EVALUATING PREDICTORS OF AN INDIVIDUAL’S DIETARY INTAKE LATENT VALUE UNDER DIFFERENT MIXED MODELS

Shuli Yu
School of Public Health and Health Sciences

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_2

Part of the Biostatistics Commons, Nutrition Commons, and the Public Health Commons

Recommended Citation
Yu, Shuli, "EVALUATING PREDICTORS OF AN INDIVIDUAL’S DIETARY INTAKE LATENT VALUE UNDER DIFFERENT MIXED MODELS" (2014). Doctoral Dissertations. 131.
https://scholarworks.umass.edu/dissertations_2/131

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
EVALUATING PREDICTORS OF AN INDIVIDUAL’S DIETARY INTAKE
LATENT VALUE UNDER DIFFERENT MIXED MODELS

A Dissertation Presented

by

SHULI YU

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2014

Public Health
Biostatistics
EVALUATING PREDICTORS OF AN INDIVIDUAL’S DIETARY INTAKE LATENT VALUE UNDER DIFFERENT MIXED MODELS

A Dissertation Presented

by

SHULI YU

Approved as to style and content by:

______________________________
Edward J. Stanek III, Chair

______________________________
Penelope S. Pekow, Member

______________________________
Hui K. Hsieh, Member

Edward J. Stanek III, Department Chair
Department of Public Health
DEDICATION

To my husband Bo, and my daughters Ava and Olivia.
ACKNOWLEDGMENTS

I am extremely grateful to my advisor, Dr. Edward J. Stanek III, for his many years of guidance with patience, support and strong encouragement. His guidance not only made this dissertation research successful, but also will benefit my future career and development. I will always cherish the many invaluable experiences I obtained while working closely with him.

I would especially like to thank the members of my committee, Dr. Penelope S. Pekow and Dr. Hui K. Hsieh, for their helpful comments and suggestions. I truly appreciate Dr. Rongheng Lin, for his helpful suggestions.

A special thank you to all the members of the Cluster Project (R01 HL071828-01A2) since this study is an extension of the research team’s work. I want to express my appreciation to Dr. Ira S. Ockene for access to the Seasons study data (NHLBI, R01-HL52745). My gratitude also to Dr. Yunsheng Ma for providing the literature of the Seasons study.
ABSTRACT

EVALUATING PREDICTORS OF AN INDIVIDUAL’S DIETARY INTAKE LATENT VALUE UNDER DIFFERENT MIXED MODELS

MAY 2014

SHULI YU

B.Med., HARBIN MEDICAL UNIVERSITY, CHINA

Ph.D., FUDAN UNIVERSITY, CHINA

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Edward J. Stanek III

The accurate estimation of an individual’s usual dietary intake is important since the estimates are essential to uncover the diet-disease relationships, and are valuable to establish a diagnosis, determine a patient’s treatment, or prevent diseases and health problems related to the individual diet. This study explores a more accurate method to estimate an individual’s latent value of usual dietary intake when it is repeatedly measured using a 24-hour dietary recall (24HR) and seven day dietary recall (7DDR), accounting for random measurement error and bias. We use a finite population mixed model (FPMM) to link identifiable subjects in evaluating the predictor at the individual (cluster) level. The performance of the (empirical) predictor of subject’s latent value obtained under the FPMM framework is compared with those obtained under the usual mixed model and the measurement error model through a simulation study. The performance of (empirical) predictors based on the 24HR and 7DDR combined data are compared with those based on the 24HR data. We analyze the predictor of latent value in
two cases – for a randomly selected subject and for a specific subject. The approach is illustrated by using dietary intake data from the Seasons study.

The simulation results reveal the predictor based on a FPMM is optimal for a randomly selected subject, but not uniformly optimal for a specific subject. For the subjects with different latent values and different within-subject variances, the distribution of optimal predictors forms a special cluster pattern.

For a randomly selected subject, only when variances are known, the predictor from the combined data is better than the predictor from the 24HR data. For a specific subject, the (empirical) predictor from the combined data is better except for those subjects with relatively larger bias in 7DDR measures.

The contributions of this study are that it provided guidance for predicting subject’s latent value of usual saturated fat intake using two source data (24HR, 7DDR) via a mixed model framework; examined the performance of predictors conditional on sampled subject; and showed that WLS estimator is a biased estimator of the average latent value of the population when within-subject variances vary.
CONTENTS

ACKNOWLEDGMENTS .................................................................................................................. v

ABSTRACT ................................................................................................................................. vi

LIST OF TABLES ......................................................................................................................... xi

LIST OF FIGURES ....................................................................................................................... xiv

CHAPTER

1. INTRODUCTION ..................................................................................................................... 1

1.1 Motivation ............................................................................................................................. 1

1.2 Objectives ............................................................................................................................ 6

1.3 Context — Seasons Study .................................................................................................... 7

1.3.1 24-hour Dietary Recall (24HR) ......................................................................................... 8

1.3.2 Seven-day Dietary Recall (7DDR) ...................................................................................... 9

1.3.3 Preliminary Descriptive Statistics .................................................................................... 10

2. LITERATURE REVIEW ......................................................................................................... 14

2.1 Dietary Assessment Instruments .......................................................................................... 14

2.2 Studies in Estimation of Individual Dietary Intake ............................................................... 17

2.3 Studies on Multiple Source Data ......................................................................................... 21

3. SPECIFICATIONS OF MODELS ............................................................................................ 24

3.1 The Population and Population Parameters ........................................................................ 24

3.2 Measurement Error Model (ME) ........................................................................................ 26

3.3 Usual Mixed Model (MM) .................................................................................................... 29

3.4 Finite Population Mixed Model (FPMM) ............................................................................ 36

3.4.1 Two-stage Sampling ........................................................................................................ 37

3.4.2 Finite Population Mixed Model (FPMM) ........................................................................ 40

3.5 7DDR Model ......................................................................................................................... 46

3.5.1 7DDR Model under the Usual Mixed Model .................................................................. 54

3.5.2 7DDR Model under the FPMM ....................................................................................... 57
3.6 Simultaneous Model ..................................................................................................................61

3.6.1 Simultaneous Model under the Usual Mixed Model ..............................................................62
3.6.2 Simultaneous Model under the FPMM ..................................................................................64

4. ESTIMATOR/EMPIRICAL PREDICTORS OF AN INDIVIDUAL’S DIETARY INTAKE LATENT VALUE IN THE SEASONS STUDY .................................................................67

4.1 Estimating $\sigma_s^2$, $\sigma_r^2$ and $\sigma_{ir}^2$ Using the 24HR Data ...........................................69
4.2 Estimating $\sigma_s^2$, $\sigma_r^2$, $\sigma_{ir}^2$, $\sigma_w^2$ and $\sigma_{iw}^2$ Using the 24HR+7DDR Data ......71
4.3 Estimator/Empirical Predictor of Subject’s Latent Value .......................................................73

5. SIMULATION STUDY ...................................................................................................................75

5.1 Determining Subject’s Parameters ............................................................................................75
5.2 Selecting Subjects ......................................................................................................................78
5.3 Generating 24HR Data ............................................................................................................79
5.4 Generating 7DDR Data ............................................................................................................81

6. COMPARING ESTIMATOR/PREDICTORS OF SUBJECT’S LATENT VALUE (SIMULATED 24HR DATA) ..................................................................................................................84

6.1 Known Variance Components ..................................................................................................84

6.1.1 Estimator/Predictor of the Latent Value for a Randomly Selected Subject ......................................................85

6.1.1.1 Relation of the Predictor $\hat{P}_2^{(i)}$ (MM) and the Weighted Population Mean $\mu^*$ .....................91

6.1.2 Estimator/Predictor of the Latent Value of a Randomly Selected Subject, Given the Realized Subject is Subject $s$ .................................................................................................96

6.1.2.1 Bias of the Predictor $\hat{P}_3^{(i)}$ (FPMM) ....................................................................................97

6.2 Unknown Variance Components ...............................................................................................113

6.2.1 Estimator/Empirical Predictor of the Latent Value for a Randomly Selected Subject ......................................................114

6.2.2 Estimator/Empirical Predictor of the Latent Value of a Randomly Selected Subject, Given the Realized Subject is Subject $s$ .................................................................................................115

7. COMPARING PREDICTORS OF SUBJECT’S LATENT VALUE (SIMULATED 24HR AND 7DDR DATA) ..................................................................................................................124
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1  Estimated overall means and variance components of daily saturated fat intake based on observed 24HR and 7DDR measures (mixed model results)</td>
<td>11</td>
</tr>
<tr>
<td>1.2 Frequency of 24HR recalls and 7DDR recalls by recall occasions</td>
<td>13</td>
</tr>
<tr>
<td>4.1 Estimates of the estimated variance parameters using the 24HR data</td>
<td>70</td>
</tr>
<tr>
<td>4.2 Estimates of $\hat{\sigma}_{ir}^2$ for the 127 subjects using the 24HR data</td>
<td>71</td>
</tr>
<tr>
<td>4.3 Variance and covariance matrix of 16 repeated measures</td>
<td>71</td>
</tr>
<tr>
<td>4.4 Structure of parametric variance and covariance matrix of the 16 repeated measures</td>
<td>72</td>
</tr>
<tr>
<td>4.5 Estimates of the estimated variance parameters using the 24HR+7DDR data</td>
<td>72</td>
</tr>
<tr>
<td>4.6 Estimates of $\hat{\sigma}<em>{ir}^2$, $\hat{\sigma}</em>{iw}^2$ using the 24HR+7DDR data</td>
<td>73</td>
</tr>
<tr>
<td>4.7 Estimates of the estimator/empirical predictor of subject’s latent value</td>
<td>74</td>
</tr>
<tr>
<td>5.1 Frequency of 24HR recalls and 7DDR recalls by recall occasions</td>
<td>75</td>
</tr>
<tr>
<td>5.2 Parameter values of each subject for 24HR and 7DDR</td>
<td>77</td>
</tr>
<tr>
<td>5.3 Population parameter values of the 7 subjects (24HR)</td>
<td>78</td>
</tr>
<tr>
<td>5.4 Parameter values of 3 selected subjects in 4 samples</td>
<td>79</td>
</tr>
<tr>
<td>5.5 Simulated 12 24HR recalls of selected subject $i$</td>
<td>81</td>
</tr>
<tr>
<td>5.6 Simulated 4 7DDR recalls of selected subject $i$</td>
<td>82</td>
</tr>
<tr>
<td>6.1 Characteristics of a population of $N = 7$</td>
<td>87</td>
</tr>
<tr>
<td>6.2 Estimates of the estimator/predictor and the squared errors of $\hat{P}^{(i)}$ by $c$ and $i$</td>
<td>88</td>
</tr>
<tr>
<td>6.3 Average of the estimator/predictor $\hat{P}^{(i)}$ and MSE of $\hat{P}^{(i)}$ by $i$</td>
<td>89</td>
</tr>
</tbody>
</table>
6.4 Average of the estimator/predictor $\hat{P}^{(i)}$ and MSE of $\hat{P}^{(i)}$ ........................................... 89

6.5 Population mean $\mu$ and the weighted population mean $\mu^*$ of $N = 7$ ....................... 90

6.6 Example of $\tilde{w}_s^*$ computation based on 15 possible sample sets when subject $s = 1$ ........................................................................................................................................... 92

6.7 Parameter $\tilde{\mu}^*$ — weighted mean based on the possible sample sets when $N = 7$ and $n = 3$ ..................................................................................................................................................... 93

6.8 Average of the estimator/predictor $\hat{P}^{(i)}$ and MSE of $\hat{P}^{(i)}$ 
($N = 50, n = 3, C = 20,000, \mu = 27.45$) ................................................................. 94

6.9 Comparisons of the $mse^{(i)}$ and the RPI for different sampling settings............... 95

6.10 Estimates of the estimator/predictor $\hat{P}^{(i)}$ and the squared error of $\hat{P}^{(i)}$ sorted
by $s$ ............................................................................................................................................... 96

6.11 Estimates of the estimator/predictor $\hat{P}^{(s)}$ and MSE of $\hat{P}^{(s)}$ by $s$ ...................... 97

6.12 Comparing two biases, $B_3^{(s)}$ and $\tilde{B}_3^{(i)}$ ................................................................. 98

6.13 Descriptive statistics of cluster profile based on the $N = 444$ subjects .................. 112

6.14 Average of the estimator/empirical predictor $\hat{P}^{(i)}$ and MSE/empirical
MSE of $\hat{P}^{(i)}$ ......................................................................................................................... 114

6.15 Comparisons of the $mse^{(i)}/mse_{\epsilon}^{(i)}$ and the RPI for different sampling settings..... 115

6.16 Estimates of the estimator/empirical predictor $\hat{P}^{(s)}$ and MSE/empirical
MSE of $\hat{P}^{(s)}$ by $s$ .................................................................................................................. 116

7.1 Average of the predictor $\hat{P}^{(i)}$ and MSE of $\hat{P}^{(i)}$ ................................................. 125

7.2 Comparisons of the $mse^{(i)}$ and the RPI for different sampling settings............... 126

7.3 Estimates of the predictor $\hat{P}^{(s)}$ and MSE of $\hat{P}^{(s)}$ by $s$
($N = 7, n = 3, C = 20,000$) .................................................................................................. 127
7.4 Average of the empirical predictor $\hat{P}^{(i)}$ and the empirical MSE of $\hat{P}^{(i)}$
\hspace{1em} ($N = 7$, $n = 3$, $C = 1,000$) ..................................................................................140

7.5 Comparisons of the empirical $mse_e^{(i)}$ and the RPI for different sampling settings
\hspace{1em} ($C = 1,000$) ........................................................................................................141

7.6 Estimates of the empirical predictor $\hat{P}^{(s)}$ and the empirical MSE of $\hat{P}^{(s)}$ by
\hspace{1em} $s$ ($N = 7$, $n = 3$, $C = 1,000$) ..................................................................................142

8.1 Summary of conclusions when variance components are unknown..................155
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Distributions of saturated fat intake for average 24HR recalls and 7DDR recalls...11</td>
</tr>
<tr>
<td>1.2</td>
<td>Scatter plot of average 7DDR vs. average 24HR saturated fat intake..............12</td>
</tr>
<tr>
<td>1.3</td>
<td>Scatter plots of standard deviation verse mean of 12 measures from 24HR and 4 measures from 7DDR.................................................................13</td>
</tr>
<tr>
<td>3.1</td>
<td>7DDR weeks relative to 24HR days..........................................................50</td>
</tr>
<tr>
<td>5.1</td>
<td>Distributions of subject’s mean and standard deviation by 24HR recall occasions...............................................................76</td>
</tr>
<tr>
<td>5.2</td>
<td>Distributions of subject’s mean and standard deviation by 7DDR recall occasions...............................................................76</td>
</tr>
<tr>
<td>6.0</td>
<td>Regression plot for $B_j^{(s)}$ (Bias</td>
</tr>
<tr>
<td>6.1</td>
<td>Plot of mse difference (mse2-mse1) vs. subject’s latent value (N=200, n=50, c=20000, 24HR, mse1 and mse2).............................................100</td>
</tr>
<tr>
<td>6.1</td>
<td>Plot of mse difference (mse2-mse1) vs. subject’s latent value (N=444, n=200, c=20000, 24HR, mse1 and mse2).............................................101</td>
</tr>
<tr>
<td>6.1</td>
<td>Enlarged Figure 6.1a without the dot with large MSE difference (N=200, n=50, c=20000, 24HR, mse1 and mse2).............................................102</td>
</tr>
<tr>
<td>6.1</td>
<td>Plot of subject’s true StdDev vs. latent value by mse (N=200, n=50, c=20000, 24HR, mse1 and mse2).............................................103</td>
</tr>
<tr>
<td>6.2</td>
<td>Plot of mse difference (mse3-mse1) vs. subject’s latent value (N=200, n=50, c=20000, 24HR, mse1 and mse3).............................................103</td>
</tr>
<tr>
<td>6.2</td>
<td>Plot of subject’s true StdDev vs. subject’s latent value by mse (N=200, n=50, c=20000, 24HR, mse1 and mse3).............................................104</td>
</tr>
<tr>
<td>6.3</td>
<td>Plot of mse difference (mse3-mse2) vs. subject’s latent value (exclude one mse, N=200, n=50, c=20000, 24HR, mse2 and mse3).................105</td>
</tr>
</tbody>
</table>
6.3b Plot of subject’s true StdDev vs. subject’s latent value by mse (N=200, n=50, c=20000, 24HR, mse2 and mse3)..................................................105

6.4a Plot of subject’s true StdDev vs. subject’s latent value by estimator/predictor with smallest mse (N=444, n=200, c=20000, 24HR).................................106

6.4b Plot of subject’s true StdDev vs. estimator P1 by estimator/predictor with smallest mse (N=444, n=200, c=20000, 24HR)...........................................107

6.4c Plot of subject’s true StdDev vs. predictor P2 by estimator/predictor with smallest mse (N=444, n=200, c=20000, 24HR)...........................................107

6.4d Plot of subject’s true StdDev vs. predictor P3 by estimator/predictor with smallest mse (N=444, n=200, c=20000, 24HR)...........................................108

6.5 Plot of subject’s true StdDev vs. estimator P1 by the best or equivalent to the best estimator/predictor (N=444, n=200, c=20000, 24HR)...............................110

6.6 Plot of percentile of the subject’s true StdDev vs. percentile of the predictor P1 by estimator/predictor with smallest mse (N=444, n=200, c=20000, 24HR).......111

6.7 Plot of percentile of the subject’s true StdDev vs. percentile of the predictor P1 by the best or equivalent to the best estimator/predictor (N=444, n=200, c=20000, 24HR).........................................................111

6.8a Plot of subject’s estimated StdDev vs. latent value by estimator/empirical predictor with smallest mse (N=444, n=200, c=20000, 24HR, unknown variance)...........................................................................117

6.8b Plot of subject’s estimated StdDev vs. estimator P1 by estimator/empirical predictor with smallest mse (N=444, n=200, c=20000, 24HR, unknown variance)...........................................................................117

6.8c Plot of subject’s estimated StdDev vs. empirical predictor P2 by estimator/empirical predictor with smallest mse (N=444, n=200, c=20000, 24HR, unknown variance)...........................................................................118

6.8d Plot of subject’s estimated StdDev vs. empirical predictor P3 by estimator/empirical predictor with smallest mse (N=444, n=200, c=20000, 24HR, unknown variance)...........................................................................118

6.9 Plot of subject’s estimated StdDev vs. estimator P1 by the best or equivalent to the best estimator/empirical predictor (N=444, n=200, c=20000, 24HR, unknown variance)...........................................................................119
6.10 Plot of percentile of the subject’s estimated StdDev vs. percentile of the estimated P1 by estimator /empirical predictor with smallest mse (N=444, n=200, c=20000, 24HR, unknown variance)…………………………120

6.11 Plot of percentile of the subject’s estimated StdDev vs. percentile of the estimated P1 by the best or equivalent to the best estimator /empirical predictor (N=444, n=200, c=20000, 24HR, unknown variance)…………………121

7.1 Plot of subject’s true StdDev vs. subject’s latent value by the better predictor (N=444,n=200,C=20000,24HR+7DDR,mixed model)…………………………128

7.2 Plot of subject’s week bias StdDev vs. subject's latent value by the better predictor (N=444,n=200,C=20000,24HR+7DDR,mixed model)………………128

7.3 Plot of average daily saturated fat intake (7DDR) vs. subject's latent value by the better predictor (N=444,n=200,C=20000,24HR+7DDR,mixed model)………129

7.4 Plot of average daily saturated fat intake (7DDR) vs. subject's latent value by the better or equivalent to the better predictor (N=444,n=200,C=20000,24HR+7DDR,mixed model)…………………………129

7.5 Plot of subject’s average response ratio vs. subject's latent value by the better predictor (N=444,n=200,C=20000,24HR+7DDR,mixed model)…………………………130

7.6 Plot of subject’s average response ratio vs. subject's latent value by the better or equivalent to the better predictor (N=444,n=200,C=20000,24HR+7DDR,mixed model)…………………………130

7.7 Plot of subject's average response ratio vs. estimator P1 by the better predictor (N=444,n=200,C=20000,24HR+7DDR,mixed model)……………………………131

7.8 Plot of subject's average response ratio vs. estimator P1 by the better or equivalent to the better predictor (N=444,n=200,C=20000,24HR+7DDR,mixed model)…………………………131

7.9 Plot of subject’s average response ratio vs. percentile of the estimator P1 by the better predictor (N=444,n=200,C=20000,24HR+7DDR,mixed model)………………132

7.10 Plot of subject’s average response ratio vs. percentile of the estimator P1 by the better or equivalent to the better predictor (N=444,n=200,C=20000,24HR+7DDR,mixed model)…………………………132

7.11 Plot of average daily saturated fat intake (7DDR) vs. subject's latent value by the better predictor (N=444,n=200,C=1000,24HR+7DDR, FPMM model)………133
7.12 Plot of average daily saturated fat intake (7DDR) vs. subject's latent value by
the better or equivalent to the better predictor
(N=444,n=200,C=1000,24HR+7DDR,FPMM model).................................I34

7.13 Plot of subject's average response ratio vs. subject's latent value by the better
predictor (N=444,n=200,C=1000,24HR+7DDR, FPMM model)......................I34

7.14 Plot of subject's average response ratio vs. subject's latent value by the better or
equivalent to the better predictor
(N=444,n=200,C=1000,24HR+7DDR, FPMM model)........................................I35

7.15 Plot of subject's average response ratio vs. estimator P1 by the better predictor
(N=444,n=200,C=1000,24HR+7DDR, FPMM model)........................................I35

7.16 Plot of subject's average response ratio vs. estimator P1 by the better or
equivalent to the better predictor
(N=444,n=200,C=1000,24HR+7DDR, FPMM model)........................................I36

