-imputed estimate of error associated with the multiply-imputed estimate of $\theta$ is computed in two steps as discussed in Section 2.5.2: first, we obtain the within variance; and second we obtain the between variance which accounts for error due to missing data. The within variance of an RMNI-imputed data set is the usual variance associated with the complete-data LSE for $\theta$, given
### Table 5.8: RMNI-Imputed Within Variance and Multiply-Imputed Within Variance

<table>
<thead>
<tr>
<th></th>
<th>Variance</th>
<th>Covariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMNI-1</td>
<td>3.992404</td>
<td>1.703911</td>
</tr>
<tr>
<td>RMNI-2</td>
<td>3.969707</td>
<td>1.681840</td>
</tr>
<tr>
<td>RMNI-3</td>
<td>3.991478</td>
<td>1.684098</td>
</tr>
<tr>
<td>RMNI-4</td>
<td>3.934418</td>
<td>1.705071</td>
</tr>
<tr>
<td>RMNI-5</td>
<td>3.995067</td>
<td>1.700320</td>
</tr>
<tr>
<td>Combined</td>
<td>3.976615</td>
<td>1.695048</td>
</tr>
</tbody>
</table>

The multiply-imputed within variance is the arithmetic mean of the $m=5$ RMNI-imputed within variances. The RMNI-imputed within variance for each completed data set is provided in Table 5.8, along with the combined within variance. Additionally in Table 5.8, we provide the covariance of the treatment effect estimates, which is computed separately with each completed data set.

**Between Variance and Covariance**  With a single imputation, the RMNI-imputed variance and covariance of a single row of Table 5.8 would be used to construct the variance-covariance matrix, $V(\hat{\theta})$. In this way, the single imputation necessarily under-estimates the variance of the estimator $\hat{\theta}$, because it treats the imputed values as real observed data. To adjust $V(\hat{\theta})$ for the use of imputed values, Rubin (1987) develops the Between Variance, $B_m$, with $m \geq 2$, as provided in Section 2.5.2. This component of variance accounts for the use of imputed values in the computation of $\hat{\theta}$. Successive RMNI-imputed between variances are provided in Table 5.9. The between variance computed using the first $m$ imputations is provided in each row of Table 5.9, along with the corresponding value for $m$. Additionally, we provide the between covariance of the treatment effect estimates in series.
Table 5.9: Multiply-Imputed Between Variances and Covariances

<table>
<thead>
<tr>
<th>$m$</th>
<th>$V(\hat{\theta}_1)$</th>
<th>$V(\hat{\theta}_2)$</th>
<th>$V(\hat{\theta}_3)$</th>
<th>$COV(\hat{\theta}_1, \hat{\theta}_2)$</th>
<th>$COV(\hat{\theta}_1, \hat{\theta}_3)$</th>
<th>$COV(\hat{\theta}_2, \hat{\theta}_3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>0.028772</td>
<td>0.003338</td>
<td>0.048473</td>
<td>-0.006139</td>
<td>0.012457</td>
<td>0.006688</td>
</tr>
<tr>
<td>3</td>
<td>0.012787</td>
<td>0.004932</td>
<td>0.021674</td>
<td>-0.002715</td>
<td>0.005533</td>
<td>0.002301</td>
</tr>
<tr>
<td>4</td>
<td>0.008023</td>
<td>0.004644</td>
<td>0.013819</td>
<td>-0.001918</td>
<td>0.003366</td>
<td>0.002091</td>
</tr>
<tr>
<td>5</td>
<td>0.006665</td>
<td>0.004140</td>
<td>0.009980</td>
<td>-0.002222</td>
<td>0.002259</td>
<td>0.001661</td>
</tr>
</tbody>
</table>

with $m$, as updated with each completed data set. Notice that the between variance varies with pattern of missingness observed under each treatment effect. Also, notice that the between variances and between covariances do decrease with an increase in the number of imputations.

**Multiply-Imputed Total Variance**  Ultimately, we obtain the Total Variance (Covariance) as the sum of the Within Variance (Covariance) and the Between Variance (Covariance), for each treatment effect estimate (pair of estimates). Details of this construction are provided in Section 2.5.2. For the pilot run with $m=5$ RMNI imputations, we obtain the total variance

$$T_5 = \begin{pmatrix}
3.98328 & 1.69282 & 1.69730 \\
3.98076 & 1.69671 & \\
3.98660 & & \\
\end{pmatrix} \tag{5.1}$$

where $T_5$ is the multiply-imputed total variance of $\hat{\theta}$, based on $m=5$ imputations.
5.3 RMNI Convergence Test Run with $m=1000$

5.3.1 Stable Multiply-Imputed Estimates with $m=400$

**Convergence with $m=400$ RMNI Imputations**  A primary advantage of the RMNI imputation procedure, in comparison to iterative methods such as Gibbs Sampling, is that there is no need to manually assess convergence of the algorithm. Nevertheless, it is necessary to determine a number of imputations, $m_{RMNI}$, which will guarantee the acquisition of a stable multiply-imputed (combined) estimate of each fixed effect, and each associated component of error. To accomplish this, we observe the behavior of the combined RMNI-imputed estimates for values of $m$ up to 1000. With a missingness rate of ten percent, we establish that $m=400$ imputations is sufficient to achieve stable multiply-imputed treatment means, with associated estimates of error.

**Recommended Number of Imputations**  In the context of Survey Sampling, it is generally accepted that three to five imputations is sufficient to obtain a stable multiply-imputed estimate of a mean. In the context of Experimental Design, it is not reasonable to expect such a large sample of observed data as would be typical of the sample survey. Advancement of computing speed and capacity, and the small sample size of a typical experiment, enable $m=400$ imputations to be accomplished in an amount of time which is suitable for desktop computing. Our algorithm for performing RMNI accomplishes this task in roughly 20 minutes of real time. We suspect that a SAS Macro could perform the task in a much shorter period of time. As recently as 2004, this task would have required hours of computing time on a typical Sun Workstation. In sum, given the recent advances to computing power, we believe this is a timely introduction of RMNI.
5.3.2 Box and Tiao Posterior Draws

Variance Component Posterior Distribution Draws The procedure for RMNI is model-specific at two critical stages: Step Three and Step Five, as discussed in Sections 4.7.2 and 4.7.3. We begin with attention to the results of Step Three of the procedure: the variance components as sampled from their posterior distributions, given observed data (Section 4.7.2). The draws themselves are displayed in Figures 5.1, 5.2, and 5.3. We are particularly interested in the behavior of these draws with successive imputations. Figure 5.4 and Figure 5.5 display the cumulative means and standard deviations, respectively, for each variance component, as successively sampled from the respective approximate
posterior distribution. Notice with Figure 5.5, that the process seems to stabilize with roughly 200 imputations.

**Histograms of Variance Components** Figures 5.1, 5.2, and 5.3 summarize the approximate posterior distribution for each variance component (subject, interaction, and error, respectively). Initial estimates for the variance components are provided in Table 5.2, with simulated parameter values provided in Table 5.1. The values displayed are standard deviations. The true parameter value as simulated is 100.0 for each variance component; that is, each true standard deviation is 10.0.
Figure 5.3: Initial Values for Error Standard Deviation
Figure 5.4: Cumulative Means of Initial Values for Each Variance Component

**Runs Charts for Variance Components**  Figures 5.4 and 5.5 are mean and standard deviation runs charts, with a display for each of the three variance components. The cumulative mean of the repeated draws stabilizes with roughly 200 imputations; the cumulative standard deviation stabilizes with roughly 50 imputations. It is an artifact of the Box and Tiao (1973) approximation algorithm that the subject variance component should stabilize the least rapidly. (Note the scale of the plots to recognize that the subject variance component stabilizes slowly in comparison with the other variance components.)
Figure 5.5: Cumulative Standard Deviations of Initial Values for Each Variance Component
5.3.3 Multiply-Imputed Treatment Means Using RMNI

Estimates of Treatment Effects  With Figures 5.6, 5.7, and 5.8, we display the RMNI-imputed Least Squares Estimates, \( \hat{\theta}_1 \), \( \hat{\theta}_2 \), and \( \hat{\theta}_3 \), respectively. Initial Estimates for the three treatment effects are provided in Table 5.2, with simulated parameter values provided in Table 5.1. The LSEs are normally distributed, as expected, but their means are not expressly determined by the incomplete-data estimates. For example, the incomplete-data estimate for the mean effect of Treatment One is approximately 95.67, with its value centrally located within Figure 5.6; but the incomplete-data estimate for the mean effect of Treatment Two is approximately 197.63, with its value nowhere located within Figure 5.7.
Comparison with the General Exam Method  

With the General Exam Method (Section 4.9), the multiply-imputed means are each normally distributed about the incomplete-data LSE for the respective treatment mean: this begs the question of whether any value is gained by the use of such multiply-imputed treatment means, as the exercise arguably validates the direct use of the incomplete-data estimators. In contrast, the RMNI-imputed treatment means do not suggest use of the incomplete-data means in tandem with multiply-imputed estimates of associated variances and covariances. The RMNI-imputed treatment means are sensible in that they are ostensibly very near the true parameter value, as guided by the observed data. Nevertheless, it is remarkable that the RMNI imputation procedure does have the potential to shift the values of estimates for treatment means, whereas the General Exam Method does not. Notice with Figure 5.7, that the incomplete-data estimate
of $\theta_2$ is off the chart just below at 197.63, and the complete-data estimate computed with $Y_{inc}$ is off the chart just above at 198.44.

**Stabilization of the Multiply-Imputed Treatment Mean** The RMNI-imputed treatment means are computed using each completed data set. The multiply-imputed treatment means using RMNI are the combined means which summarize the $m$ RMNI-imputed means with an arithmetic average, for each treatment effect. In simple terms: the multiply-imputed treatment effect estimate is the average of the $m$ RMNI-imputed treatment means. The cumulative value of the combined mean for Treatment One is provided in Figure 5.9, in series with $m$.  

---

Figure 5.8: Completed-Data LSE for $\theta_3$ Using RMNI
Defunct RMNI Imputations These runs charts for the LSE of each treatment effect, Figures 5.9, 5.10, and 5.11, display the cumulative path of the multiply-imputed treatment mean up to \( m = 986 \) imputations. Though we attempted \( m = 1000 \) imputations, we found that 14 of these are corrupted. A mathematical anomaly causes the IMSL LSGRR subroutine to fail to compute the inverse of a matrix as required by Mardia, Kent, and Bibby’s conditional distribution, \( \left[ Y_{\text{mis}} | Y_{\text{obs}} \right] \), and as specified with Equation (4.5). Given the rarity of this corruption, we have otherwise overlooked the anomaly in presentation of these results; the anomaly will be resolved prior to publication with further simulation.

Treatment Means Run for Treatment Two With three treatment effects means runs to review, we find the means run for Treatment Two to be the most enlightening (Figure
Figure 5.10: Multiply-Imputed LSE for $\theta_2$ Using RMNI

5.10). The plot scale for Figure 5.10 amplifies the process. It is apparent that the pattern of missing data is capable of causing the RMNI procedure to stabilize at different rates for different treatment effects within the same experiment. While the process seems to stabilize with $m=200$ imputations for Treatment One and Treatment Three, the process under Treatment Two seems to require more fully the $m=400$ runs we ultimately recommend. Close inspection of the charts, with comparison to scale, reveals that the estimator of Treatment Two actually stabilizes the most quickly, as it has the most real observed data contributing to the calculation (see Table 5.10). Nevertheless, it is with Figure 5.7 that we clearly display the behavior of the process up to $m=400$ imputations. Given the microscopic scale of Figure 5.7, and other displays supporting the adoption of $m_{RMNI}=400$, we consider our recommendation of 400 imputations to be conservative.
5.3.4 Multiply-Imputed Total Variance Using RMNI

RMNI Completed-Data Variance Components With each RMNI-imputed data set, we compute a set of variance component estimates. As discussed in Section 3.4, we use Type III Sums of Squares, and the associated Expected Mean Squares, as tabulated by the SAS GLM Procedure. The RMNI-imputed observation standard deviation is displayed in Figure 5.12 for each of \( m = 986 \) imputations. The incomplete-data estimate of \( \sigma_y \) is 16.65 (Table 5.2).

Multiply-Imputed Within Variance The multiply-imputed within variance is computed as the arithmetic mean of \( m \) RMNI-imputed values of \( V(\hat{\theta}) \). The details of this computation are provided in Sections 2.5.2 and 3.3.1. The sequence of multiply-imputed within variance
Completed-Data Standard Deviation for a Single Observation

Figure 5.12: RMNI-Imputed $\sigma_y$ with each of 986 Imputations
estimates, $\tilde{U}_m$, is displayed in Figure 5.13, where each plotted value of $\tilde{U}_m$ is calculated using within variances from imputations $\ell=1, 2, \ldots, m$. The scale of the plot amplifies the behavior of the within variance across successive values of $m$: the estimator stabilizes with 200 imputations, and arguably with 100 imputations. Conservatively, the run beyond $m=400$ imputations indicates the most clearly stable estimate of the within variance.

**Multiply-Imputed Within Covariance Using RMNI**  With Figure 5.14, we see that the estimator of the within covariance seems to drift, and stabilizes with roughly $m=400$ imputations. Notice the magnitude of the within variance ($\approx 4.0$) and the within covariance ($\approx 1.7$). Our goal is to establish a model-based imputation procedure (RMNI) which is
Figure 5.14: Multiply-Imputed Within Covariance Using RMNI

proper for performance of multiple comparisons of the treatment effects. It is critical that the within variance dominate the variance-covariance matrix associated with the multiply-imputed estimates of the treatment effects. That is, the between variance and the between covariance (Figures 5.15 and 5.16) should be much smaller than the within variance and the within covariance, respectively. Also, the variability of the between variance need be much smaller than the variability of the within variance (Condition 2.61). (See Section 2.9.2 for details.) Consequently, we conservatively seek the value of $m_{RMNI}$ for which the least possible amount of RMNI process noise is evident.
The computation of the between variance is discussed in Section 2.5.2. In Figure 5.15, we plot the sequence of combined values for \((1 + 1/m)B_m\) for \(m=1, 2, \ldots, 986\). With the panel for Treatment Two (center plot), we see the need for \(m=400\) imputations to obtain a stable estimate of between variance. It should be noted that the scale of these plots is microscopic, and it may serve one’s purposes well to use only \(m=200\) RMNI-imputed data sets. However, we suspect that using \(m=400\) imputations would resolve unnecessary noise in the combined estimate of total variance.
Figure 5.16: Multiply-Imputed Between Covariance Using RMNI
Multiply-Imputed Between Covariance Using RMNI  The combined between covariance, updated with sequential imputations, is displayed in series with \( m \) in Figure 5.16. It is notable that the between covariance for each treatment effect stabilizes with as few as \( m=100 \) imputations. While the within variance and the within covariance are the same value for each treatment effect (see Section 3.3.1), the between variance and the between covariance vary with treatment effect. More precisely, the between variance (covariance) varies with the observed pattern of missing data from treatment to treatment. If the treatment effect is fully observed as intended by design, the multiply-imputed between variance of that treatment effect is 0.0: the multiply-imputed treatment mean will not vary from imputation to imputation. However, the multiply-imputed between covariance will be non-zero wherever this covariance involves a treatment which is partially unobserved.

Multiply-Imputed Total Variance Using RMNI  The multiply-imputed total variance is the sum of the within variance and the between variance, for each treatment effect estimator. An adjustment factor of \( (1 + 1/m) \) is applied to the between variance, as discussed with Equation (2.24). With Figure 5.17, we display total variance associated with each multiply-imputed treatment effect estimate. It is evident that this variance stabilizes with \( m=200 \) imputations, and arguably with \( m=100 \) imputations.

Multiply-Imputed Total Covariance Using RMNI  The multiply-imputed total covariance is the sum of the within covariance and the between covariance for each pair of treatment effect estimators. The adjustment factor of \( (1 + 1/m) \) is applied to the between covariance, as specified with Equation (2.24). Noting the apparent drift in the series plotted
Figure 5.17: Multiply-Imputed Total Variance Using RMNI
Figure 5.18: Multiply-Imputed Total Covariance Using RMNI

for each pair of treatment effect estimators (Figure 5.18), we determine that \( m = 400 \) imputations are necessary to obtain a stable estimate of total covariance for the multiply-imputed treatment effect estimates using RMNI-imputed data.

### 5.4 Reasonable Extension of Simulation Results to Practice

**Generalization Beyond Simulation Parameters** It is reasonable to assume that \( m = 400 \) imputations would be sufficient to obtain stable multiply-imputed estimates of treatment effects and the associated variance-covariance matrix, with experimentwise rates of missingness up to ten percent. It is natural to assume that a larger rate of missingness would
<table>
<thead>
<tr>
<th>Treatment</th>
<th>Missingness Count</th>
<th>Missingness Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>18</td>
<td>12%</td>
</tr>
<tr>
<td>Two</td>
<td>12</td>
<td>8%</td>
</tr>
<tr>
<td>Three</td>
<td>15</td>
<td>10%</td>
</tr>
<tr>
<td>Experimentwise</td>
<td>45</td>
<td>10%</td>
</tr>
</tbody>
</table>

Table 5.10: Rates of Missingness for Each Treatment Effect

demand a larger number of imputations to obtain such a stable estimate. For example, the multiply-imputed treatment mean stabilizes the most quickly for Treatment Two of our simulation (e.g.: Figure 5.15 and Figure 5.17); with Table 5.10, it is clear that Treatment Two is the most completely observed treatment effect of the three under study.

A Note on the Simulated Rate of Missingness The multiply-imputed estimate will be the less variable in series with $m$ with the smaller amount of missing data. However, with the smaller amount of missing data, the impact of the imputed values upon the multiply-imputed estimate of each treatment effect will require more imputations to ascertain. Consequently, we perceive a missingness rate of ten percent to be an instructive vantage point for discussion of the RMNI imputation process, and a reasonable basis for comparison with any natural occurrence of missingness under the simulated model.

5.4.1 Proper Imputation Related with Number of Imputations

Prescribing a Standard $m_{RMNI}$ No procedure for Multiple Imputation is proper per se: such a procedure is potentially proper for an intended use of the multiply-imputed data sets. The determination of a standard number of imputations, $m_{RMNI}$, is accurately described as
the determination of a value for $m$ at which the RMNI imputation procedure is *proper* for a given intended use of the RMNI-imputed data, such as with specification of a standard procedure for Multiple Comparisons. To declare a practicable standard number of RMNI imputations, we must investigate the behavior of RMNI-imputed estimates in the practice of their intended use. Ultimately, we will need to assess the performance of RMNI with a prescribed standard number of imputations $m=m_{RMNI}$, across many patterns of missing data, and with varying potential values of $Y_{obs}$. See Section 2.9.2 for specifications of a *proper* multiple imputation procedure.

**Declaration of $m_{RMNI}=400$** It is our intuition that a prescribed value of $m_{RMNI}=400$ is conservative, based on the simulation study presented in this chapter. We believe that exploration of RMNI with a given multiple comparisons procedure (MCP) and $m=400$ imputations will be instructive. A protocol for an assessment of RMNI-imputed multiple comparisons of treatment effects is provided in Chapter Six.

**Intuition versus Demonstration** It is clearly demonstrated within our simulation study that $m=400$ RMNI imputations is sufficient to yield a stable estimate for each treatment effect and each associated gauge of error in this particular simulation. However, it is not clear whether repeated simulation with varied pattern of missingness, or varied value of $Y_{obs}$ would indicate a different value for $m_{RMNI}$. Given the rigor of our determination of $m_{RMNI}$ to be 400, in terms of the variability of the multiply-imputed estimates in series with $m$ beyond 400 imputations, we believe that 400 imputations should easily suffice for investigative application of RMNI with a standard MCP for our model. It is likely that the process is sufficiently quiet to yield *proper* multiply-imputed estimates with many fewer
imputations, though we are not prepared to defend such a statement rigorously. In particular, we are not able to compute a bound for acceptable noise in the process, pending application with a multiple comparisons procedure (see Chapter Six), and pending further simulation of an estimate of the variability of the between variances and the between covariances (see Condition Three of Section 5.4.1).

A Note on Proper Imputation In Section 2.9.2, the conditions for a proper imputation procedure indicate a fixed value for $Y$. The reader should carefully distinguish the usage of $Y$ in Survey Sampling from the usage of the same with Experimental Design. Fixing $Y$ in Section 2.9.2 is fixing the population from which a sample value of $Y_{inc}$ may be drawn; it is not fixing the value of $Y_{inc}$, and it is not fixing the value of $Y_{obs}$. Having said this, we assert that with repeated simulation of RMNI, using $m_{RMNI}=400$ imputations, with varied pattern of missingness, and with varied values of $Y_{inc}$ (complete data) and $Y_{obs}$ (incomplete data), we have substantive reason to believe that Rubin’s (1987) conditions would be clearly observable. We present the extent of our findings regarding the randomization-validity (Section 2.9.1) of RMNI, with particular respect to its requirement that RMNI be proper for its intended use.

5.4.2 Heuristic Argument That RMNI is Proper for $\hat{\theta}$

Condition One In Section 2.9.2, we specify the conditions under which RMNI would be a proper multiple imputation procedure for the estimation of treatment effects and the associated error. Condition (2.57) is satisfied by construction of the RMNI procedure, and specification of the model, given that the LSE is used to estimate the treatment effect. Condition (2.58) is not verified by demonstration, as we have no simulated distribution for the between variance $B$. However, it is for this condition, in part, that we recommend the
conservative value $m_{\text{RMNI}}=400$, which corresponds with an exceedingly stable estimate of the between covariance, as well as the treatment effects themselves. The between variance (covariance) is slight [$\approx 0.01 (\approx 0.001)$], requiring an exceedingly stable estimate which varies an order of magnitude less than its mean value. It is apparent that with sufficiently large $m$, Condition (2.58) is achievable with RMNI. Condition One of a proper multiple imputation procedure comprises Conditions (2.57) and (2.58).

**Condition Two** Though we do not prove this directly, it is evident that the combined within variance is an unbiased estimate of the complete-data estimate of $U=V(\hat{\theta})$. The within variance is a mean of normally distributed and unbiased quantities, as specified in Section 3.3.1. The question is whether the variability of the within variance is much smaller than the between variance. With repeated simulation of the within variance, this quality might be established: Condition (2.60). Notice that the tiny magnitude of the between variance requires an exceedingly stable estimate for each treatment effect: hence the conservative requirement of $m_{\text{RMNI}}=400$.

**Condition Three** The requirement of Condition (2.61) is that the between variance be much less variable than the complete-data estimator of $\theta$. Though we do not prove this directly, we can assert that with a conservative number of imputations, $m_{\text{RMNI}}=400$, the estimates of the between variance and the between covariance are so stable, and so much smaller in magnitude than the within variance and the within covariance, that Condition (2.61) must be satisfied (Figures 5.13, 5.14, 5.15, and 5.16).

**Bayesianly Proper** Rubin’s (1987) conclusion, that procedures based on Bayesian methodology tend to be proper (Section 2.9.2), supports our intuition that RMNI is proper for the
estimation of treatment effects with Least Squares Estimators (Section 3.3.1), and the associated error. The obtainable stability of the RMNI process, as demonstrated with our simulation study, emphasizes the potential for RMNI to satisfy the said conditions for a proper multiple imputation.

**A Note on Satisfying the Conditions of Proper** In our review of the literature, we find no instance of a rigorous proof that a practicable multiple imputation procedure is proper. The requirements are sufficiently theoretically or computationally intensive as to be prohibitive of direct proof. Schafer (1997) expands on this difficulty with the practice of multiple imputation, and provides heuristic guidelines, as discussed in Section 2.9.3. For further discussion of this matter, we refer the reader to Rubin (1993), which answers the criticisms of Fay (1991) with a focus on the conditions of proper multiple imputation procedures.

**Randomization-Valid Multiple Imputation Procedure** It is evident that the RMNI imputation procedure is randomization-valid, as defined in Section 2.9.1. Condition (2.55) is met by construction, Condition (2.56) is met by requiring sufficiently large $m_{RMNI}$, and the condition of being proper is affirmed as discussed above. In simple terms, it is our impression that the conditions for a proper and randomization-valid multiple imputation procedure are conservative as set by Rubin (1987) and simply require this:

1. **Shape**: The multiply-imputed estimator must be normally distributed.

2. **Center**: The multiply-imputed estimators of the fixed treatment effect and of the associated gauges of error must be unbiased.
3. **Spread**: The process should be as stable as computationally possible so as to detect and capture the contribution of the imputed values to the total variance of the multiply-imputed estimator of each treatment effect.

It is well established that a multiple imputation procedure is not readily verifiable as proper for any given usage of the completed data. The discussion of the topic is cumbersome and difficult: such discussion is avoided in the literature, with rare confessions of the impossibility of verifying this property. However, we do believe that the above synopsis is accurate, and that the complexity of theoretical exposition derives from the fact that the process of multiple imputation involves so many layers requiring tight control in practice. We do hope that the display of each of these component layers with our simulation serves to illuminate the intent of Rubin’s (1987) condition that such a procedure be proper, as he so expertly detailed.
Simulated Demonstration of Multiple Comparisons Using RMNI

Scope  In this chapter, we consider the problem of performing pairwise comparisons of treatment effects under the two-way mixed interaction model with repeated measures. With the simulation study presented in Chapter Five, we determine that $m_{RMNI}=400$ imputations should yield a stable multiply-imputed combined estimate for each treatment mean; as well as a stable multiply-imputed estimate of the associated variance-covariance matrix, $V(\hat{\theta})$. It is expected that each fixed treatment mean is estimated using the standard complete-data Least Squares Estimator, as provided in Section 3.3.1 along with the associated gauge for error. We apply the repeated RMNI-imputed treatment means (with variance-covariance matrix) under the known standard complete-data method for obtaining 100(1-$\alpha$)% confidence bounds for each pairwise comparison: a Tukey procedure with the Studentized Range reference distribution is performed, per direction of Hochberg and Tamhane (1983), and as computed using the SAS GLM Procedure. We investigate the performance of these multiple comparisons with RMNI-imputed data for the requisite $m=400$ imputations; and we compare this performance to that of a Tukey-Kramer Multiple Comparisons Procedure, as computed by the SAS MIXED Procedure.
**Data** For the simulated presentation of this chapter, we assume the identical environment prescribed with the simulation study of Chapter Five. The RMNI-imputed data employed in this chapter are the same RMNI-imputed data assessed and displayed in Chapter Five. The simulated parameter value for each treatment effect and each variance component is provided in Table 5.1.

**Conclusions** We find that the exact procedure of Hochberg and Tamhane (1983) can be obtained using the SAS GLM Procedure, but not with the SAS Mixed Procedure. Also, we discover that the unbalanced-data procedures allowed by the GLM Procedure and the Mixed Procedure yield wider confidence intervals than the complete-data procedures; these unbalanced-data options, based upon Tukey- and Tukey-Kramer procedures, are not recommended and not explored beyond the example in this dissertation. We consider completed-data procedures for obtaining confidence intervals for treatment comparisons under three approaches: Single Imputation, Repeated Imputation, and Multiple Imputation. Ultimately, we find that of these three approaches to imputed inference using RMNI data, the multiply-imputed inference appears to be the most reasonable. For the performance of multiply-imputed inference with treatment contrasts, we propose multiply-imputed statistics for estimating each contrast and estimating the standard error associated with the Studentized Range Distribution. To close the chapter, we discuss the influence of the imputed values upon practical results, and we discuss the potential that our method is randomization-valid.

### 6.1 Selection of the Multiple Comparisons Procedure

**Standard Complete-Data Method** The value of Multiple Imputation is that it enables the use of standard complete-data methods with incomplete data, while incorporating a component of variance designed to account for the presence of the imputed values in each
completed data set. With the computation of statistics under the two-way interaction model, we have encountered time and again the lack of consensus amongst statisticians regarding what the so-called standard complete-data method should be, in theory or in practice (e.g.: Voss, 1999). With the application of Multiple Comparisons Procedures under the two-way interaction model, we encounter some relief in the theoretical development of an exact MCP (Hochberg and Tamhane, 1983), yet we encounter difficulties with the offerings of standard commercial software packages. SAS does provide options within the MIXED Procedure, the ANOVA Procedure, and the GLM Procedure. The MIXED Procedure and the GLM Procedure do both offer variations of the correct MCP: Tukey pairwise comparisons with a studentized range reference distribution.

**Theoretical Complete-Data Method**  With Hochberg and Tamhane (1983), we find a proof for a formulation of an exact 100(1-α)% Confidence Interval for each component pairwise treatment contrast of \( \theta_{\text{contrast}} \). For reference, we provide this expression for \( \theta_{\text{contrast}} \):

\[
\theta_{\text{contrast}} = \begin{pmatrix}
\theta_1 - \theta_2 \\
\theta_1 - \theta_3 \\
\theta_2 - \theta_3
\end{pmatrix}
\]  

(6.1)

Hochberg and Tamhane (1983) provide the following theorem as guidance to the performance of multiple comparisons under Model (3.4):

1. Under (Model 3.4), the random variable

\[
1 \leq i < i' \leq n_t \quad |\bar{Y}_{i.} - \bar{Y}_{i'.} - (\mu_i - \mu_{i'})| \quad \frac{\sqrt{MS_{int}/n_s n_r}}{\sqrt{MS_{int}/n_s n_r}}
\]

(6.2)

is distributed as a Studentized range variable \( Q_{n_t,(n_t-1)(n_s-1)} \).
2. The simultaneous 100(1-\(\alpha\))% confidence intervals for \(\mu_i - \mu_{i'}\) \((1 \leq i < i' \leq n_t)\) are given by the probability statement

\[
P\{\mu_i - \mu_{i'} \in [\bar{Y}_{i.} - \bar{Y}_{i'.} \pm Q_{n_t,(n_t-1)(n_t-1)}^{(\alpha)} \sqrt{MS_{int}/n_xn_r}], \forall i < i'\} = 1 - \alpha. \quad (6.3)
\]

Hochberg and Tamhane (1983) also provide a statement of proof for an exact Scheffe F test for all treatment comparisons under Model (3.4). Note that the statement of the theorem above is paraphrased in terms of the notation and the context of this dissertation.

**Practical Complete-Data Method**  
There is no definitive standard complete-data method for performing multiple comparisons under Model (3.4); there are options which are variously defensible. We attempt to apply standard complete data procedures to RMNI-imputed data sets, and defer to those methods which are standardly available in SAS. To be clear, we believe it is the most accurate statement of procedure to present the precise SAS code used for production of our results. With Table 6.1, we provide the MIXED Procedure statements; and with Table 6.2, we provide the GLM Procedure statements, as used for the simulation study of this chapter. With the MIXED Procedure, we have the advantage of explicit specification of the variance-covariance matrix of the observations (e.g.: type=vc), with the deficit of having no option for obtaining the Tukey procedure of Hochberg and Tamhane (1983). With the GLM Procedure (and the ANOVA Procedure), we have the advantage of an option to obtain the Studentized range critical value requisite of Equation (6.3), and the advantage to specify the correct studentizing Mean Square, the interaction Mean Square (e.g.: /E=trt*sub), as an option to the MEANS statement.

**Data**  
The data set for input, sasuser.RMNIdata, is an RMNI-imputed data set. That is, it is a completed data set, which has the appearance, in isolation, of being a complete and
Table 6.1: PROC MIXED Statements for Performance of Pairwise Comparisons

balanced data set observed under the two-way mixed interaction model (Model 3.4). The observations are stored as the vector-valued variable, y. The variables trt (treatment) and sub (subject) are vector-valued identity labels corresponding to each observation, and are declared as CLASS variables to distinguish them from continuous response values. It is not necessary to specify an identity label for each repetition; this structure is implicit with the REPEATED statement as used in Table 6.1.

Table 6.1 With this code, we make explicit two default options within MIXED Procedure. First, the MIXED Procedure treats the Variance Components structure for the variance-covariance matrix of the observed data as the default type of structure: /type=vc. The SUBJECT option to the REPEATED statement is used here to instruct the MIXED Procedure to create a block-diagonal variance-covariance matrix, with blocks corresponding to subjects (e.g.: /subject=sub). The alpha level of 0.05 is a second default option within the MIXED Procedure, made explicit with the option /ALPHA=0.05.
Studentizing Factor: Interaction Mean Square As coded per Table 6.1 and Table 6.2, as well as with a MEANS statement with the ANOVA Procedure, the SAS procedure is directed to perform pairwise comparisons of treatment effects with the option /ADJUST=Tukey. However, this directive does not yield the same result with each SAS procedure. With the Mixed Procedure, this directive yields the performance of a Tukey-Kramer procedure with the Student’s \( t \) reference distribution for hypothesis tests, yet studentized correctly by the interaction mean square. We would expect a test statistic computed in this way to be referred to the Studentized Range Distribution. It is broadly recognized (Hochberg and Tamhane, 1983 and 1987; Hsu, 1999) that the residual mean square is the incorrect denominator of the desired (corresponding) F statistic in the presence of the interaction effect: the MSE is not independent of linear combinations of the treatment mean square (usually written MSA) and the subject mean square (usually MSB). The appropriate studentizing factor is the interaction mean square (usually MSAB), which is independent of linear combinations involving \( MS_{sub} \), such as the numerator of the desired F statistic.

Studentized Range Distribution Critical Value As written in Equation (6.3), the correct and exact 100(1-\( \alpha \))% Confidence Region for \( \theta_{contrast} \) relies upon a critical value, \( Q \),
which is computed using the Studentized Range Distribution. The Tukey-Kramer method of PROC MIXED does not appear to utilize this reference distribution for hypothesis tests, but does for confidence intervals of treatment contrasts. With the GLM Procedure (and the ANOVA Procedure), and with the MEANS statement option given by /Tukey, we direct the SAS procedure to compute the Studentized range critical value.

**Multiple Comparisons with Unbalanced Data** Using the MIXED Procedure, one may obtain pairwise comparisons of fixed treatment effects under Model (3.4) by using the /ADJUST=Tukey option to the LSMEANS statement. *We do not recommend this.* The properties of such pairwise comparisons are unknown, and the procedure is ill-advised in the presence of missing data under our model. No valid unbalanced-data method of Multiple Comparisons exists for our model. For options with handling unbalanced data under Model (3.4) to obtain multiple comparisons of treatment effects, we refer the reader to Voss and Hsu (1998).

**Meaning of Standard Procedure** Briefly, we note an awareness that the terms *Standard Procedure* and *Correct Procedure* are used interchangeably in the field. It is apparent that a statistical procedure is often accepted equally as *standard* and *correct*, if the procedure generates no ERROR statements in a SAS Procedure Log file, or the analogous indication of a program fault in other commercial software packages. Perhaps it is *good enough* to apply such a criterion for *correct procedure* in practice. We intend to consider the performance of multiple comparisons with RMNI-imputed data using such *standard* procedures as would be performed in the field.
Table 6.3: Complete-Data Multiple Comparisons of Treatment Effects

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Y_{i, \cdot} - Y_{\ell, \cdot}</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROC GLM</td>
<td>\theta_1 - \theta_2</td>
<td>-102.7090</td>
<td>-107.9640</td>
<td>-97.4538</td>
</tr>
<tr>
<td></td>
<td>\theta_1 - \theta_3</td>
<td>-203.0940</td>
<td>-208.3490</td>
<td>-197.8390</td>
</tr>
<tr>
<td></td>
<td>\theta_2 - \theta_3</td>
<td>-100.3850</td>
<td>-105.6400</td>
<td>-95.1298</td>
</tr>
<tr>
<td>PROC Mixed</td>
<td>\theta_1 - \theta_2</td>
<td>-102.7100</td>
<td>-107.9600</td>
<td>-97.4609</td>
</tr>
<tr>
<td></td>
<td>\theta_1 - \theta_3</td>
<td>-203.0900</td>
<td>-208.3400</td>
<td>-197.8500</td>
</tr>
<tr>
<td></td>
<td>\theta_2 - \theta_3</td>
<td>-100.3800</td>
<td>-105.6300</td>
<td>-95.1367</td>
</tr>
</tbody>
</table>

6.2 Simulated Assessment of Multiple Comparisons Using RMNI

Synopsis We consider two procedures for the performance of multiple comparisons of treatment effects using RMNI-imputed data. The procedures considered are the standard procedures made explicit with Table 6.1 and Table 6.2. The correct procedure is made explicit with Equation (6.2) and Equation (6.3), and is accomplished with the SAS code of Table 6.2. We begin with a survey of results from these procedures using complete data (Y_{inc}) and incomplete data (Y_{obs}). In Section 6.3, we present our results using completed data (Y_{RMNI, \ell}, \ell = 1, \ldots, m).

6.2.1 Complete-Data Results

Table 6.3 With Table 6.3, we display the components and associated confidence intervals for \( \hat{\theta}_{\text{contrast}} \), as produced with SAS PROC GLM and SAS PROC MIXED, using the statements provided in Table 6.1 and Table 6.2. For each treatment effect, the simulated parameter value is displayed in Table 5.1, and the complete-data Least Squares Estimate is displayed in Table 5.2. The simulated value of \( \theta_{\text{contrast}} \) is \((-100, -200, -100)^\prime\). By
Table 6.4: Incomplete-Data Multiple Comparisons of Treatment Effects

<table>
<thead>
<tr>
<th>Contrast</th>
<th>$Y_{i..} - Y_{i'}$</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Dubious Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROC GLM</td>
<td>$\theta_1 - \theta_2$</td>
<td>-101.963</td>
<td>-107.245</td>
<td>-96.681</td>
</tr>
<tr>
<td></td>
<td>$\theta_1 - \theta_3$</td>
<td>-203.445</td>
<td>-208.756</td>
<td>-198.134</td>
</tr>
<tr>
<td></td>
<td>$\theta_2 - \theta_3$</td>
<td>-101.483</td>
<td>-106.735</td>
<td>-96.230</td>
</tr>
<tr>
<td>PROC Mixed</td>
<td>$\theta_1 - \theta_2$</td>
<td>-102.3023</td>
<td>-107.61</td>
<td>-96.9923</td>
</tr>
<tr>
<td></td>
<td>$\theta_1 - \theta_3$</td>
<td>-202.7723</td>
<td>-208.10</td>
<td>-197.4500</td>
</tr>
<tr>
<td></td>
<td>$\theta_2 - \theta_3$</td>
<td>-100.4700</td>
<td>-105.78</td>
<td>-95.1668</td>
</tr>
</tbody>
</table>

construction, each contrast is significantly different from zero, and so no significance test is presented here; we focus on the behavior of computed confidence bounds for each treatment contrast. Notice, in particular, that the computed bounds in Table 6.3 do vary from SAS PROC GLM (Hochberg and Tamhane, 1983) to SAS PROC MIXED.

Table 6.4 With Table 6.4, we present our findings using SAS PROC GLM and SAS PROC MIXED with incomplete data, $Y_{obs}$. These SAS procedures yield no messages of warning or error in the associated log files. It is evident that both SAS procedures recognize the data to be unbalanced, and adjust their computational method and output display accordingly.

Table 6.5 We use the width of the confidence interval for each treatment contrast to compare procedures, and register the impact of unbalanced data upon computed bounds for each interval. Notice that the unbalanced-data confidence intervals are wider than the complete-data confidence intervals; and notice that the complete-data confidence interval computed
6.3 Confidence Intervals for Treatment Contrasts Using RMNI

**Goal** Referring to Equation (6.3), we see that the exact 100(1-\(\alpha\))% confidence interval for each treatment contrast under Model 3.4 depends upon two statistics:

1. The estimate of the contrast, given by \(\bar{Y}_{i} - \bar{Y}_{i'}\), \(1 \leq i < i' \leq n_t\); and

2. The estimate of the standard error of the contrast estimate, given by \(\sqrt{MS_{int}/n_sn_r}\),

where \(MS_{int}\) denotes the interaction mean square, as computed using a complete and balanced data set under Model (3.4).

Given these two statistics for each treatment contrast, we may construct an exact 100(1-\(\alpha\))% confidence interval for each treatment contrast. Hochberg and Tamhane (1983) have proven that this confidence interval may be referred to the Studentized Range Distribution, on the appropriate degrees of freedom, as specified with Equation (6.3). This fact is restated here to emphasize that the correctness of the Studentized Range critical value is

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Complete Data, (Y_{inc})</th>
<th>Incomplete Data, (Y_{obs})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\theta_1 - \theta_2)</td>
<td>10.5102 PROC GLM</td>
<td>10.4991 PROC MIXED</td>
</tr>
<tr>
<td>(\theta_1 - \theta_3)</td>
<td>10.5100 PROC GLM</td>
<td>10.4900 PROC MIXED</td>
</tr>
<tr>
<td>(\theta_2 - \theta_3)</td>
<td>10.5102 PROC GLM</td>
<td>10.4933 PROC MIXED</td>
</tr>
</tbody>
</table>

Table 6.5: Interval Widths for Complete-Data and Incomplete-Data MCP

using a Tukey-Kramer procedure in SAS PROC MIXED is narrower than the exact 100(1-\(\alpha\))% confidence interval produced by SAS PROC GLM.
not proven for any case other than the complete observance of $Y_{inc}$. That is, the confidence interval of Equation (6.3) is proven to be exact if $Y_{inc} = Y_{obs}$. We begin to consider, in this section, the performance of Confidence Interval (6.3) in the case that $Y_{inc} \neq Y_{obs}$.

### 6.3.1 Completed-Data Estimates of Treatment Contrasts

**Unbiased Normal Estimator for Each Treatment Contrast** The confidence interval consists of a point estimate, a standard error, and a critical value to be obtained from the appropriate reference distribution. In this vein, we begin with the consideration of the point estimate of each treatment contrast under Model (3.4), as computed using RMNI-imputed data. Let $\hat{\theta}_{\text{contrast}}$ be the estimator of $\theta_{\text{contrast}}$ of Expression (6.1). The construction of $\hat{\theta}_{\text{contrast}}$ is as follows:

$$\hat{\theta}_{\text{contrast}} = \left( \begin{array}{c} \hat{\theta}_1 - \hat{\theta}_2 \\ \hat{\theta}_1 - \hat{\theta}_3 \\ \hat{\theta}_2 - \hat{\theta}_3 \end{array} \right),$$

(6.4)

where $\hat{\theta}_i$ is given by Equation (3.11) for each treatment $i=1, \ldots, n_{trt}$.