7.17 Plot of subject's average response ratio vs. percentile of the estimator P1 by the
better predictor (N=444,n=200,C=1000,24HR+7DDR, FPMM model)...............I36

7.18 Plot of subject's average response ratio vs. percentile of the estimator P1 by the
better or equivalent to the better predictor
(N=444,n=200,C=1000,24HR+7DDR, FPMM model)........................................I37

7.19 Plot of subject's estimated StdDev vs. subject's latent value by the better
empirical predictor (N=444,n=200,C=1000,24HR+7DDR,mixed model).........142

7.20 Plot of subject's estimated week bias StdDev vs. subject's latent value by the
better empirical predictor (N=444,n=200,C=1000,24HR+7DDR,mixed
model).......................................................................................................143

7.21 Plot of average daily saturated fat intake (7DDR) vs. subject's latent value by
the better empirical predictor (N=444,n=200,C=1000,24HR+7DDR,mixed
model).......................................................................................................144

7.22 Plot of average daily saturated fat intake (7DDR) vs. subject's latent value by
the better or equivalent to the better empirical predictor
(N=444,n=200,C=1000,24HR+7DDR,mixed model)........................................144

7.23 Plot of subject's average response ratio vs. subject's latent value by the better
empirical predictor (N=444,n=200,C=1000,24HR+7DDR,mixed
model).......................................................................................................145
7.24 Plot of subject's average response ratio vs. subject's latent value by the better or equivalent to the better empirical predictor (N=444, n=200, C=1000, 24HR+7DDR, mixed model)..................................................145

7.25 Plot of subject's average response ratio vs. estimator P1 by the better empirical predictor (N=444, n=200, C=1000, 24HR+7DDR, mixed model).................................146

7.26 Plot of subject's average response ratio vs. estimator P1 by the better or equivalent to the better empirical predictor (N=444, n=200, C=1000, 24HR+7DDR, mixed model)........................................146

7.27 Plot of subject's average response ratio vs. percentile of the estimator P1 by the better empirical predictor (N=444, n=200, C=1000, 24HR+7DDR, mixed model).....................................................................147

7.28 Plot of subject's average response ratio vs. percentile of the estimator P1 by the better or equivalent to the better empirical predictor (N=444, n=200, C=1000, 24HR+7DDR, mixed model)................................................147

7.29 Plot of average daily saturated fat intake (7DDR) vs. subject's latent value by the better empirical predictor (N=444, n=200, C=1000, 24HR+7DDR, FPMM model)........................................................................148

7.30 Plot of average daily saturated fat intake (7DDR) vs. subject's latent value by the better or equivalent to the better empirical predictor (N=444, n=200, C=1000, 24HR+7DDR, FPMM model)................................................148

7.31 Plot of subject's average response ratio vs. subject's latent value by the better empirical predictor (N=444, n=200, C=1000, 24HR+7DDR, FPMM model)..............149

7.32 Plot of subject's average response ratio vs. subject's latent value by the better or equivalent to the better empirical predictor (N=444, n=200, C=1000, 24HR+7DDR, FPMM model)................................................149

7.33 Plot of subject's average response ratio vs. estimator P1 by the better empirical predictor (N=444, n=200, C=1000, 24HR+7DDR, FPMM model)...............................150

7.34 Plot of subject's average response ratio vs. estimator P1 by the better or equivalent to the better empirical predictor (N=444, n=200, C=1000, 24HR+7DDR, FPMM model)................................................150

7.35 Plot of subject's average response ratio vs. percentile of the estimator P1 by the better empirical predictor (N=444, n=200, C=1000, 24HR+7DDR, FPMM model).....................................................................151
7.36 Plot of subject's average response ratio vs. percentile of the estimator P1 by the better or equivalent to the better empirical predictor
(N=444,n=200,C=1000,24HR+7DDR, FPMM model).................................151
CHAPTER 1

INTRODUCTION

1.1 Motivation

The initial dietary intake data were collected nationwide by the United States government in 1930s. In the 1980s, the application of a measurement error model for dietary intake data proposed by National Research Council (NRC) was developed and provided a statistical model for analyzing the dietary intake data (National Research Council 1986). Since then, the statistical modeling involved in dietary intake data evolved gradually because dietary intake data are difficult to analyze. Currently, compared with the rapidly developing statistical methods for analyzing individual genetic profiles at the micro level, there is a significant lag when it comes to investigate the interaction of diet and genetics. The accurate estimation of an individual’s usual dietary intake is still an open research problem. However, it is essential to uncover the relationship between diet and disease. Knowing the individual dietary intake may help to establish a diagnosis, determine a patient’s treatment, or prevent diseases and health problems related to the individual diet. Although many studies pointed to the significance of future endeavors in the estimation of an individual’s usual dietary intake, surprisingly, there are very limited literatures on the estimation of an individual’s usual dietary intake, especially on how to take advantage of the multiple source data to estimate an individual’s usual dietary intake.

The characteristics that individual dietary intake varies from day to day and cannot be obtained without measurement errors make the dietary data analysis
challenging. Further, nutrients can be stored, and some foods are episodically consumed and some foods are commonly consumed. Usually, researchers’ interest is the long-term average daily intake rather than the daily intake. In nutritional studies, usual dietary intake indicates the long-term average daily dietary or nutrient intake over a certain period, usually over one year. Dietary data are collected by dietary assessment instruments. Currently, all feasible dietary assessment instruments used in epidemiological and nutritional studies result in data that involve systematic and/or random measurement errors. Repeated measures and different dietary assessment instruments are applied to one subject in order to reduce variability and measurement errors. Even though the developing innovative electronic technologies can improve dietary assessment, the potential for inherent individual bias still exists because of the self-reported dietary intake. Many factors contribute to the bias, such as the social desirability and the social approval (Hebert et al. 1997). Moreover, the number of repeated measures for one subject’s dietary intake is limited due to the cost, subject acceptance and reliability. The dietary intake data are often skewed with relatively larger day-to-day variation compared with the variation between individuals. It is common that the within-individual variance is correlated with the individual usual dietary intake. To accurately estimate an individual’s usual dietary intake based on the repeated measures with measurement errors, statistical models are important to account for the random measurement error and adjust for the bias. Commonly used statistical models for analyzing dietary intake data in the literature are the measurement error model, the mixed effect model and the structural equation model. Most studies aim to estimate the
population usual dietary intake, the accuracy of dietary assessment instrument or the disease relative risk.

Different from the measurement error in covariate in the errors-in-variables model which many studies have investigated, this study examines the measurement error in response (continuous variable) in a sample survey. Discussion about this topic can be found in recent literature about measurement error (Carroll et al. 2006; Buonaccorsi 2010). We call the individual true value of daily dietary intake “the subject’s latent value of daily dietary intake” since the true value can never be observed directly. This study focuses on the estimation of the latent value based on different models at an individual level rather than at the population level when a subject is repeatedly measured with measurement error.

A measurement error model is the simplest model to estimate a subject’s latent value. The best linear unbiased estimator (BLUE) of a subject’s latent value from a measurement error model is the average of the repeated measures. Previous studies proved that the best linear unbiased predictor (BLUP) from a mixed model is a more accurate estimator and hence is often used to estimate a subject’s latent value (Robinson 1991). However, when examining the relationship of sampling, mixed model and model assumptions, Stanek and Singer (Stanek and Singer 2011) argued that a portion of the mixed model sample space is artificial and not potentially realizable, i.e. the artificial sample space is one that assigns positive probabilities to the potentially unrealizable responses. Such artificial points are due to the assumption of exchangeability of subject’s latent values. When the subject effect is treated as a random effect, subject’s label is
disregarded; however, subject’s label is used for measurement error of the realized subject.

To explore the role of subject label in the estimation of a subject’s latent value, an alternative model called finite population mixed model (FPMM) was developed by Stanek and Singer (2004). Different from the usual mixed model, the FPMM model assigns positive probabilities only to potentially realizable responses. Thus, artificial responses do not exist in the FPMM model. The information of labeling is used in the FPMM model through incorporating sampling into the mixed model framework. A subject’s label is linked to both the subject effect and the measurement error.

The two predictors of the subject’s latent value based on the usual mixed model and the FPMM model are different. When all variance components are known and within-subject variances are equal, the predictor of the subject’s latent value based on the FPMM model have smaller MSE than the predictor of the subject’s latent value based on the usual mixed model (Stanek and Singer 2004). When all variance components are unknown and within-subject variances are equal, a simulation study conducted by San Martino et al. (2008) revealed that the empirical predictor based on the FPMM model has better performance than the empirical predictor based on the usual mixed model.

However, when all variance components are known and the within-subject variance is heterogeneous, Stanek and Singer (2011) compared the MSEs of the BLUPs based on the usual mixed model and the FPMM model by using a simple example, and found the BLUP from the usual mixed model is often (but not always) more accurate. This result holds, even though an artificial sample space is used in the usual mixed model. But their results are based on a very small example. Thus, one of the objectives of this study is to
compare the MSEs of the BLUPs based on the usual mixed model and the FPMM model when there is heterogeneous within-subject variance by using real data and through a simulation study.

The individual dietary intake data collected in Seasons study (a large observational longitudinal study, NHLBI: R01-HL52745) are ideal for us to demonstrate the method and results. Moreover, this evaluation contributes to the literature in the area of nutritional epidemiological studies. The fact that the FPMM does not require assumptions about parameter distributions of the observed data makes the FPMM appealing to fit to the dietary intake data. Because the distribution of dietary intake data is often skewed, other modeling approaches make transformations to eliminate this skewness. The resulting back-transformation is necessary to report results on the original scale, involving complex calculation (Dodd et al. 2006). The FPMM avoids these complexities since it is distribution free. This advantage further motivates us to investigate the performance of predictor of subject’s latent value in the context of the FPMM framework.

Another objective of this study aims to improve the performance of predictor of subject’s latent value by taking the advantage of multiple source data. Since the dietary intake data were collected through both 24 hour dietary recall (24HR) and seven day dietary recall (7DDR) instruments in the Seasons study, we make use of the two different source data to compare the performance of BLUPs based on the 24HR data only and the combined data from 24HR recall and 7DDR recall. The combined data from two different instruments might be superior to either one. The exploration of statistical method for the combined data to estimate individual usual dietary intake is promising and challenging.
Illner and his colleagues (2011) reported a preliminary study on the feasibility of using the combined data from different dietary assessment instruments and stated that they were working towards an approach of combining different dietary assessment instruments on an individual level (Illner et al. 2010). This research contributes to the undergoing study of the utilization of multiple source data in estimation of an individual’s usual dietary intake.

1.2 Objectives

The main aim of this study is to evaluate the performance of estimator/predictors of subject’s latent value under the measurement error model, the usual mixed model and the FPMM model in the setting of the repeatedly measured response with heterogeneous within-subject variance or bias in a finite population. We perform the analysis in two cases: known variance components and unknown variance components. In each case, we target the evaluation at an individual level in two situations: estimator/predictor of latent value for a randomly selected subject and estimator/predictor of latent value for a randomly selected subject, given a realized subject. These investigations are conducted in a setting where repeated measures are available from two sources, where the data from one source are biased.

There is no normality assumption for the observed response in this study.

Specifically, our goal is to:

1) evaluate and obtain an optimal estimator/predictor of subject’s latent value in different settings in order to guide the use of an appropriate estimator/predictor in practice;
2) examine whether combination of two instruments provides more accurate prediction than repeated measures from one instrument;
3) better understand the mixed model assumptions, the role of labeling in statistical inference.

1.3 Context — Seasons Study

The primary aim of the Seasons study is to identify the causes of seasonal variation in serum cholesterol. One of the specific aims is to quantify the effects of main factors determining the variation. Seasonal differences in diet may partially explain the seasonal differences observed in serum lipids; however, it has not been adequately tested (Ockene 1993). The Seasons study followed a cohort of individuals and obtained serial individual data. The Seasons study collected multiple measures of serum lipid levels, diet, physical activity and other important parameters on study subjects.

The cohort of individuals recruited in the Seasons study are a volunteer sample of enrollees between the ages of 20 and 70 in the Fallon Health Plan, a Massachusetts’s health maintenance organization. Subjects were enrolled between 1994 and 1998. Measures of each subject’s cholesterol level, dietary intake, physical activity, light exposure, etc. were collected five times at consecutive three-month intervals. Blood was collected for serum cholesterol level at each clinic visit for five quarters over one year. In each quarter, the subject’s reported dietary intake data were collected by computer-assisted evening telephone interview by a trained nutritionist using a 24 hour dietary recall instrument and self-reported via a written questionnaire for a seven-day dietary recall period.
1.3.1 24-hour Dietary Recall (24HR)

At the beginning of the interview, the interviewer describes the nature of the 24-hour recall, and then has the subject recall the foods, beverages and dietary supplements consumed during the prior day for a quick list. The interviewer and the subject then review each food in detail following the Nutrition Data System for Research (NDSR) software (version 2.6) instruction. 24HR data were entered and analyzed using the NDSR software. The system converts the reported food items into nutritional intake, does logic checks for corrections, and produces an analytic data file. The NDSR software is one of the leading resource for nutrient database and analysis systems for research in the U.S. (Ockene 1993).

Ideally, three 24HR recalls were collected in each quarter over five quarters; fifteen 24HR recalls were obtained in five consecutive quarters over one year follow-up period for each subject. In each quarter, the three recall days were randomly spread out across seven days with two weekdays and one weekend day. Thus, the recall day varies from quarter to quarter, and subject to subject. The study design was based on previous studies (Liu et al. 1978; El Lozy 1983; St. Jeor et al. 1983; Nelson et al. 1989) that concluded that three days are sufficient and a reasonable number to catch commonly consumed dietary intake in one season. The weekday and weekend day were included in each quarter since dietary intake was thought to be different between these two periods (Tarasuk and Beaton 1992).
1.3.2 Seven-day Dietary Recall (7DDR)

The 7DDR questionnaire was designed to measure short-term dietary intake, especially for fat (Hebert et al. 1997). The questionnaire is a self-administered instrument; it is mailed to subjects prior to each clinic visit. The questionnaire asks subjects to recall their meals and snacks over the last seven days ending yesterday. Subjects recall how many times of each food items consumed and portion sizes. The questionnaire mainly includes 119 kinds of foods and 13 types of beverages. Similar to the food frequency questionnaires, the subjects are asked the typical or average intake when it is impossible for them to recall their dietary intake for a certain day.

The first day of the seven recall days could be any day of the week. Similar to the 24HR, the start of the seven days varies from quarter to quarter, and subject to subject. In fact, in the Seasons study, the recall days of both 24HR and 7DDR were limited within a 42-day window time around each clinic visit day, i.e. before 28 days or after 14 days of the date of blood draw.

The 7DDR nutrient scores are also computed using the NDSR software. The final 7DDR data correspond to one value of average daily intake for each nutrient over the seven days. For example, the saturated fat intake is the average value of daily saturated fat intake reported over consecutive seven days for a subject.

This study uses the individual saturated fat intake (continuous variable) from 24HR recall and 7DDR recall as the subject’s response to illustrate the method and results. Saturated fat intake is an important predictor of cholesterol levels, a risk factor for heart disease. For each subject, we define the subject’s latent value of saturated fat intake as the average daily saturated fat intake over a one-year time period (called subject’s
usual saturated fat intake). In the Seasons study, data arise following two stage sampling. At the first stage, subjects are sampled. For each selected subject, a sample of days or weeks is selected. Finally, the saturated fat intake is measured on a selected day or week indirectly using the 24HR instrument or the 7DDR instrument for the selected subject. As in previous studies (Dodd et al. 2006), we assume that measure from the 24HR recall is unbiased, but with measurement error. The measure from the 7DDR recall is biased and with measurement error. The 24HR recall serves as a reference instrument when we adjust the measure from the 7DDR recall for bias.

1.3.3 Preliminary Descriptive Statistics

The basic dataset of the Seasons study includes a total of 641 subjects who completed the baseline questionnaire and at least one blood draw. This study uses two subsets of the basic dataset. One subset includes 127 subjects and another subset includes 444 subjects. The 127 subjects used for analysis in Chapter 4 include those subjects who completed 12 24HR recalls and 4 7DDR recalls in the consecutive four quarters during one year period. The 444 subjects used for simulation study in Chapters 5, 6 and 7 include those subjects who completed 12 24HR recalls or more and 4 7DDR recalls or more in the five quarters.

Each of the 127 subjects completed three 24HR recalls and one 7DDR recall in each quarter. The 127 subjects totally completed 1,524 24HR recalls and 508 7DDR recalls in a year. Among the 127 subjects, 62(48.8%) are male. The average age is 50.5 (SD = 12.7) years old with the range 20-70. The range of 24HR measures of saturated fat intake is 0.77-150.41g/day, the range of 7DDR measures is 3.95-104.36g/day. We think
the large amount of saturated fat intake for a certain day is possible and include all of the data in the analysis. We obtain average 24HR recall and average 7DDR recall for each subject by averaging 12 24HR recalls and 4 7DDR recalls, respectively. Figure 1.1 shows the distributions of average 24HR recalls and average 7DDR recalls of 127 subjects. We see both distributions are skewed to the right.

Figure 1.1 Distributions of saturated fat intake for average 24HR recalls and 7DDR recalls, \( n = 127 \)

![Histograms of average 24HR and 7DDR recalls.](image)

From the mixed model results in Table 1.1, we see that the estimated overall mean of 7DDR measures (27.91g/day) is greater than the estimated overall mean of 24HR measures (24.96g/day). Compared with the variances of 7DDR measures, the between-subject variance of 24HR measures is smaller, and the residual variance of 24HR measures is larger.

Table 1.1 Estimated overall means and variance components of daily saturated fat intake based on observed 24HR and 7DDR measures (mixed model results)

<table>
<thead>
<tr>
<th>Daily saturated fat intake (g/day)</th>
<th>Overall mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>24HR</td>
<td>24.96</td>
<td>Subject 118.85</td>
</tr>
<tr>
<td>7DDDR</td>
<td>27.91</td>
<td>Residual 142.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Subject 133.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Residual 87.03</td>
</tr>
</tbody>
</table>
Figure 1.2 shows the scatter plot of average 7DDR verse average 24HR saturated fat intake of 127 subjects. The correlation coefficient of average 7DDR and average 24HR saturated fat intake is 0.77 (p < 0.05). Ideally, the dots should be distributed on or randomly around the diagonal line, which means the observed subject’s average daily saturated fat intake should be equal using 24HR instrument and 7DDR instrument. In fact, using both instruments, due to the variability and bias, some dots are distributed away from the diagonal line. The dashed line is the regression line, we see relatively higher reported value from 7DDR at the lower level of 24HR saturated fat intake and relatively lower reported value from 7DDR at the higher level of 24HR saturated fat intake.

Figure 1.2 Scatter plot of average 7DDR vs. average 24HR saturated fat intake (n = 127)

Figure 1.3 is the scatter plots of standard deviation verse mean of 12 measures from 24HR and 4 measures from 7DDR of 127 subjects, respectively. Both plots show a horn shaped pattern, which reveals that the constant error variance across subjects is violated. Thus, we think that subjects have heteroscedastic error variances.
Figure 1.3 Scatter plots of standard deviation verse mean of 12 measures from 24HR and 4 measures from 7DDR, $n = 127$

The subset of 444 subjects is used when we conduct the simulation study. The 444 subjects are those who completed not less than 12 24HR recalls and not less than 4 7DDR recalls. Table 1.2 shows the frequencies of 24HR recalls and 7DDR recalls of the 444 subjects by recall occasions.

<table>
<thead>
<tr>
<th></th>
<th>24HR</th>
<th></th>
<th>7DDR</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall #</td>
<td>Subject #</td>
<td>Cum. subject #</td>
<td>Recall #</td>
<td>Subject #</td>
</tr>
<tr>
<td>12</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>4</td>
<td>57</td>
</tr>
<tr>
<td>13</td>
<td>40</td>
<td>80</td>
<td></td>
<td>5</td>
<td>387</td>
</tr>
<tr>
<td>14</td>
<td>77</td>
<td>157</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>287</td>
<td>444</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Among the 444 subjects, 221 (49.8%) are male. The average age is 49.6 (SD = 11.8) years old with the range 20-70. The range of 24HR measures is 0-150.41g/day, the range of 7DDR measures is 2.15-287.94g/day.
CHAPTER 2
LITERATURE REVIEW

This chapter focuses on the review of dietary assessment instruments, individual dietary intake studies, and studies on multiple source data.

2.1 Dietary Assessment Instruments

Statistical methods for estimating the individual latent value of dietary intake are closely related to the development of the dietary intake survey methods. Accurate measurement of individual dietary intake is needed to estimate the individual usual dietary intake and its relationship with disease. Currently, conventional dietary assessment instruments used in surveillance and epidemiological research are inevitably involved in measurement error and/or bias.

With the development of electronic techniques, innovative technologies to improve dietary assessment have been explored in order to overcome the limitations of conventional dietary assessment method. Illner et al. (2012) reviewed published studies year between 1995 and 2011 that evaluated new technologies for measuring diet in nutritional epidemiology. The new technologies include personal digital assistants, mobile phone, interactive computer, web, camera and tape recorder, scan and sensor-based methods. Compared with the traditional food records, such as 24 hour dietary recall and food frequency questionnaires, etc., they concluded that the main improvements of new methods are the efficient data collection and higher subject acceptance. However, the bias related to self-reported dietary intake still exists. The validity of new dietary
assessment technologies and their application in large epidemiological studies need more investigation.

Since an individual usually does not consume similar food every day. Some foods are commonly consumed, while some foods are episodically consumed. The well-known conventional dietary assessment methods include long-term dietary assessment instruments (such as food frequency questionnaires) and short-term dietary assessment instruments (such as food record, 24 hour dietary recall, seven day dietary recall). The 24 hour dietary recall (24HR) and the food frequency questionnaires (FFQ) are two of the major dietary data collection instruments.