**Figure 6.1** With Figure 6.1, we display the componentwise point estimates of Equation (6.4), as obtained with each successive completed data set of the simulation study of Chapter Five. In Chapter Five, we showed that the Repeated Measures Normal Imputation procedure is effective with the production of least squares estimates: each treatment effect estimate is unbiased and distributed as a Normal variate, with the correct variance-covariance matrix as expected under Model (3.4). For further investigation, see Figure 5.6, and Result (3.12). It is easily verifiable that the componentwise estimator of each treatment contrast of $\theta_{\text{contrast}}$ is normally distributed and unbiased.

**Stable Multiply-Imputed Estimates of Treatment Contrasts** With Figure (6.2), we display the combined mean of $m$ estimates for each treatment contrast. The combined
mean is formulated in the manner of Equation (2.21), and is given by

$$
\hat{\theta}_{\text{contrast}}^{(m)} = \frac{\sum_{\ell=1}^{m} \hat{\theta}_{\text{contrast},\ell}}{m}.
$$

(6.5)

With Figure (6.1), we see each point estimate of each treatment contrast which would serve to construct a confidence interval using Equation (6.3) under the approach of Single Imputation. For specification of the computation of the combined mean, we refer you to Equation (2.21). Notice that this multiply-imputed estimate stabilizes with as few as \( m = 100 \) imputations, for each treatment contrast (Figure 6.2). Briefly, we note that the combined estimate for each treatment contrast is the same for Repeated Imputation as Multiple Imputation.
Figure 6.2: Multiply-Imputed Treatment Contrasts Using RMNI in Series with $m$

### 6.3.2 Completed-Data Estimates of Tukey Standard Error

**Goal**  Given a satisfactory point estimate of the treatment contrast, we proceed to obtain a satisfactory estimate of the associated standard error. With Hochberg and Tamhane (1983) providing guidance through Equation (6.3), we seek a multiply-imputed standard error of the form

$$SE_{contrast} = \sqrt{MS_{int}/n_r} = \sqrt{MS_{int}/150}.$$  \hspace{1cm} (6.6)

We begin with an assessment of the RMNI-imputed values of $MS_{int}$, as obtained with $m$ successive imputations.
Under Model (3.4), with complete and balanced data, the statistic $MS_{int}$ is unbiased for a linear combination of variance components. We have that

$$E(MS_{int}) = \sigma_e^2 + 3\sigma_c^2,$$  \hspace{1cm} (6.7)

where $\sigma_c^2$ is defined to be the interaction variance component. The simulated parameter value for each variance component is 100.0, so that the simulated parameter value for $E(MS_{int})$ is 400.0.

Figure 6.3 With Table 6.3, we display the $MS_{int} = 365.708$, as computed with complete data, $Y_{inc}$. With Table 6.4, we display $MS_{int} = 332.3651$, as computed with incomplete data, $Y_{obs}$. With Figure 6.3, we display a histogram of the values of $MS_{int}$ as computed
with RMNI-imputed data, $Y_{RMNI,\ell}$, $\ell = 1, \ldots, m$. The RMNI-imputed values of $MS_{int}$ have a mean of 345.412, with a standard deviation of 4.386. Notice with Figure 6.3 that the complete-data and incomplete-data values of $MS_{int}$ are just above and below the body of RMNI-imputed values of $MS_{int}$. It is notable that analogous behavior of RMNI-imputed data is also observed with the completed-data estimates of $\theta_2$, as displayed in Figure 5.7. That is, the completed-data estimates of $MS_{int}$ do not appear to be distributed about either the complete- or incomplete-data estimate of $MS_{int}$. This phenomenon may be an artifact of the simulated data set ($Y_{inc}$), or of the simulated pattern of missingness ($Y_{obs}$).
Construction of a Multiply-Imputed Tukey Within Variance  With \( m \) repeated RMNI-imputed data sets, we explore the natural option for an estimator of \( MS_{int}/n_sn_r \) in

\[
\hat{MS}_{int} = \frac{1}{n_sn_r} \sum_{\ell=1}^{m} MS_{int,\ell}, \tag{6.8}
\]

the arithmetic mean of the RMNI-imputed values \( MS_{int,\ell}, \ell=1, \ldots, m \). This extends the construction known as Within Variance, as formulated by Rubin (1987) and given by Equation (2.22), toward the multiply-imputed standard error of an estimated treatment contrast, \( SE_{contrast} \). (See Equations 6.3 and 6.5.) With Figure 6.4, we display the behavior of \( \hat{MS}_{int} \) in series with \( m \); single RMNI-imputed values are displayed in the upper chart, and the combined values for multiple (and repeated) imputation are displayed in the lower chart. Notice that a stable combined value of \( \hat{MS}_{int} \) is obtained with as few as 100 imputations.

Construction of a Multiply-Imputed Tukey Between Variance  Given a summary estimate of within variance, we seek to employ Rubin’s (1987) construction of a Between Variance, which accounts for variability in \( \hat{\theta}_{contrast} \) due to the use of RMNI-imputed values. The between variance of Chapter Five is provided as Equation (2.23), and adjusted by a factor of \((1 + 1/m)\) as shown in Equation (2.24). We extend this formulation of a between variance to the problem of estimating treatment contrasts by the following:

\[
B_m = \frac{1}{m-1} \sum_{\ell=1}^{m} (\hat{\theta}_{contrast,\ell} - \bar{\theta}_{\ell}^{(m)})' (\hat{\theta}_{contrast,\ell} - \bar{\theta}_{\ell}^{(m)}) \tag{6.9}
\]

and

\[
B_{m, adjusted} = (1 + \frac{1}{m})B_m. \tag{6.10}
\]

Figure 6.5  With Figure 6.5, we exhibit the behavior of the combined between variance, as computed with Equation (6.10). Most notably, the between variance is a degree of
Figure 6.5: Multiply-Imputed Between Variance for Each Treatment Contrast Using RMNI

magnitude smaller than the within variance of Figure 6.4. This result encourages us to create a multiply-imputed analog to the Total Variance of Equation (2.24): Total Variance = Within Variance + Adjusted Between Variance.

**Multiply Imputed Total Variance**  With Figure 6.6, we display the multiply-imputed total variance for each estimate of a treatment contrast. Note that the total variance and the between variance each do stabilize with as few as 100 imputations, and clearly with 200 imputations. We use the term variance here in terms of Equation (6.6). Literally, we propose construction of Tukey pairwise confidence intervals as instructed with Equation (6.3), where each contrast estimate is given by Equation (6.4), and the estimator for $SE_{\text{contrast}}$ of
Equation (6.5) is given by

\[ T_{m_{\text{RMNI}}}^{(\text{RMNI})} = \hat{M}\text{S}_{\text{int}} + B_{m, \text{adjusted}}. \quad (6.11) \]

Our formulation of \( \hat{M}\text{S}_{\text{int}} \) is provided with Equation (6.8).

### 6.4 Tukey Confidence Bounds with Imputed Values

**Goal** We intend to assess the performance of RMNI-imputed Tukey Confidence Intervals, as constructed using Equation (6.3), under each of three approaches to imputation: Single Imputation, Repeated Imputation, and Multiple Imputation. It is with multiple imputation that the between variance is incorporated into the estimate of standard error of each treatment contrast estimate. Since the between variance accounts for variability in the
estimate of a treatment contrast due to the presence of the imputed values (missing data),
we consider this to be the most reasonable approach of the three considered here.

**Single Imputation** Single Imputation refers to the use of one RMNI-imputed data set to
compute the confidence intervals specified by Equation (6.3). That is, the RMNI-imputed
data is treated as real and fully-observed, and the standard complete-data method is ap-
plied. Insodoing, one cannot possibly assess the impact of missing data upon the imputed
confidence region for $\theta_{\text{contrast}}$. With Figure 6.7, we display pointwise the half-width of a
confidence interval computed in this way. The Single Imputation Margins of Error (half
widths) are computed with the complete-data Studentized range critical value of 3.36564.
The value of the singly-imputed half width ranges from 5.012 to 5.212, with a mean of
5.107 and a standard deviation of 0.032.

**Repeated Imputation** Repeated Imputation refers to the use of an imputation procedure
such as RMNI, and popularly the Hot Deck procedure, over and over, to obtain repeated
values for inferential statistics. With Figure 6.8, we display the Runs Chart for the process
of combining repeated observations of $\hat{\theta}_{\text{contrast}}$ and $MS_{\text{int}}$, in series with $m$. Note that this
is simply Multiple Imputation without the incorporation of a Between Variance.

**Multiple Imputation** Multiple Imputation refers to repeated imputation with the require-
ment of a randomization-valid imputation procedure. We believe RMNI to be randomization-
valid, as discussed in Section 5.4. In contrast, the Hot Deck procedure is appropriate
for Repeated Imputation, but is not randomization-valid for use with Multiple Imputation.
With Figure 6.9, we present the combined and multiply-imputed half width associated with
Equation (6.3), and computed using Equation (6.5) with Equation (6.11), with the Studentized range complete-data critical value of 3.36564.

6.5 **Influence of the Imputed Value Using RMNI**

**Significance of Treatment Contrasts** The computation of the multiply-imputed confidence interval for each treatment contrast is dominated by $Y_{obs}$. Though we simulate the impact here of ten percent missingness, we do not recognize any chance that a multiply-imputed confidence interval would fail to cover the true value of a treatment contrast if the complete-data confidence interval covers the true value. In short, each RMNI-imputed data set yields a confidence interval (Equation 6.3) which covers the true value of the treatment contrast, for each treatment contrast. Further comment will require further simulation.
Randomization-Validity of the Multiply-Imputed MCP  We suspect that our proposed construction of a multiply-imputed Multiple Comparisons Procedure is randomization-valid. We cannot prove this without extensive simulation. One critical issue needs to be addressed: the validity of the Studentized range reference distribution in the context of application of Equation (6.11). It is simply not clear whether the complete-data reference distribution remains correct under the multiply-imputed procedure.

Half Widths  It is apparent that the half widths of confidence intervals based upon Equation (6.3) and computed using RMNI-imputed data are consistently narrower than the confidence intervals produced with the corresponding complete data, and the corresponding
incomplete data (Table 6.5). Further comment will require further simulation. Further simulation is needed to explore the impacts of the pattern of missing data, the value of the observed data, and the value of the complete data itself.

6.6 Additional Simulations

Table 6.6 Although a simulation of the coverage probability for the RMNI-imputed confidence interval for a treatment contrast is postponed to further work, we do present three RMNI-imputed confidence intervals for each of the three treatment contrasts considered in this chapter. These examples are produced with a simulated run of the entire RMNI demonstration process displayed here in Chapter Five and Chapter Six. That is, we generated three complete data sets, three patterns of missingness, and consequently three incomplete data
<table>
<thead>
<tr>
<th>Contrast</th>
<th>Point Estimate</th>
<th>Margin of Error</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1 - \theta_2$</td>
<td>-102.2913</td>
<td>5.1270</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>-97.3718</td>
<td>5.8084</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>-101.5052</td>
<td>5.6749</td>
<td>3</td>
</tr>
<tr>
<td>$\theta_1 - \theta_3$</td>
<td>-202.7926</td>
<td>5.1311</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>-192.4031</td>
<td>5.8086</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>-199.5277</td>
<td>5.6765</td>
<td>3</td>
</tr>
<tr>
<td>$\theta_2 - \theta_3$</td>
<td>-100.5012</td>
<td>5.1260</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>-95.0313</td>
<td>5.8081</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>-98.0226</td>
<td>5.6695</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 6.6: RMNI-Imputed Multiple Comparisons of Treatment Effects

ultimately we perform multiple comparisons of treatment effects with each of the three incomplete data sets under Model (3.4).

**Stable RMNI-Imputed Estimates** We found that with three values for $Y_{inc}$, and three patterns of missingness, that the RMNI procedure stabilized easily within our suggested number of $m_{RMNI} = 400$ imputations. While we expect that the minimal number of imputations required for stable multiply-imputed estimates will vary with $Y_{inc}$ and with pattern of missingness, we suspect that a value of $m_{RMNI} = 400$ is sufficiently conservative to yield stable RMNI-imputed estimates across values of $Y_{inc}$, and across patterns of missingness. This assumption will be tested with further simulation.

**RMNI-Imputed Confidence Intervals** The true parameter values under Model (3.4) are given by Table 5.1: $\theta_1 = 100.0$, $\theta_2 = 200.0$, and $\theta_3 = 300.0$. Table 6.6 is a display of the multiply-imputed point estimates and margins of error for each treatment contrast, as
simulated three times independently. Note that with Simulation Two, the margin of error differs only in the fourth decimal place from one contrast to the next. With each treatment contrast, the margin of error is adjusted separately to reflect variability due to the presence of missing data; missingness occurs at a different rate and in a different pattern under each treatment.
CHAPTER 7

Model-Based Multiple Imputation

7.1 Two Paradigms for Multiple Imputation

The Notion of Practical Paradigms  Generally speaking, any valid Multiple Imputation Procedure has basis in a model for the incomplete data. Noting Rubin’s (1987) Condition (2.51), it is expected that the said model is Multivariate Normal: that is, the population of potentially sampled data, \( \mathbf{Y} \), can be modeled with

\[
\mathbf{Y} \sim \mathcal{N}(\mathbf{\mu}, \Sigma),
\]

(7.1)

where \( \mathbf{\mu} \) is the vector-valued mean of \( \mathbf{Y} \), and \( \Sigma \) is the variance-covariance matrix for the data. Data obtained by a sample survey can be modeled in this way, as can experimental data following the two-way mixed interaction model. We characterized the practical approaches to modeling of the data as following two paradigms: one, collected by sample survey; and two, collected by experimentation.

7.1.1 The Multivariate Normal Approach

Survey Sampling  Multiple Imputation was developed in the context of Survey Sampling. In practice, we have found that survey data may be processed with commercial software such as the SAS GLM Procedure and the SAS MI Procedure, while the practitioner is
unable to explicate the model being fitted, and while the practitioner may be unaware that there is such a thing as a model being applied with such procedures. We have observed that popular software for imputation, based on the procedure of Data Augmentation (Schafer, 1997), such as NORM (Schafer, 1997) and the MI Procedure, is designed to process data which are multivariate and normal: that is, these modules expect multiple variables, each of which is normally distributed. (NORM offers a flexibility which is discussed in Section 7.2.1.)

Data Augmentation  Schafer (1997) develops the procedure of Data Augmentation, which is a close cousin of the EM Algorithm. For details of these procedures, we refer the reader to the text. Data Augmentation is employed by commercial software including the SPlus PAN function (for panel data) and the SAS MI Procedure; it first became available as NORM (Schafer, 1997). This iterative Monte Carlo Markov Chain algorithm requires manual assessment of convergence; the manual assessment may be accomplished using diagnostic tools provided with NORM and PAN, or with the SAS ARIMA Procedure. We have found the application of Data Augmentation to be cumbersome and time-consuming with all such software.

7.1.2 The Model-Based Approach

Experimental Design  RMNI was developed in the context of Experimental Design. In particular, we begin with a model specification of the form of Model (3.4). We consider whether each factor included in the model is fixed or random. We consider whether an interaction term is present, and we adapt RMNI for inclusion of the interaction effect: Yamashita (1995) develops RMNI for the One-Way Layout with repeated measures but no
interaction. The RMNI-imputed value is tailored specifically to be a reasonable imputation for the cell of the Two-Way Layout into which it is imputed, with consideration of the presence or absence of an interaction in the imputed model. While data following a hierarchical model do follow the form of Expression (7.1), this data is not the focus of development of commercial software for the performance of multiple imputations. The RMNI procedure explicitly differentiates between fixed-, random-, and mixed effects models. Like the Multivariate Normal approaches of Schafer (1997) and Rubin (1987), RMNI generates Bayesianly Proper (Section 2.9.3) imputed values, for estimation of treatment effects and associated error.

**RMNI**  
RMNI is the only method for generating an imputed value tailored to preserve the sensitive variance-covariance matrix structure of a hierarchical model. If Data Augmentation is capable of doing so, we have not witnessed this with the documentation, or any published example, under any software application. In Section 7.2 below, we successfully employ NORM to obtain imputed values for our data, assuming the absence of interactions; we are unable to employ the SAS MI Procedure to generate imputed values using the incomplete data set of Chapter Five and Chapter Six.

### 7.2 Comparison of Imputed Values: RMNI and Data Augmentation

#### 7.2.1 Procedures

**RMNI**  
For the demonstration of RMNI in Section 7.2.2, we use the same data, $Y_{obs}$, and the same imputed values, $Y_{RMNI,\ell}$, $\ell = 1, \ldots, m$, as we used in the demonstrations of Chapter Five and Chapter Six. The RMNI procedure employs the two-way mixed interaction model as the imputed model. The RMNI procedure can be adapted to employ any one-way
model, any two-way model, any three-way model, any fixed effects model, any random effects model, and any no-interaction model as the imputed model. The work of developing RMNI procedural application modules, tailored to each variation of the hierarchical model, is listed in Chapter Eight as an option for further work. We present results of 100 RMNI imputations in comparison with 20 NORM-imputed values.

**NORM** For the demonstration of NORM in Section 7.2.2, we use the same data, \( Y_{\text{obs}} \), as used for the demonstrations of Chapter Five and Chapter Six. Schafer’s software (1997, NORM), offers the flexibility of declaring *dummy variables*, so as to specify indicators of treatment, subject, and interaction. This feature, however, is not adequately supported for our data. NORM generates Warning messages for the large number of dummy variables to be created (49 for the subject effects and 149 for the interaction effects). With the inclusion of the indicators of interaction effect, the Data Augmentation procedure fails at Iteration One. With exclusion of the interaction effect, NORM produces sensible imputed values, which do seem to be adjusted for each level of the treatment factor, and plausibly adjusted for each level of the subject factor. With the exclusion of the interaction effect from the imputed model, the multiple imputation produced by NORM is not proper for any analysis which includes the interaction effect in the model. We note that NORM does not ostensibly recognize our data as following a mixed effects model; the same imputation would be performed under a one-way layout with repeated measures and with fixed subject effects.

**PROC MI** For the demonstration of Section 7.2.2, we attempt to include imputed values from the SAS MI Procedure. We are unable to obtain reasonable imputed values for our model with the MI Procedure. To the best of our knowledge, the MI Procedure offers no
utility for recognizing the indicator variables for the treatment, the subject, and the interaction effects. Inclusion of these indicator variables in the VAR statement causes the MI Procedure to fail: Warning messages in the SAS Log file indicate a singular starting value for the variance-covariance matrix of the data. Upon investigation, it appears that the MI Procedure cannot distinguish a vector of treatment labels from a normally distributed response variable; that is, we found no utility for specifying the distinction, as is available with the SAS GLM Procedure and with NORM. We attempt to supply the MI Procedure with a variance-covariance matrix for the data as an instructive initial value for its Data Augmentation algorithm; this approach fails. The MI Procedure requires at least two variables to be supplied in the VAR statement: two normally distributed response variables, we assume. We did not attempt to supply binary indicator variables to the MI Procedure (3 indicators of treatment, 50 indicators of subject, and 150 indicators of interaction). It is apparent that the MI Procedure was not developed for use with hierarchical model data. Since the MI Procedure utilizes the same Data Augmentation procedure as NORM (Schafer, 1997), we are content to present a comparison of RMNI with Data Augmentation based upon the results from NORM.

**A Note on PROC MI** We did hope that the computing power of SAS would enable the Data Augmentation procedure to function per demands of our data. The use of the MI Procedure became labor-intensive, and with no results. We acknowledge that the MI Procedure may be capable of imputing our data under our model; however, we recognize that no example of such application is known to us through publication, SAS documentation, or response from SAS support. We did consider supplying the MI Procedure with the data under each treatment as a separate variable, in the hope of obtaining suitable imputed values;
Table 7.1: NORM-Imputed Values with Five Imputations Under Treatment One

<table>
<thead>
<tr>
<th>NORM-1</th>
<th>NORM-2</th>
<th>NORM-3</th>
<th>NORM-4</th>
<th>NORM-5</th>
<th>Trt</th>
<th>Sub</th>
<th>Rep</th>
</tr>
</thead>
<tbody>
<tr>
<td>92.648</td>
<td>109.919</td>
<td>116.617</td>
<td>109.653</td>
<td>106.941</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>82.439</td>
<td>77.708</td>
<td>77.184</td>
<td>76.666</td>
<td>61.185</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>60.557</td>
<td>98.064</td>
<td>100.648</td>
<td>65.775</td>
<td>101.973</td>
<td>1</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>114.262</td>
<td>101.687</td>
<td>104.758</td>
<td>92.748</td>
<td>107.149</td>
<td>1</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>94.544</td>
<td>96.321</td>
<td>91.387</td>
<td>97.704</td>
<td>79.683</td>
<td>1</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>86.173</td>
<td>88.706</td>
<td>79.916</td>
<td>91.055</td>
<td>108.150</td>
<td>1</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>97.356</td>
<td>110.423</td>
<td>93.170</td>
<td>80.327</td>
<td>94.788</td>
<td>1</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>119.874</td>
<td>124.550</td>
<td>128.680</td>
<td>118.010</td>
<td>99.020</td>
<td>1</td>
<td>26</td>
<td>3</td>
</tr>
<tr>
<td>55.787</td>
<td>85.082</td>
<td>71.376</td>
<td>85.299</td>
<td>91.214</td>
<td>1</td>
<td>28</td>
<td>2</td>
</tr>
<tr>
<td>93.848</td>
<td>103.252</td>
<td>86.853</td>
<td>98.196</td>
<td>105.195</td>
<td>1</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>111.689</td>
<td>88.406</td>
<td>98.538</td>
<td>79.811</td>
<td>95.951</td>
<td>1</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>92.536</td>
<td>82.438</td>
<td>99.873</td>
<td>75.440</td>
<td>90.225</td>
<td>1</td>
<td>33</td>
<td>2</td>
</tr>
<tr>
<td>110.641</td>
<td>105.383</td>
<td>117.626</td>
<td>111.018</td>
<td>104.842</td>
<td>1</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>107.318</td>
<td>116.467</td>
<td>99.876</td>
<td>98.123</td>
<td>79.071</td>
<td>1</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>97.093</td>
<td>109.644</td>
<td>118.510</td>
<td>120.183</td>
<td>111.934</td>
<td>1</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>107.921</td>
<td>109.739</td>
<td>95.262</td>
<td>87.579</td>
<td>128.972</td>
<td>1</td>
<td>42</td>
<td>2</td>
</tr>
<tr>
<td>105.075</td>
<td>102.500</td>
<td>107.419</td>
<td>111.439</td>
<td>112.752</td>
<td>1</td>
<td>46</td>
<td>1</td>
</tr>
<tr>
<td>97.337</td>
<td>99.353</td>
<td>94.853</td>
<td>114.615</td>
<td>83.844</td>
<td>1</td>
<td>50</td>
<td>3</td>
</tr>
</tbody>
</table>

the results do not warrant discussion here because this approach deletes both the subject effect and the interaction effect from the imputed model. We do regret the need to abandon the MI Procedure from our comparison in Section 7.2.2. For this dissertation, the MI Procedure is reviewed under SAS Version 8.2 and SAS Version 9.0.

7.2.2 Comparison

The Imputed Values  We begin our comparison by looking at the imputed values themselves. For our simulation study of Chapter Five, $Y_{mis}$ comprises 45 values which are removed from the simulated value of $Y_{inc}$ to simulate an MCAR pattern of missingness.
The RMNI-imputed values supplied to $Y_{mis}$, and completing the data set, are displayed in Table 5.4 for Treatment One, in Table 5.5 for Treatment Two, and in Table 5.6 for Treatment Three. The simulated parameter values for $Y_{inc}$ are provided in Table 5.1. The NORM-imputed values, $Y_{NORM,\ell}$, $\ell = 1, \ldots, 5$, are given with Table 7.1 for Treatment One, Table 7.2 for Treatment Two, and Table 7.3 for Treatment Three.

**Quality of NORM-Imputed Values**  NORM (Schafer, 1997) produces sound and reasonable imputed values under our model. If the interaction effect were excluded from analysis, then the imputed values of Table 7.1, Table 7.2, and Table 7.3, would presumably constitute a proper imputation for computation of treatment effect estimates and associated error. Notice that the imputed values of Table 7.1 neighbor the parameter value for Treatment One of 100.0, and likewise under Treatments Two and Three for parameter values of 200.0

<table>
<thead>
<tr>
<th>NORM-1</th>
<th>NORM-2</th>
<th>NORM-3</th>
<th>NORM-4</th>
<th>NORM-5</th>
<th>Trt</th>
<th>Sub</th>
<th>Rep</th>
</tr>
</thead>
<tbody>
<tr>
<td>182.275</td>
<td>190.363</td>
<td>199.684</td>
<td>194.475</td>
<td>185.295</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>220.797</td>
<td>222.142</td>
<td>204.965</td>
<td>222.485</td>
<td>215.916</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>183.027</td>
<td>204.517</td>
<td>198.421</td>
<td>201.415</td>
<td>197.673</td>
<td>2</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>151.752</td>
<td>195.031</td>
<td>175.268</td>
<td>183.432</td>
<td>200.487</td>
<td>2</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>220.943</td>
<td>222.827</td>
<td>190.593</td>
<td>200.676</td>
<td>230.405</td>
<td>2</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>212.543</td>
<td>216.808</td>
<td>208.623</td>
<td>197.035</td>
<td>214.716</td>
<td>2</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>211.080</td>
<td>203.813</td>
<td>217.946</td>
<td>209.992</td>
<td>203.073</td>
<td>2</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>194.279</td>
<td>198.886</td>
<td>221.375</td>
<td>212.070</td>
<td>203.439</td>
<td>2</td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td>196.189</td>
<td>216.922</td>
<td>177.979</td>
<td>165.784</td>
<td>173.671</td>
<td>2</td>
<td>38</td>
<td>2</td>
</tr>
<tr>
<td>221.505</td>
<td>205.785</td>
<td>222.983</td>
<td>208.271</td>
<td>207.414</td>
<td>2</td>
<td>39</td>
<td>2</td>
</tr>
<tr>
<td>219.185</td>
<td>181.029</td>
<td>186.482</td>
<td>200.299</td>
<td>180.146</td>
<td>2</td>
<td>43</td>
<td>2</td>
</tr>
<tr>
<td>204.481</td>
<td>199.505</td>
<td>207.625</td>
<td>165.898</td>
<td>175.305</td>
<td>2</td>
<td>44</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 7.2: NORM-Imputed Values with Five Imputations Under Treatment Two
and 300.0, respectively. Although NORM allows no specification of the subject effect as a random effect, it may be true that Data Augmentation properly utilizes $Y_{obs}$ in such a way as to preserve the assumption that the subject effect has a mean of zero. It is unclear whether the underlying structure of the variance-covariance matrix is properly preserved, even with exclusion of the interaction effect from the model. For further investigation of this, we refer the reader to Schafer (1997). The initial estimate of the subject variance is determined by a run of an EM algorithm. (For RMNI, the same is computed by the SAS GLM Procedure using Type III Sums of Squares and the associated Expected Mean Squares.) We produced twenty completed data sets using NORM in eight minutes, with additional time for assessment of convergence. Successive imputed values from the Data Augmentation procedure are highly correlated; we sampled one from every 1000 iterations. Successive RMNI imputed values are mutually independent of one another by design, conditioning on $Y_{obs}$.

**Summary of NORM Imputed Values** We use summary statistics as a basis of comparison of the NORM imputed value with the RMNI imputed value. The results presented with Table 7.4 are comparable to the same under Treatments Two and Three. The final column of Table 7.4, Mean - Actual, is tabulated as the difference between the imputed value and the complete-data value of $Y_{inc}$. While the mean difference of the imputed value from actual is nearly zero (Table 7.5), the Range of an imputed value for a single missing value is quite large: the standard deviation of a natural observation would be 17.32 (Table 5.1). With the standard deviation of the imputed value, holding the missing value fixed, ranging from roughly 9 to roughly 17, we question whether the observed data are being adequately utilized to update the posterior distribution of $Y_{mis}$, $[Y_{mis}|Y_{obs}]$. 

167
<table>
<thead>
<tr>
<th>NORM-1</th>
<th>NORM-2</th>
<th>NORM-3</th>
<th>NORM-4</th>
<th>NORM-5</th>
<th>Trt</th>
<th>Sub</th>
<th>Rep</th>
</tr>
</thead>
<tbody>
<tr>
<td>291.669</td>
<td>287.123</td>
<td>298.923</td>
<td>322.401</td>
<td>275.723</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>261.136</td>
<td>287.278</td>
<td>275.915</td>
<td>247.586</td>
<td>279.932</td>
<td>3</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>287.040</td>
<td>280.136</td>
<td>282.571</td>
<td>287.000</td>
<td>273.995</td>
<td>3</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>287.884</td>
<td>295.881</td>
<td>269.667</td>
<td>297.899</td>
<td>280.235</td>
<td>3</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>284.712</td>
<td>294.328</td>
<td>274.096</td>
<td>309.163</td>
<td>279.670</td>
<td>3</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>320.978</td>
<td>314.894</td>
<td>332.225</td>
<td>328.181</td>
<td>317.396</td>
<td>3</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>293.825</td>
<td>312.658</td>
<td>291.671</td>
<td>310.739</td>
<td>310.554</td>
<td>3</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>296.969</td>
<td>308.098</td>
<td>253.657</td>
<td>289.725</td>
<td>309.131</td>
<td>3</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>319.504</td>
<td>300.448</td>
<td>300.597</td>
<td>315.790</td>
<td>289.135</td>
<td>3</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>307.922</td>
<td>308.488</td>
<td>296.248</td>
<td>303.891</td>
<td>301.805</td>
<td>3</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>297.611</td>
<td>315.523</td>
<td>298.423</td>
<td>302.526</td>
<td>281.264</td>
<td>3</td>
<td>31</td>
<td>2</td>
</tr>
<tr>
<td>282.762</td>
<td>287.055</td>
<td>287.675</td>
<td>311.925</td>
<td>283.921</td>
<td>3</td>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>288.603</td>
<td>303.186</td>
<td>312.310</td>
<td>300.627</td>
<td>301.811</td>
<td>3</td>
<td>36</td>
<td>2</td>
</tr>
<tr>
<td>304.593</td>
<td>335.265</td>
<td>316.470</td>
<td>322.878</td>
<td>296.458</td>
<td>3</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>287.499</td>
<td>290.214</td>
<td>315.203</td>
<td>337.354</td>
<td>303.475</td>
<td>3</td>
<td>41</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 7.3: NORM-Imputed Values with Five Imputations Under Treatment Three
<table>
<thead>
<tr>
<th>Y NORM.1</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Range</th>
<th>Mean - Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>104.63</td>
<td>10.63</td>
<td>82.69</td>
<td>123.86</td>
<td>41.17</td>
<td>15.36</td>
<td></td>
</tr>
<tr>
<td>Y NORM.2</td>
<td>65.50</td>
<td>16.85</td>
<td>27.32</td>
<td>92.17</td>
<td>64.85</td>
<td>-21.51</td>
</tr>
<tr>
<td>Y NORM.3</td>
<td>85.27</td>
<td>15.44</td>
<td>60.55</td>
<td>106.98</td>
<td>46.42</td>
<td>13.66</td>
</tr>
<tr>
<td>Y NORM.4</td>
<td>103.49</td>
<td>11.25</td>
<td>77.22</td>
<td>116.99</td>
<td>39.76</td>
<td>6.39</td>
</tr>
<tr>
<td>Y NORM.5</td>
<td>94.44</td>
<td>9.27</td>
<td>77.25</td>
<td>112.41</td>
<td>35.15</td>
<td>13.56</td>
</tr>
<tr>
<td>Y NORM.6</td>
<td>84.50</td>
<td>14.40</td>
<td>59.56</td>
<td>108.15</td>
<td>48.58</td>
<td>-11.67</td>
</tr>
<tr>
<td>Y NORM.7</td>
<td>98.34</td>
<td>11.24</td>
<td>80.32</td>
<td>122.30</td>
<td>41.97</td>
<td>33.79</td>
</tr>
<tr>
<td>Y NORM.8</td>
<td>112.91</td>
<td>13.75</td>
<td>89.78</td>
<td>141.65</td>
<td>51.87</td>
<td>-16.28</td>
</tr>
<tr>
<td>Y NORM.9</td>
<td>68.30</td>
<td>12.58</td>
<td>45.06</td>
<td>91.21</td>
<td>46.15</td>
<td>-11.52</td>
</tr>
<tr>
<td>Y NORM.10</td>
<td>90.71</td>
<td>14.24</td>
<td>61.94</td>
<td>120.77</td>
<td>58.82</td>
<td>-18.16</td>
</tr>
<tr>
<td>Y NORM.11</td>
<td>97.57</td>
<td>11.70</td>
<td>73.03</td>
<td>119.40</td>
<td>46.37</td>
<td>-6.82</td>
</tr>
<tr>
<td>Y NORM.12</td>
<td>96.36</td>
<td>10.53</td>
<td>75.44</td>
<td>124.93</td>
<td>49.49</td>
<td>-9.13</td>
</tr>
<tr>
<td>Y NORM.13</td>
<td>106.83</td>
<td>13.96</td>
<td>75.77</td>
<td>147.82</td>
<td>72.05</td>
<td>-4.90</td>
</tr>
<tr>
<td>Y NORM.14</td>
<td>95.62</td>
<td>11.55</td>
<td>70.61</td>
<td>116.46</td>
<td>45.85</td>
<td>12.56</td>
</tr>
<tr>
<td>Y NORM.15</td>
<td>105.28</td>
<td>11.31</td>
<td>84.82</td>
<td>124.41</td>
<td>39.58</td>
<td>4.39</td>
</tr>
<tr>
<td>Y NORM.16</td>
<td>98.20</td>
<td>15.02</td>
<td>73.01</td>
<td>128.97</td>
<td>55.96</td>
<td>-16.18</td>
</tr>
<tr>
<td>Y NORM.17</td>
<td>112.46</td>
<td>9.97</td>
<td>94.39</td>
<td>134.37</td>
<td>39.97</td>
<td>8.46</td>
</tr>
<tr>
<td>Y NORM.18</td>
<td>94.22</td>
<td>15.52</td>
<td>67.80</td>
<td>125.73</td>
<td>57.93</td>
<td>-9.59</td>
</tr>
</tbody>
</table>

Table 7.4: Summary Statistics for Twenty NORM-Imputed Values Under Treatment One
<table>
<thead>
<tr>
<th></th>
<th>RMNI</th>
<th>NORM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>190.84</td>
<td>191.18</td>
</tr>
<tr>
<td><strong>Standard Deviation</strong></td>
<td>3.66</td>
<td>13.02</td>
</tr>
<tr>
<td><strong>Range</strong></td>
<td>18.11</td>
<td>49.23</td>
</tr>
<tr>
<td>$Mean_{Imputed} - Actual$</td>
<td>-1.83</td>
<td>-1.49</td>
</tr>
</tbody>
</table>

Based on 100 Imputations | Based on 20 Imputations

Table 7.5: Performance Comparison of RMNI with NORM

**RMNI and NORM**  With Table 7.5, we provide a summary of Table 7.4 alongside a summary of Table 7.6. Notice with Table 5.5 and Table 7.4 that the imputed values are ostensibly sensible for Treatment One under both the RMNI procedure and the Data Augmentation procedure. The typical standard deviation and the typical range of the imputed values (across missing values) is much larger with NORM (Data Augmentation) than with RMNI. With Table 7.4 and Table 7.6, we omit data under Treatments Two and Three for their redundancy; for Table 7.5, we include imputed values under each of the three treatment effects.

**Table 7.5** To tabulate results for Table 7.5, we use 100 RMNI-imputed data sets, and 20 NORM-imputed data sets. The parameter mean value for the data is simulated to be 200.0: the mean imputed value of Table 7.5 is computed across treatment value, as are the other statistics contributing to the table. RMNI is comparable to NORM in mean imputed value, as well as in mean difference of the imputed value from the actual value. RMNI differs from NORM in its ability to incorporate knowledge of the variance component values, given $Y_{obs}$; notice the difference in value of the typical standard deviation of an imputed
<table>
<thead>
<tr>
<th>$y_{\text{RMNI},i}$</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
<th>Mean - Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_{\text{RMNI},1}$</td>
<td>104.09</td>
<td>3.45</td>
<td>14.64</td>
<td>14.83</td>
</tr>
<tr>
<td>$y_{\text{RMNI},2}$</td>
<td>76.41</td>
<td>3.15</td>
<td>16.33</td>
<td>-10.59</td>
</tr>
<tr>
<td>$y_{\text{RMNI},3}$</td>
<td>88.41</td>
<td>3.63</td>
<td>18.44</td>
<td>16.80</td>
</tr>
<tr>
<td>$y_{\text{RMNI},4}$</td>
<td>101.45</td>
<td>3.81</td>
<td>20.19</td>
<td>4.36</td>
</tr>
<tr>
<td>$y_{\text{RMNI},5}$</td>
<td>99.60</td>
<td>3.65</td>
<td>17.21</td>
<td>18.73</td>
</tr>
<tr>
<td>$y_{\text{RMNI},6}$</td>
<td>91.91</td>
<td>3.57</td>
<td>19.24</td>
<td>-4.25</td>
</tr>
<tr>
<td>$y_{\text{RMNI},7}$</td>
<td>81.79</td>
<td>3.71</td>
<td>19.08</td>
<td>17.24</td>
</tr>
<tr>
<td>$y_{\text{RMNI},8}$</td>
<td>113.18</td>
<td>3.68</td>
<td>17.41</td>
<td>-16.01</td>
</tr>
<tr>
<td>$y_{\text{RMNI},9}$</td>
<td>84.14</td>
<td>4.04</td>
<td>22.78</td>
<td>4.31</td>
</tr>
<tr>
<td>$y_{\text{RMNI},10}$</td>
<td>95.41</td>
<td>3.81</td>
<td>20.74</td>
<td>-13.46</td>
</tr>
<tr>
<td>$y_{\text{RMNI},11}$</td>
<td>97.60</td>
<td>3.35</td>
<td>15.76</td>
<td>-6.80</td>
</tr>
<tr>
<td>$y_{\text{RMNI},12}$</td>
<td>90.15</td>
<td>3.84</td>
<td>18.60</td>
<td>-15.34</td>
</tr>
<tr>
<td>$y_{\text{RMNI},13}$</td>
<td>106.34</td>
<td>3.50</td>
<td>16.33</td>
<td>-5.39</td>
</tr>
<tr>
<td>$y_{\text{RMNI},14}$</td>
<td>95.97</td>
<td>3.40</td>
<td>17.76</td>
<td>12.91</td>
</tr>
<tr>
<td>$y_{\text{RMNI},15}$</td>
<td>103.27</td>
<td>3.96</td>
<td>18.30</td>
<td>2.37</td>
</tr>
<tr>
<td>$y_{\text{RMNI},16}$</td>
<td>112.02</td>
<td>3.72</td>
<td>18.33</td>
<td>-2.36</td>
</tr>
<tr>
<td>$y_{\text{RMNI},17}$</td>
<td>101.25</td>
<td>3.23</td>
<td>14.53</td>
<td>-2.74</td>
</tr>
<tr>
<td>$y_{\text{RMNI},18}$</td>
<td>99.14</td>
<td>4.01</td>
<td>19.85</td>
<td>-4.67</td>
</tr>
</tbody>
</table>

Table 7.6: Summary Statistics for 100 RMNI-Imputed Values Under Treatment One

value, and of the typical range of imputed values for a given missing value. Notably, the performance of NORM is comparable to the performance of the General Exam Method discussed in Section 4.9: the General Exam Method performs the imputation identically to RMNI, except it does not utilize the posterior distribution of $Y_{mis}$ (Equation 4.5) to update the imputed value with observed data.

**Table 7.6** With Table 7.6, we provide a summary of the RMNI-imputed values which are used to substitute for a single missing value, for each of the eighteen missing values manifest under Treatment One. The actual imputed values are exhibited in Table 5.4 for
the first five imputations. Table 7.6 is based upon results with 100 RMNI-imputed data sets. With comparison to NORM, it is notable that even with 100 imputations, we do not achieve the range in imputed value typical of the NORM imputed value, nor do we with 986 imputations using RMNI. While RMNI and NORM both impute values sampled from a distribution of possibilities for the missing value, it is apparent that RMNI succeeds in updating the possible values for a missing datum, given $Y_{obs}$, by honoring the structure of the particular hierarchical model which our data is simulated to follow.

### 7.3 Conclusion

**Potential for RMNI** In conclusion, we consider the problem of performing multiple comparisons of treatment effects under the two-way mixed interaction model. The underlying structure of the variance-covariance matrix for the data (and the model) is sensitive. With the small sample size of the typical set of experimental data, we suspect that imputed values could easily distort the estimate of the variance-covariance matrix, if the imputed values are generated without particular respect to the model for analysis. The procedure of RMNI is proper and randomization-valid for estimation of treatment effects and the associated variance-covariance matrix; the procedure is also proper for the estimation of treatment contrasts, and appears to be proper for the estimation of the associated error provided with Confidence Interval (6.3). We hope to pursue simulation study of RMNI to establish the coverage rate of the RMNI-imputed confidence region for pairwise treatment contrasts under Model (3.4). Additionally, we hope to extend RMNI for use across the family of hierarchical models.