The 24HR is the primary individual dietary survey instrument used in surveillance and the nutritional epidemiological survey (Ferro-Luzzi 2002; Willett 1998). Although the 24HR reduces the systematic error by following up 24 hours’ dietary intake, a single day’s reported dietary intake inadequately reflects the long-term average daily intake due to the variability of dietary intake from day to day. To overcome this limitation, the early solution is to obtain multiple 24HR recalls on per subject and average the observed responses. Based on the results for energy in Observing Protein and Energy Nutrition study, the 24HR has a general tendency toward underestimation (Kipnis et al. 2003). But for some foods or food groups, the 24HR may overestimate true usual intake (Dodd et al. 2006). Some studies (Dodd et al. 2006) assume that original 24HR measure or transformed 24HR measure is unbiased, but has measurement error. The unbiased 24HR can serve as a reference instrument to adjust for bias of observed measure from other dietary assessment instrument, such as FFQ. Studies of diet-disease relationship often use regression calibration method to correct relative risk estimate by collecting additional
measure obtained from reference instrument in a substudy (Kaaks and Riboli 1997; Coulston et al. 2013). Dietary biomarker, food record or 24HR could serve as reference instrument with their own limitations (Kaaks 1997; Hedrick et al. 2012; Kipnis et al. 1999). The calibration that corrects relative risk using a calibration study is different from the calibration that intends to obtain more accurate estimate of individual dietary intake. If studies consider the 24HR measure is biased or random errors between two instruments and between repeated measures are correlated, they use biochemical markers in their analyses (Rosner et al. 2008). However, a few dietary biomarkers can be used as reference instruments.

The 24HR, Automated Multiple Pass Method (AMPM) of the United States Department of Agriculture (USDA), is used in the National Health and Nutrition Examination Survey (NHANES) in the United States (Steinfeldt et al. 2013). An EPIC- SOFT computer program for 24HR is used in Europe (Slimani et al. 1999). More recently, a web-based automated self-administered 24-hour dietary recall (ASA24) is developed and evaluated by the United States National Cancer Institute (NCI) (Subar et al. 2012).

Usually, multiple 24HR measures are limited due to higher costs. Long-term dietary assessment instruments, such as FFQ, in large epidemiological or nutritional studies are commonly used due to its reasonable costs and easy administration, especially for the estimation of episodically consumed foods (Midthune et al. 2011). The food frequency questionnaires (FFQ) is designed to measure long-term dietary behavior. Subjects need to recall the frequency and portion size of food items over a certain time period, usually over one year. The retrospective recalls over a long time period cause inaccurate report of individual dietary intake. As a result, such reported intake may
distort the relationship between diet and disease. Carroll et al. (2012) suggested to take advantage of the strengths of 24HR and FFQ by combining 24HR and FFQ data in the regression calibration analysis. Subar et al. (2006) concluded that FFQ data may offer important covariate information when using 24HR to estimate usual intake of episodically consumed foods. Tooze and his colleagues (2006) proposed the National Cancer Institute (NCI) method, which uses a two-part model to estimate the usual intake of episodically consumed foods and adds the FFQ information as a covariate when using 24HR data. Kipnis et al. (2009) extended the NCI method to predict individual usual intake of episodically consumed foods using 24HR data. They also included the FFQ information as a covariate in their regression calibration model.

In the Seasons study, similar to the FFQ, a self-administered 7-day dietary recall (7DDR) instrument is designed to measure short-term behavior in dietary intake, especially for fat (Hebert 1997). Hebert et al. (2002) found that the bias of 7DDR is associated with social approval in estimates of women’s energy intake.

### 2.2 Studies in Estimation of Individual Dietary Intake

In 1986, National Research Council (NRC) first suggested the application of a statistical model to estimate the distribution of usual dietary intake (National Research Council 1986). The NRC and the Institute of Medicine (IOM) proposed a measurement error model for the estimation of usual dietary intake using 24HR data,

\[ R_{ij}^* = R_{ij} + e_{ij}. \]  

(2.1)

The observed dietary intake data are often skewed. In order to use the properties of the normal distribution to estimate the usual dietary intake, they transformed the original
observed 24HR measure using a power or log transformation for normality. Thus, \( \tilde{R}_{ij} \) represents the transformed 24HR dietary intake for subject \( i \) at time \( j \), \( \tilde{R}_i \) is the usual intake for the subject on the transformed scale, and the \( e_{ij} \) represents within-subject variation including day-to-day variation and random measurement error. At the population level, the \( \tilde{R}_i \) is assumed to have mean \( \mu_i \) and variance \( \sigma_i^2 \); the \( e_{ij} \) has mean 0, variance \( \sigma_e^2 \) (assuming it is identical for different subjects), and is independent of the \( \tilde{R}_i \) (Institute of Medicine 2003; Tooze et al. 2010). Subsequently, the Iowa State University (ISU) method and a simplified ISU method called Best-Power method were proposed as an extension of the NRC method (Nusser et al. 1996; Guenther et al. 1997). The ISU method considers the heterogeneous within-subject variance, however, the application of the semiparametric transformation and back-transformation is complex (Nusser et al. 1996). Nusser et al. (1997) proposed the Iowa State University Food (ISUF) method applied to episodically consumed foods. ISUF method deals with the data with some zero observations for some of 24HR recalls. It separates zero 24HR recalls and nonzero 24HR recalls, and combines the information of the distribution of consumption probability (a ratio of the number of days on which the food is observed to be consumed to the number of days observed). The ISUF method is based on an assumption that the consumption probability and the usual consumption-day intake are not correlated. However, this assumption may not be held in reality. Dodd et al. (1996) showed a positive correlation between the consumption probability and the usual consumption-day intake. Although the NRC and ISU methods generally get better estimates of the standard
deviation for the 24HR by adjusting for within-subject variation, they do not improve the accuracy of estimates of the population mean (Freedman et al. 2004).

In order to allow for the correlation between the consumption probability and the usual consumption-day intake, Tooze and his colleagues (2006; 2010) proposed National Cancer Institute (NCI) method to overcome the limitation of the ISUF method which assumes that the consumption probability and the usual consumption-day intake are not correlated, by using a two-part model. The two-part model is a two-part mixed effects measurement error model with correlated random effects. They applied a Box-Cox transformation for the original observed 24HR data \((R_{ij})\) with the right skewness distribution, and defined \(R^*_y = g(R_y, \lambda)\) to be the transformed 24HR data for subject \(i\) at time \(j\) where \(\lambda\) is the power parameter of Box-Cox transformation. Then

\[
R^*_y = \mu^*_i + e_{ij}, \quad e_{ij} \sim N(0, \sigma^2_e)
\]

where \(\mu^*_i\) is the individual mean for subject \(i\), \(e_{ij}\) is within-subject random error, \(\mu^*_i\) and \(e_{ij}\) are independent. Further,

\[
\mu^*_i = X_i^T \beta + u_i, \quad u_i \sim N(0, \sigma^2_u)
\]

where \(X_i\) is a vector of covariates, the random effect \(u_i\) represents subject-specific deviation of usual intake from the mean of a particular population of subjects with the same covariates. Therefore, a mixed effects model for 24HR can be expressed as

\[
g(R_y, \lambda) = X_i^T \beta + u_i + e_{ij}.
\]

Although the NCI method aims to estimate the distribution of usual dietary intake for the population, it may be valuable for exploring the method of estimating an individual’s usual intake.
For the purpose of predicting individual usual dietary intake, Kipnis et al. (2009) proposed an extension of NCI method by using regression calibration model. This model follows the regression calibration approach for measurement error correction, individual usual dietary intake is predicted as the conditional mean intake given 24HR reported intake and other covariates in the model. For individual \( i \) on day \( j, i = 1,...,n, j = 1,...,J \), let \( T_{ij} \) denote the subject’s latent value of dietary intake on day \( j \), and \( R_{ij} \) denote the observed 24HR intake. True individual usual intake \( T_i \) is the expectation of \( T_{ij} \). Based on model (2.2) and the same assumptions as the NCI method, regression calibration requires evaluation of the best predictor \( E(T_i | R_i, X_i') \) given the reported intake \( R_i \) and covariates \( X_i' \) for subject \( i \). The usual intake of subject \( i \) is

\[
T_i = E(R_i | X_i', u_i) = E\{g^{-1}(X_i' \beta + u_i + e_{ij}, \lambda) | X_i', u_i\} \approx g(X_i' \beta + u_i, \lambda).
\]

The best predictor of individual usual intake is given by

\[
E(T_i | R_i, X_i') \approx E\{g*(X_i' \beta + u_i, \lambda) | R_i, X_i'\}.
\]

This study also quantified the increased accuracy of the predicted usual intake by including the FFQ as a covariate in the calibration model using data from the Eating at America’s Table Study (EATS).

In studies that aim to evaluate the validity of self-report dietary assessment instrument, researchers investigated the measurement error structure and developed measurement error models including bias and random measurement error. Typically, as described by Freedman (2011), using a measurement error model, for reported intake of individual \( i \) at measure \( j \), \( R_{ij} \) can be expressed as

\[
R_{ij} = \beta_0 + \beta_i T_{ij} + r_i + e_{ij}
\]
where $T_i$ is a parameter representing a subject’s true intake; $e_{ij}$ is a random measurement error of subject $i$ at measure $j$. The parameters $\beta_0$, $\beta_i$ and $r_i$ are related to bias. Under the assumption of linear bias, $\beta_0$ is the intercept and called additive bias, $\beta_i$ is the slope and called multiplicative bias; these two types of bias are systematic bias. Systematic bias is due to the instrument and is common to all subjects. The additional parameter $r_i$ is a bias occurring at an individual level, it is specific to a subject. It is called the subject-specific bias, and is equal to the deviation of the subject-specific bias from the average bias of the population. The bias $r_i$ varies for each subject and is defined to average to zero over all the subjects in the population. Model (2.3) is helpful to quantify the absolute amount of biases in the observed measure and remove them when we want to estimate a subject’s latent value. The parameters in (2.3) may depend on the type of dietary assessment instrument and the population or the data used (Carroll et al. 1998; Kaaks et al. 1994). Based on different study designs and assumptions, slightly different measurement error models have been used in different studies (Plummer and Clayton 1993; Kaaks et al. 1994; Spiegelman et al. 2005). Based on the measurement error model, a structural equation model has also been used to estimate the error parameters in some studies (Kaaks et al. 1994; Ocké and Kaaks 1997; Rosner et al. 2008).

2.3 Studies on Multiple Source Data

For the multiple source data, the response of interest is measured by two or more instruments. Some researchers perform separate analysis for each source. Some researchers combine the responses from multiple sources into a single outcome which usually excludes the individual with missing data in one source. A correlated data model,
such as regression model with correlated data, can handle the missing data. The primary interest of fitting a correlated data model is in comparing the means and measuring the association of different instruments in a unified framework. Separate regression models correspond to each instrument. An indicator variable is used to identify the different source. The generalized estimating equations (GEE) approach can be used to fit the marginal models (Fitzmaurice et al. 2004). Depending on the response and the study design, different regression model can be used. The literature related to multiple source data is common in application to psychological studies. Fitzmaurice et al. (1995) proposed bivariate logistic regression analysis for the study using multiple informants. Horton et al. (2008) presented a latent variable regression model for multiple informants. O'Brien and Fitzmaurice (2005) analyzed the longitudinal multiple source Gaussian data using a regression model. In a different way, using Bayesian approach, Best et al. (2005) developed Bayesian graphical model for combining multiple source data with bias. Based on graphs of a series of local submodels that relate to the different sources of data, the model incorporates the submodels into a global analysis. The analysis can be performed using the WinBUGS software.

Plummer and Clayton (1993) suggested that a combination of different instruments may bear more information than repeated application of the same instrument based on the measurement error analysis for different dietary assessment instruments. Haubrock et al. (2011), a European research group for food, reported a study using the multiple source method to estimate individual usual dietary intake. They included multiple source data as a covariate in the model. In many settings, data of multiple sources are used as nuisance variable.
There is very limited literature on how to analyze the multiple source dietary intake data based on the “fused” data, i.e. combining the responses from multiple sources into a single outcome to investigate if the combined data provide more accurate estimate than the data obtained from a single instrument. Fokianos et al. (1998) reported their study on how to quantify the bias of the data from one instrument, and then combine data from two instruments to obtain more accurate estimate by using a semiparametric model for meteorological data. To a certain extent, the idea is similar to our analysis plan for the combined dietary data.
CHAPTER 3
SPECIFICATIONS OF MODELS

Dietary intake survey includes both systematic and random measurement errors and as a consequence, statistical models are utilized to estimate subject’s latent value of usual dietary intake based on observed data. This chapter reviews and defines the measurement error model, the usual mixed model, and the finite population mixed model for estimating a subject’s latent value. We use the Seasons study as the context, and the observed saturated fat intake from 24HR and 7DDR recalls as an example to present different predictors of a subject’s latent value under the three models. We first define a population and population parameters, then the model expressions, assumptions, the BLUEs and the BLUPs of the subject’s latent value. These are illustrated for each model.

3.1 The Population and Population Parameters

We define a population in space and time as a list of \( s = 1, \ldots, N \) subjects, where each subject is potentially observed over \( d = 1, \ldots, D \) days. We define \( \mu_{sd} \) as the latent value for subject \( s \) on day \( d \) for a particular parameter, saturated fat intake. We also define the average daily intake of saturated fat for subject \( s \) as

\[
\mu_s = \frac{1}{D} \sum_{d=1}^{D} \mu_{sd},
\]

(3.1)

and the population mean is

\[
\mu = \frac{1}{N} \sum_{s=1}^{N} \mu_s.
\]

(3.2)
We define the subject effect $\beta_s = \mu_s - \mu$ as the deviation of subject $s$’s latent value from the population mean, and the day effect $\delta_{sd} = \mu_{sd} - \mu_s$ as the deviation of subject $s$’s latent value on day $d$ from the subject’s latent value. The definitions of the subject effect and the day effect result in two constraints:\[ \sum_{s=1}^{N} \beta_s = 0, \quad \text{and} \quad \sum_{d=1}^{D} \delta_{sd} = 0 \] for all $s = 1,\ldots,N$.

We define the between-subject variance as\[ \sigma_S^2 = \frac{1}{N-1} \sum_{s=1}^{N} (\mu_s - \mu)^2, \quad (3.3) \]
and the within-subject variance $\text{Var}_D(\delta_{sd}) = \sigma_{sd}^2$ where\[ \sigma_{sd}^2 = \frac{1}{D-1} \sum_{d=1}^{D} (\mu_{sd} - \mu_s)^2 \quad (3.4) \]
for all $s = 1,\ldots,N$ (Note: we use uppercases $S$ and $D$ indicating the variance over all subjects $s$ and over all days $d$, respectively). We define the average within-subject variance as\[ \sigma_D^2 = \frac{1}{N} \sum_{s=1}^{N} \sigma_{sd}^2. \quad (3.5) \]

Based on these notations, we can express $\mu_{sd}$ as\[ \mu_{sd} = \mu_s + \delta_{sd} = \mu + \beta_s + \delta_{sd}. \quad (3.6) \]

Typically, $\mu_{sd}$ cannot be directly observed since the true value of saturated fat intake on a day is not directly observable. When dietary assessment instruments are used, the observed dietary intake data include measurement error. An additive measurement error model can relate the observed value to the subject’s latent value.
3.2 Measurement Error Model (ME)

We define the observed saturated fat intake for subject $s$ on day $d$ from the 24HR instrument as $Y_{sdk}^*$, where $k^*$ indicates a “measurement occasion” on day $d$.

Based on equation (3.6), a parametric measurement error model for $Y_{sdk}^*$ can be written as

$$Y_{sdk}^* = \mu_s + \delta_{sd} + E_{sdk},$$

(3.7)

where $E_{sdk}$ is the random measurement error using the 24HR instrument for subject $s$ on day $d$. We define the random variable such that $E_M(E_{sdk}) = 0$, and

$$\text{Var}_M(E_{sdk}) = \sigma_{sde}^2,$$

(3.8)

for any $s \neq s^*$, $d \neq d^*$, $k^* \neq k'^*$, $\text{Cov}_M(E_{sdk}, E_{s'd'k'^*}) = 0$. The subscript $M$ denotes the expectation with respect to the distribution of the random measurement error. We define the subject’s average measurement error variance as

$$\sigma_{sE}^2 = \frac{1}{D} \sum_{d=1}^{D} \sigma_{sde}^2,$$

(3.9)

We also define the overall average measurement error variance as

$$\sigma_E^2 = \frac{1}{N} \sum_{s=1}^{N} \sigma_{sE}^2,$$

(3.10)

When the days are randomly selected from the $D$ days by using simple random sampling without replacement for each subject $s$, we index the selections of the days by $j = 1, \ldots, m_s$. For simplicity, we assume that for each subject, the $m_s = m$ is the same, such that $j = 1, \ldots, m$. 

26
In (3.7), when the observed response from the 24HR recall is collected on the \( j^{th} \) randomly selected day, the day effect \( \delta_{sd} \) becomes random. We represent it as \( D_{sj} \) and have \( E_D(D_{sj}) = 0 \) and

\[
\text{Var}_D(D_{sj}) = \frac{D-1}{D} \sigma^2_{sd},
\]

(3.11)

where the subscript \( D \) denotes the expectation with respect to the distribution of the day effect \( D_{sj} \) (Note: for a usual mixed model, \( \frac{D-1}{D} \equiv 1 \)). For any \( j \neq j^* \),

\[
\text{Cov}_D(D_{sj}, D_{sj'}) = 0.
\]

(3.12)

In practice, when \( k^* = 1 \), i.e. one 24HR recall is collected on the \( j^{th} \) selected recall day, the random measurement error \( E_{sdk^*} \) changes to \( E_{sj} \). We assume \( \text{Cov}_M(E_{sj}, E_{sj'}) = 0 \) for any \( j \neq j^* \). It is not possible to separate day-to-day variability and the random measurement error in data analysis without additional information. As a result, we define response error as

\[
E^*_sj = D_{sj} + E_{sj}
\]

(3.13)

corresponding to the sum of day-to-day variability and the random measurement error, where \( E_R(E^*_sj) = 0 \),

\[
\text{Var}_R(E^*_sj) = \sigma^2_{sr}.
\]

(3.14)

The subscript \( R \) denotes the expectation with respect to the distribution of the response error.

Subsequently, we derive \( E_R(E^*_sj) = 0 \) and \( \text{Var}_R(E^*_sj) = \sigma^2_{sr} \). We have
\[ E_R(\varepsilon^*_j) = E_{DM}(D_j + E_j) = E_D(D_j) + E_M(E_j). \] Since \( E_D(D_j) = 0 \) and \( E_M(E_j) = 0 \), we have \( E_R(\varepsilon^*_j) = 0 \). We use the conditional expectation to evaluate the variance where

\[
\text{Var}_R(\varepsilon^*_j) = \text{Var}_{DM}(D_j + E_j) \\
= \text{Var}_D\left[E_M(D_j + E_j)\right] + E_D\left[\text{Var}_M(D_j + E_j)\right] \\
= \text{Var}_D\left[E_M(D_j) + E_M(E_j)\right] + E_D\left[\text{Var}_M(D_j) + \text{Var}_M(E_j)\right].
\]

Since \( E_M(E_j) = 0 \) and \( \text{Var}_M(D_j) = 0 \), we have

\[
\text{Var}_R(\varepsilon^*_j) = \text{Var}_D\left[E_M(D_j)\right] + E_D\left(\sigma^2_{\text{wde}}\right) \\
= \frac{D-1}{D} \sigma^2_{\text{sD}} + \sigma^2_{\text{sE}}.
\]

We define

\[
\sigma^2_{\text{sr}} = \frac{D-1}{D} \sigma^2_{\text{sD}} + \sigma^2_{\text{sE}}. \tag{3.15}
\]

Therefore, \( \text{Var}_R(\varepsilon^*_j) = \sigma^2_{\text{sr}} \).

Thus, the random variable representing the observed saturated fat intake for subject \( s \) on the \( j^{th} \) selected recall day using the 24HR instrument, \( Y_j \) can be written as

\[
Y_j = \mu_s + \varepsilon^*_j, \tag{3.16}
\]

where \( E_R(Y_j) = \mu_s \) and \( \text{Var}_R(Y_j) = \sigma^2_{\text{sr}} \).

We represent the best linear unbiased estimator (BLUE) of the subject’s latent value \( \mu_s \) as \( \hat{P}_1 \), where \( \hat{P}_1 \) is the ordinary least squares estimate given by the average of the observed responses for subject \( s \)

\[
\hat{P}_1 = \overline{Y}_s \tag{3.17}
\]
where \( \bar{Y}_s = \frac{1}{m} \sum_{j=1}^{m} Y_{ij} \).

### 3.3 Usual Mixed Model (MM)

In a mixed model, treating the subject effect as a random effect, the best linear unbiased predictor (BLUP) approach can be used to predict the random effect and obtain an estimate of subject’s latent value. The BLUP approach was proposed by Goldberger in 1962 (Goldberger 1962), but developed earlier to predict the unobserved value of random variable (Henderson et al. 1959). There is an extensive literature on these methods. Papers with an emphasis on finite population sampling include those written by Scott and Smith (1969), Royall (1976), and Stanek and Singer (2004). A review is given by Robinson (1991), and extensive discussion is given in the books by Searle et al. (1992) and McCulloch and Searle (2001).