**Advantage of RMNI** The procedure of RMNI for multiple imputations of an experimental data set is fast, randomization-valid, and free of the need for manual assessment.
of convergence. RMNI satisfies the need for an imputation procedure of Paradigm Two (Section 7.1.2): RMNI cannot be used in Paradigm One with typical sample survey data; RMNI is for use with experimental data following a hierarchical model by design. Given the advances of statistical computing over the past decade, RMNI can now be performed with a SAS Macro or an SPlus function, and is suitable for desktop computing.
CHAPTER 8

Prospects for Continued Research

**Intention**  This dissertation is an exposition of Applied Research in statistical methods developed for the handling of missing data. Nevertheless, we have encountered a litany of theoretical and technical issues in the application of what began as a seemingly straightforward approach to managing missing data under a specific interaction model. While we have explored and resolved those concerns most pertinent to the simulated demonstration of RMNI, a few topics remain incompletely researched. With this chapter, we selectively present those issues with the investigation of which, we believe, the performance and interpretive value of RMNI could be advanced.

8.1 **Theoretical Statistics**

**Fraction of Missing Information**  With this dissertation, we consider the *percentage* of data which is missing. However, we have reason to believe that it is the Fraction of Missing Information (FMI) which may be the relevant characteristic of an incomplete data set worthy of further research. In particular, we intuit a correspondence between the pattern of missing data and the FMI. We suspect that it may be fruitful to explore this relationship on a theoretical level, especially in terms of re-arrangement of incomplete data with the goal of optimizing procedures for multiple imputation. (We have seen, for example,
that re-arranging observed data into a monotonic pattern of missingness enables or enhances performance of multiple imputations within the context of Survey Sampling: Little & Rubin (1987), Rubin (1987).) Certainly, we can see with our model that the impact of random concentrations of missingness on a given subject or treatment/subject combination can have dramatic effects in comparison to a uniform dispersion of missing values: each scenario does occur in repeated simulation of MCAR data with a constant rate of missingness. Clearly, one key difference between two patterns of missing data, MCAR with a common rate of missingness, is the Fraction of Missing Information.

**Model Selection** With this dissertation, our focus has been the Two-Way Mixed Effects Interaction Model. An initial criticism of the model-specific procedure we have developed is that tailoring imputed values to fit a speculative theoretical model would likely bias multiply-imputed inference in favor of the imputed model. We have reason to believe that characterization of the impact of model selection on a theoretical level could inform judicious use of procedures for multiple imputation in general. Donald Rubin (1987) advises that the imputed model should be used as the analytic model to be generative of proper inference. This may not be comprehensively advisable: with mild missingness, and an imputed Two-Way Interaction Model, Analysis of Variance may yield an insignificant interaction effect; consequently, analysis need progress under a reduced model. The question remains whether a subsequent multiple imputation under the reduced model need be performed. Ideally, a stochastic metric for the impact of an imputed model upon multiply-imputed inference would present itself.
8.2 Applied Statistics

RMNI for the Family of Hierarchical Models  With this dissertation, we developed Yamashita’s (1995) initial exploration of RMNI into a fully functional Multiple Imputation Procedure, free of binding caveats. It is evident that RMNI can be utilized under any mixed hierarchical model. In fact, RMNI is practicable with any hierarchical model for which Bayesian Statistics has provided an expression for the joint posterior distribution of model parameters, given unbalanced data. The listing of designs for which Box and Tiao (1973) prescribes approximation algorithms for the individual posterior distributions of each parameter is comprehensive. Consequently, the replication of our RMNI protocol under each possible hierarchical model is a reasonable Applied Statistics research project. Should it prove advisable to repeat a multiple imputation under a reduced model, such as in the case of an insignificant interaction term in the full model, an RMNI procedure tailored to the reduced model will be required.

Non-Ignorable Missingness  With this dissertation, we opted to focus on the event of observing ignorable missingness, specifically MCAR. The assumption of an ignorable response mechanism is not realistic if missingness is observed to be systematically dependent upon the observable response value, or known covariates of the response. With experimental data, such may be the case with observance of toxicity, or the observance of wear trend from a machine producing measurements or experimental units. In the event of non-ignorable missingness, a response mechanism modeling the systematic nonresponse need be incorporated into the imputed model. It is not immediately clear that RMNI is resilient to such an adjustment of the imputed model. That is, it is not clear whether Box and Tiao’s (1973) posterior distribution approximations, nor Mardia, Kent, and Bibby’s (1979)
posterior distribution for $Y_{mis}$, would be resilient to such an adjustment to the prescribed hierarchical model for the design of the experiment. We suspect that inasmuch as the model for missingness is independent of the design of the experiment, RMNI may be adjusted to facilitate incorporation of the non-ignorable response mechanism.

### 8.3 Statistical Computing

**Modular RMNI** With this dissertation, algorithmic programs have been developed which perform RMNI under the Two-Way Mixed Interaction Model. As developed, these programs reflect the learning curve through development of RMNI itself. Now functional, the algorithm for RMNI could easily be produced as a portable SAS Macro. Additionally, model-specific RMNI modules could readily be produced in the form of integrated SAS Macros, the body of which could constitute viable application software.
APPENDIX A

Master Program Files

Description  The simulation of incomplete data and performance of RMNI are conducted by the Master Program file miauto_rmni.ksh. This Korn Shell script manages the integration of FORTRAN programs with calls to SAS. It was necessary to use shell script to automate calls to SAS for Type III Sums of Squares and associated Expected Means Squares. Once we used shell script to activate SAS and alternate between SAS and FORTRAN, we chose to take advantage of Korn’s strengths to perform other tasks throughout the process of RMNI. It should be noted that IMSL FORTRAN subroutines are called by FORTRAN programs and subroutines as needed.

A.1  RMNI Master File: miauto_rmni.ksh

Description  The function of miauto_rmni.ksh is to perform a Repeated Measures Normal Imputation as described in Chapter IV. This performance of RMNI utilizes the approximation procedures of Box and Tiao (Section 4.7.2), and the posterior $[Y_{mis}|Y_{obs}]$ provided by Mardia, Kent, and Bibby (Section 4.7.3).

#!/bin/ksh
date > start.tim
$F90 $F90FLAGS compdata.f refmat.f datgen.f $LINK_F90
mv a.out compdata.out
compsas.out
batchsas.ksh batchglm_comp.sas compas.out compas.log
parseglm.ksh compas.out compas.par
parseglm2.ksh compas.par compas.ems
$F90 $F90FLAGS compstat.f refmat.f datgen.f lsmns.f varcmp.f $LINK_F90
mv a.out compstat.out
compstat.out
$F90 $F90FLAGS missdata.f refmat.f datgen.f miss.f subchk.f
reduce.f $LINK_F90
mv a.out missdata.out
missdata.out
seq_file.ksh < minum.dat
batchsas.ksh batchglm_misdat.sas misdat_sas.out misdat_sas.log
parseglm.ksh misdat_sas.out misdat_sas.par
parseglm2.ksh misdat_sas.par misdat_sas.ems
A.2 Master File: miauto.ksh

Description  The function of miauto.ksh is to perform a Repeated Measures Normal Imputation as described in Section 4.9 as the General Exam Method. This procedure utilizes the approximation procedures provided by Box and Tiao (Section 4.7.2). However, it does not utilize the posterior algorithm based upon the posterior distribution for $Y_{mis}$ described in Section 4.7.3.

#!/bin/ksh
date > start.tim
f90 $F90FLAGS compdata.f refmat.f datgen.f $LINK_F90
mv a.out compdata.out
compdata.out
batchsas.ksh batchglm_comp.sas compsas.out compsas.log
parsegml.ksh compsas.out compsas.par
parsegml2.ksh compsas.par compsas.ems
f90 $F90FLAGS compstat.f refmat.f datgen.f
lsmns.f varcmp.f $LINK_F90
mv a.out compstat.out
compstat.out
integer n=1 #initializing loop counter
integer seed=32397
until ((n > 1000)) #n is compared to nsim
do
    ((seed = seed + 1))
    print "$seed" >| seed.dat
cat seed.dat >> seedstore.dat
cp blank.fil mi_totvar.dat
cp blank.fil mi_totlsm.dat
#GEf90 $F90FLAGS missdata.f lsmns.f refmat.f datgen.f miss.f
subchk.f $LINK_F90
#redf90 $F90FLAGS missdata.f lsmns.f refmat.f datgen.f miss.f
subchk.f reduce.f #$LINK_F90
#mv a.out missdata.out
missdata.out
seq_file.ksh < minum.dat
batchsas.ksh batchglm_unbh3.sas misdat_sas.out misdat_sas.log
parsegml.ksh misdat_sas.out misdat_sas.par
parsegml2.ksh misdat_sas.par misdat_sas.ems
#batchsas.ksh batchglm_misdat.sas misdat_sas.out misdat_sas.log
#parsegml.ksh misdat_sas.out misdat_sas.par
#parsegml2.ksh misdat_sas.par misdat_sas.ems

179
#batchsas.ksh batchglm_red.sas datareduced_sas.out datareduced_sas.log
#parseglm.ksh datareduced_sas.out datareduced_sas.par
#parseglm2.ksh datareduced_sas.par datareduced_sas.ems
rm sigmas.foo
rm datcom.foo
rm jnew.dat
#f90 $F90FLAGS multimp.f reftime.f datgen.f miss.f subchk.f
sigmas.f lsmns.f datcom.f milsmns.f $LINK_F90
#mv a.out multimp.out
multimp.out
midata_sas.ksh
rm mivarcmp.dat
#f90 $F90FLAGS mistat.f mivarcmp.f milsevar.f mifinal.f
#mv a.out mistat.out
mistat.out
((n = n + 1)) #n is incremented
rm *mil*.dat
#for file in ./1[0-9][0-9][0-9][0-9].dat
#do
#rm $file
#done
cat mi_totvar.dat >> totvar.dat
cat mi_totlsm.dat >> totlsm.dat
done #simulation loop ended
date > end.tim
APPENDIX B

Programs: Utility

B.1 Code: batchsas.ksh

Description  The function of batchsas.ksh is to access SAS Proc GLM for the computation of Type III Sums of Squares and associated Expected Mean Squares. This script requires prior production of an input file, e.g. compdata.dat or missdata.dat, which is supplied to SAS. This script also requires SAS batch code as input, in a file with extension .sas.

#!/bin/ksh
if (($# > 0))
  #if at least one command line argument follows
  #the command calling this script
  then
    sas_src=$1
    #$1,$2,$3 are default positional parameters
    #for first three arguments
    sas_out=$2
    sas_log=$3
  else
    print -n "error: script argument(s) missing;
    batchsas.ksh src_file.sas out_file.sas log_file.sas"
  fi
exec 2> $sas_log
(sas -stdio < $sas_src > $sas_out) > $sas_log
exec 2>&-

B.2 Code: fcdatain.sas

Description  The function of fcdatain.sas is to create a SAS data set in the format of the Two-Way Mixed Effects Model.

data sasuser.tryfc01;
filename ytuc '/home/lsefc_ytuc.dat';
filename xl '/home/lsefc_xl.dat';
B.3 Code: miautoclean.ksh

Description The function of miautoclean.ksh is to delete unnecessary files from the host directory at the end of a simulation run.

#!/bin/ksh
rm compsas.dat
rm compsas.ems
rm compsas.log
rm compsas.par
rm initstat.dat
rm misdat_sas.dat
rm misdat_sas.ems
rm misdat_sas.log
rm misdat_sas.par
rm datared_sas.dat
rm datared_sas.ems
rm datared_sas.log
rm datared_sas.out
rm datared_sas.par
rm jnew.dat
rm sigmas.foo
rm datcom.foo

B.4 Code: midata_sas.ksh

Description The function of midata_sas.ksh is to supply a sequence of data files to SAS for processing with Proc GLM. That is, Type III Sums of Squares and Expected Mean Squares must be computed for each of the $m$ completed data sets. This script automates this detail for $m$ successively named data files.

#!/bin/ksh
integer index
index=1
for file in mi1[0-9][0-9][0-9][0-9].dat
do
cp $file midata.dat
batchsas.ksh batchglm_midata.sas out_$file log_$file
parseglm.ksh out_$file par_$file

182
parseglm2.ksh par_$file ems_$file
if (( index == 1 ))
then
cp ems_$file mi_ems.dat
index=0
else
cat ems_$file >> mi_ems.dat
fi
done

B.5 Include File: model.par

Description  The function of model.par is to enable changes in model parameter values to be accomplished in one location for the family of programs.

C MODEL.PAR

    INTEGER NSIM
    PARAMETER( NSIM=2 )

    INTEGER MINUM
    PARAMETER( MINUM=5 )

    INTEGER I
    PARAMETER( I=3 )

    INTEGER J
    PARAMETER( J=50 )

    INTEGER K
    PARAMETER( K=3 )

    REAL PCTMIS
    PARAMETER( PCTMIS=.10 )

    INTEGER IJ
    INTEGER IK
    INTEGER JK
    INTEGER IJK

    INTEGER NMISS
    PARAMETER(IJ=I*J,JK=J*K,IK=I*K,IJK=I*J*K,NMISS=I*J*K*PCTMIS)

    REAL MISVAL
    PARAMETER(MISVAL=999999.0)

    REAL SDc
    PARAMETER( SDc=1.0 )

    REAL SDt

183
PARAMETER( SDt=1.0 )

REAL SDe
PARAMETER( SDe=1.0 )

REAL THETA(I)
DATA THETA(1),THETA(2),THETA(3)/10.0,20.0,30.0/

C integer comparisons

INTEGER ZERO,ONE,TWO
PARAMETER(ZERO=0,ONE=1,TWO=2)

B.6 Code: parseglm.ksh

Description  The function of parseglm.ksh is to parse the standard output file produced by SAS Proc GLM and stored in file compsas.out, and produce a file of Type III Sums of Squares and Expected Mean Squares readable by varcmp.f. This script works in conjunction with parseglm2.ksh to produce the final product: compsas.ems. An interim file, compsas.par, is produced and supplied to the second parsing script.

#!/bin/ksh
typeset -l line
#typeset declares variable line to be a string by default;
#-l forces lowercase text

if (($# > 0))
    glm_output=$1
    glm_tokens=$2
else
    print -n "error: arguments missing;
glmparse.ksh sas_output.foo parsed_file.foo"
    fi

IFS="
   \\
"

#IFS is space-tab-newline by default;
#the first element of IFS is the default separator
#of the print command;
#$glm_tokens will hold one parsed goodie per line

while read line
#this while-do-done loop will execute until the read command
#returns a boolean failure status
    do
        if [[ ("$line" = error*) ]] ||
( ("$line" = subj*) && ("$line" = *[.v]*) ) || ("$line" = theta\*subj*) || ("$line" = uncorr*) ]]
# if [[ ("$line" = Error*) ||
# ("$line" = subj*) && ("$line" = *[.v]*) ) ||
# ("$line" = "theta\*subj"*) ]]

#if [[ C1 || ( C2 && C3 ) || C4]]
#C1: catches the line beginning with ‘error’
#C2: catches the lines beginning with ‘subj’
#C3: ensures its the line with decimals
#C4: catches the lines beginning with ‘theta\*subj’
#C5: catches the line beginning with ‘uncorrected total’

then
    set $line
    #set clears all positional parameters and then
    #assigns new ones sequentially
    #to each IFS-element delimited word in variable line
    print "$*"
    #the wildcard causes print to operate on each positional
    #parameter sequentially;
    #each value $* is a word, and each word is printed to
    #$glm_tokens
    #(done statement);
    #the first element of IFS is the separator
fi

done < $glm_output >> $glm_tokens
#done specifies the input for read first and the
destination for print

B.7 Code: parseglm2.ksh

Description  The function of parseglm2.ksh is to parse the input file compsas.par which is produced by
parseglm.ksh. The output of this script is compsas.ems, which is required by varcmp.f to produce estimates
of the variance components. See the description given with code in Section B.6.
# for the first two arguments
ems_data=$2
else
    print -n "error: missing command line arguments"
fi

count=0
while read line # while-do-done will loop until the read command
# returns a boolean failure status
do
    ((count = count + 1))

    line2=$line

    case $line2 in
        "+") line2="1.0";;
        "-" ) (line2="-1.0");;
    esac

    case $count in
        2|4|7|10|12|16|18|24|25|28|29|35|36) print "$line2";;
        #2|4|6|8|12|14|20|21|24|25|31|32) print "$line2";;
        #print -n " ";;
    esac

done < $parseglm_out >> $ems_data # done specifies the input for read
# and the destination for print

B.8 Code: parseglm_mi.ksh

Description The function of parseglm_mi.ksh is to parse the output file of SAS Proc GLM, which holds
the coefficients necessary to perform Scheffe’s ANOVA Method (Section 3.5.2) for obtaining Expected Mean
Squares.

#!/bin/ksh
typeset -l line
# typeset declares variable line to be a string by default; -l forces lowercase text
integer count

if (($# > 0))
    if at least one command line argument follows
    # the command calling this script
    then
        parseglm_out=$1
        #$1,$2 are the default positional parameters
        # for the first two arguments
        ems_data=$2
    else
        print -n "error: missing command line arguments"
    fi
count=0
while read line
# while-do-done will loop until the read command
# returns a boolean failure status
do
  ((count = count + 1))
  line2=$line
  case $line2 in
    "+") line2="1.0";;
    "-") (line2="-1.0");;
  esac
  case $count in
    4|8|14) print "$line2";;
      #print -n " ";
  esac
done < $parseglm_out >> $ems_data
# done specifies the input for read and the destination for print

B.9 Code: parseglm2_mi.ksh

Description This code operates in conjunction with parseglm_mi.ksh. See the description provided with code in Section B.8.

#!/bin/ksh
typeset -l line # typeset declares variable line to be a # string by default; -l forces lowercase text
integer count
if (($# > 0)) # if at least one command line argument
  # follows the command calling this script
then
  parseglm_out=$1 #$1, $2 are the default positional
  parameters for the first two arguments
  ems_data=$2
else
  print -n "error: missing command line arguments"
fi

count=0
while read line
# while-do-done will loop until the read command
# returns a boolean failure status
do
  ((count = count + 1))
  line2=$line
case $count in
  3|6|9) print "$line2";;
  esac

done < $parseglm_out >> $ems_data
#done specifies the input for read and the destination for print

B.10 Code: parsevc.ksh

Description The function of parsevc.ksh is to parse the output file generated by SAS Proc VARCOMP. The product is a text file listing of variance component estimates.

#!/bin/ksh
typeset -l line #typeset declares variable line
  #to be a string by default; -l forces lowercase text
if (($# > 0)) #if at least one command line argument
  #follows the command calling this script
  then
    glm_output=$1 #$1,$2 are the default positional i
      parameters for the first two arguments
    glm_tokens=$2
  else
    print -n "error: arguments missing;
      glmparse.ksh sas_output.foo parsed_file.foo"
  fi
IFS="
  \(\)"
#IFS is space-tab-newline by default;
#the first element of IFS is the default separator i
#of the print command;
#$glm_tokens will hold one parsed goodie per line
while read line #this while-do-done loop will
    execute until the read command returns a boolean failure status
    do
      if [[ ("$line" = var\(subj*) || ("$line" = var\(theta*)
        || ("$line" = var\(error*) )
      )
      #if [[ ("$line" = error*) ||
  #if ( ("$line" = subj*) && ("$line" = *\.[v]*) ) ||
  #("$line" = theta\*subj*) ]]
    #  if [[ ("$line" = Error*) ||
    #("$line" = subj*) && ("$line" = *\.[v]*) )
    #|| ("$line" = "theta\*subj"*) ]]
      #if [[ C1 || ( C2 && C3 ) || C4]]
      #C1: catches the line beginning with 'error’
then

set $line

#set clears all positional parameters and then
#assigns new ones sequentially
#to each IFS-element delimited word in variable line

print "$*"

#the wildcard causes print to operate on each
#positional parameter sequentially;
#each value $* is a word, and each word is
#printed to $glm_tokens (done statement);
#the first element of IFS is the separator

fi

done < $glm_output >> $glm_tokens

#done specifies the input for read first
#and the destination for print

B.11 Code: seq_file.ksh

Description The function of seq_file.ksh is to create the flat file fn.dat which holds the sequence numbers for files to be created by upcoming programs. Note that the sequence is 5 digits and counts from 10000 up to 1000m, whatever the number of imputations m might be.

#!/bin/ksh
integer minum
integer index
integer num
num=10000
index=1
while read minum
do
while ((index <= minum))
done
print $num
index=index+1
num=num+1
done
done >> fn.dat
done
APPENDIX C

Programs: Simulation and Subject Deletion

Step Zero  Note that these programs correspond generally to Step Zero of the RMNI process (Section 4.2.2). For simulation, this includes generation of a complete data set, generation of an MCAR pattern of missingness, and construction of an incomplete data set. This step also includes deletion of completely unobserved subjects from the incomplete data set.

C.1  Main Program: compdata.f

Description  The function of compdata.f is to produce a complete data set following the Two-Way Mixed Interaction Model. Subroutines refmat.f, datgen.f, and IMSL RNNOR are called; include file model.par is utilized. The output is compsas.dat which holds three columns in a flat file: response, treatment ID, and subject ID.

PROGRAM COMPDATA

C Model:  two-way mixed-effects repeated-measures interaction

C This program is intended to produce a single complete data set.
C The output file is to be supplied to shell routines which call
C SAS and ultimately produce a file of values needed to calculate
C estimates of the variance components for this model in a
C subsequent FORTRAN program.

IMPLICIT NONE

INCLUDE ’model.par’

C counters
INTEGER INDEX

EXTERNAL REFMAT,DATGEN

C refmat.f-complete data
INTEGER ORVEC1(IJK),ORVEC2(IJK)

C datgen.f
REAL ORIG(IJK)
REAL GOODc(J),TEMPt(IJ),GOODE(IJK)
RNSET initializes the random generator seed to the argument value.

CALL RNSET(32397)

CALL REFMAT(I,J,K,IJK,ORVEC1,ORVEC2)

CALL DATGEN(I,J,IJ,IJK,SDc,SDt,SDe,THETA,ORIG,ORVEC1,ORVEC2, & GOODc,TEMPt,GODe)

OPEN(UNIT=6,FILE='compsas.dat',STATUS='NEW')

DO index=1,IJK
   WRITE(6,*) & ORIG(index),ORVEC1(index),ORVEC2(index)
END DO

STOP
END

C.2 Main Program: compstat.f

Description The function of compstat.f is to produce a flat file containing parameter values for the model being simulated as well as the parameter estimates which are calculated using the complete simulated data set. The result is a file with two rows. This program calls refmat.f, datgen.f, lsmns.f, and varcmp.f. This program requires prior production of Type III Mean Squares and associated Expected Mean Squares.

PROGRAM COMPSTAT

IMPLICIT NONE

INCLUDE 'model.par'

INTEGER L

EXTERNAL REFMAT,DATGEN,LSMNS,VARCMP

INTEGER ORVEC1(IJK),ORVEC2(IJK)
REAL ORIG(IJK)
REAL GOODc(IJK),TEMt(IJK),GOODe(IJK)
REAL LSMN(I)
REAL SIGRET(3)

C RNSET initializes the random generator seed to the argument value.
CALL RNSET(32397)
CALL REFMAT(I,J,K,IJK,ORVEC1,ORVEC2)
CALL DATGEN(I,J,IJ,IJK,SDc,SDt,SDe,THETA,ORIG,ORVEC1,ORVEC2,
& GOODc,TEMt,GOODe)
CALL LSMNS(I,IJK,LSMN,ORIG,MISVAL,ORVEC1)
CALL VARCMP(SIGRET)
OPEN(UNIT=6,FILE='initstat.dat',STATUS='NEW')
C fix-to automate, must loop 1 to I printing row of lsmn(), then sigrets
WRITE(6,*)
& (THETA(L),L=1,I),SDc**2,SDt**2,SDe**2
WRITE(6,*)
& (LSMN(L),L=1,I),SIGRET(1),SIGRET(2),SIGRET(3)
STOP
END

C.3 Code: datgen.f

Description The function of datgen.f is to generate standard normal draws, rescale them by supplied variance component values, and then sum them according to the specification of Model (3.4). Subroutine IMSL RNNOR is utilized. The result is data simulated to follow a Two-Way hierarchical model.

SUBROUTINE DATGEN(I,J,Nt,Ne,SDc,SDt,SDe,THETA,DATA,VEC1,VEC2,
& GOODc,TEMt,GOODe)
C Model: two-way mixed-effects repeated-measures interaction
C DATGEN generates a complete data set given the parameter values
C SD = standard deviation for c=subject, t=interaction, e=error
C THETA = I treatment effects parameter values
C Nt = I*J = number of interaction terms
C Ne = I*J*K = number of error terms (observations)
IMPLICIT NONE

C counters
INTEGER L1, L2, L3

C model dimensions
INTEGER I, J

C model variance components (standard deviations)
REAL SDc, SDt, SDe

C model treatment effects
REAL THETA(I)

C design indicators
INTEGER VEC1(Ne), VEC2(Ne)

C DATGEN return data
REAL DATA(Ne)

EXTERNAL RNNOR

C IMSL RNNOR
INTEGER Nt, Ne
REAL RNNOR
REAL GOODc(J), TEMPt(Nt), GOODe(Ne)

C intermediate calculations
REAL GOODt(I, J)

C RNNOR generates standard normal variates and stores them in its second argument:
CALL RNNOR(J, GOODc)  
CALL RNNOR(Nt, TEMPt)  
CALL RNNOR(Ne, GOODe)

C Each variate is rescaled by its associated standard deviation:
DO L1 = 1, J
   GOODc(L1) = GOODc(L1) * SDc
END DO

L2 = 1
L3 = 1
DO L1 = 1, Nt
   GOODt(L2, L3) = TEMPt(L1) * SDt
IF (L3.EQ.J) THEN
   L2 = L2 + 1
   L3 = 1
ELSE
L3 = L3 + 1
END IF
END DO

DO L1=1,Ne
GOODe(L1) = GOODe(L1) * SDe
END DO

C DATA holds a complete balanced data set of combined effects indexed by
C VEC1 and VEC2, which are created by refmat.f:

DO L1=1,Ne
DATA(L1) = THETA(VEC1(L1)) + GOODc(VEC2(L1)) +
& GOODt(VEC1(L1),VEC2(L1)) + GOODe(L1)
END DO
RETURN
END

C.4 Code: miss.f

Description The function of miss.f is to replace missing data with the value of 999999.0, and to generate
vectors which serve as the labels for the missing values.

SUBROUTINE MISS(MISVLd,IJKd,NMISSd,ORIGd,DATAd,INDMIS)
C MISS returns DATAd, the data set with missing values indicated by
C MISVLd, along with INDMIS, the indices of the missing values.
INTEGER IJKd,L,NMISSd,INDMIS(NMISSd)
REAL MISVLd,ORIGd(IJKd),DATAd(IJKd)
EXTERNAL RNSRI

C RNSRI is an IMSL subroutine that generates NMISSd uniform random
C variates between 1 and IJKd. These values are stored in INDMIS.
C All arguments must be declared INTEGER.
CALL RNSRI(NMISSd,IJKd,INDMIS)

C The following loop replicates the original data set.
DO L=1,IJKd
DATAd(L)=ORIGd(L)
END DO

C Each value labelled by INDMIS is replaced by MISVLd.
DO L=1, NMISSd
DATAd(INDMIS(L))=MISVLd
END DO
C.5 Main Program: missdata.f

Description The function of missdata.f is to produce an incomplete data set, MCAR, along with initial estimates for the treatment effects, which are calculated using incomplete data (Section 3.3.2). This program calls refmat.f, datgen.f, miss.f, subchk.f, and reduce.f; and also relies on calls to IMSL subroutines RNNOR and RNSRI.

PROGRAM MISSDATA

C Model: two-way mixed-effects repeated-measures interaction

C This program is intended to produce 1 incomplete data set. C The output file is to be supplied to shell routines which call C SAS and ultimately produce a file of values needed to calculate C estimates of the variance components for this model in a C subsequent FORTRAN program.

IMPLICIT NONE

INCLUDE 'model.par'

C counters
INTEGER L1

EXTERNAL REFMAT,DATGEN,MISS,SUBCHK,REDUCE

C lsmns.f:
REAL LSMN(I)

C refmat.f: complete data
INTEGER ORVEC1(IJK),ORVEC2(IJK)

C refmat.f: reduced-balanced data
INTEGER REDVEC1(IJK),REDVEC2(IJK)

C datgen.f
REAL ORIG(IJK),
  & GOODC(IJK),TEMPT(IJK),
  & GOODE(IJK)

C miss.f
INTEGER INDMIS(NMISS)
REAL MISDAT(IJK)

C subchk.f
INTEGER Jnew,IJKnew
INTEGER NMISS2, INDMIS2(NMISS)
INTEGER RDVEC1(IJK), RDVEC2(IJK)
REAL MISDT2(IJK)

C reduce.f
C   INTEGER Jred
C   REAL DATAre(IJK)

C imsl: RNSET
INTEGER SEED

C IMSL RNSET initializes the random generator seed
to the argument value:

OPEN(UNIT=5, FILE='seed.dat', STATUS='OLD')
READ(UNIT=5, FMT=623) SEED
CLOSE(UNIT=5)

C Use seed=32397 for original dataset;
C Read in SEED for simulation,
C the value increments in a loop in the master korn shell.

CALL RNSET(32397)

OPEN(UNIT=6, FILE='minum.dat', STATUS='OLD')
WRITE(6, *) MINUM
CLOSE(UNIT=6)

CALL REFMAT(I, J, K, IJK, ORVEC1, ORVEC2)

CALL DATGEN(I, J, IJ, IJK, SDC, SDT, SDE, THETA, ORIG, ORVEC1, ORVEC2,
&    goodc, tempt, goode)

CALL RNSET(SEED)

CALL MISS(MISVAL, IJK, NMISS, ORIG, misdat, indmis)

CALL SUBCHK(I, J, K, IJK, ORVEC1, ORVEC2, MISDAT, Jnew, IJKnew,
&    MISDT2, MISVAL, NMISS, INDMIS, NMISS2, INDMIS2, RDVEC1, RDVEC2)

C   CALL REDUCE(J, MISVAL, IJK, ORVEC2, MISDAT, Jred, DATAre)
C   CALL REFMAT(I, Jred, K, IJK, RDVEC1, RDVEC2)

OPEN(UNIT=6, FILE='misdat_sas.dat', STATUS='OLD')

DO L1=1, IJKnew
   IF (MISDT2(L1).NE.MISVAL) THEN
      WRITE(6, *) MISDT2(L1), RDVEC1(L1), RDVEC2(L1)
   END IF
END DO
C.6 Main Program: multimp.f

Description  The function of multimp.f is to perform a multiple imputation and produce multiply-imputed data sets along with a file of multiply-imputed LSE’s for the treatment effects. This program employs the General Exam Method discussed in Section 4.9.

PROGRAM MULTIMP

C Model:  two-way mixed-effects repeated-measures interaction
C This program is intended to produce 1 incomplete data set.
C The output file is to be supplied to shell routines which call
C SAS and ultimately produce a file of values needed to calculate
C estimates of the variance components for this model in a
C subsequent FORTRAN program.

IMPLICIT NONE

INCLUDE ’model.par’

EXTERNAL REFMAT,DATGEN,MISS,SUBCHK,SIGMAS,DATCOM,MILSMNS

C refmat.f: complete data
INTEGER ORVEC1(IJK),ORVEC2(IJK)

C datgen.f
REAL ORIG(IJK)
REAL GOODC(IJK), TEMPT(IJK), GOODE(IJK)

C miss.f
INTEGER INDMIS(NMISS)
REAL MISDAT(IJK)

C subchk.f
INTEGER Jnew, IJKnew
INTEGER NMISS2, INDMIS2(NMISS)
INTEGER RDVEC1(IJK), RDVEC2(IJK)
REAL MISDT2(IJK)

C sigmas.f
REAL SIGMAE(MINUM), SIGMAT(MINUM), SIGMAC(MINUM)
REAL CHIe(MINUM), CHIt(MINUM), CHIC(MINUM)

C datcom.f
REAL IMPDAT(IJK, MINUM)
REAL JEFF(NMISS, MINUM), IJEFF(NMISS, MINUM), IJKEFF(NMISS, MINUM)

C milsmns.f
REAL MILSMN(I)

C rnset
INTEGER SEED
OPEN(UNIT=5, FILE='seed.dat', STATUS='OLD')
READ(UNIT=5, FMT=623) SEED
CLOSE(UNIT=5)

623 FORMAT(I6)

C RNSET initializes the random generator seed to the argument value.
CALL RNSET(32397)

CALL REFMAT(I, J, K, IJK, orvec1, orvec2)

CALL DATGEN(I, J, IJK, SDC, SDT, SDE, THETA, ORIG, ORVEC1, ORVEC2,
& goodc, tempt, goode)

CALL RNSET(SEED)

CALL MISS(MISVAL, IJK, NMISS, ORIG, misdat, indmis)

CALL SUBCHK(I, J, K, IJK, ORVEC1, ORVEC2, MISDAT,
& Jnew, IJKnew, MISDT2, MISVAL, NMISS, INDMIS, NMISS2, INDMIS2,
& RDVEC1, RDVEC2)

CALL SIGMAS(MINUM, I, K, SIGMAe, SIGMAT, SIGMAC, CHIe, CHIt, CHIC)
C.7 Code: refmat.f

Description The function of refmat.f is to produce column vectors which serve as labels for treatment effects and subject effects.

```
SUBROUTINE REFMAT(I,J,K,IJK,VEC1,VEC2)
C REFMAT creates 2 indicators to identify an observation
C using numeric labels;
C VEC1 holds the treatment label, 1,...,I;
C VEC2 holds the subject label, 1,...,J

IMPLICIT NONE
C counters
INTEGER L1,L2,L3
C model parameters
INTEGER I,J,K,IJK
C return data
INTEGER VEC1(IJK),VEC2(IJK)
C VEC1 holds J*K 1’s, then J*K 2’s, ....
C with each treatment label
C repeated J*K times:
DO L1=1,I
    DO L2=*((L1-1)*J*K+1),((L1-1)*J*K+1)
        VEC1(L2)=L1
    END DO
END DO
C VEC2 holds 1,2,...,J,1,2,...,J, ....
C with each subject label repeated
C K times:
DO L1=1,I
    DO L2=1,J
        L3=*((L1-1)*J*K+L2-1)*K+1),((L1-1)*J*K+L2*K)
    END DO
END DO
```
C.8 Code: subchk.f

Description  The function of subchk.f is to remove any subject who is completely unobserved from the MCAR data set.

SUBROUTINE SUBCHK(I,J,K,IJK,VEC1,VEC2,MISDAT,Jnew,IJKnew, & MISDT2,MISVAL,NMISS,INDMIS,nmiss2,indmis2,VEC1new,VEC2new)

IMPLICIT NONE

C counters:
INTEGER L1,L2,L3,INDEX

C integer comparison:
INTEGER ZERO
PARAMETER(ZERO=0)

C data set supplied by user
REAL MISDAT(IJK)

C MISDAT parameters
INTEGER I,J,K,IJK
INTEGER INDMIS(NMISS),NMISS
INTEGER VEC1(IJK),VEC2(IJK)
REAL MISVAL

C SUBCHK interim calculations
INTEGER DELETE(J,I),JKEEP(J)

C return data set and parameters
REAL MISDT2(IJK)

C INTEGER Junb(I)
INTEGER Jnew,IJKnew
INTEGER INDMIS2(NMISS),NMISS2
INTEGER VEC1new(IJK),VEC2new(IJK)

C DELETE is an indicator of whether a subject j is nowhere observed
C on treatment i:

INDEX=0

DO 100, L1=1,J
   DO 90, L2=1,I

VEC2(L3)=L2
END DO
END DO
END DO
RETURN
END

INDEX=0

DO 100, L1=1,J
   DO 90, L2=1,I

DELETE(J,I)=INDEX
JKEEP(J)=""
DO L3=(((L2-I)*J+L1-1)*K+1),(((L2-I)*J+L1)*K)
    IF (MISDAT(L3).EQ.MISVAL) THEN
        INDEX=INDEX+1
    END IF
END DO
    IF (INDEX.EQ.K) THEN
        DELETE(L1,L2)=1
    ELSE
        DELETE(L1,L2)=0
    END IF
    INDEX=0
90   CONTINUE
100  CONTINUE
C Junb holds the number of subjects observed at least once under
C treatment i, for each treatment i; use Junb if final data analysis
C is appropriate for unbalanced data:
C
C    DO L1=1,I
C        Junb(L1)=0
C    END DO
C
C    DO L1=1,I
C        DO L2=1,J
C            Junb(L1) = Junb(L1) + (1 - DELETE(L2,L1))
C        END DO
C    END DO
C
C We intend to perform standard complete-data analysis on MISDT2 once
C an imputation has been conducted. Therefore, we remove a subject
C from the data set only if that subject is nowhere observed.
C JKEEP = I where a subject is nowhere observed.
DO L1=1,J
    JKEEP(L1)=0
END DO
DO L1=1,I
    DO L2=1,J
        JKEEP(L2) = JKEEP(L2) + DELETE(L2,L1)
    END DO
END DO
C Jnew is the number of subjects somewhere observed:
Jnew=0
DO L1=1,J
    IF (JKEEP(L1).NE.I) THEN
Jnew = Jnew + 1
END IF
END DO

INDEX=0
DO L1=1,IJK
   IF (JKEEP(VEC2(L1)).NE.I) THEN
      INDEX = INDEX + 1
      MISDT2(INDEX)=MISDAT(L1)
      VEC1new(INDEX)=VEC1(INDEX)
      VEC2new(INDEX)=VEC2(INDEX)
   END IF
END DO

C The following note and loop is designed to remove subjects from C inclusion with treatments under which no observation is obtained:

C IJKnew is the number of observations for MISDT2:

IJKnew=INDEX

C NMISS2 holds the number of missing values in MISDT2:

NMISS2=0
DO L1=1,IJKnew
   IF (MISDT2(L1).EQ.MISVAL) THEN
      NMISS2 = NMISS2 + 1
      INDMIS2(NMISS2)=L1
   END IF
END DO

RETURN
END
APPENDIX D

Programs: Calculating Initial Estimates

Step One  Note that these programs correspond generally to Step One of the RMNI process (Section 4.2.2). This includes calculation of initial estimates using incomplete- and unbalanced-data methods. This step also includes the optional use of Experimentwise Subject Deletion (Section 3.4.2).

D.1 Code: batchglm_comp.sas

Description  The function of batchglm_comp.sas is to run SAS Proc GLM to produce Type III Sums of Squares and Expected Mean Squares. The output file is stored as compas.out, and the log file is stored as compas.log, as written in miauto.ksh. The output file, compas.out, is fed to parseglm.ksh to be parsed into component parts. The result is a flat file with a listing of only desired input for varcmp.f: the SS’s and E(MS)’s.

data sasuser.compdata;
filename ytuc ‘/home/compsas.dat’;
infile ytuc;
input y theta subj;
run;
proc glm data=sasuser.compdata;
class theta subj;
model y=theta subj theta*subj/noint ss3;
random subj theta*subj;
run;
quit;

D.2 Code: batchglm_misdat.sas

Description  The function of batchglm_misdat.sas is to run SAS Proc GLM to produce Type III Sums of Squares and Expected Mean Squares. The output file is stored as misdat_sas.out, and the log file is stored as misdat_sas.log, as written in miauto.ksh. This program is identical to batchglm_comp.sas, except that the computation is with unbalanced MCAR data.

data sasuser.missdata;
filename ytuc ‘/home/misdat_sas.dat’;
infile ytuc;
input y theta subj;
run;
proc glm data=sasuser.missdata;
class theta subj;
model y=theta subj theta*subj/noint ss3;
random subj theta*subj;
run;
quit;

D.3 Code: clsevar.f

Description  The function of clsevar.f is to compute the variances and covariances which constitute \( \text{Var}(\hat{\theta}) \).

SUBROUTINE CLSEVAR(VARCMP,LSEVAR)

C CLSEVAR will calculate the scalar variance of each least squares
C estimate of a treatment effect, 1...I; also, scalar covariance
C of any ordered pair of lse’s is calculated; the variance-covariance
C matrix of the Ix1 vector of lse’s is constructed by using the
C variances as diagonal elements, and the covariance for each off-
C diagonal element.

IMPLICIT NONE

INCLUDE 'model.par'

C VARCMP is calculated by varcmp.f
C VARCMP(1) = subject
C VARCMP(2) = interaction
C VARCMP(3) = error

REAL VARCMP(3)

C LSEVAR is returned:
C The first column holds the common variance of the treatment effect
C estimates; the second column holds the common covariance of
C pairs of effect estimates.

REAL LSEVAR(2)

C LSEVAR is calculated, variance in col 1, covariance in col 2:

LSEVAR(1) =
& ( VARCMP(3) + (VARCMP(1) + VARCMP(2))*K ) / (J*K)

LSEVAR(2) = VARCMP(1)/J

RETURN
END
D.4 Code: reduce.f

Description The function of reduce.f is to create a balanced data set from an incomplete data set by removing any subject with a missing value.

SUBROUTINE REDUCE(J,MISVAL, IJK, VEC2, MISDAT, Jred, DATAred)

C Model: Two-Way Mixed-Effects Repeated-Measures Interaction:

C REDUCE produces the reduced data set DATAred wherein all subjects with incomplete data have been deleted, leaving a balanced data set.

IMPLICIT NONE

C counters
INTEGER L1, INDEX

C integer comparison
INTEGER ZERO
PARAMETER(ZERO=0)

C model parameters
INTEGER J, IJK

C input data set
INTEGER VEC2(IJK)
REAL MISDAT(IJK)
REAL MISVAL

C intermediate calculations
INTEGER ISMISS(J)

C return data
INTEGER Jred
REAL DATAred(IJK)

C ISMISS(j) is a binary indicator of incomplete subjects j=1,...,J:

DO L1=1, J
   ISMISS(L1)=0
END DO

DO L1=1, IJK
   IF (MISDAT(L1).EQ.MISVAL) THEN
      ISMISS(VEC2(L1))=1
   END IF
END DO

C Jred is the number of fully-observed subjects:

Jred=0
DO L1=1, J
Jred = Jred + (1 - ISMISS(L1))
END DO

C DATAre = the reduced-balanced data set:

INDEX = 1
DO L1 = 1, IJK
  IF (ISMISS(VEC2(L1)).EQ.ZERO) THEN
    DATAre(INDEX) = MIDSAT(L1)
    INDEX = INDEX + 1
  END IF
END IF
RETURN
END

D.5 Code: varcmp.f

Description  The function of varcmp.f is to perform Scheffe’s ANOVA method for estimating variance components using Type III Sums of Squares and the associated Expected Mean Squares, as computed by SAS Proc GLM. See Section 3.5.2 for details.