In the simplest case, the BLUP consists of the estimated sample mean plus a deviation of the subject’s mean from an estimate of the overall mean multiplied by a shrinkage factor. In the mixed model, the term “predictor” is used for random effects and “estimator” is used for fixed effects. Correspondingly, we call the predictor of a random effect “the best linear unbiased predictor (BLUP)”. It is derived as a linear function of the data that is unconditionally unbiased and minimizes the expected mean squared error (MSE). The “best linear unbiased estimator (BLUE)” is obtained in a fixed effect model. It is a linear function of the data, unconditionally unbiased and minimizes the expected MSE. The expected MSE of the BLUP can be used to evaluate the accuracy of the predictor. When comparing predictors, the better predictor has smaller expected MSE. This can be evaluated conditionally on the random effect or unconditionally.
In the measurement error model (3.16), $\mu_s = \mu + \beta_s$ where the subject effect $\beta_s$ is fixed. For repeated measures data from one subject where our interest is on the subject’s latent value, the subject effect parameter typically varies from one subject to another. Therefore, accounting for the natural heterogeneity of subjects in the population, we expect subject effects to vary. If we randomly select a subject, the subject effect is random. We treat the subject effect $\beta_s$ as a random effect $B_i$ in the model (3.18), and the model becomes a mixed model. The mixed model accounts for the correlation among measures on the same subject and possibly heterogeneous measurement error variances across subjects. A usual mixed model contains both fixed effect and random effect. Usually, the estimator of fixed effect relates to population parameter; the predictor of random effect relates to the latent value for realized subject.

We present the usual mixed model following the definition used by Searle et al. (1992). When a sample of $n$ subjects is randomly selected, we index the selected subject by $i$, where $i = 1, ..., n$. We denote response (i.e. the observed 24HR saturated fat intake) on the $j^{th}$ recall day for the $i^{th}$ selected subject as $Y_{ij}$,

$$Y_{ij} = \mu + B_i + E_{ij}^*$$

(3.18)

where $\mu$ is a fixed effect corresponding to the population parameter. The subject effect $B_i$ is a random effect. Usual assumptions are that $E_s (B_i) = 0$ and

$$Var_s (B_i) = \frac{N - 1}{N} \sigma^2_s,$$

(3.19)
for any \( i \neq i^* \), \( \text{Cov}_S(B_i, B_{i^*}) = 0 \). The subscript \( S \) indicates the expectation with respect to the distribution of the subject effects. \( E_{ij}^* \) is the random response error, assuming that 
\[
E_R(E_{ij}^*) = 0, \\
\text{Var}_R(E_{ij}^*) = \sigma_{ir}^2, \\
\text{Cov}_R(E_{ij}^*, E_{ij'}^*) = 0 \quad \text{for any} \ j \neq j^*, \text{and} \ \text{Cov}_{S,R}(B_i, E_{ij}^*) = 0.
\]

Suppose our interest is on the \( i^\text{th} \) subject’s latent value \( P_{2^{(i)}} = \mu + B_i \). This value is random since we have not specified which subject \( s \), and corresponds to the \( i^\text{th} \) subject. The \( i^\text{th} \) subject’s latent value can be predicted using the BLUP derived under the usual mixed model. The BLUP of \( P_{2^{(i)}} \) (Searle et al. 1992; Stanek and Singer 2011) is

\[
\hat{P}_{2^{(i)}} = \hat{\mu}^* + k_i \left( \bar{Y}_i - \hat{\mu}^* \right)
\]

where \( \hat{\mu}^* \) is the weighted least squares estimate of the \( \mu \), \( \hat{\mu}^* = \sum w_i^* \bar{Y}_i \) where
\[
w_i^* = \frac{1}{v_i^*}, \quad v_i^* = \sigma_s^2 + \frac{\sigma_{ir}^2}{m} \quad \text{and the shrinkage factor}
\]
\[
k_i = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_{ir}^2/m},
\]

which is different for different subjects.

When variance components are unknown, we replace \( \sigma_s^2 \) and \( \sigma_{ir}^2 \) by the estimated variance components \( \hat{\sigma}_s^2 \) and \( \hat{\sigma}_{ir}^2 \). The estimates of \( \hat{\sigma}_s^2 \) and \( \hat{\sigma}_{ir}^2 \) can be obtained from the observed data and the details are described in chapters 4 and 6. Using

31
the estimates of $\hat{\sigma}_S^2$ and $\hat{\sigma}_{iR}^2$, we obtain the empirical predictor under the usual mixed model

$$\hat{P}^{(i)*}_{2'} = \hat{\mu}^* + \hat{k}_i \left( \bar{Y}_i - \hat{\mu}^* \right)$$

(3.23)

where

$$\hat{\mu}^* = \frac{\sum_{i=1}^n \hat{w}_i^* \bar{Y}_i}{\sum_{i=1}^n 1/\hat{v}_i^*}, \quad \hat{v}_i^* = \hat{\sigma}_S^2 + \frac{\hat{\sigma}_{iR}^2}{m} \quad \text{and} \quad \hat{k}_i = \frac{\hat{\sigma}_S^2}{\hat{\sigma}_S^2 + \hat{\sigma}_{iR}^2 / m}.$$

The mixed model (3.18) can be expressed in matrix form as

$$Y = X\mu + ZB + E$$

(3.24)

where $Y$ is an $nm \times 1$ column vector, $Y = (Y_1' \quad Y_2' \quad \cdots \quad Y_n')'$; $Y_i = (Y_{i1} \quad Y_{i2} \quad \cdots \quad Y_{im})'$; $X = I_{nm}$, an $nm \times 1$ column vector with all elements equal to 1; $Z = I_n \otimes I_m$, an $nm \times n$ matrix, $I_n$ is an $n \times n$ identity matrix, and $\otimes$ denotes the Kronecker product. We denote a vector or matrix in bold and non-italic letter. $B = (B_1 \quad B_2 \quad \cdots \quad B_n)'$ is a vector of random effects. Response errors are given by $E = (E_1' \quad E_2' \quad \cdots \quad E_n')'$ where

$$E_i = (E_{i1}^* \quad E_{i2}^* \quad \cdots \quad E_{im}^*)'$$

and $E$ is an $nm \times 1$ column vector. Typically, it is assumed that $B \sim (0_n, G)$ and $E \sim (0_{nm}, R)$. The covariance matrix of $B$ is $G = \sigma_S^2 I_n$; $R$ is the covariance matrix of $E$, and $R = \bigoplus_{i=1}^n \left( \sigma_{iR}^2 I_m \right)$. Thus,

$$V = \text{Var}(Y) = ZGZ' + R = \bigoplus_{i=1}^n \left( \sigma_S^2 J_m + \sigma_{iR}^2 I_m \right)$$

(3.25)
where $\bigoplus_{i=1}^{n} A$ denotes a block-diagonal matrix with blocks $A$ and $J_m = I_m I_m'$. Based on Henderson’s mixed model equations (Searle et al. 1992), the best linear unbiased estimator (BLUE) of $\mu$ is

$$\bar{\mu} = \left( X'V^{-1}X \right)^{-1} X'V^{-1}Y,$$

(3.26)

where $\text{Var}(\bar{\mu}) = \left( X'V^{-1}X \right)^{-1}$. The best linear unbiased predictor (BLUP) of $B$ is

$$\bar{B} = GZ'V^{-1} \left( Y - X \left( X'V^{-1}X \right)^{-1} X'V^{-1}Y \right).$$

(3.27)

We obtain the estimate of subject’s latent value by the predictor

$$\hat{\mu}_{(2)} = \bar{\mu} + \bar{B}_i. \quad (3.28)$$

We can express this predictor using similar notation to that in expression (3.21) for the $i^{th}$ subject, where $\bar{\mu} = \hat{\mu}^*$ and $\bar{B}_i = k_i \left( \bar{Y}_i - \hat{\mu}^* \right)$. To do so, we use an expression of

$$V_i^{-1} = \left( \sigma^2_s J_m + \sigma^2_{ir} I_m \right)^{-1} \text{ based on a corollary given by Harville (Harville 1997).}$$

**Corollary** (Harville 1997 p424). Let $Q$ be a $m \times m$ nonsingular matrix, and $s$ and $u$ be $m \times 1$ column vectors. Then, $(Q + su')^{-1}$ is nonsingular if and only if $u'Q^{-1}s \neq -1$, where

$$(Q + su')^{-1} = Q^{-1} - \left(1 + u'Q^{-1}s\right)^{-1} Q^{-1}su'Q. \quad (3.29)$$
Since \( V_i = \sigma_S^2 J_m + \sigma_{ir}^2 I_m = \sigma_S^2 \left( I_m + \frac{\sigma_S^2}{\sigma_{ir}^2} J_m \right), \) let \( Q = I_m, \) \( s = \frac{\sigma_S^2}{\sigma_{ir}^2} I_m \) and \( u = I_n. \)

With these definitions \( u'Q^{-1}s = \frac{m\sigma_S^2}{\sigma_{ir}^2} \neq -1 \) since all elements are greater than 0. Using Harville’s corollary,

\[
V_i^{-1} = \left( \sigma_S^2 J_m + \sigma_{ir}^2 I_m \right)^{-1} = \frac{1}{\sigma_{ir}^2} \left( I_m - \frac{1}{1 + I_m' \frac{\sigma_S^2}{\sigma_{ir}^2} J_m} \right) = \frac{1}{\sigma_{ir}^2} \left( I_m - \frac{\sigma_S^2}{m\sigma_S^2 + \sigma_{ir}^2} J_m \right)
\]

Then, since \( V = \bigoplus_{i=1}^n V_i, \) \( V^{-1} = \bigoplus_{i=1}^n V_i^{-1}. \) Using \( X = I_{nm}, \) \( XV^{-1}X = \sum_{i=1}^n (I_m' V_i^{-1} I_m) \) where

\[
I_m' V_i^{-1} I_m = \frac{m \sigma_S^2}{m \sigma_S^2 + \sigma_{ir}^2} \frac{1}{V_i} \Rightarrow \frac{m \sigma_S^2}{m \sigma_S^2 + \sigma_{ir}^2} \frac{1}{V_i}
\]

Similarly, \( XV^{-1}Y = \sum_{i=1}^n (I_m' V_i^{-1} Y_i) \) where

\[
I_m' V_i^{-1} Y_i = \frac{1}{\sigma_S^2} \left( I_m' - \frac{m\sigma_S^2}{m\sigma_S^2 + \sigma_{ir}^2} I_m' \right) Y_i = \frac{1}{m\sigma_S^2 + \sigma_{ir}^2} mY_i = \frac{1}{V_i} Y_i
\]

As a result,
\[\hat{\mu} = (X'V^{-1}X)^{-1}X'V^{-1}Y\]
\[= \left(\sum_{i=1}^{n} (X'_iV^{-1}X_i)\right)^{-1} \sum_{i=1}^{n} (X'_iV^{-1}Y_i)\]
\[= \left(\sum_{i=1}^{n} \frac{1}{v_i^*}\right)^{-1} \sum_{i=1}^{n} \frac{1}{v_i^*} \bar{Y}_i\]
\[= \sum_{i=1}^{n} w_i^* \bar{Y}_i\]
\[= \hat{\mu}^*\]

where \(w_i^* = \frac{1}{\sum_{i=1}^{n} \frac{1}{v_i^*}}\). We evaluate \(B_i\) in a similar manner. From \(B_i = \sigma^2 Z'_i V^{-1} (Y_i - X_i \hat{\mu}^*)\), we have

\[Z'_i V^{-1} Y_i = 1'_m V^{-1} Y_i\]
\[= \left(1/m\sigma^3_s + \sigma^3_{ir}\right) m\bar{Y}_i\]

and

\[Z'_i V^{-1} X_i \hat{\mu}^* = 1'_m V^{-1} 1_m \hat{\mu}^*\]
\[= \frac{m}{m\sigma^2_s + \sigma^2_{ir}} \hat{\mu}^*\]

Thus,

\[B_i = \sigma^2_s Z'_i V^{-1} (Y_i - X_i \hat{\mu}^*)\]
\[= \sigma^2_s \left[\left(\frac{m}{m\sigma^2_s + \sigma^2_{ir}}\right) \bar{Y}_i - \left(\frac{m}{m\sigma^2_s + \sigma^2_{ir}}\right) \hat{\mu}^*\right]\]
\[= \sigma^2_s \frac{\bar{Y}_i - \hat{\mu}^*}{m\sigma^2_s + \sigma^2_{ir}}\]
\[= k_i (\bar{Y}_i - \hat{\mu}^*)\]
When we do not know the variance components $\sigma^2_S$ and $\sigma^2_{ir}$, the estimated variance components $\hat{\sigma}^2_S$ and $\hat{\sigma}^2_{ir}$ are used. $V$ and $G$ change to $\hat{V}$ and $\hat{G}$, and the empirical BLUE of $\mu$ is

$$\hat{\mu} = \left( X'\hat{V}^{-1}X \right)^{-1} X'\hat{V}^{-1}Y,$$

where $Var(\hat{\mu}) = \left( X'\hat{V}^{-1}X \right)^{-1}$. The empirical BLUP of $B$ is

$$\hat{B} = \hat{G}Z'\hat{V}^{-1} \left( Y - X \left( X'\hat{V}^{-1}X \right)^{-1} X'\hat{V}^{-1}Y \right)$$

$$= \hat{G}Z'\hat{V}^{-1} \left( Y - X\hat{\mu} \right).$$

(3.30)

We obtain the empirical estimate of the subject’s latent value by the empirical predictor

$$\hat{P}_{zr}^{(i)} = \hat{\mu} + \hat{B}_i.$$

(3.31)

We can express this predictor using similar notation to that in expression (3.23) for one subject.

### 3.4 Finite Population Mixed Model (FPMM)

Different from the model-based usual mixed model in which the subject’s latent value follows a given distribution, the FPMM is a design-based model based on the finite population sampling and random permutation ideas without additional assumption. Thus, it does not require assumption about parametric distribution of responses. The usual random permutation model (Cassel et al. 1977; Mukhopadhyay 1984; Rao 1984; Padmawar and Mukhopadhyay 1985) is actually a random permutation superpopulation model, which collapses the random variable to a lower dimensional space with the loss of subject label (Zhang 2010). Different from the usual random permutation model, the
FPMM model proposed by Stanek and Singer (2004) includes the information of both subject label and position in a sample, similar to Godambe’s (1955) representation of a sample. The finite population may be considered as a realization of superpopulation random variable. It is the random variable in the finite population that generates permutations of subjects. The usual assumption is that the probability of each permutation occurs equally likely. The random permutation provides a probabilistic “link” to relate a sample to its parent population (Stanek and Singer 2004). This model has been used to predict random effects for the response with response error (Stanek and Singer 2004) and to estimate rate and standardization (Li 2003). Based on the context of the Seasons study, we first describe the two-stage sampling from a finite population, and then we define the FPMM. The representation via random permutations is general since it enables representing unequal probability sampling.

3.4.1 Two-stage Sampling

The two-stage sampling uses simple random sampling without replacement (Hedayat and Sinha 1991; Krishnaiah and Rao 1988; Valliant et al. 2000), and can be associated with the random permutation process. The first stage is to select \( n \) subjects from a finite population of size \( N \); the second stage is to select \( m \) days from a set of \( D \) days for each selected subject.

At the first stage, in order to clearly describe the sampling process based on the random permutation, we introduce the term of “subject label” which is a classical concept used in the literature on sampling finite populations (Godambe 1955; Royall 1970). We
refer to the order of a subject in a vector by its position. For the finite population $N$, we define a vector $\lambda = ((\lambda_s))$ as a listing of subject labels, such as

\[
\begin{pmatrix}
\text{Mary} \\
\text{Tom} \\
\vdots \\
\text{Lily}
\end{pmatrix} \Leftrightarrow \begin{pmatrix}
\lambda_1 \\
\lambda_2 \\
\vdots \\
\lambda_N
\end{pmatrix}
\]

where $\lambda_s$ (i.e., $\lambda_1 = \text{Mary}$) is the subject label for the subject in position $s$ in the listing, $s = 1,...,N$. Each time we permute the $N$ subjects, the subjects could be in a different order. We define the subjects occupying the first $n$ positions in a permutation to be the sampled subjects, so one random sample corresponds to one of the random permutations of subjects.

In order to identify subject $s$ in the finite population corresponding to the $i^{th}$ position (selected subject) in a sample, we use the “sampling indicator random variable, $U_{is}$”, used by Stanek and Singer (2004). It serves as a “bridge” between the population and sample. The value of $U_{is}$ indicates a selection of subject $s$ in position $i$, $s = 1,...,N$ and $i = 1,...,N$ where $U_{is} = 1$ when the $i^{th}$ position (selected subject) is subject “$s$” and $U_{is} = 0$ otherwise. For the population, we define a matrix of indicator random variables, $U$, as

\[
U_{N \times N} = \begin{pmatrix}
U_{11} & U_{12} & \cdots & U_{1N} \\
U_{21} & U_{22} & \cdots & U_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
U_{N1} & U_{N2} & \cdots & U_{NN}
\end{pmatrix} = \begin{pmatrix}
U'_1 \\
U'_2 \\
\vdots \\
U'_N
\end{pmatrix} = (U'_i) = (U_{is})
where $\mathbf{U}_i = (U_{i1} \quad U_{i2} \quad \cdots \quad U_{in})'$. In this way, the latent value of the subject in the $i^{th}$ position in a sample can be represented as $M_i = \sum_{s=1}^{N} U_{is} \mu_s$ for $i = 1, ..., n$. A particular permutation is a realization of $\mathbf{U}$ given by $\mathbf{u}_{N \times N} = ((u_{is}))$, where $u_{is}$ identifies the subject, $l_i = \sum_{s=1}^{N} u_{is} \lambda_s$, in the finite population in the $i^{th}$ position.

At the second stage, a “day” is sampled for each selected subject. We describe the process for subject $s$. For each selected subject, we index days by $d = 1, ..., D$ and imagine a listing of days. We permute the days in the listing, and define the first $m$ days in a permutation to be in a sample. We use the sampling indicator random variable $U_{jd}^{(s)}$ for selection of day $d$ in $j^{th}$ position, $j = 1, ..., D$, where $U_{jd}^{(s)} = 1$ when the $j^{th}$ position (selected day) is day “$d$” and $U_{jd}^{(s)} = 0$ otherwise. We define a random variable

$$M_{ij} = \sum_{s=1}^{N} U_{is} \sum_{d=1}^{D} U_{jd}^{(s)} \mu_{sd}$$

(3.32)

to indicate the latent value on the $j^{th}$ day for the $i^{th}$ subject. Similar to $\mathbf{U}$, we define $\mathbf{U}^{(s)} = \left( (U_{jd}^{(s)}) \right)$ as a $D \times D$ matrix of random variables. Also, let $\mu = (\mu_1' \quad \mu_2' \quad \cdots \quad \mu_N')'$ and $\mu_s = (\mu_{s1} \quad \mu_{s1} \quad \cdots \quad \mu_{sD})'$, where $\mu$ is an $ND \times 1$ parameter vector for the population. We define

$$\mathbf{M} = (\mathbf{U} \otimes \mathbf{I}_D) \left( \bigotimes_{s=1}^{N} \mathbf{U}^{(s)} \right) \mu$$

(3.33)

as the equation (13) in Stanek and Singer (2004) as a random permutation of parameters in the population, where $\mathbf{M}$ is an $ND \times 1$ vector.
3.4.2 Finite Population Mixed Model (FPMM)

We describe a finite population mixed model similar to the one defined in Stanek and Singer (2004). A finite population mixed model can be developed directly from the two-stage sampling of a finite population. Using equations (3.6) and (3.32), we develop a stochastic model for the finite population parameters as

\[
M_{ij} = \sum_{s=1}^{N} U_{is} \sum_{d=1}^{D} U_{jd}^{(s)} \mu_{sd}
\]

\[
= \sum_{s=1}^{N} U_{is} \sum_{d=1}^{D} U_{jd}^{(s)} (\mu + \beta_s + \delta_{sd})
\]

\[
= \mu + \sum_{s=1}^{N} U_{is} \beta_s + \sum_{s=1}^{N} U_{is} \sum_{d=1}^{D} U_{jd}^{(s)} \delta_{sd}
\]

The model is defined for \( N \times D \) random variables \( M_{ij} \). We make the simplifying assumption that all permutations are equally likely. We define a finite population mixed model as

\[
M_{ij} = \mu + B_i^* + D_{ij}^*
\]

(3.34)

where \( B_i^* = \sum_{s=1}^{N} U_{is} \beta_s \), \( D_{ij}^* = \sum_{s=1}^{N} U_{is} \sum_{d=1}^{D} U_{jd}^{(s)} \delta_{sd} \), \( i = 1, ..., N \) and \( j = 1, ..., D \). The population mean \( \mu \) is fixed. The subject effect \( B_i^* \) and day effect \( D_{ij}^* \) are random since \( U_{is} \) and \( U_{jd}^{(s)} \) are random. We define \( D_{ij}^* = \sum_{d=1}^{D} U_{jd}^{(s)} \delta_{sd} \), so that \( D_{ij}^* = \sum_{s=1}^{N} U_{is} D_{ij}^* \).

Since \( \mu_{sd} \) cannot be observed directly, we express the observed response of 24HR measure for the \( i^{th} \) subject on the \( j^{th} \) day as \( Y_{ijk}^* \), where \( k^* \) indicates the “measurement occasion” on the \( j^{th} \) day. We express \( Y_{ijk}^* \) as

40
\[ Y_{ijk}^{*} = M_{ij} + \sum_{s=1}^{N} U_{is} \sum_{d=1}^{D} U_{jd}^{(s)} E_{sdk}^{*} \]
\[ = \mu + B_{ij}^{*} + D_{ij}^{*} + E_{ijk}^{*}, \] (3.35)

where the random measurement error \( E_{ijk}^{*} = \sum_{s=1}^{N} U_{is} E_{gjk}^{*} \) and \( E_{gjk}^{*} = \sum_{d=1}^{D} U_{jd}^{(s)} E_{skd}^{*} \).

When one \( (k^* = 1) \) 24HR recall is collected on one recall day, we define the response error as \( E_{ij}^{**} = \sum_{s=1}^{N} U_{is} E_{ij}^{**} \) and

\[ E_{ij}^{**} = D_{ij}^{*} + E_{ijk}^{*}. \] (3.36)

corresponding to the sum of day-to-day variability and the random measurement error.

Therefore, the FPMM model (3.35) can further be expressed as

\[ Y_{ij} = \mu + B_{ij}^{*} + E_{ij}^{**}. \] (3.37)

Consistent with the FPMM defined in Stanek and Singer (2004), we have \( E_{s} (B_{ij}^{*}) = 0, \)

\[ \text{Var}_{s} (B_{ij}^{*}) = \frac{N-1}{N} \sigma_{s}^{2}, \] (3.38)

and for any \( i \neq i^{*}, \)

\[ \text{Cov}_{s} (B_{ij}^{*}, B_{i^{*}j}^{*}) = \frac{1}{N} \sigma_{s}^{2}. \] (3.39)

For the response error, we assume that \( \text{Cov}_{r} (E_{ij}^{**}, E_{i^{*}j}^{**}) = 0 \) for any \( i \neq i^{*} \) and the random measurement error \( E_{ijk}^{*} \) within subject \( s \) is uncorrelated with each other. Then, we have

\[ E_{r} (E_{ij}^{**}) = 0, \]

\[ \text{Var}_{r} (E_{ij}^{**}) = \frac{D-1}{D} \sigma_{D}^{2} + \sigma_{E}^{2}. \] (3.40)
where $\sigma^2_{\delta}$ is defined by (3.5) and $\sigma^2_{\delta}$ is defined by (3.10). For any $i \neq i$ or $j \neq j$,

$$ \text{Cov}_D \left( E_{ij}^{ss}, E_{ij}^{ss} \right) = -\frac{1}{D} \sigma^2_{\delta}. \tag{3.41} $$

Model (3.35) can be expressed in matrix form for the population as

$$ \mathbf{Y} = \mathbf{M} + \mathbf{W}^\prime. \tag{3.42} $$

where $\mathbf{W}^\prime = (\mathbf{U} \otimes \mathbf{I}_D) \left( \bigoplus_{s=1}^{N} \mathbf{U}^{(s)} \right) \mathbf{W}$, $\mathbf{W} = (\mathbf{W}_1 \quad \mathbf{W}_2 \quad \ldots \quad \mathbf{W}_N)'$ and

$$ \mathbf{W}_s = \left( (E_{sk}^\prime) \right) = (E_{s11} \quad E_{s21} \quad \ldots \quad E_{s1D})' \quad \text{(as the equation (14) in Stanek and Singer (2004))}. $$

$\mathbf{Y}$ and $\mathbf{M}$ are $ND \times 1$ vectors. $\mathbf{M}$ is defined in (3.33) where

$$ \mathbf{\mu} = \begin{pmatrix} \mu_1' \\ \mu_2' \\ \vdots \\ \mu_N' \end{pmatrix} = \mathbf{1}_{ND} \mathbf{\mu} + \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{pmatrix} \otimes \mathbf{1}_D + \begin{pmatrix} \delta_1' \\ \delta_2' \\ \vdots \\ \delta_N' \end{pmatrix} $$

and the day effect vector

$$ \mathbf{\delta}_s = (\delta_{s1} \quad \delta_{s2} \quad \ldots \quad \delta_{sd})'. $$

We refer to the equations (23) and (24) in Stanek and Singer (2004) for the expressions of the expected value and the variance of $\mathbf{Y}$. Using $S$ to indicate expectation with respect to permutations of the subjects, using $D$ to indicate expectation with respect to permutations of days of the subject $s$, using $M$ to indicate expectation with respect to measurement error,

$$ E_{SDM}(\mathbf{Y}) = \mathbf{1}_{ND} \mathbf{\mu} \tag{3.43} $$

and

$$ \text{Var}_{SDM}(\mathbf{Y}) = (\sigma^2_E + \sigma^2_{\delta}) \mathbf{1}_{ND} + \sigma^2(\mathbf{I}_N \otimes \mathbf{J}_D) - \frac{\sigma^2_{\delta}}{N} \mathbf{J}_{ND} \tag{3.44} $$

where
\[ \sigma^2 = \sigma_i^2 - \frac{\sigma_D^2}{D}. \] (3.45)

When a sample is obtained, the elements of \( Y \) are partitioned into two portions: the sample (\( Y_I \)) and the remainder (\( Y_{II} \)). This can be expressed as

\[
\begin{pmatrix}
Y_I \\
Y_{II}
\end{pmatrix} = \begin{pmatrix}
K_I Y \\
K_{II} Y
\end{pmatrix}
\] (3.46)

where \( K_I = (I_n \quad 0_{n \times (N-n)}) \otimes (I_m \quad 0_{m \times (D-m)}) \) and \( K_{II} = \left( \begin{pmatrix} I_n & 0_{n \times (N-n)} \otimes (0_{(D-m) \times m} | I_{D-m}) \\ 0_{(N-n) \times m} & I_{N-n} \otimes I_D \end{pmatrix} \right) \).

Based on (3.46), Stanek and Singer (2004) derived the variance of \( Y_I \), \( V^* \) where

\[ V^* = \text{Var}_{SDM}(Y_I) = \sigma_D^2 \left( I_n - \frac{1}{N} J_n \right) \otimes J_m + \left( \sigma_D^2 + \sigma_E^2 \right) I_{nm} - \frac{1}{D} \sigma_D^2 (I_n \otimes J_m) \] (3.47)

Based on (3.37), the FPMM model for the sample can be expressed in matrix form as

\[ Y_i = X_i \mu + Z_i \beta^* + E_i \] (3.48)

where \( Y_i \) is an \( nm \times 1 \) column vector, \( Y_i = (Y_{i1} \quad Y_{i2} \quad \cdots \quad Y_{im})' \), \( Y_i = (Y_{i1} \quad Y_{i2} \quad \cdots \quad Y_{im})' \); \( X_i = I_{nm} \); \( Z_i = I_n \otimes I_m \) and \( \beta^* = (\beta_1^* \quad \beta_2^* \quad \cdots \quad \beta_n^*)' \) is a vector of random effects. Response errors are given by \( E_i = (E_{i1}^{**} \quad E_{i2}^{**} \quad \cdots \quad E_{im}^{**})' \) where

\[ E_i^{**} = (E_{i1}^{**} \quad E_{i2}^{**} \quad \cdots \quad E_{im}^{**})' \] and \( E_i \) is an \( nm \times 1 \) column vector. Typically, it is assumed
that $\mathbf{B}^* \sim (0_n, \mathbf{G}^*)$, $\mathbf{E}_i \sim (0_{n_m}, \mathbf{R}^*)$. The covariance matrix of $\mathbf{B}^*$ is $\mathbf{G}^* = \sigma_B^2 \left( \mathbf{I}_n - \frac{1}{N} \mathbf{J}_n \right)$; $\mathbf{R}^*$ is the covariance matrix of $\mathbf{E}_i$, and $\mathbf{R}^* = \left( \sigma_D^2 + \sigma_E^2 \right) \mathbf{I}_{n_m} - \frac{1}{D} \sigma_D^2 \left( \mathbf{I}_n \otimes \mathbf{J}_m \right)$.

The latent value of the $i^{th}$ subject is $P_s^{(i)} = \mu + B_i^*$. It is random and can be predicted using the BLUP derived under the FPMM. Based on the Henderson’s mixed model equations (Searle et al. 1992), the BLUE of $\mu$ is

$$\hat{\mu} = \left( \mathbf{X}_i' \hat{\mathbf{V}}^{-1} \mathbf{X}_i \right)^{-1} \mathbf{X}_i' \hat{\mathbf{V}}^{-1} \mathbf{Y}_i , \quad (3.49)$$

where $\operatorname{Var}(\hat{\mu}) = \left( \mathbf{X}_i' \hat{\mathbf{V}}^{-1} \mathbf{X}_i \right)^{-1}$. The BLUP of $\mathbf{B}^*$ is

$$\hat{\mathbf{B}}^* = \mathbf{G}^* \mathbf{Z}_i \hat{\mathbf{V}}^{-1} \left( \mathbf{Y}_i \cdot \mathbf{X}_i \left( \mathbf{X}_i' \hat{\mathbf{V}}^{-1} \mathbf{X}_i \right)^{-1} \mathbf{X}_i' \hat{\mathbf{V}}^{-1} \mathbf{Y}_i \right) = \mathbf{G}^* \mathbf{Z}_i \hat{\mathbf{V}}^{-1} \left( \mathbf{Y}_i \cdot \mathbf{X}_i \hat{\mu} \right) \quad (3.50)$$

We obtain the estimate of subject’s latent value by the predictor

$$\hat{P}_s^{(i)} = \hat{\mu} + \hat{B}^* . \quad (3.51)$$

When variance components are unknown, we compute the empirical predictor based on the estimated variance components. We replace $\sigma_S^2$, $\sigma_D^2$ and $\sigma_E^2$ by the estimated variance components $\hat{\sigma}_S^2$, $\hat{\sigma}_D^2$ and $\hat{\sigma}_E^2$. The estimates of $\hat{\sigma}_S^2$, $\hat{\sigma}_D^2$ and $\hat{\sigma}_E^2$ are obtained from the observed data and the details are described in chapters 4 and 6. After knowing $\hat{\sigma}_S^2$, $\hat{\sigma}_D^2$ and $\hat{\sigma}_E^2$, we obtain the estimated $\hat{\mathbf{V}}^*$ and $\hat{\mathbf{G}}^*$. The empirical BLUE of $\mu$ is

$$\hat{\mu} = \left( \mathbf{X}_i' \hat{\mathbf{V}}^{-1} \mathbf{X}_i \right)^{-1} \mathbf{X}_i' \hat{\mathbf{V}}^{-1} \mathbf{Y}_i , \quad (3.52)$$

where $\operatorname{Var}(\hat{\mu}) = \left( \mathbf{X}_i' \hat{\mathbf{V}}^{-1} \mathbf{X}_i \right)^{-1}$. The empirical BLUP of $\mathbf{B}^*$ is
\[
\hat{B}^* = \hat{G}^* Z_i \hat{V}^{-1} \left( Y_i - X_i \left( X_i' \hat{V}^{-1} X_i \right)^{-1} X_i' \hat{V}^{-1} Y_i \right).
\]

We obtain the empirical estimate of subject’s latent value by the empirical predictor
\[
\hat{P}_n^{(i)} = \hat{\mu} + \hat{B}_i^*.
\]

Corresponding with (3.21), Stanek and Singer (2004) derived the predictor \( \hat{P}_3^{(i)} \)
under the FPMM as
\[
\hat{P}_3^{(i)} = f \left( \bar{Y} + k_r \left( \bar{Y}_i - \bar{Y} \right) \right) + (1 - f) \left( \bar{Y} + k_r \left( \bar{Y}_i - \bar{Y} \right) \right)
\]
where \( f \) is the unit (day) sampling fraction, \( f = \frac{m}{D} \), \( k_r = \frac{m \sigma_r^2 + \sigma_D^2}{m \sigma_r^2 + \sigma_D^2 + \sigma_E^2} \) and
\[
k_r = \frac{m \sigma_r^2}{m \sigma_r^2 + \sigma_D^2 + \sigma_E^2}.
\]
In our study, we assume \( D \) is large, \( D \gg m \) so that \( f = \frac{m}{D} \approx 0 \),
and \( \frac{1}{D} \approx 0 \) so that \( \sigma_r^2 \approx \sigma_D^2 \). Thus, (3.55) is simplified as
\[
\hat{P}_3^{(i)} = \bar{Y} + k \left( \bar{Y}_i - \bar{Y} \right)
\]
where \( \bar{Y}_i = \frac{1}{m} \sum_{j=1}^{m} Y_{ij} \), \( \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} \bar{Y}_i \), and the shrinkage factor
\[
k = \frac{\sigma_s^2}{\sigma_s^2 + \left( \sigma_D^2 + \sigma_E^2 \right) / m}.
\]
Different from \( k_i \) in (3.22), \( k \) is calculated by using \( \left( \sigma_D^2 + \sigma_E^2 \right) \) in place of \( \sigma_{ir}^2 \). \( k \) is
constant across different samples.

In practice, when \( k^* = 1 \), i.e. one 24HR recall is collected in the \( j^{th} \) selected
recall day, it is not possible to separate \( \sigma_r^2 \) (day-to-day variability) and \( \sigma_E^2 \) (random
measurement error). We define an average response error
\[ \sigma^2_R = \sigma^2_D + \sigma^2_E. \]  

(3.58)

This is comparable to

\[ \sigma^2_R = \frac{1}{N} \sum_{i=1}^{N} \sigma^2_{sr}. \]  

(3.59)

using (3.15).

Similarly, when variance components are unknown, we replace \( \sigma^2_S \), \( \sigma^2_D \) and \( \sigma^2_E \) by the estimated variance components \( \hat{\sigma}^2_S \), \( \hat{\sigma}^2_D \) and \( \hat{\sigma}^2_E \), and obtain the empirical predictor under the FPMM as

\[ \hat{P}^{(i)}_{3e} = \bar{Y} + \hat{k} (\bar{Y}_i - \bar{Y}) \]  

(3.60)

where the estimated shrinkage factor \( \hat{k} = \frac{\hat{\sigma}^2_S}{\hat{\sigma}^2_S + (\hat{\sigma}^2_D + \hat{\sigma}^2_E) / m} \).

### 3.5 7DDR Model

As a short-term dietary assessment instrument, the 7DDR requires subjects to recall their meals and snacks over seven consecutive days, and the final 7DDR data correspond to a single reported value of average daily intake for each nutrient over the seven days. In this study we assume that there is no bias in the 24HR measure. Ideally, without bias in the 7DDR measure, the estimate obtained from a 7DDR should have the same expected value as would an estimate obtained from the average 24HR over the same seven consecutive days. But previous studies have shown that the 7DDR is biased (Hebert et al. 2002). The 7DDR bias may be smaller than the bias resulting from long-term dietary assessment instruments, such as food frequency questionnaires, due to the relatively short recall time period.
Our scheme is to quantify the bias of the 7DDR measure and set up a simultaneous model that uses the 7DDR data and the 24HR data. Comparing the predictors obtained from the 24HR data with the predictors obtained from the combined data, we explore the possibility of obtaining a more accurate predictor of a subject’s latent value. Since the 24HR serves as a reference instrument, the 7DDR model is formed by relating the subject’s 7DDR measure to the subject’s 24HR measure and including the possible biases. We first define a measurement error model for the 7DDR, and then define a 7DDR model under the usual mixed model and a 7DDR model under the FPMM framework.

We define $\mu_{sw}$ as subject $s$’s latent value of average daily intake in week $w$, $w = 1,\ldots,W$. We have

$$
\mu_s = \frac{1}{W} \sum_{w=1}^{W} \mu_{rw},
$$

(3.61)

where we assume $D = 7W$, so that the intake period is identical for the subject regardless of whether intake is measured via a 24HR or 7DDR instrument. We define the week effect $\omega_{sw} = \mu_{sw} - \mu_s$ as the deviation of subject $s$’s latent value of average daily intake in week $w$ from the subject’s latent value. The definition of the week effect results in the constraint $\sum_{w=1}^{W} \omega_{sw} = 0$ for all $s = 1,\ldots,N$. We also define $\sigma_{sw}^2 = \frac{1}{W - 1} \sum_{w=1}^{W} \omega_{sw}^2$ for all $s = 1,\ldots,N$.

We define the observed saturated fat intake in week $w$ for subject $s$ using the 7DDR instrument as $\tilde{Y}_{swk}$, where $k^*$ indicates the “measurement occasion” in week $w$. 
(Note: we use tilde “~” to denote the terms related to the bias). Similar to (3.7), a parametric measurement error model for $\tilde{Y}^{*}_{swk}$ can be written as

$$\tilde{Y}^{*}_{swk} = \mu_{sw} + \tilde{F}^{*}_{swk},$$

where $\tilde{F}^{*}_{swk}$ denotes measurement error term which includes random measurement error using 7DDR instrument and bias, given by $E_{M}(\tilde{F}^{*}_{swk}) = \tilde{\delta}_{sw}$, where the subscript $M$ denotes the expectation with respect to the distribution of the measurement error. Let

$$\tilde{\delta}_{s} = \frac{1}{W} \sum_{w=1}^{W} \tilde{\delta}^{*}_{sw}$$

and define $\tilde{\delta}^{*}_{sw} = \tilde{\delta}^{*}_{sw} - \tilde{\delta}_{s}$, so $\tilde{\delta}^{*}_{sw} = \tilde{\delta}_{s} + \tilde{\delta}^{*}_{sw}$ so that

$$\sum_{w=1}^{W} \tilde{\delta}^{*}_{sw} = 0$$

for all $s = 1,...,N$. We define

$$\tilde{\sigma}^{2}_{sw} = \frac{1}{W-1} \sum_{w=1}^{W} \tilde{\delta}^{2}_{sw}$$

for all $s = 1,...,N$ and

$$\tilde{\sigma}^{2}_{w} = \frac{1}{N} \sum_{s=1}^{N} \tilde{\sigma}^{2}_{sw}.$$ 

Now let $F_{swk} = \tilde{F}^{*}_{swk} - \tilde{\delta}^{*}_{sw}$, then $E_{M}(F_{swk}) = 0$. Using this notation,

$$\tilde{F}^{*}_{swk} = \tilde{\delta}^{*}_{sw} + F_{swk},$$

$$= \tilde{\delta}_{s} + \tilde{\delta}^{*}_{sw} + F_{swk}.$$ 

where $\tilde{\delta}_{s}$ and $\tilde{\delta}^{*}_{sw}$ are associated with bias. The term $\tilde{\delta}_{s}$ denotes subject-specific deviation in 7DDR measure which may be associated with subject’s certain
characteristics such that the subject-specific bias will be reproduced when response is repeatedly observed on the same subject. We define

$$\tilde{\mu} = \frac{1}{N} \sum_{s=1}^{N} \tilde{\delta}_s$$  \hspace{1cm} (3.67)

and

$$\tilde{\beta}_s = \tilde{\delta}_s - \tilde{\mu},$$  \hspace{1cm} (3.68)

where $\tilde{\beta}_s$ is called subject effect bias – a deviation of the subject-specific bias from the mean $\tilde{\mu}$. These definitions imply that $\sum_{s=1}^{N} \tilde{\beta}_s = 0$. We define the variance in the subject effect bias as

$$\tilde{\sigma}_s^2 = \frac{1}{N-1} \sum_{s=1}^{N} \tilde{\beta}_s^2.$$  \hspace{1cm} (3.69)

The term $\tilde{\delta}_{sw}$ is nested in $\tilde{\delta}_s$ and is called week bias. Based on (3.62), we can further express $\tilde{Y}_{swk}^*$ as

$$\tilde{Y}_{swk}^* = \mu_s + \tilde{\delta}_s + \omega_{sw} + \tilde{\delta}_{sw} + F_{swk}^*$$

$$= \mu + \tilde{\mu} + \beta_s + \tilde{\beta}_s + \omega_{sw} + \tilde{\delta}_{sw} + F_{swk}^*,$$  \hspace{1cm} (3.70)

where $F_{swk}^*$ is the random measurement error using the 7DDR instrument in week $w$ for subject $s$; $E_M(F_{swk}^*) = 0$ and we define $Var_M(F_{swk}^*) = \sigma_{sw}^2$. We assume $F_{swk}^*$ is uncorrelated with $F_{swk'}^*$ for any $s \neq s'$, $w \neq w'$ or $k \neq k'$. The subscript $M$ denotes the expectation with respect to the distribution of the random measurement error.
The $\text{Var}_M(F_{swk})$ from the 7DDR can be associated with the $\text{Var}_M(E_{sdk})$ from the 24HR by linking the 7DDR recall to the average of seven 24HR recalls. Figure 3.1 shows the 7DDR weeks relative to the 24HR days.

![Figure 3.1 7DDR weeks relative to 24HR days](image)

Let $D_w = \{d = 7w - 6, d = 7w - 5, \ldots, d = 7w\}$ a set of days. From (3.7), we have the average $Y_{sdk}$ of seven 24HR recalls

$$\frac{1}{7} \sum_{d \in D_w} Y_{sdk} = \mu_w + \frac{1}{7} \sum_{d \in D_w} E_{sdk}. \tag{3.71}$$

and define the average measurement error of seven 24HR recalls as

$$E_{swk} = \frac{1}{7} \sum_{d \in D_w} E_{sdk}. \tag{3.72}$$

Since $E_M(E_{sdk}) = 0$ and $\text{Var}_M(E_{sdk}) = \sigma_{sde}^2$, we have $E_M(E_{swk}) = 0$ and

$$\text{Var}_M(E_{swk}) = \frac{1}{7} \left( \frac{1}{7} \sum_{d \in D_w} \sigma_{sde}^2 \right). \tag{3.73}$$

We define

$$\sigma_{swe}^2 = \frac{1}{7} \sum_{d \in D_w} \sigma_{sde}^2, \tag{3.73}$$

so that

$$\text{Var}_M(E_{swk}) = \frac{1}{7} \sigma_{swe}^2. \tag{3.74}$$

In order to link the variances of random measurement errors of average seven-24HR and 7DDR, we assume
\[ \text{Var}_M(F_{swk}) = \text{Var}_M(E_{swk}), \quad (3.75) \]

i.e.

\[ \sigma_{sw}^2 = \frac{1}{7} \sigma_{swc}^2, \quad (3.76) \]

When the weeks are randomly selected from the \( W \) weeks by using simple random sampling without replacement for each subject \( s \), we index the selected week by \( t = 1, \ldots, q_s \), where we assume that there are \( q_s \) weeks selected for subject \( s \). For simplicity, we assume that \( q_s = q \) for all subjects.