SUBROUTINE VARCMP(SIGINI)

C VARCMP is used to solve the system of linear equations produced
C by the SAS GLM expected mean squares procedure.
C SIGINI is returned with unbiased variance component estimates
C for subject-c (1), interaction-t (2), and error-e (3).
C SIGINI does not hold standard deviations!

IMPLICIT NONE

INTEGER I, DFe, Dft, DFc, IJK
REAL ZERO
PARAMETER(ZERO=0.0)

C The following variables are read in from external file ems.sas.
C The file is created by glmparse2.ksh with glmparse.ksh.
C The file contains mean squares and coefficients from SAS GLM
C Type III expected mean squares output.

REAL MSE, MST, MSC, SIGN1, COEFF1, SIGN2, COEFF2, SIGN3, COEFF3

C SIGINI is the return variable.

real SIGINI(3)

OPEN(UNIT=5,FILE='compsas.ems',STATUS='OLD')
READ(UNIT=5,FMT=623) DFe
READ(UNIT=5,FMT=523) MSe

206
C The following loop initializes SIGINI with values of 0.0.

DO I=1,3
   SIGINI(I)=0.0
END DO

C SIGINI(3)=MSE from ANOVA table, unbiased for the error variance.

SIGINI(3)=MSE

C SIGINI(2) is unbiased for the interaction variance.

SIGINI(2)=(MST-MSE)/(SIGN3*COEFF3)

C SIGINI(1) is unbiased for the subject variance.

SIGINI(1)=(MSC-MSE-SIGN1*COEFF1*SIGINI(2))/(SIGN2*COEFF2)

DO I=1,3
   IF (SIGINI(I).LT.ZERO) THEN
      SIGINI(I)=0.0
   END IF
END DO

323 FORMAT(F4.0)
523 FORMAT(F20.0)
623 FORMAT(I6)

RETURN
END
APPENDIX E

Programs: Posterior Draws for Parameter Values

Step Three  Note that these programs correspond generally to Step Three of the RMNI process (Section 4.2.2). This algorithm is inspired by Box and Tiao (1973), and remains model-specific. This program is intended for use with only the Two-Way Mixed Interaction Model; any variation on this model requires adjustment to the approximation algorithm. See Section 4.7.2.

E.1 Code: sigmas.f

Description  The function of sigmas.f is to compute the approximation to the posterior distribution of each treatment effect per guidelines of Box and Tiao (1973). This subroutine returns random draws from the approximated posteriors as standard deviations for each random effect. See code comments for details.

```fortran
SUBROUTINE SIGMAS(MINUM,I,K,SIGMAe,SIGMAt,SIGMAc,CHIe,CHIt,CHIc)


C SIGMAS calculates estimates of component variances according to the
c two stage inverse chi squared approximation method proposed in
C Box and Tiao (1973,p287)
C Each SIGMA* is returned, and holds a random sample of size MINUM
C from the posterior distribution of one variance component--the values
C are standard deviations.
C e = error
C t = interaction
C c = subject

IMPLICIT NONE

C counters
INTEGER L1

C integer comparison
INTEGER ZERO,ONE
PARAMETER(ZERO=0,ONE=1)

C model/imputation parameters
INTEGER I,K
INTEGER MINUM
```

208
C Box and Tiao (1973) variables:

C degrees of freedom
REAL Ve,Vt,Vc
REAL Ve2,Vt2,Vt3

C mean squares
REAL Me,Mt,Mc
REAL Me2,Mt2,Mt3

C intermediate calculations
REAL X1,X2,X3
REAL p,q
REAL A1,A2,A3

C truncation boundaries for chi-squared variates
REAL TRUNCc,TRUNCT

C SIGMAS return data
REAL SIGMae(MINUM),SIGMat(MINUM),SIGMac(MINUM)

EXTERNAL BETAI,RNCHI,RNSET

C IMSL BETAI
REAL BETAI
REAL I1,I2,I3

C IMSL RNCHI
REAL RNCHI
C REAL CHIe(MINUM),CHIt(MINUM),CHIc(MINUM)

C IMSL RNSET
INTEGER SEED

OPEN(UNIT=5,FILE='misdat_sas.ems',STATUS='OLD')
READ(UNIT=5,FMT=323) Ve
READ(UNIT=5,FMT=523) Me
READ(UNIT=5,FMT=323) Vc
READ(UNIT=5,FMT=523) Mc
READ(UNIT=5,FMT=323) Vt
READ(UNIT=5,FMT=523) Mt

323 FORMAT(F20.0)
523 FORMAT(F20.0)
OPEN(UNIT=5,FILE='seed.dat',STATUS='OLD')
READ(UNIT=5,FMT=524) SEED
CLOSE(UNIT=5)
524 FORMAT(I6)

C The details of the following calculations are in
Box and Tiao (1973)
in section 5.3.5.
Formulas: bot 287, top 289 (5.3.24)

\[ X_1 = \frac{(V_c M_c)}{(V_t M_t + V_c M_c)} \]
\[ p = 0.5 \times V_c \]
\[ q = 0.5 \times V_t \]

The following are incomplete beta variates obtained from the IMSL subroutine BETAI for which the arguments are real. Multiplication by (0.5) above ensures the result is real.

\[ I_1 = \text{BETAI}(X_1, p, (q + 2.0)) \]
\[ I_2 = \text{BETAI}(X_1, p, (q + 1.0)) \]
\[ I_3 = \text{BETAI}(X_1, p, q) \]
\[ A_1 = (q + 1.0) \times \frac{I_1}{I_2} - q \times \frac{I_2}{I_3} \]
\[ V_{t_2} = \frac{V_t}{A_1} \times \frac{I_2}{I_3} \]
\[ M_{t_2} = \frac{V_t M_t}{A_1 V_{t_2}} \]

Formulas: (5.3.27)

\[ X_2 = \frac{(V_t^2 M_{t_2})}{(V_t^2 M_{t_2} + V_e M_e)} \]
\[ p = 0.5 \times V_{t_2} \]
\[ q = 0.5 \times V_e \]

\[ I_1 = \text{BETAI}(X_2, p, (q + 2.0)) \]
\[ I_2 = \text{BETAI}(X_2, p, (q + 1.0)) \]
\[ I_3 = \text{BETAI}(X_2, p, q) \]
\[ A_2 = (q + 1.0) \times \frac{I_1}{I_2} - q \times \frac{I_2}{I_3} \]
\[ V_{e_2} = \frac{V_e}{A_2} \times \frac{I_2}{I_3} \]
\[ M_{e_2} = \frac{V_e M_e}{A_2 V_{e_2}} \]

The following is discussed in Box and Tiao (1973) in section 5.3.7.
Formulas (5.3.33)

\[ X_3 = \frac{(V_t M_t)}{(V_e M_e + V_t M_t)} \]
\[ p = 0.5 \times V_t \]
\[ q = 0.5 \times V_e \]

\[ I_1 = \text{BETAI}(X_3, (p + 2.0), q) \]
\[ I_2 = \text{BETAI}(X_3, (p + 1.0), q) \]
I3 = BETAI(X3,p,q)
A3 = (p+1.0)*(I1/I2) - p*(I2/I3)
Vt3 = (Vt/A3)*(I2/I3)
Mt3 = (Vt*Mt)/(A3*Vt3)

C RNCHI is an IMSL subroutine which generates chi-squared variates. Its first argument is integral, the second real, and the third stores the output. The TRUNCation values are provided in sections 5.3.5 (error, unconstrained), 5.3.6 (interaction), and 5.3.7 (subject). Table 5.3.3 on page 292 summarizes the whole process. Formulas: TRUNCC (bot 291); TRUNCD (Section 5.3.6 in text)

TRUNCC = (Vc*Mc)/Mt3
TRUNCT = (Vt2*Mt2)/Me

CALL RNCHI(MINUM,Vt2,CHIt)

DO L1=1,MINUM
623 IF (CHIt(L1).GT.TRUNCT) THEN
   CALL RNCHI(ONE,Vt2,CHIt(L1))
   GOTO 623
END IF
END DO

CALL RNCHI(MINUM,Vc,CHIc)

DO L1=1,MINUM
723 IF (CHIc(L1).GT.TRUNCC) THEN
   CALL RNCHI(ONE,Vc,CHIc(L1))
   GOTO 723
END IF
END DO

CALL RNCHI(MINUM,Ve2,CHIe)

DO L1=1,MINUM
SIGMAe(L1) = (Ve2*Me2/CHIe(L1))**(0.5)
SIGMAT(L1) = ( (Vt2*Mt2/CHIt(L1) - Me) / real(K) )**(0.5)
SIGMAC(L1) = ( (Vc*Mc/CHIc(L1) - Mt3) / (real(K)*real(I)) )**(0.5)
END DO

open(unit=6,file='sigmas.dat',status='NEW')
do 11=1,minum
write(6,*) sigmae(11),sigmat(11),sigmac(11)
end do
RETURN
END
APPENDIX F

Programs: Posterior Draws for $Y_{mis}$

Step Five  Note that these programs correspond generally to Step Five of the RMNI process (Section 4.2.2). We utilized two approaches to imputation: one which constructs imputed values using model assumptions, the General Exam Method of Section 4.9, and another method (preferred) which utilizes the result of Mardia, Kent, and Bibby (1979) for $[Y_{mis}, Y_{obs}]$ which is provided in Section 4.2.2 by Equations (4.4) and (4.5). These programs are coded for use with only the Two-Way Mixed Interaction Model. We recommend adjustment to the algorithm with any variation in model assumption. It is this step which is most critical to preserving the structure of the underlying variance-covariance matrix of the data (or the model). See Section 4.7.3.

F.1  Code: datcom.f

Note  This code is used with the Simulation Master File miauto.ksh as provided in Appendix A.2. The properties of Multiple Comparisons using this approach (General Exam Method, Section 4.9) have not been thoroughly explored. Nevertheless, the imputed values are valid and do possess model-specific distributional properties; this algorithm may be preferred for its simplicity and do-ability.

Description  The function of datcom.f is to perform an imputation by using posterior draws of the variance components obtained in Step Three of RMNI. It is observable that this method is effective in the presence of mild to moderate missingness for obtaining completed data sets and computing estimates of treatment effects and the associated variance-covariance matrix.

SUBROUTINE DATCOM(MISDAT,SIGMAC,SIGMAT,SIGMAE,MISVAL,MINUM,
  & I,J,K,VEC1,VEC2,NE,NMISS,IMPDAT,INDMIS,
  & jeff,ijeff,ijkeff)

C DATCOM uses the variance component estimates derived by SIGMAS
C to create values for imputation, and then conducts the imputation.
C The final product of DATCOM is an array of MINUM completed data
C sets which are stored in IMPDAT().

IMPLICIT NONE

C counters
  INTEGER L,M,P
C integer comparison
   INTEGER ONE
   PARAMETER(ONE=1)

C DATCOM arguments-model/imputation parameters
   INTEGER I,J,K,NE
   INTEGER MINUM,NMISS,INDMIS(NMISS)
   REAL MISVAL

C DATCOM arguments-user supplied data
   REAL MISDAT(NE)
   INTEGER VEC1(NE),VEC2(NE)
   REAL SIGMAC(MINUM),SIGMAT(MINUM),SIGMAE(MINUM)

C DATCOM arguments-returned variables
   REAL IMPDAT(NE,MINUM)

C indicator variables
   INTEGER IJLST1(NE),IJLST2(NE),JLIST(J),IJKLS1(NE),IJKLS2(NE)
   INTEGER ISMISS(J),ISIJOK(J,I)

C random effects for imputed values
   INTEGER Jnum,IJnum,IJKnum
   REAL JEFF(NMISS,MINUM),IJEFF(NMISS,MINUM),IJKEFF(NMISS,MINUM)

C LSMNS arguments
   REAL TRTMNS(I)

C intermediate calculations
   REAL JVAL,IJVAL
   REAL IMPVAL(NMISS,MINUM)

C output filename array
   CHARACTER*5 FN(MINUM)

EXTERNAL LSMNS,RNNOR

C LSMNS calculates the least squares estimate for each treatment effect.
   CALL LSMNS(I,NE,TRTMNS,MISDAT,MISVAL,VEC1)

C The following loops create lists of specific effects required to
C create an imputed value.
C IJKLS1 and IJKLS2 hold the treatment and subject labels corresponding
C to each missing datum in MISDAT:
   IJKnum=0
   DO L=1,NE
IF (MISDAT(L).EQ.MISVAL) THEN
   IJKnum = IJKnum + 1
   IJKLS1(IJKnum)=VEC1(L)
   IJKLS2(IJKnum)=VEC2(L)
END IF
END DO

C ISIJOK is an indicator of which trt/sub combinations are missing
data; ultimately, IJLST1 and IJLST2 hold the treatment and subject
c labels that correspond to each unique interaction effect that is
c somewhere unobserved:

DO L=1,I
   DO M=1,J
      ISIJOK(M,L)=0
   END DO
END DO

DO L=1,NE
   IF (MISDAT(L).EQ.MISVAL) THEN
      ISIJOK(VEC2(L),VEC1(L))=1
   END IF
END DO

IJnum=0
DO L=1,I
   DO M=1,J
      IF (ISIJOK(M,L).EQ.ONE) THEN
         IJnum = IJnum + 1
         IJLST1(IJnum)=L
         IJLST2(IJnum)=M
      END IF
   END DO
END DO

C ISMISS is an indicator of which subjects are somewhere unobserved;
C JLIST holds the labels for each subject somewhere not observed:

DO L=1,J
   ISMISS(L)=0
END DO

DO L=1,IJnum
   ISMISS(IJLST2(L))=1
END DO

Jnum=0
DO L=1,J
   IF (ISMISS(L).EQ.ONE) THEN
      Jnum = Jnum + 1
      JLIST(Jnum)=L
   END IF
END DO
C JEFF holds Jnum subject effects;
C IJEFF holds IJnum interaction effects; and
C IJKEFF holds IJKnum error effects:

\[
L = NMISS \times MINUM
\]

CALL RNNOR(L,JEFF)
CALL RNNOR(L,IJEFF)
CALL RNNOR(L,IJKEFF)

DO L=1,NMISS
  DO M=1,MINUM
    JEFF(L,M) = JEFF(L,M) * SIGMAC(M)
    IJEFF(L,M) = IJEFF(L,M) * SIGMAT(M)
    IJKEFF(L,M) = IJKEFF(L,M) * SIGMAE(M)
  END DO
END DO

C IMPVAL holds an array of imputed values:

DO 610, P=1,MINUM
  DO 600, L=1,NMISS
    DO M=1,IJNUM
      IF ((IJLST1(M).EQ.IJKLS1(L)).AND.(IJLST2(M).EQ.IJKLS2(L))) THEN
        IJVAL=IJEFF(M,P)
      END IF
    END DO
    DO M=1,Jnum
      IF (JLIST(M).EQ.IJKLS2(L)) THEN
        JVAL=JEFF(M,P)
      END IF
    END DO
    IMPVAL(L,P)=TRTMNS(IJKLS1(L))+JVAL+IJVAL+IJKEFF(L,P)
  END DO
  600 CONTINUE
  610 CONTINUE

C IMPDAT holds MINUM completed data sets, and is returned by DATCOM:

DO L=1,MINUM
  DO M=1,NE
    IMPDAT(M,L)=MISDAT(M)
  END DO
END DO

DO L=1,MINUM
DO M=1,NMISS
   IMPDAT(INDMIS(M),L)=IMPVAL(M,L)
END DO
END DO

C To be analyzed by SAS GLM, each completed data set is written to an
C external file; the array FN allows sequential names for the output
C files:

OPEN(UNIT=5,FILE=`fn.dat’,STATUS=’OLD’)
DO L=1,MINUM
   READ(UNIT=5,FMT=923) FN(L)
END DO

923 FORMAT(A5)

OPEN(UNIT=6,FILE=`jnew.dat’,STATUS=’NEW’)
WRITE(6,*') J

DO P=1,MINUM
   OPEN(UNIT=6,FILE=`mi’//FN(P)//’.dat’,STATUS=’NEW’)
   DO M=1,NE
      WRITE(6,*') IMPDAT(M,P),VEC1(M),VEC2(M)
   END DO
   END DO

OPEN(unit=6,file=`datcom.foo’,status=’NEW’)
DO L=1,NMISS
   WRITE(6,*') vec1(indmis(l)),vec2(indmis(l)),misdat(indmis(l)),
   & (impdat(indmis(l),m),m=1,minum)
END DO

RETURN
END

F.2 Code: rmni.f

Description The function of rmni.f is to perform a Repeated Measures Normal Impuation procedure:
random samples from the posterior distributions of the variance components are produced by sigmas.f; they
are used in conjunction with the posterior distribution \( Y_{mis} | Y_{obs} \) provided by Mardia, Kent, and Bibby (1979),
to generate imputed values. The output of this subroutine is multiply-imputed data (\( m \) completed data sets).

SUBROUTINE RMNI(SIGMAe,SIGMAT,SIGMAC,
   & YinMIS,YinID1,YinID2,RMNIIdt)

IMPLICIT NONE

INCLUDE 'model.par'

C Loop Counters:
INTEGER L,M,N,P,MIcnt,L1

C Array Dimension: Nfull=IJKnew of subchk.f
C INTEGER Nfull,Nmis,Nobs !Declared in model.par after
C I look at the data in EXCEL. Automating dynamically
C changing matrix sizes was overwhelming to complete.

C Box and Tiao posterior draws:
REAL SIGMAe(MINUM),SIGMAT(MINUM),SIGMAC(MINUM)
REAL BtSIG(MINUM,THREE),BTtmp(THREE)

C Data Vectors: NMISS.NE.Nmis; Nfull.NE.IJK (subject deletion)
C YinMIS = MISDT2, YinID1=VEC1new, YinID2=VEC2new from subchk.f.
   REAL YinMIS(IJK)
   INTEGER YinID1(IJK),YinID2(IJK)
C YinID1 holds treatment labels 1,2,...,I;
C YinID2 holds subject labels 1,2,...,50 excepting unobserved subjects;
C YinID3 is intended to hold repetition labels.
   INTEGER YinID3(Nfull)

C YinMI is the complet-ed data in form of [Yobs,Y-imputed];
C RMNIIdt is the complet-ed data in form of YinMIS.
   REAL YmkbMI(Nfull,MINUM)
   REAL RMNIIdt(Nfull,MINUM)

C Means Storage: TRTmns is LSMNS output;
C MNS* internal holds for intermediate
C calculation with correct vector dimensions;
C Qvec is an intermediate MKB calc
   REAL TRTmns(I)
   REAL MNSinc(Nfull),MNSmis(Nmis),MNSobs(Nobs)
   REAL Qvec(Nobs)

C Variance Covariance Matrices:
   REAL INCVAR(Nfull,Nfull)
   REAL MKB11(Nobs,Nobs),MKB12(Nobs,Nmis)
   REAL MKB21(Nmis,Nobs),MKB22(Nmis,Nmis)

C LSGRR: TOL declared in model.par:
   REAL IRANK
   REAL GINV11(Nobs,Nobs)

C CHFAC & RNMVN:
INTEGER IRANK2
REAL RSIG(Nmis,Nmis)

C Imputed Value Storage and File Management:
REAL IMPtmp(Nmis,1),IMPval(Nmis,MINUM)
CHARACTER*5 FN(MINUM)

C MKB Interim Calculations:
C MRRRR: M2111I = MKB21 * MKB11-Inverse = MKB21 * GINV11
REAL M2111I(Nmis,Nobs)
C MRRRR: CMAT = MKB21 * inv(MKB11) * MKB12 = M2111I * MKB12
REAL Cmat(Nmis,Nmis)
C Dmat = MKB22 - Cmat
REAL Dmat(Nmis,Nmis)
C Emat = M2111I * Qvec
REAL Emat(Nmis)
C Fmat = MNSmis + Emat
REAL Fmat(Nmis)

C Calls external modules:
EXTERNAL LSMNS

C Call to IMSL:
EXTERNAL LSGRR,MRRRR,CHFAC,RNMVN

C Start imputation counter: 1,2,...,MINUM
MIcnt=0

C Create Yinc, Ymis, Yobs, Ymkb = [Yobs,Ymis]

  do L=1,Nfull,K
    do M=1,K
      YinID3(L+M-1)=M
    end do
  end do
  do L=1,Nfull
    Yinc(L,1)=YinMIS(L)
    Yinc(L,2)=YinID1(L)
    Yinc(L,3)=YinID2(L)
    Yinc(L,4)=YinID3(L)
  end do