In (3.70), when the observed response from the 7DDR recall is collected in the \( t^{th} \) randomly selected week, the observed saturated fat intake for subject \( s \) in the \( t^{th} \) selected recall week using the 7DDR instrument, \( \tilde{Y}_{stk}^* \), can be written as

\[ \tilde{Y}_{stk}^* = \mu_s + \delta_s + W_{st} + \tilde{D}_{st} + F_{stk}, \quad (3.77) \]

where \( W_{st} \) is the random week effect. We assume \( E_w(W_{st}) = 0 \) and

\[ \text{Var}_w(W_{st}) = \frac{W - 1}{W} \sigma_{sW}^2, \quad (3.78) \]

where the subscript \( W \) denotes the expectation with respect to the distribution of the week effect \( W_{st} \). For any \( t \neq t^* \),

\[ \text{Cov}_w(W_{st}, W_{st'}) = 0. \quad (3.79) \]

\( \tilde{D}_{st} \) is the random week bias. We assume \( E_w(\tilde{D}_{st}) = 0 \) and

\[ \text{Var}_w(\tilde{D}_{st}) = \frac{W - 1}{W} \tilde{\sigma}_{iW}^2, \quad (3.80) \]
where the subscript $W$ denotes the expectation with respect to the distribution of the week bias $\bar{D}_{st}$. For any $t \neq t^*$,

$$\text{Cov}_W(\bar{D}_{st}, \bar{D}_{st'}) = 0.$$  \hfill (3.81)

In practice, when $k^* = 1$, i.e. one 7DDR recall is collected in the $t^{th}$ selected recall week, it is not possible to separate week-to-week variability and the random measurement error due to the 7DDR instrument in data analysis. We define a response error

$$F_{st}^* = W_{st} + F_{stk},$$  \hfill (3.82)

corresponding to the sum of week-to-week variability and the random measurement error. We define the observed saturated fat intake for subject $s$ in $t^{th}$ selected week from the 7DDR instrument as $\bar{Y}^*_s$, a measurement error model for $\bar{Y}^*_s$ can be written as

$$\bar{Y}^*_s = \mu_s + \bar{\delta}_s + \bar{D}_{st} + F_{st}^*$$  \hfill (3.83)

where $F_{st}^*$ is the random response error, we define $E_R(F_{st}^*) = 0$ and $\text{Var}_R(F_{st}^*) = \sigma_{stk}^2$, the subscript $R$ denotes the expectation with respect to the distribution of the random response error.

Next, we derive the relationship of the $\text{Var}_R(F_{st}^*)$ from 7DDR and the $\text{Var}_R(E_{sj}^*)$ from 24HR. We use the conditional expectation to evaluate the variance $\text{Var}_R(F_{st}^*)$. The variance is given by

$$\text{Var}_R(F_{st}^*) = \text{Var}_W(F_{st}^*)$$
$$= E_W(\text{Var}_M(F_{st}^*)) + E_M(\text{Var}_W(F_{st}^*))$$
$$= E_W(\text{Var}_M(F_{stk}^*)) + \text{Var}_W(W_{st}).$$
Then, from (3.73) to (3.76), and (3.9),

\[ E_w \left( \text{Var}_W \left( F_{st'} \right) \right) = E_w \left( \sigma^2_{sw} \right) \]

\[ = \frac{1}{W} \sum_{w=1}^{W} \left( \frac{1}{7} \left( \frac{1}{D} \sum_{d=1}^{D} \sigma^2_{sde} \right) \right) \]

\[ = \frac{1}{7} \left( \frac{1}{D} \sum_{d=1}^{D} \sigma^2_{sde} \right) \]

\[ = \frac{1}{7} \sigma^2_{sk} \]

Next, we evaluate the variance \( \text{Var}_W \left( W_{st} \right) \). Since the day effect vector \( \mathbf{s} \) is a \( D \times 1 \) column vector with element \( \delta_{sd} \), for a random permutation of day \( d \), we define

\[ \mathbf{D}_s = \mathbf{U}^{(s)} \mathbf{s} \], so \( E_D \left( \mathbf{D}_s \right) = 0 \) and \( \text{Var}_D \left( \mathbf{D}_s \right) = \sigma^2_{sD} \left( \mathbf{I}_D - \frac{1}{D} \mathbf{J}_D \right) \) where \( \sigma^2_{sD} = \frac{1}{D-1} \sum_{d=1}^{D} \delta^2_{sd} \).

Let the week effect vector \( \omega_s = (\omega_{s1} \ \omega_{s2} \ \cdots \ \omega_{sw})' \), a \( W \times 1 \) column vector with element \( \omega_{sw} \). Since \( \omega_{sw} = \frac{1}{7} \sum_{d=7w-6}^{7w} \delta_{sd} \), for a random permutation of week \( w \), we define

\[ \mathbf{W}_s = \mathbf{U}^{(s)} \omega_s \]. Thus, we have \( \mathbf{W}_s = \left( \frac{1}{7} \mathbf{1}' \mid 0_{D-7} \right) \mathbf{D}_s \) and

\[ \text{Var} \left( \mathbf{W}_s \right) = \left( \frac{1}{7} \mathbf{1}' \mid 0_{D-7} \right) \text{Var}_D \left( \mathbf{D}_s \right) \left( \frac{1}{7} \mathbf{1}' \mid 0_{D-7} \right)' \]

\[ = \sigma^2_{sD} \left( \frac{1}{7} \mathbf{1}' \mid 0_{D-7} \right) \left( \mathbf{1}_D - \frac{1}{D} \mathbf{J}_D \right) \left( \frac{1}{7} \mathbf{1}' \mid 0_{D-7} \right)' \]

\[ = \sigma^2_{sD} \left( \frac{1}{7} - \frac{1}{D} \right) \]

\[ = \sigma^2_{sD} \left( \frac{1}{7} - \frac{1}{7 \times W} \right) \]

\[ = \frac{1}{7} \left( \frac{W-1}{W} \right) \sigma^2_{sD} \]

Therefore,
\[
\text{Var}_w(W_{st}) = \text{Var}(W_{st}) = \frac{1}{7} \left( \frac{W-1}{W} \right) \sigma_{sd}^2 \\
= \frac{1}{7} \sigma_{sd}^2 - \frac{1}{D} \sigma_{sd}^2 
\]

and

\[
\text{Var}_r(F_{st}^*) = E_w \left( \text{Var}_m(F_{stk}) \right) + \text{Var}_w(W_{st}) \\
= \frac{1}{7} \sigma_{sD}^2 + \frac{1}{7} \sigma_{sd}^2 - \frac{1}{D} \sigma_{sd}^2 \\
= \frac{1}{7} \left( \sigma_{sD}^2 + \sigma_{st}^2 \right) - \frac{1}{D} \sigma_{sd}^2
\]

Since \( \text{Var}_r(E_{st}^*) = \sigma_{sr}^2 = \sigma_{sd}^2 + \sigma_{st}^2 \), as a result, we have

\[
\text{Var}_r(F_{st}^*) = \sigma_{sr}^2 \\
= \frac{1}{7} \left( \sigma_{sD}^2 + \sigma_{st}^2 \right) - \frac{1}{D} \sigma_{sd}^2 \\
= \frac{1}{7} \sigma_{sr}^2 - \frac{1}{D} \sigma_{sd}^2
\]

For a usual mixed model, \( \frac{1}{D} \approx 0 \),

\[
\text{Var}_r(F_{st}^*) = \frac{1}{7} \text{Var}_r(E_{st}^*) .
\]

### 3.5.1 7DDR Model under the Usual Mixed Model

We define the observed saturated fat intake for the \( i^{th} \) randomly selected subject from the population in the \( t^{th} \) randomly selected week from the 7DDR instrument as \( \tilde{Y}_{it} \), where \( i = 1, \ldots, n \), \( n \leq N \); \( t = 1, \ldots, q \), \( q \leq W \). A usual mixed model can be expressed as

\[
\tilde{Y}_{it} = \mu + B_i + \tilde{\mu} + \tilde{B}_i + \tilde{D}_i + F_{it}^*.
\]
where $\tilde{B}_i$ is the subject effect bias, and is a random effect. We assume that $E_S(\tilde{B}_i) = 0$,

$$Var_S(\tilde{B}_i) = \frac{N-1}{N} \sigma^2_s,$$  \hspace{1cm} (3.88)

and for any $i \neq i^*$, $Cov_S(\tilde{B}_i, \tilde{B}_{i^*}) = 0$. The subscript $S$ denotes the expectation with respect to the distribution of the subject effect bias $\tilde{B}_i$. For the random week bias, we have $E_W(\tilde{D}_u) = 0$,

$$Var_W(\tilde{D}_u) = \frac{W-1}{W} \sigma^2_{iw}.$$  \hspace{1cm} (3.89)

The subscript $W$ denotes the expectation with respect to the distribution of the week bias $\tilde{D}_u$. For a usual mixed model, $\frac{1}{W} \equiv 0$, we have

$$Var_W(\tilde{D}_u) = \tilde{\sigma}^2_{iw}.$$  \hspace{1cm} (3.90)

We assume, for any $i \neq i^*$ or $t \neq t^*$, $Cov_W(\tilde{D}_u, \tilde{D}_{i,t}) = 0$. For the random response error, we have $E_R(F^*_u) = 0$,

$$Var_R(F^*_u) = \sigma^2_{ir}$$

$$= \frac{1}{7} \sigma^2_{ir}$$  \hspace{1cm} (3.91)

and assume for any $i \neq i^*$ or $t \neq t^*$, $Cov_R(F^*_u, F^*_{i,t}) = 0$. We also assume

$Cov_S(B_i, \tilde{B}_{i}) = 0$.

The mixed model (3.87) is expressed in matrix form as

$$\tilde{Y} = X_\mu \mu + Z_\tilde{B} \tilde{B} + \tilde{F}$$  \hspace{1cm} (3.92)
where $\tilde{Y}$ is an $nq \times 1$ column vector, $\tilde{Y} = (\tilde{Y}_1' \ \tilde{Y}_2' \ \cdots \ \tilde{Y}_n')'$, $\tilde{Y}_i = (\tilde{Y}_{i1} \ \tilde{Y}_{i2} \ \cdots \ \tilde{Y}_{iq})'$; $X_w = I_n \otimes (1_q' \ 1_q')$, an $nq \times 2$ matrix with all elements equal to 1; $\mu_c = \begin{pmatrix} \mu' \\ \tilde{\mu} \end{pmatrix}$ is fixed effects vector; $Z_w = I_n \otimes (1_q' \ 1_q')$, an $nq \times 2n$ block diagonal matrix with blocks

$$
(1_q' \ 1_q')' \tilde{B} \text{ is a vector of random effects, } \tilde{B} = \left( \bigoplus_{i=1}^n \begin{pmatrix} B_i' \\ B_i \end{pmatrix} \right) 1_n; \tilde{F} \text{ is an } nq \times 1 \text{ column vector, } \tilde{F} = \left( \tilde{F}_1' \ \tilde{F}_2' \ \cdots \ \tilde{F}_n' \right)' \text{ and } \tilde{F}_i = \left( \tilde{D}_{i1} + F_{i1}^* \ \tilde{D}_{i2} + F_{i2}^* \ \cdots \ \tilde{D}_{iq} + F_{iq}^* \right)' .
$$

Assuming that $\tilde{B} \sim (0_{2n}, \tilde{G})$ and $\tilde{F} \sim (0_{nq}, \tilde{R})$. The covariance matrix of $\tilde{B}$ is $\tilde{G} = I_n \otimes \begin{pmatrix} \sigma_s^2 & 0 \\ 0 & \tilde{\sigma}_s^2 \end{pmatrix}$; $\tilde{R}$ is the covariance matrix of $\tilde{F}$, and

$$
\tilde{R} = \bigoplus_{i=1}^n \left( \tilde{\sigma}_{iw}^2 I_q + \frac{1}{7} \sigma_{ik}^2 I_q \right) .
$$

Thus,

$$
\tilde{V} = \text{Var}(\tilde{Y})
= Z_w \tilde{G} Z_w' + \tilde{R}
= \left( \sigma_s^2 + \tilde{\sigma}_s^2 \right) \left( I_n \otimes I_q \right) + \bigoplus_{i=1}^n \left( \tilde{\sigma}_{iw}^2 I_q + \frac{1}{7} \sigma_{ik}^2 I_q \right).
$$

With the observed 7DDR data only, we cannot obtain the BLUE of $\mu$ based on the expression $\tilde{\mu} = \left( X_w' \tilde{V}^{-1} X_w \right)^{-1} X_w' \tilde{V}^{-1} \tilde{Y}$ since $\left( X_w' \tilde{V}^{-1} X_w \right)^{-1}$ is singular. However, this helps us to build a simultaneous model by combining 24HR and 7DDR data in Section 3.6.
3.5.2 7DDR Model under the FPMM

We use sampling indicator random variable $U^{(s)}_{nw}$ for selection of week $w$ in position $t$, where $U^{(s)}_{nw} = 1$ when the $t^{th}$ position (selected week) is week “$w$” and $U^{(s)}_{nw} = 0$ otherwise. We define the observed saturated fat intake of 7DDR measure for the $i^{th}$ subject in the $t^{th}$ week as $\tilde{Y}_{itk}$, a FPMM for $\tilde{Y}_{itk}$ can be written as

$$\tilde{Y}_{itk} = \mu + B^{*}_t + \tilde{B}^{*}_t + \left( D^{*}_t + W^{*}_t + F^{*}_{itk} \right), \quad (3.95)$$

where $i = 1, ..., N, t = 1, ..., W$. We define the random week effect

$$W^{*}_t = \sum_{s=1}^{N} U^{(s)}_{is} W_{st} \quad (3.96)$$

and

$$W_{st} = \sum_{w=1}^{W} U^{(s)}_{nw} \omega_{pw}. \quad (3.97)$$

We define the random measurement error

$$F^{*}_{itk} = \sum_{s=1}^{N} U^{(s)}_{is} F^{*}_{stk} \quad (3.98)$$

and

$$F^{*}_{stk} = \sum_{w=1}^{W} U^{(s)}_{nw} F^{*}_{stk}. \quad (3.99)$$

Similar to 24HR, when $k^{*} = 1$, one 7DDR recall is collected in one week and it is not possible to separate week-to-week variability and the random measurement error due to the 7DDR instrument in data analysis. We define the response error as

$$F^{**}_{it} = \sum_{s=1}^{N} U^{(s)}_{is} F^{*}_{st} \quad (3.100)$$
where $F_{st}^* = W_{st} + F_{stk}^*$, corresponding to the sum of week-to-week variability and the random measurement error. The FPMM model (3.95) can further be expressed as

$$\tilde{Y}_i = (\mu + \tilde{\mu})\left(\tilde{B}_i^* + \tilde{B}_t^*\right) + (\tilde{D}_a^* + F_{st}^*).$$  

(3.101)

Similar to the model (3.37), consistent with the FPMM defined in Stanek and Singer (2004), we define the random subject effect bias

$$\tilde{B}_i^* = \sum_{i=1}^{N} U_{is} \tilde{\beta}_s,$$  

(3.102)

and have $E_S (\tilde{B}_i^*) = 0$ and

$$Var_S (\tilde{B}_i^*) = \frac{N - 1}{N} \tilde{\sigma}_S^2,$$  

(3.103)

the subscript $S$ denotes the expectation with respect to the distribution of the subject effect bias $\tilde{B}_i^*$. For any $i \neq i^*$,

$$Cov_S (\tilde{B}_i^*, \tilde{B}_t^*) = -\frac{1}{N} \tilde{\sigma}_S^2.$$  

(3.104)

We also have $Cov_S (\tilde{B}_i^*, \tilde{B}_t^*) = 0$. Since $E_W (\tilde{D}_a) = 0$ and $Var_W (\tilde{D}_a) = \frac{W - 1}{W} \tilde{\sigma}_W^2$, for the random week bias $\tilde{D}_a^* = \sum_{i=1}^{N} U_{is} \tilde{D}_a$, we have $E_W (\tilde{D}_a^*) = 0$ and

$$Var_W (\tilde{D}_a^*) = \frac{W - 1}{W} \tilde{\sigma}_W^2,$$  

(3.105)

the subscript $W$ denotes the expectation with respect to the distribution of the week bias. For any $i \neq i^*$ or $t \neq t^*$,

$$Cov_W (\tilde{D}_a^*, \tilde{D}_{it}^*) = -\frac{1}{W} \tilde{\sigma}_W^2.$$  

(3.106)
Since \( E_R(F^*_{u}) = 0 \), from (3.85), \( Var_R(F^*_{u}) = \frac{1}{7} \left( \sigma_{sD}^2 + \sigma_{sE}^2 \right) - \frac{1}{D} \sigma_{sD}^2 \), and from (3.100), we have \( E_R(F^*_{it}) = 0 \) and

\[
Var_R(F^*_{it}) = \frac{1}{7} \left( \sigma_{sD}^2 + \sigma_{sE}^2 \right) - \frac{1}{D} \sigma_{sD}^2 .
\] (3.107)

For any \( i \neq i^* \) or \( t \neq t^* \),

\[
Cov_R(F^*_{it}, F^*_{it^*}) = -\frac{1}{D} \sigma_{sD}^2 .
\] (3.108)

Since model (3.7) is very similar to model (3.7) using the parameters in (3.6) given by

\[
Y^*_{sdk} = \mu + \beta_s + \delta_{sd} + E_{sdk^*} ,
\]

where model (3.70) is given by

\[
\tilde{Y}^*_{mk^*} = (\mu + \tilde{\mu}) + (\beta_s + \tilde{\beta}_s) + (\omega_s + \tilde{\delta}_s) + F_{mk^*} ,
\]
similar to the model (3.35) is expressed as (3.42), the model (3.95) can be expressed in matrix form for the population as

\[
\tilde{Y} = \tilde{M} + F^* ,
\] (3.109)

where \( F^* \) is the random measurement error vector, \( F^* = (U \otimes I_W) \left( \bigoplus_{s=1}^{N} U^{(s)} \right) F \),

\[
F = \begin{pmatrix} F'_1 & F'_2 & \ldots & F'_N \end{pmatrix}' \text{ and } F_2 = \begin{pmatrix} (F_{s1})' \end{pmatrix} = F_{s1} \ F_{s2} \ldots \ F_{sw1}' . \]

\( \tilde{Y} \) and \( \tilde{M} \) are \( NW \times 1 \) vectors. Let \( \tilde{\mu} = (\tilde{\mu}'_1 \ \tilde{\mu}'_2 \ \ldots \ \tilde{\mu}'_N)' \) where \( \tilde{\mu}_s = (\tilde{\mu}_s_1 \ \tilde{\mu}_s_2 \ \ldots \ \tilde{\mu}_sw)' \) and

\[
\tilde{\delta} = (\tilde{\delta}'_1 \ \tilde{\delta}'_2 \ \ldots \ \tilde{\delta}'_N)' \text{ where } \tilde{\delta}_s = (\tilde{\delta}_s_1 \ \tilde{\delta}_s_2 \ \ldots \ \tilde{\delta}_sw)' , \]

\( \tilde{M} \) is defined as

\[
\tilde{M} = (U \otimes I_W) \left( \bigoplus_{s=1}^{N} U^{(s)} \right) \tilde{\mu}
\] (3.110)

59
where \( \tilde{\mu} = \begin{pmatrix} \tilde{\mu}_1' \\ \tilde{\mu}_2' \\ \vdots \\ \tilde{\mu}_N' \end{pmatrix} = 1_{N \times W} (\mu + \tilde{\mu}) + \begin{pmatrix} \beta_1 + \tilde{\beta}_1 \\ \beta_2 + \tilde{\beta}_2 \\ \vdots \\ \beta_N + \tilde{\beta}_N \end{pmatrix} \otimes 1_w + \begin{pmatrix} \omega_1' + \tilde{\omega}_1' \\ \omega_2' + \tilde{\omega}_2' \\ \vdots \\ \omega_N' + \tilde{\omega}_N' \end{pmatrix}.

Similar to (3.43) and (3.44), for the expressions of the expected value and the variance of \( \tilde{Y} \). Using \( S \) to indicate expectation with respect to permutations of the subjects, using \( W \) to indicate expectation with respect to permutations of weeks of the subject \( s \), using \( M \) to indicate expectation with respect to measurement error,

\[
E_{SWM}(\tilde{Y}) = 1_{N \times W}\left( \mu + \tilde{\mu} \right)
\]

and

\[
Var_{SWM}(\tilde{Y}) = \left( \frac{1}{7} \sigma^2_S + \frac{1}{7} \sigma^2_D + \tilde{\sigma}^2_D \right) 1_{N \times W} + \left( \sigma^2_S + \tilde{\sigma}^2_D - \frac{1}{W} \left( \frac{1}{7} \sigma^2_D + \tilde{\sigma}^2_D \right) \right) (I_N \otimes J_W) - \frac{1}{N} \left( \sigma^2_S + \tilde{\sigma}^2_D \right) J_{N \times W}
\]

where \( \sigma^2_S \) is defined by (3.3), \( \sigma^2_D \) is defined by (3.5), \( \sigma^2_E \) is defined by (3.10), \( \tilde{\sigma}^2_D \) is defined by (3.69) and \( \tilde{\sigma}^2_S \) is defined by (3.65).

Similar to (3.46), when a sample is obtained, where \( i = 1, \ldots, n \), \( t = 1, \ldots, q \), \( \tilde{Y} \) can be expressed as

\[
\begin{pmatrix} \tilde{Y}_I \\ \tilde{Y}_H \end{pmatrix} = \begin{pmatrix} \tilde{K}^*_I \tilde{Y} \\ \tilde{K}^*_H \tilde{Y} \end{pmatrix}
\]

where \( \tilde{K}^*_I = (I_n \mid 0_{n \times (N-n)}) \otimes (I_q \mid 0_{q \times (W-q)}) \) and \( \tilde{K}^*_H = \left( \begin{pmatrix} I_n \mid 0_{n \times (N-n)} \end{pmatrix} \otimes \begin{pmatrix} 0_{(W-q) \times q} \mid I_{W-q} \end{pmatrix} \right)\).

Similar to (3.47), the variance of \( \tilde{Y}_I, \tilde{V}^* \) is
\[ \tilde{V}^* = \text{Var}(\tilde{Y}_i) \]
\[ = \left( \sigma_S^2 + \tilde{\sigma}_S^2 \right) \left( I_n - \frac{1}{N} J_n \right) \otimes J_q + \left( \frac{1}{7} \sigma_E^2 + \frac{1}{7} \tilde{\sigma}_D^2 + \tilde{\sigma}_W^2 \right) I_{nq} - \frac{1}{W} \left( \frac{1}{7} \sigma_D^2 + \tilde{\sigma}_W^2 \right) \left( I_n \otimes J_q \right) \, . \]

(3.114)

Based on (3.101), when \( i = 1, \ldots, n \), \( t = 1, \ldots, q \), the FPMM model for the sample can be expressed in matrix form as

\[ \tilde{Y}_i = X_{nl} \mu_c + Z_{nl} \tilde{B}^* + \tilde{F}_i \]  

(3.115)

where \( \tilde{Y}_i \) is an \( nq \times 1 \) column vector, \( \tilde{Y}_i = (\tilde{Y}_{i1} \tilde{Y}_{i2} \cdots \tilde{Y}_{in})' \) and \( \tilde{Y}_i = (\tilde{Y}_{i1} \tilde{Y}_{i2} \cdots \tilde{Y}_{in})' \); \( X_{nl} = 1_n \otimes \left( \begin{array}{c} 1_q \\ 1_q \end{array} \right) \), an \( nq \times 2 \) matrix with all elements equal to 1; \( \mu_c = \begin{pmatrix} \mu \\ \bar{\mu} \end{pmatrix} \); \( Z_{nl} = 1_n \otimes \left( \begin{array}{c} 1_q \\ 1_q \end{array} \right), \) an \( nq \times 2n \) block diagonal matrix with blocks \( \left( \begin{array}{c} 1_q \\ 1_q \end{array} \right); \) \( \tilde{B}^* \) is a vector of random effects, \( \tilde{B}^* = \left( \bigotimes_{i=1}^{n} \tilde{B}_i^* \right) 1_n \); \( \tilde{F}_i \) is an \( nq \times 1 \) column vector,

\[ \tilde{F}_i = (\tilde{F}_{i1} \tilde{F}_{i2} \cdots \tilde{F}_{in})' \]  

and \( \tilde{F}_i = (\tilde{D}_{i1} + F_{i1}^{**} \tilde{D}_{i2} + F_{i2}^{**} \cdots \tilde{D}_{in} + F_{in}^{**})' \). Assuming that \( \tilde{B}^* \sim (0_{2n}, \tilde{G}^*) \) and \( \tilde{F}_i \sim (0_{nq}, \tilde{R}^*) \). The covariance matrix of \( \tilde{B}^* \) is

\[ \tilde{G}^* = \left( I_n - \frac{1}{N} J_n \right) \otimes \begin{pmatrix} \sigma_S^2 & 0 \\ 0 & \tilde{\sigma}_S^2 \end{pmatrix} \]; \( \tilde{R}^* \) is the covariance matrix of \( \tilde{F}_i \), and

\[ \tilde{R}^* = \left( \frac{1}{7} \sigma_E^2 + \frac{1}{7} \sigma_D^2 + \tilde{\sigma}_W^2 \right) I_{nq} - \frac{1}{W} \left( \frac{1}{7} \sigma_D^2 + \tilde{\sigma}_W^2 \right) \left( I_n \otimes J_q \right) \, . \]

3.6 Simultaneous Model

We build a simultaneous model by combining 24HR and 7DDR data. It helps us to explore whether a combination of 24HR and 7DDR measures can obtain more accurate
predictor than repeated measures from a single instrument. It is easier to express the simultaneous model using the matrix format. Corresponding to the 24HR and 7DDR models under the usual mixed model and under the FPMM model, we define the simultaneous model under the usual mixed model and under the FPMM model separately.

3.6.1 Simultaneous Model under the Usual Mixed Model

Referring to (3.24) and (3.92), the simultaneous model under the usual mixed model can be written as

\[
Y_c = X_c \mu_c + Z_c \tilde{B} + E_c
\]

(3.116)

where \( X_c \) and \( Z_c \) are matrices, the remaining terms are vectors. \( Y_c \) is a vector of combined observed data, \( Y_c = \left( Y'_{c1} \quad Y'_{c2} \quad \cdots \quad Y'_{cn} \right)' \) and

\[
Y_{ci} = (Y_{i1} \quad Y_{i2} \quad \cdots \quad Y_{im} \quad \tilde{Y}_{i1} \quad \tilde{Y}_{i2} \quad \cdots \quad \tilde{Y}_{iq})'; \quad E_c \text{ is a vector including the response errors and the week biases. Each term is defined in (3.117),}
\]

\[
Y_c = X_c \mu_c + Z_c \tilde{B} + E_c
\]

\[
\begin{pmatrix}
Y'_{c1} \\
Y'_{c2} \\
\vdots \\
Y'_{cn}
\end{pmatrix} = \begin{pmatrix} 1_n \otimes \begin{pmatrix} 1_m \\
1_q \\
1_q \end{pmatrix} \end{pmatrix} \begin{pmatrix} \mu \\
\tilde{\mu} \end{pmatrix} + \begin{pmatrix} I_n \otimes \begin{pmatrix} 1_m \\
1_q \\
1_q \end{pmatrix} \end{pmatrix} \left( \bigoplus_{i=1}^{n} \begin{pmatrix} \tilde{B}_i \\
\tilde{B}_i \end{pmatrix} \right) \begin{pmatrix} 1_n \\
1_n \end{pmatrix} + \begin{pmatrix} E'_1 \\
\tilde{D}_1 + F'_1 \\
\vdots \\
E'_n \\
\tilde{D}_n + F'_n
\end{pmatrix}
\]

(3.117)
where vectors $Y_c$ and $E_c$ have $n(m+q)$ rows. $E_c^* = (E_{i1}^* \ E_{i2}^* \ \cdots \ E_{im}^*)'$ and

$$\tilde{D}_i + \tilde{F}_i^* = \left(\tilde{D}_{i1} + F_{i1}^* \ \tilde{D}_{i2} + F_{i2}^* \ \cdots \ \tilde{D}_{iq} + F_{iq}^*\right)'.$$

Matrix $X_c$ has $(n+q)$ rows and 2 columns, and matrix $Z_c$ has $(n+m)+q$ rows and 2 columns. Assuming that

$$\mathbf{\tilde{B}} \sim (0_{2n}, \mathbf{\tilde{G}}) \text{ and } E_c \sim (0_{n(m+q)}, \mathbf{R}_c).$$

The covariance matrix of $\mathbf{\tilde{B}}$ is $\mathbf{\tilde{G}} = I_n \otimes \left(\sigma_s^{-2} \ 0 \ 0 \ \sigma_s^{-2}\right);$ and

$$\mathbf{R}_c \text{ is the covariance matrix of } E_c \text{ and we assume } Cov\left(E_{ij}^*, \tilde{D}_{it} + F_{it}^*\right) = 0,$$

then

$$\mathbf{R}_c = \bigoplus_{i=1}^{n} \begin{pmatrix} \sigma_{ir}^{-2} \mathbf{I}_m & 0_{m \times m} \\ 0_{m \times m} & \left(\tilde{\sigma}_{ir}^{-2} + \frac{1}{7} \sigma_{ir}^{-2}\right) \mathbf{I}_q \end{pmatrix},$$

where $\sigma_{ir}^{-2}$ is given by (3.20) and $\tilde{\sigma}_{ir}^{-2}$ is given by (3.90). Thus,

$$\mathbf{V}_c = \text{Var}(Y_c) = \mathbf{Z}_c \tilde{G} \mathbf{Z}_c' + \mathbf{R}_c = \bigoplus_{i=1}^{n} \begin{pmatrix} \sigma_{ir}^{-2} \mathbf{J}_m + \sigma_{ir}^{-2} \mathbf{I}_m & \sigma_{ir}^{-2} \mathbf{J}_{m \times q} \\ \sigma_{ir}^{-2} \mathbf{J}_{q \times m} & (\sigma_{ir}^{-2} + \tilde{\sigma}_{ir}^{-2}) \mathbf{J}_q + (\tilde{\sigma}_{ir}^{-2} + \frac{1}{7} \sigma_{ir}^{-2}) \mathbf{I}_q \end{pmatrix}.$$  \hspace{1cm} (3.118)

The BLUE of $\mu_c$ is

$$\mathbf{\bar{\mu}}_c = (\mathbf{X}_c' \mathbf{V}_c^{-1} \mathbf{X}_c)^{-1} \mathbf{X}_c' \mathbf{V}_c^{-1} \mathbf{Y}_c,$$ \hspace{1cm} (3.119)

where $\text{Var}(\mathbf{\bar{\mu}}_c) = (\mathbf{X}_c' \mathbf{V}_c^{-1} \mathbf{X}_c)^{-1}$. The BLUP of $\mathbf{\tilde{B}}$ is

$$\mathbf{\bar{B}} = \tilde{G} \mathbf{Z}_c' \mathbf{V}_c^{-1} \left(\mathbf{Y}_c - \mathbf{X}_c (\mathbf{X}_c' \mathbf{V}_c^{-1} \mathbf{X}_c)^{-1} \mathbf{X}_c' \mathbf{V}_c^{-1} \mathbf{Y}_c\right)$$

$$= \tilde{G} \mathbf{Z}_c' \mathbf{V}_c^{-1} (\mathbf{Y}_c - \mathbf{X}_c \mathbf{\bar{\mu}}_c).$$ \hspace{1cm} (3.120)

Since $\mathbf{\bar{\mu}}_c = \left(\begin{array}{c} \mathbf{\bar{\mu}}_1 \\ \mathbf{\bar{\mu}}_2 \\ \vdots \\ \mathbf{\bar{\mu}}_n \end{array}\right)$ and $\mathbf{\bar{B}} = \left(\begin{array}{c} \mathbf{\bar{B}}_1 \\ \mathbf{\bar{B}}_2 \\ \vdots \\ \mathbf{\bar{B}}_n \end{array}\right),$ we obtain the estimate of subject’s latent value by

the predictor

$$\hat{P}_{c2} = \mathbf{\bar{\mu}} + \mathbf{\bar{B}}_i.$$ \hspace{1cm} (3.121)
When the estimated variance components are used, the estimates of \( \hat{\sigma}_s^2 \), \( \hat{\sigma}_s^2 \), \( \hat{\sigma}_{ir}^2 \) and \( \hat{\sigma}_{iw}^2 \) are obtained from the observed data and the details are described in chapters 4 and 7.

After knowing \( \hat{\sigma}_s^2 \), \( \hat{\sigma}_s^2 \), \( \hat{\sigma}_{ir}^2 \) and \( \hat{\sigma}_{iw}^2 \), we replace \( V_c \) and \( G \) by \( \hat{V}_c \) and \( \hat{G} \), the empirical BLUE of \( \mu_c \) is

\[
\hat{\mu}_c = \left( X_c' \hat{V}_c^{-1} X_c \right)^{-1} X_c' \hat{V}_c^{-1} Y_c,
\]

where \( \text{Var}(\hat{\mu}_c) = \left( X_c' \hat{V}_c^{-1} X_c \right)^{-1} \). The empirical BLUP of \( \hat{B} \) is

\[
\hat{B} = \hat{G}Z_c' \hat{V}_c^{-1} \left( Y_c - X_c \left( X_c' \hat{V}_c^{-1} X_c \right)^{-1} X_c' \hat{V}_c^{-1} Y_c \right) = \hat{G}Z_c' \hat{V}_c^{-1} \left( Y_c - X_c \hat{\mu}_c \right).
\]

Since \( \hat{\mu}_c = \begin{pmatrix} \hat{\mu} \\ \hat{\mu} \end{pmatrix} \) and \( \hat{B} = \begin{pmatrix} \hat{B}_1 \\ \vdots \\ \hat{B}_n \end{pmatrix} \), we obtain the empirical estimate of subject’s latent value by the empirical predictor

\[
\hat{\mu}_{c2e} = \hat{\mu} + \hat{B}_c.
\]

### 3.6.2 Simultaneous Model under the FPMM

Referring to (3.48) and (3.115), the simultaneous model under the FPMM can be written as

\[
Y_{el} = X_{el} \mu_c + Z_{el} \tilde{B}^* + E_{el},
\]

where \( X_{el} \) and \( Z_{el} \) are matrices, the remaining terms are vectors; \( Y_{el} \) is a vector of combined observed data, \( Y_{el} = \left( Y_{e1}' \quad Y_{e2}' \quad \cdots \quad Y_{en}' \right)' \) and
\( \mathbf{Y}_{cl} = \left( Y_{i_1} \ Y_{i_2} \ \cdots \ Y_{i_m} \ \tilde{Y}_{i_1} \ \tilde{Y}_{i_2} \ \cdots \ \tilde{Y}_{i_q} \right)' \); \( \mathbf{E}_{cl} \) is a vector including the response errors and the week biases. Each term is defined in (3.126),

\[
\begin{align*}
\left( \begin{array}{c}
\mathbf{Y}_{cl}^1 \\
\mathbf{Y}_{cl}^2 \\
\vdots \\
\mathbf{Y}_{cl}^n
\end{array} \right) = (I_n \otimes \begin{pmatrix}
1_m & 0_m \\
1_q & 1_q
\end{pmatrix}) \left( \mu \otimes \begin{pmatrix}
1_m & 0_m \\
1_q & 1_q
\end{pmatrix}\right) + \begin{pmatrix}
\mathbf{B}_i^* \\
\mathbf{E}_{cl}^*
\end{pmatrix} + \\
\begin{pmatrix}
\mathbf{E}_{i_1}^* \\
\mathbf{E}_{i_2}^* \\
\vdots \\
\mathbf{E}_{i_q}^*
\end{pmatrix}
\end{align*}
\]

(3.126)

where \( \mathbf{E}_{i_l}^* = (E_{i_1}^* \ E_{i_2}^* \ \cdots \ E_{i_m}^*)' \) and \( \mathbf{B}_i^* + \mathbf{F}_i^* = (\tilde{B}_i^* + F_{i_1}^* \ \tilde{B}_i^* + F_{i_2}^* \ \cdots \ \tilde{B}_i^* + F_{i_q}^*)' \).

Assuming that \( \mathbf{B}_i^* \sim (\mathbf{0}_{2n^*}, \mathbf{G}_i^* ) \) and \( \mathbf{E}_{cl} \sim (\mathbf{0}_{(m+q)}, \mathbf{R}_c^* ) \). The covariance matrix of \( \mathbf{B}_i^* \) is

\[
\mathbf{G}_i^* = \begin{pmatrix}
I_n - \frac{1}{N} \mathbf{J}_n \\
\sigma_D^2 & 0 \\
0 & \sigma_S^2
\end{pmatrix} ; \quad \mathbf{R}_c^* \text{ is the covariance matrix of } \mathbf{E}_{cl} \text{ and we assume}
\]

\[
\text{Cov}(E_{ij}^*, \tilde{D}_i^* + F_{ij}^*) = 0 , \text{ then}
\]

\[
\mathbf{R}_c^* = I_n \otimes \begin{pmatrix}
\sigma_D^2 + \sigma_E^2 \\
\frac{1}{D} \sigma_D^2 \mathbf{I}_m - \frac{1}{D} \sigma_D^2 \mathbf{I}_m \\
\frac{1}{7} \left( \sigma_D^2 + \sigma_E^2 \right) + \sigma_W^2 \\
\frac{1}{7} \left( \sigma_D^2 + \sigma_E^2 \right) + \sigma_W^2
\end{pmatrix} . \text{ Thus,}
\]

\[
\mathbf{V}_c^* = \text{Var}(\mathbf{Y}_{cl}) \\
= \mathbf{Z}_{cl} \mathbf{G}_c^* \mathbf{Z}_{cl} + \mathbf{R}_c^* 
\]

(3.127)

The BLUE of \( \mu_c \) is

\[
\hat{\mu}_c = \left( \mathbf{X}_{cl} \mathbf{V}_c^{-1} \mathbf{X}_{cl} \right)^{-1} \mathbf{X}_{cl} \mathbf{V}_c^{-1} \mathbf{Y}_{cl} . 
\]

(3.128)

where \( \text{Var}(\hat{\mu}_c) = \left( \mathbf{X}_{cl} \mathbf{V}_c^{-1} \mathbf{X}_{cl} \right)^{-1} \). The BLUP of \( \mathbf{B}_i^* \) is
\[
\tilde{\mathbf{B}}^* = \mathbf{G}^* \mathbf{Z}' \mathbf{V}^{-1} \left( \mathbf{Y}_{cl} - \mathbf{X}_{cl} \left( \mathbf{X}' \mathbf{V}^{-1} \mathbf{X}_{cl} \right)^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{Y}_{cl} \right),
\]

(3.129)

\[
= \mathbf{G}^* \mathbf{Z}' \mathbf{V}^{-1} \left( \mathbf{Y}_{cl} - \mathbf{X}_{cl} \mathbf{\mu}_c \right)
\]

Since \( \mathbf{\mu}_c = \left( \bar{\mu} \right) \) and \( \tilde{\mathbf{B}}^* = \left( \oplus_{i=1}^{n} \left( \mathbf{\hat{B}}_{i}^* \right) \right) \mathbf{1}_n \), we obtain the estimate of subject’s latent value by the predictor

\[
\hat{P}_{ci}^{(i)} = \bar{\mu} + \tilde{\mathbf{B}}^*_i.
\]

(3.130)

When the estimated variance components are used, the estimates of \( \hat{\sigma}_s^2 \), \( \hat{\sigma}_d^2 \), \( \hat{\sigma}_e^2 \) and \( \hat{\sigma}_w^2 \) are obtained from the observed data and the details are described in chapters 4 and 7.

After knowing \( \hat{\sigma}_s^2 \), \( \hat{\sigma}_d^2 \), \( \hat{\sigma}_e^2 \) and \( \hat{\sigma}_w^2 \), we replace \( \mathbf{V}_c^* \) and \( \mathbf{G}^* \) by \( \mathbf{\hat{V}}_c^* \) and \( \mathbf{\hat{G}}^* \), the empirical BLUE of \( \mathbf{\mu}_c \) is

\[
\hat{\mathbf{\mu}}_c = \mathbf{X}' \mathbf{\hat{V}}_c^{-1} \mathbf{X}_{cl} \mathbf{\hat{V}}_c^{-1} \mathbf{Y}_{cl},
\]

(3.131)

where \( \text{Var}(\hat{\mathbf{\mu}}_c) = \left( \mathbf{X}' \mathbf{\hat{V}}_c^{-1} \mathbf{X}_{cl} \right)^{-1} \). The empirical BLUP of \( \tilde{\mathbf{B}}^* \) is

\[
\hat{\tilde{\mathbf{B}}}^* = \mathbf{G}^* \mathbf{Z}' \mathbf{\hat{V}}_c^{-1} \left( \mathbf{Y}_{cl} - \mathbf{X}_{cl} \left( \mathbf{X}' \mathbf{\hat{V}}_c^{-1} \mathbf{X}_{cl} \right)^{-1} \mathbf{X}' \mathbf{\hat{V}}_c^{-1} \mathbf{Y}_{cl} \right),
\]

(3.132)

\[
= \mathbf{G}^* \mathbf{Z}' \mathbf{\hat{V}}_c^{-1} \left( \mathbf{Y}_{cl} - \mathbf{X}_{cl} \mathbf{\hat{\mu}}_c \right)
\]

Since \( \mathbf{\hat{\mu}}_c = \left( \hat{\mu} \right) \) and \( \hat{\tilde{\mathbf{B}}}^* = \left( \oplus_{i=1}^{n} \left( \mathbf{\hat{B}}_{i}^* \right) \right) \mathbf{1}_n \), we obtain the empirical estimate of subject’s latent value by the empirical predictor

\[
\hat{P}_{ci}^{(i)} = \bar{\mu} + \hat{\tilde{\mathbf{B}}}^*_i.
\]

(3.133)
CHAPTER 4

ESTIMATOR/EMPIRICAL PREDICTORS OF AN INDIVIDUAL’S DIETARY INTAKE LATENT VALUE IN THE SEASONS STUDY

This chapter illustrates the analysis of the estimator/empirical predictor of subject’s latent value of saturated fat intake through an example of 127 subjects in the Seasons study. We consider the 127 subjects to be a sample of a finite population of size \( N \), i.e. \( n = 127 \). We use simple random sampling without replacement method to select 127 subjects from the \( N \) subjects. When we permute the subject IDs of the \( N \) subjects, there are \( N! \) permutations. We randomly choose one permutation from the \( N! \) permutations, and define the 127 subjects occupying the first 127 positions of the selected permutation as a sample of size \( n = 127 \).