L1=1
  do L=1,Nfull
    if (Yinc(L,1).EQ.MISVAL) then
      Ymis(L1,1)=Yinc(L,1)
      Ymis(L1,2)=Yinc(L,2)
      Ymis(L1,3)=Yinc(L,3)
      Ymis(L1,4)=Yinc(L,4)
    end if
  end do
L1=L1+1
end if
end do

L1=1
do L=1,Nfull
  if (Yinc(L,1).NE.MISVAL) then
    Yobs(L1,1)=Yinc(L,1)
    Yobs(L1,2)=Yinc(L,2)
    Yobs(L1,3)=Yinc(L,3)
    Yobs(L1,4)=Yinc(L,4)
    L1=L1+1
  end if
end do

do L=1,Nmis
  Ymkb(L+Nobs,1)=Ymis(L,1)
  Ymkb(L+Nobs,2)=Ymis(L,2)
  Ymkb(L+Nobs,3)=Ymis(L,3)
  Ymkb(L+Nobs,4)=Ymis(L,4)
end do

do L=1,Nobs
  Ymkb(L,1)=Yobs(L,1)
  Ymkb(L,2)=Yobs(L,2)
  Ymkb(L,3)=Yobs(L,3)
  Ymkb(L,4)=Yobs(L,4)
end do

C Use lsmns.f to obtain LSmns(1,2,...,I); and create MNS* vectors:
C MNSobs = vector of means matching treatment ID for Yobs
C MNSmis = vector of means matching treatment ID for Ymis
call lsmns(I,Nfull,TRTmns,YinMIS,MISVAL,YinID1)

do L=1,Nobs
  do M=1,I
    if (Yobs(L,2).EQ.real(M)) then
      MNSobs(L) = TRTmns(M)
    end if
  end do
end do

do L=1,Nmis
  do M=1,I
    if (Ymis(L,2).EQ.real(M)) then
      MNSmis(L) = TRTmns(M)
    end if
  end do
end do

C GOTO sends here after completion of an imputation.
C Create vector number MICTN: updates with iteration.
   BTtmp(1) = SIGMAc(MICTN)
   BTtmp(2) = SIGMAa(MICTN)
   BTtmp(3) = SIGMAe(MICTN)

C Create the variance-covariance matrix of Ymbk: INCVAR(nfullxnfull)
   do L=1,Nfull
      do M=1,Nfull
         INCVAR(L,M)=0.0
      end do
   end do

   do L=1,Nfull
      do M=1,Nfull
         if (Ymbk(L,3).EQ.Ymbk(M,3)) then
            INCVAR(L,M) = INCVAR(L,M) + BTtmp(1)
         end if
      end do
   end do

   do L=1,Nfull
      do M=1,Nfull
         if ((Ymbk(L,2).EQ.Ymbk(M,2)).AND.
             (Ymbk(L,3).EQ.Ymbk(M,3))) then
            INCVAR(L,M) = INCVAR(L,M) + BTtmp(2)
         end if
      end do
   end do

   do L=1,Nfull
      do M=1,Nfull
         if ((Ymbk(L,2).EQ.Ymbk(M,2)).AND.
             (Ymbk(L,3).EQ.Ymbk(M,3)).AND.
             (Ymbk(L,4).EQ.Ymbk(M,4))) then
            INCVAR(L,M) = INCVAR(L,M) + BTtmp(3)
         end if
      end do
   end do

C Create partitions of the Var(Ymbk):
C MKB11 = Var(Yobs)
C MKB12 = COV(Yobs,Ymis)
C MKB21 = COV(Ymis,Yobs)
C MKB22 = Var(Ymis)
   do L=1,Nobs
      do M=1,Nobs
         MKB11(L,M) = INCVAR(L,M)
      end do
   end do
end do

do L=1,Nobs
    do M=1,Nmis
        MKB21(L,M) = INCVAR(L,Nobs,M)
    end do
end do

do L=1,Nmis
    do M=1,Nobs
        MKB22(L,M) = INCVAR(L,M+Nobs)
    end do
end do

do L=1,Nmis
    do M=1,Nmis
        MKB22(L,M) = INCVAR(L,M+Nobs)
    end do
end do

C Prepare for MKB calculations: [Ymis|Yobs] is N(A,B);
C must calculate A and B; MKB pg 63, dissertation Section 4.2.2.
C Obtain mu-1 (obs) and mu-2 (mis):
    do L=1,Nobs
        YobVEC(L) = Yobs(L,1)
    end do
    do L=1,Nobs
        Qvec(L) = YobVEC(L) - MNSobs(L)
    end do

C Obtain Inverse of MKB11:
call LSGRR(Nobs,Nobs,MKB11,Nobs,TOL,IRANK,GINV11,Nobs)

C Obtain MKB21 * inv(MKB11)
call MRRRR(Nmis,Nobs,MKB21,Nmis,Nobs,Nobs,GINV11,Nobs,
    & Nmis,Nmis,Cmat,Nmis)

C Obtain MKB22 - MKB21*inv(MKB11)*MKB12 = MKB22 - Cmat:
    do L=1,Nmis
        do M=1,Nmis
            Dmat(L,M) = MKB22(L,M) - Cmat(L,M)
        end do
    end do

C Obtain MKB21*inv(MKB11)*Qvec:
    do L=1,Nmis
        Emat(L) = 0.0
    end do
    do L=1,Nmis

221
do M=1,Nobs
    Emat(L) = Emat(L) + M211I(L,M)*Qvec(M)
end do
end do

do L=1,Nmis
    Fmat(L) = MNSmis(L) + Emat(L)
end do

C MKB Theorem: Ymis(1) | Yobs(1) is MVN(Fmat,Dmat)
C This is done using IMSL RNMVN which requires a
C Cholesky factorization of the Var-Cov (Dmat) as input.
call CHFAC(Nmis,Dmat,Nmis,TOL,IRANK2,RSIG,Nmis)
call RNMVN(Nmis,One,RSIG,Nmis,IMPtmp,Nmis)
do L=1,Nmis
    IMPval(L,MICnt) = IMPtmp (L,1) + Fmat(L)
end do

if (MICnt.LT.MINUM) then
    goto 123
else
    goto 223
end if

223 do L=1,Nobs
do M=1,MINUM
    YmkbMI(L,M) = Ymkb(L,1)
end do
end do
do L=1,Nmis
do M=1,MINUM
    YmkbMI(L+Nobs,M) = IMPval(L,M)
end do
end do
do L=1,Nfull
do N=1,Nfull
    if ((Yinc(L,2).EQ.Ymkb(N,2)).AND.
        & (Yinc(L,3).EQ.Ymkb(N,3)).AND.
        & (Yinc(L,4).EQ.Ymkb(N,4))) then
        do M=1,MINUM
            RMNIdt(L,M) = YmkbMI(N,M)
        end do
    end if
end do
end do
end do

222
C To be analyzed by SAS GLM, each completed data set is written to an
C external file; the array FN allows sequential names
C for the output files:
OPEN(UNIT=5,FILE='fn.dat',STATUS='OLD')
DO L=1,MINUM
   READ(UNIT=5,FMT=923) FN(L)
END DO

923 FORMAT(A5)

OPEN(UNIT=6,FILE='jnew.dat',STATUS='NEW')
WRITE(6,*) J

DO P=1,MINUM
   OPEN(UNIT=6,FILE='mi'//FN(P)//'.dat',STATUS='NEW')
   DO M=1,Nfull
      WRITE(6,*) RMNIdt(M,P),YinID1(M),YinID2(M)
   END DO
END DO

RETURN
END
APPENDIX G

Programs: Completed-Data Analysis

Steps Six, Seven and Eight  Note that these programs correspond generally to the final steps of RMNI as well as subsequent analysis. Some may be utilized in Step One to obtain initial estimates. These modules are designed to compute multiply-imputed estimates using completed data per specifications of Rubin (1987). The methods and estimators are described throughout Chapters II and III of this dissertation.

G.1 Code: lsmns.f

Description  The function of lsmns.f is to compute the Least Squares Means for treatment effects under specifications of Model (3.4). See Section 3.3.1 for the calculating formula and distributional properties of each estimator.

```fortran
SUBROUTINE LSMNS(I,IJK,LSMN,MISDAT,MISVAL,VEC1)

IMPLICIT NONE

INTEGER I,IJK,L,M,VEC1(IJK),COUNT

REAL MISVAL,MISDAT(IJK),LSMN(I)

COUNT=0

DO L=1,I
    LSMN(L)=0.0
END DO

DO L=1,I
    DO M=1,IJK
        IF ((VEC1(M).EQ.L).AND.(MISDAT(M).NE.MISVAL)) THEN
            LSMN(L)=LSMN(L)+MISDAT(M)
            COUNT=COUNT+1
        END IF
    END DO
    LSMN(L)=LSMN(L)/COUNT
    COUNT=0
END DO

RETURN
END
```

224
G.2  Code: mifinal.f

Description  The function of mifinal.f is to compute the combined multiply-imputed estimates of treatment effects, and the associated variance-covariance matrix. In particular, this subroutine computes the *between variance* noted as the component of variability due to missingness. See Section 2.5.2.

```fortran
SUBROUTINE MIFINAL(I,MINUM,VARCOV)

IMPLICIT NONE

C Loop counters:
INTEGER L1,L2,L3

C Integer arithmetic:
INTEGER ONE
PARAMETER(ONE=1)

C Model/imputation parameters:
C I=number of treatments;
C MINUM=number of completed data sets
INTEGER I,MINUM

C MIFINAL arguments supplied by user:
REAL VARCOV(MINUM,2)

C Read in from external file:
REAL TRTMNS(MINUM,I)

C Returned by MIFINAL:
REAL MIMEANS(I),MITOTVAR(I,I)

C Intermediate calculations:
REAL CTRMNS(MINUM,I),AVGVAR(2),BETVAR(I,I)

C Real-valued vector and array cells are initialized with values of 0.0:
AVGVAR(1) = 0.0
AVGVAR(2) = 0.0

DO L1=1,I
   MIMEANS(L1) = 0.0
   DO L2=1,I
      BETVAR(L1,L2) = 0.0
   END DO
END DO
```

225
C Treatment effect estimates:

TRTMNS holds MINUM rows of I treatment effect estimates, which are calculated using the preferred estimator; MIMEANS holds the I multiply-imputed treatment effect estimates which are obtained as the average of the MINUM completed-data estimates stored in TRTMNS for each effect i=1,...,I.

OPEN(UNIT=5,FILE='mi_lsm.dat',STATUS='OLD')

DO L1=1,MINUM
   DO L2=1,I
      READ(UNIT=5,FMT=323,END=999) TRTMNS(L1,L2)
   END DO
END DO

323 FORMAT(F20.0)

DO L1=1,I
   DO L2=1,MINUM
      MIMEANS(L1) = MIMEANS(L1) + TRTMNS(L2,L1)
   END DO
   MIMEANS(L1) = MIMEANS(L1)/MINUM
END DO

C Average variance-covariance:

VARCOV holds MINUM pairs (variance,covariance) where it is expected that the var-cov matrix of the preferred estimator is of the form (quantity-1)(identity matrix) + (quantity-2)(IxI matrix of ones), so that the TRTMNS have a common variance (q-1 + q-2) and each ordered pair has a common covariance (q-2). AVGVAR holds the two values necessary to construct the mean var-cov matrix over the MINUM completed-data var-cov estimates; see U sub-m bar on pg 76 of Rubin (1987).

DO L1=1,2
   DO L2=1,MINUM
      AVGVAR(L1) = AVGVAR(L1) + VARCOV(L2,L1)
   END DO
   AVGVAR(L1) = AVGVAR(L1)/MINUM
END DO

C Between variance among the MINUM completed-data trt effect estimates:

BETVAR holds an array of elements of the IxI var-cov matrix of the MINUM completed-data estimates of I treatment effects; see B sub m on pg 76 of Rubin (1987).

C CTRMNS = TRTMNS adjusted by MIMEANS.
DO L1=1,I
   DO L2=1,MINUM
      CTRMNS(L2,L1) = TRTMNS(L2,L1) - MIMEANS(L1)
   END DO
END DO

C Upper triangle of BETVAR is filled with sums of products contributing
C to the variance (covariance) of the associated trt effect (pair).

DO L1=1,MINUM
   DO L2=1,I
      DO L3=L2,I
         BETVAR(L2,L3) = BETVAR(L2,L3)+CTRMNS(L1,L2)*CTRMNS(L1,L3)
      END DO
   END DO
END DO

C Var-cov completed with division by (MINUM-1); lower triangle mirrors
C upper triangle of BETVAR; BETVAR is multiplied by (1+1/MINUM) for
C its contribution to the total variance, MITOTVAR.

DO L1=1,I
   DO L2=L1,I
      BETVAR(L1,L2) = BETVAR(L1,L2) / (MINUM - ONE)
      BETVAR(L1,L2) = BETVAR(L1,L2) * (1.0 + 1.0/MINUM)
   END DO
END DO

C Total variance of the multiply-imputed estimate of treatment effects:

C MITOTVAR holds the IxI var-cov matrix of MIMEANS;
C see T sub m on pg 76 of Rubin (1987).

C MITOTVAR is initialized with values of BETVAR upper triangle.

DO L1=1,I
   DO L2=L1,I
      MITOTVAR(L1,L2) = BETVAR(L1,L2)
   END DO
END DO

C MITOTVAR diagonal elements hold variances.

DO L1=1,I
   MITOTVAR(L1,L1) = MITOTVAR(L1,L1) + AVGVAR(1)
END DO

C MITOTVAR off-diagonal elements hold covariances;
C the lower triangle is filled with upper triangle values.
DO L1=1,(I-ONE)
    DO L2=(L1+ONE),I
        MITOTVAR(L1,L2) = MITOTVAR(L1,L2) + AVGVAR(2)
        MITOTVAR(L2,L1) = MITOTVAR(L1,L2)
    END DO
END DO

OPEN(UNIT=6,FILE='mi_totvar.dat',STATUS='OLD')
DO L1=1,I
    WRITE(6,*) (MITOTVAR(L1,L2),L2=1,I)
END DO

999 CLOSE(UNIT=5)
CLOSE(UNIT=6)

RETURN
END

G.3 Code: milsevar.f

Description The function of milsevar.f is to compute the variances and covariances which constitute \(\text{Var}(\hat{\theta})\).

SUBROUTINE MILSEVAR(I,K,MINUM,VARCMP,LSEVAR)
C MILSEVAR will calculate the scalar variance of each least squares C estimate of a treatment effect, 1...I; also, scalar covariance C of any ordered pair of lse’s is calculated; the variance-covariance C matrix of the IxI vector of lse’s is constructed by using the C variances as diagonal elements, and the covariance for each off- C diagonal element.

IMPLICIT NONE

C Loop counters:

INTEGER L1

C MILSEVAR arguments supplied by user:
C VARCMP is calculated by mivarcmp.f;
C I=number of treatment effects;
C K=number of repetitions; and
C MINUM=number of imputations.

INTEGER I,K,MINUM
REAL VARCMP(MINUM,3)

C Read from external data:
C Jnew is the number of subjects assumed for the completion of
C the incomplete data set; any subject nowhere observed is
C nowhere imputed; the value of Jnew is provided by jnew.dat
C which is created by datcom.f.

INTEGER Jnew

C Returned by MILSEVAR: LSEVAR
C The first column holds the common variance of the treatment effect
C estimates; the second column holds the common covariance of
C pairs of effect estimates. The m-th row corresponds with the
C m-th multiply-imputed data set, m=1,...,MINUM.

REAL LSEVAR(MINUM,2)

C Jnew is read:
OPEN(UNIT=5,FILE='jnew.dat',STATUS='OLD')
READ(UNIT=5,FMT=323) Jnew
323 FORMAT(I4)

C LSEVAR is calculated, variance in col 1, covariance in col 2:

DO L1=1,MINUM
   LSEVAR(L1,1) = 
   & ( VARCMP(L1,3) + (VARCMP(L1,1) + VARCMP(L1,2))*K ) / (Jnew*K)
   LSEVAR(L1,2) = VARCMP(L1,1)/JNEW
END DO

RETURN
END

G.4 Code: milsmns.f

Description The function of milsmns.f is to compute the combined multiply-imputed estimate of each
treatment effect in the model, assuming Model (3.4). This combination method is introduced by Rubin (1987)
and discussed in Section 2.5.2.

SUBROUTINE MILSMNS(I,IJK,MILSMN,IMPDAT,VEC1,MINUM)

IMPLICIT NONE

INTEGER L,M,P,COUNT
INTEGER I,IJK,MINUM
INTEGER VEC1(IJK)
REAL IMPDAT(IJK,MINUM),LSMN(MINUM,I),MILSMN(I)

COUNT=0

229
DO L=1,I
    MILSMN(L)=0.0
    DO M=1,MINUM
        LSMN(M,L)=0.0
    END DO
END DO

DO 500, P=1,MINUM
  DO 400, L=1,I
    DO 300, M=1,1JK
      IF (VEC1(M).EQ.L) THEN
        LSMN(P,L)=LSMN(P,L)+IMPDAT(M,P)
        COUNT=COUNT+1
      END IF
  300 CONTINUE
  LSMN(P,L)=LSMN(P,L)/COUNT
  COUNT=0
400 CONTINUE
500 CONTINUE

OPEN(UNIT=6,FILE='mi_lsm.dat',STATUS='OLD')
DO M=1,MINUM
  DO L=1,I
    WRITE(6,*) LSMN(M,L)
  END DO
END DO

DO L=1,I
  DO M=1,MINUM
    MILSMN(L) = MILSMN(L) + LSMN(M,L)
  END DO
  MILSMN(L) = MILSMN(L) / MINUM
END DO

OPEN(UNIT=6,FILE='mi_totlsm.dat',STATUS='OLD')
WRITE(6,*) (MILSMN(L),L=1,I)
RETURN
END

G.5  Code: mivarcmp.f

Description  The function of mivarcmp.f is to perform Scheffe’s ANOVA Method for obtaining variance component estimates. The procedure is performed for each of m completed data sets, as one subroutine call. See Section 3.5.2.
SUBROUTINE MIVARCMP(VARCMP,MINUM)

C MIVARCMP is used to solve the system of linear equation produced
C by the SAS GLM expected mean squares procedure.
C VARCMP is returned with unbiased variance component estimates
C for subject (1), interaction (2), and error (3).
C VARCMP does not hold standard deviations!

IMPLICIT NONE

INTEGER I,J,MINUM
REAL ZERO
PARAMETER(ZERO=0.0)

C The following variables are read in from external file ems.sas.
C The file is created by glmparse2.ksh with glmparse.ksh.
C The file contains mean squares and coefficients from SAS GLM
C Type III expected mean squares output.

INTEGER DFe(MINUM),DFt(MINUM),DFc(MINUM)
REAL MSE(MINUM),MST(MINUM),MSC(MINUM)
REAL SIGN1(MINUM),SIGN2(MINUM),SIGN3(MINUM)
REAL COEFF1(MINUM),COEFF2(MINUM),COEFF3(MINUM)

C VARCMP is the return variable.

REAL VARCMP(MINUM,3)

OPEN(UNIT=5,FILE='mi_ems.dat',STATUS='OLD')

DO I=1,MINUM
  READ(UNIT=5,FMT=623) DFe(I)
  READ(UNIT=5,FMT=523) MSE(I)
  READ(UNIT=5,FMT=623) DFt(I)
  READ(UNIT=5,FMT=523) MST(I)
  READ(UNIT=5,FMT=623) DFc(I)
  READ(UNIT=5,FMT=523) MSC(I)
  READ(UNIT=5,FMT=523) SIGN1(I)
  READ(UNIT=5,FMT=523) COEFF1(I)
  READ(UNIT=5,FMT=523) SIGN2(I)
  READ(UNIT=5,FMT=523) COEFF2(I)
  READ(UNIT=5,FMT=523) SIGN3(I)
  READ(UNIT=5,FMT=523) COEFF3(I)
END DO

C The following loop initializes VARCMP with values of 0.0.

DO I=1,MINUM
  DO J=1,3
    VARCMP(I,J)=0.0
  ENDDO
END DO
DO I=1,MINUM
C VARCMP(3)=MSE from ANOVA table, unbiased for the error variance.

VARCMP(I,3)=MSE(I)
C VARCMP(2) is unbiased for the interaction variance.

VARCMP(I,2)=(MSC(I)-MSE(I))/(SIGN3(I)*COEFF3(I))
C VARCMP(1) is unbiased for the subject variance.

VARCMP(I,1)=(MST(I)-MSE(I)-
 & SIGN1(I)*COEFF1(I)*VARCMP(I,2))/(SIGN2(I)*COEFF2(I))

END DO
DO I=1,MINUM
DO J=1,3
   IF (VARCMP(I,J).LT.ZERO) THEN
      VARCMP(I,J)=0.0
   END IF
END DO
END DO

OPEN(UNIT=6,FILE='mivarcmp.dat',STATUS='NEW')
DO I=1,MINUM
   DO J=1,3
      WRITE(6,*) VARCMP(I,J)
   END DO
END DO

523 FORMAT(F20.0)
623 FORMAT(I6)

RETURN
END


233