The data from the 127 subjects are balanced with 12 24HR recalls and 4 7DDR recalls for each subject, i.e. \( m = 12 \), \( q = 4 \). Using the 24HR data, we compute the estimator \( \hat{P}_1^{(i)} \) (ME) and the two empirical predictors \( \hat{P}_{2e}^{(i)} \) (MM) and \( \hat{P}_{3e}^{(i)} \) (FPMM). The estimator \( \hat{P}_1^{(i)} \) (ME) is defined by (3.17) after replacing \( s \) by \( i \), where

\[
\hat{P}_1^{(i)} = \bar{Y}_i,
\]

an average of the observed responses for the \( i^{th} \) randomly selected subject. The empirical predictor \( \hat{P}_{2e}^{(i)} \) (MM) is derived under the usual mixed model and defined by (3.23), where

\[
\hat{P}_{2e}^{(i)} = \hat{\mu}^* + \hat{k}_i \left( \bar{Y}_i - \hat{\mu}^* \right).
\]
It is computed by using the weighted least squares estimate and the subject-specific response error variance. The empirical predictor $\hat{\mathcal{P}}_{3e}^{(i)}$ (FPMM) is derived under the FPMM and defined by (3.60), where

$$\hat{\mathcal{P}}_{3e}^{(i)} = \bar{Y} + \hat{k}(\bar{Y}_i - \bar{Y}).$$

It is computed by using the average response error variance over subjects in the population. Using the combined 24HR and 7DDR data, we compute the two empirical predictors $\hat{\mathcal{P}}_{2e}^{(i)}$ (MM) and $\hat{\mathcal{P}}_{3e}^{(i)}$ (FPMM). The empirical predictor $\hat{\mathcal{P}}_{2e}^{(i)}$ (MM) is derived under the usual mixed model and defined by (3.124), where

$$\hat{\mathcal{P}}_{2e}^{(i)} = \hat{\mu} + \hat{B}_i,$$

the $\hat{\mu}$ and $\hat{B}_i$ are obtained from (3.122) and (3.123). The empirical predictor $\hat{\mathcal{P}}_{3e}^{(i)}$ (FPMM) is derived under the FPMM and defined by (3.133), where

$$\hat{\mathcal{P}}_{3e}^{(i)} = \hat{\mu} + \hat{B}_i^*,$$

the $\hat{\mu}$ and $\hat{B}_i^*$ are obtained from (3.131) and (3.132).

Since the variance parameters are unknown, we need to obtain the estimated variance components in order to compute the empirical predictors of subject’s latent value. We estimate variance components using the 24HR data and the combined 24HR and 7DDR data, respectively. Since the dietary intake data are distribution free in the finite population setting and the first two central moments exist, we use method of moments to estimate the variance parameters $\sigma_s^2$, $\sigma_k^2$, $\tilde{\sigma}_s^2$, $\tilde{\sigma}_w^2$ through a one way ANOVA table by equating the observed mean squares to the expected mean squares.
4.1 Estimating $\sigma_S^2$, $\sigma_R^2$ and $\sigma_{ir}^2$ Using the 24HR Data

We estimate $\sigma_S^2$ and $\sigma_R^2$ using the 24HR data from the $n = 127$ subjects. Each subject has 12 repeated measures, $m = 12$. From the one-way ANOVA table, the observed mean squares between subjects (MSB) is

$$MSB = \frac{1}{n-1} \sum_{i=1}^{n} m (\bar{y}_i - \bar{y})^2,$$  \hspace{1cm} (4.1)

and the observed mean squared error (MSR, “R” indicates response error) is

$$MSR = \frac{1}{n} \sum_{i=1}^{n} SS_i^2$$  \hspace{1cm} (4.2)

where

$$SS_i^2 = \frac{1}{m-1} \sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2.$$  \hspace{1cm} (4.3)

The expected mean squares are the same as derived by Stanek and Singer (2004) and San Martino et al. (2008). Under the usual mixed model, the expected value of MSB is

$$E(\text{MSB}) = m\sigma_S^2 + \frac{1}{n} \sum_{i=1}^{n} \sigma_{ir}^2,$$  \hspace{1cm} (4.4)

and the expected value of MSR is

$$E(\text{MSR}) = \frac{1}{n} \sum_{i=1}^{n} \sigma_{ir}^2.$$  \hspace{1cm} (4.5)

Under the FPMM, the expected value of MSB is

$$E(\text{MSB}) = m\sigma_S^2 + \sigma_R^2,$$  \hspace{1cm} (4.6)

and the expected value of MSR is

$$E(\text{MSR}) = \sigma_R^2.$$  \hspace{1cm} (4.7)

Therefore, we have
\[ \hat{\sigma}_R^2 = MSR, \]  
\[ (4.8) \]

\[ \hat{\sigma}_S^2 = \max \left( 0, \frac{MSB - MSR}{m} \right), \]  
\[ (4.9) \]

and

\[ \hat{\sigma}_{ir}^2 = SS_i^2. \]  
\[ (4.10) \]

We obtain the estimates of \( \hat{\sigma}_S^2 \) and \( \hat{\sigma}_R^2 \) shown in Table 4.1.

**Table 4.1 Estimates of the estimated variance parameters using the 24HR data**

<table>
<thead>
<tr>
<th>Estimated variance parameter</th>
<th>Using 24HR data</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\sigma}_S^2 )</td>
<td>118.85</td>
</tr>
<tr>
<td>( \hat{\sigma}_R^2 )</td>
<td>142.57</td>
</tr>
</tbody>
</table>

In Table 4.2, subject \( i = 1 \) and subject ID \( s = 599 \) indicate the first selected subject is the subject \( \lambda_{599} \) in the population. From (3.23), we need to estimate \( \hat{\sigma}_{ir}^2 \) to compute the empirical predictor \( \hat{P}_{2e}^{(i)} \) (MM) under the usual mixed model. For each selected subject \( i \), we estimate \( \hat{\sigma}_{ir}^2 \) by \( SS_i^2 \) where \( SS_i^2 = \frac{1}{m-1} \sum_{j=1}^{m} \left( Y_{ij} - \bar{Y}_i \right)^2 \), \( m = 12 \). The estimates of \( \hat{\sigma}_{ir}^2 \) are shown in Table 4.2.
Table 4.2 Estimates of $\hat{\sigma}_{ir}^2$ for the 127 subjects using the 24HR data ($m = 12$)

<table>
<thead>
<tr>
<th>$i$</th>
<th>Subject ID</th>
<th>$\bar{Y}_i$</th>
<th>$\hat{\sigma}_{ir}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>599</td>
<td>19.29</td>
<td>151.12</td>
</tr>
<tr>
<td>2</td>
<td>775</td>
<td>31.59</td>
<td>175.05</td>
</tr>
<tr>
<td>3</td>
<td>737</td>
<td>50.31</td>
<td>217.10</td>
</tr>
<tr>
<td>4</td>
<td>450</td>
<td>27.15</td>
<td>236.50</td>
</tr>
<tr>
<td>5</td>
<td>447</td>
<td>33.73</td>
<td>198.55</td>
</tr>
<tr>
<td>6</td>
<td>767</td>
<td>18.36</td>
<td>63.20</td>
</tr>
<tr>
<td>7</td>
<td>350</td>
<td>17.32</td>
<td>111.11</td>
</tr>
<tr>
<td>8</td>
<td>170</td>
<td>10.04</td>
<td>29.54</td>
</tr>
<tr>
<td>9</td>
<td>417</td>
<td>19.70</td>
<td>97.69</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>127</td>
<td>1022</td>
<td>24.41</td>
<td>180.90</td>
</tr>
</tbody>
</table>

4.2 Estimating $\sigma_s^2$, $\sigma_R^2$, $\sigma_z^2$ and $\hat{\sigma}_{iw}^2$ Using the 24HR+7DDR Data

We estimate $\sigma_s^2$, $\sigma_R^2$, $\sigma_z^2$ and $\hat{\sigma}_{iw}^2$ using the 24HR and 7DDR data from the 127 subjects. Each subject has 16 repeated measures, $m = 12$, $q = 4$. Table 4.3 shows the variance and covariance matrix of the 16 repeated measures of 127 subjects.

Table 4.3 Variance and covariance matrix of 16 repeated measures ($n = 127$)

<table>
<thead>
<tr>
<th></th>
<th>24.1</th>
<th>24.2</th>
<th>24.3</th>
<th>24.4</th>
<th>24.5</th>
<th>24.6</th>
<th>24.7</th>
<th>24.8</th>
<th>24.9</th>
<th>24.10</th>
<th>24.11</th>
<th>24.12</th>
<th>7.1</th>
<th>7.2</th>
<th>7.3</th>
<th>7.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.1</td>
<td>222.2</td>
<td>112.3</td>
<td>55.3</td>
<td>87.4</td>
<td>119.2</td>
<td>107.2</td>
<td>137.8</td>
<td>150.1</td>
<td>121.9</td>
<td>100.2</td>
<td>100.2</td>
<td>72.1</td>
<td>116.5</td>
<td>106.2</td>
<td>74.0</td>
<td>76.7</td>
</tr>
<tr>
<td>24.2</td>
<td>186.0</td>
<td>61.1</td>
<td>88.0</td>
<td>117.5</td>
<td>106.8</td>
<td>96.1</td>
<td>77.6</td>
<td>98.2</td>
<td>79.2</td>
<td>83.9</td>
<td>88.1</td>
<td>102.8</td>
<td>105.7</td>
<td>58.0</td>
<td>85.5</td>
<td></td>
</tr>
<tr>
<td>24.3</td>
<td>222.8</td>
<td>65.5</td>
<td>90.8</td>
<td>93.4</td>
<td>84.1</td>
<td>74.2</td>
<td>88.3</td>
<td>68.1</td>
<td>57.5</td>
<td>69.3</td>
<td>77.3</td>
<td>48.9</td>
<td>39.2</td>
<td>60.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.4</td>
<td>216.9</td>
<td>124.4</td>
<td>110.1</td>
<td>112.8</td>
<td>106.6</td>
<td>112.4</td>
<td>99.1</td>
<td>108.6</td>
<td>92.7</td>
<td>97.0</td>
<td>66.4</td>
<td>100.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.5</td>
<td>329.4</td>
<td>202.6</td>
<td>140.8</td>
<td>187.2</td>
<td>169.1</td>
<td>168.9</td>
<td>147.5</td>
<td>171.2</td>
<td>125.2</td>
<td>140.2</td>
<td>119.9</td>
<td>152.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.6</td>
<td>278.3</td>
<td>143.3</td>
<td>186.8</td>
<td>177.7</td>
<td>147.8</td>
<td>151.2</td>
<td>162.5</td>
<td>100.9</td>
<td>132.8</td>
<td>132.3</td>
<td>131.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.7</td>
<td>278.9</td>
<td>156.1</td>
<td>135.9</td>
<td>128.4</td>
<td>126.3</td>
<td>129.8</td>
<td>131.0</td>
<td>158.5</td>
<td>114.5</td>
<td>117.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.8</td>
<td>305.5</td>
<td>157.7</td>
<td>149.8</td>
<td>143.6</td>
<td>132.3</td>
<td>117.7</td>
<td>137.6</td>
<td>137.6</td>
<td>132.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.9</td>
<td>309.3</td>
<td>126.5</td>
<td>116.5</td>
<td>153.0</td>
<td>91.7</td>
<td>104.5</td>
<td>117.6</td>
<td>115.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.10</td>
<td>227.4</td>
<td>146.4</td>
<td>141.3</td>
<td>99.1</td>
<td>114.8</td>
<td>99.8</td>
<td>134.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.11</td>
<td>270.0</td>
<td>130.0</td>
<td>101.8</td>
<td>114.8</td>
<td>112.3</td>
<td>143.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.12</td>
<td>288.2</td>
<td>89.7</td>
<td>124.8</td>
<td>130.2</td>
<td>139.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1</td>
<td>175.3</td>
<td>146.3</td>
<td>106.0</td>
<td>111.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.2</td>
<td>252.0</td>
<td>146.5</td>
<td>148.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.3</td>
<td>226.1</td>
<td>139.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.4</td>
<td>226.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Since we assume the population size $N$, day $D$ and week $W$ are large, $\frac{1}{N} \equiv 0$, $\frac{1}{D} \equiv 0$ and $\frac{1}{W} \equiv 0$. Based on (3.47) and (3.114), Table 4.4 shows the structure of parametric variance and covariance matrix of the 16 repeated measures.

Table 4.4 Structure of parametric variance and covariance matrix of the 16 repeated measures

<table>
<thead>
<tr>
<th></th>
<th>$24_{1}$</th>
<th>$24_{12}$</th>
<th>$24_{12}$</th>
<th>$7_{1}$</th>
<th>$7_{4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$24_{1}$</td>
<td>$\sigma^2_s + \hat{\sigma}^2_s$</td>
<td>$\hat{\sigma}^2_s$</td>
<td>$\hat{\sigma}^2_s$</td>
<td>$\sigma^2_s$</td>
<td>$\sigma^2_s$</td>
</tr>
<tr>
<td>$24_{12}$</td>
<td>$\hat{\sigma}^2_s$</td>
<td>$\hat{\sigma}^2_s + \hat{\sigma}^2_s$</td>
<td>$\hat{\sigma}^2_s + \hat{\sigma}^2_s$</td>
<td>$\hat{\sigma}^2_s + \hat{\sigma}^2_s$</td>
<td>$\hat{\sigma}^2_s + \hat{\sigma}^2_s$</td>
</tr>
<tr>
<td>$7_{1}$</td>
<td>$(\hat{\sigma}^2_s + \hat{\sigma}^2_s) + \frac{1}{7} \hat{\sigma}^2_s + \hat{\sigma}^2_w$</td>
<td>$(\hat{\sigma}^2_s + \hat{\sigma}^2_s)$</td>
<td>$(\hat{\sigma}^2_s + \hat{\sigma}^2_s) + \frac{1}{7} \hat{\sigma}^2_s + \hat{\sigma}^2_w$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$7_{4}$</td>
<td>$(\hat{\sigma}^2_s + \hat{\sigma}^2_s)$</td>
<td>$(\hat{\sigma}^2_s + \hat{\sigma}^2_s)$</td>
<td>$(\hat{\sigma}^2_s + \hat{\sigma}^2_s)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Using method of moments by equating the observed variance and covariance matrix in Table 4.3 to the parametric variance and covariance matrix in Table 4.4, we solve the equations simultaneously and obtain the estimates of $\hat{\sigma}^2_s$, $\hat{\sigma}^2_r$, $\hat{\sigma}^2_s$ and $\hat{\sigma}^2_w$ shown in Table 4.5.

Table 4.5 Estimates of the estimated variance parameters using the 24HR+7DDR data

<table>
<thead>
<tr>
<th>Estimated variance parameter</th>
<th>Using 24HR +7DDR data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\sigma}^2_s$</td>
</tr>
<tr>
<td>Estimate</td>
<td>114.64</td>
</tr>
</tbody>
</table>

From (3.118), we need to estimate $\hat{\sigma}^2_{ir}$ and $\hat{\sigma}^2_{iw}$ in order to compute the empirical predictor $\hat{P}^{(i)}_{e2e}$ (MM) under the usual mixed model when using the combined 24HR and 7DDR data. For each selected subject $i$, we estimate $\left(\frac{\hat{\sigma}^2_{iw} + \frac{1}{7} \hat{\sigma}^2_{ir}}{SS^2_i}\right)$ by $SS^2_i$ where
\[ SS_i^{*2} = \frac{1}{q-1} \sum_{t=1}^{q} (Y_{it} - \bar{Y}_i)^2, \quad q = 4 \quad \text{based on the 7DDR data.} \] 

Since we already obtained \( \hat{\sigma}_{ir}^2 \) \( (\hat{\sigma}_{ir}^2 = SS^2) \) based on the 24HR data, we have \( \hat{\sigma}_{iw}^2 = \max \left( 0, SS_i^{*2} - \frac{1}{7} \hat{\sigma}_{ir}^2 \right) \), i.e. for those subjects with \( \hat{\sigma}_{iw}^2 < 0 \), we replace the \( \hat{\sigma}_{iw}^2 \) by zero. The adjustment results in zero week bias for these subjects and keeps parameters in a valid range. The estimates of \( \hat{\sigma}_{iw}^2 \) are shown in Table 4.6.

Table 4.6 Estimates of \( \hat{\sigma}_{ir}^2, \hat{\sigma}_{iw}^2 \) using the 24HR+7DDR data

<table>
<thead>
<tr>
<th>( i )</th>
<th>( s )</th>
<th>Using 24HR+7DDR data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>599</td>
<td>151.12</td>
</tr>
<tr>
<td>2</td>
<td>775</td>
<td>175.05</td>
</tr>
<tr>
<td>3</td>
<td>737</td>
<td>217.10</td>
</tr>
<tr>
<td>4</td>
<td>450</td>
<td>236.50</td>
</tr>
<tr>
<td>5</td>
<td>447</td>
<td>198.55</td>
</tr>
<tr>
<td>6</td>
<td>767</td>
<td>63.20</td>
</tr>
<tr>
<td>7</td>
<td>350</td>
<td>111.11</td>
</tr>
<tr>
<td>8</td>
<td>170</td>
<td>29.54</td>
</tr>
<tr>
<td>9</td>
<td>417</td>
<td>97.69</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>127</td>
<td>1022</td>
<td>180.90</td>
</tr>
</tbody>
</table>

4.3 Estimator/Empirical Predictor of Subject’s Latent Value

With the estimated variance components in Sections 4.1 and 4.2, corresponding to the formulae of the estimator/empirical predictor: (3.17), (3.23), (3.60), (3.124) and (3.133), and assuming \( \frac{1}{D} \equiv 0 \) and \( \frac{1}{W} \equiv 0 \), we compute \( \hat{P}_{1}^{(i)} \) (ME), \( \hat{P}_{2e}^{(i)} \) (MM), \( \hat{P}_{3e}^{(i)} \) (FPM), \( \hat{P}_{e2e}^{(i)} \) (MM) and \( \hat{P}_{e3e}^{(i)} \) (FPMM), respectively. The empirical predictors \( \hat{P}_{e2e}^{(i)} \) (MM) and \( \hat{P}_{e3e}^{(i)} \) (FPMM) are computed using SAS PROC IML procedure.
Table 4.7 shows the estimates of the five different estimator/empirical predictors for each selected subject. When using the 24HR data or using the 24HR+7DDR data, we see that no estimator/empirical predictor is always greater or smaller than the others across subjects. As an example, for subject $i = 3$ or $i = 4$, the estimates of the subject’s latent value vary based on different estimator/empirical predictor. This raises the question as to which one is more accurate. It is important to know, especially when the estimates are around a certain critical value. Based on the Seasons study, we cannot tell which estimator/empirical predictor is more accurate. Since we do not know the subject’s true value of saturated fat intake, we cannot evaluate the performance of the estimator/empirical predictor by comparing their MSEs. Hence, in the subsequent chapters, we conduct a simulation study to evaluate the performance of these estimator/empirical predictors.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$s$</th>
<th>Using 24HR data</th>
<th></th>
<th></th>
<th>Using 24HR+7DDR data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\hat{P}^{(0)}_{1}$ (ME)</td>
<td>$\hat{P}^{(0)}_{2}$ (MM)</td>
<td>$\hat{P}^{(0)}_{3}$ (FPMM)</td>
<td>$\hat{P}^{(0)}_{4}$ (MM)</td>
</tr>
<tr>
<td>1</td>
<td>599</td>
<td>19.29</td>
<td>19.78</td>
<td>19.81</td>
<td>17.87</td>
</tr>
<tr>
<td>2</td>
<td>775</td>
<td>31.59</td>
<td>30.80</td>
<td>30.99</td>
<td>33.14</td>
</tr>
<tr>
<td>3</td>
<td>737</td>
<td>50.31</td>
<td>46.88</td>
<td>48.01</td>
<td>46.35</td>
</tr>
<tr>
<td>4</td>
<td>450</td>
<td>27.15</td>
<td>26.75</td>
<td>26.95</td>
<td>30.39</td>
</tr>
<tr>
<td>5</td>
<td>447</td>
<td>33.73</td>
<td>32.58</td>
<td>32.93</td>
<td>31.33</td>
</tr>
<tr>
<td>6</td>
<td>767</td>
<td>18.36</td>
<td>18.61</td>
<td>18.96</td>
<td>19.51</td>
</tr>
<tr>
<td>7</td>
<td>350</td>
<td>17.32</td>
<td>17.83</td>
<td>18.01</td>
<td>18.19</td>
</tr>
<tr>
<td>8</td>
<td>170</td>
<td>10.04</td>
<td>10.33</td>
<td>11.39</td>
<td>10.14</td>
</tr>
<tr>
<td>9</td>
<td>417</td>
<td>19.70</td>
<td>20.00</td>
<td>20.18</td>
<td>19.44</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>127</td>
<td>1022</td>
<td>24.41</td>
<td>24.40</td>
<td>24.46</td>
<td>23.10</td>
</tr>
</tbody>
</table>
CHAPTER 5
SIMULATION STUDY

A simulation study is conducted to evaluate the performance of different estimator/predictor of subject’s latent value. The goal is to find the optimal estimator/predictor by comparing the mean squared error (MSE) of the estimator/predictor under different models. The simulation is conducted in three steps: i) determining subject’s parameters; ii) generating 24HR and 7DDR data for each selected subject; iii) analyzing simulated data. This chapter describes the steps i) and ii), and the chapters 6 and 7 describe the step iii) for the simulated 24HR data and for the simulated 24HR+7DDR data, respectively.

5.1 Determining Subject’s Parameters

We make use of an empirical method to determine subject’s parameters by mimicking the Seasons Study data. We investigate the 641 subjects in the basic dataset of the Seasons study. After excluding those subjects with less than 12 24HR recalls and less than 4 7DDR recalls, we obtain a population of size \( N = 444 \). Table 5.1 shows the frequencies of 24HR recalls and 7DDR recalls of the 444 subjects.

<table>
<thead>
<tr>
<th>7DDR recall #</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>11</td>
<td>3</td>
<td>11</td>
<td>32</td>
<td>57</td>
</tr>
<tr>
<td>5</td>
<td>29</td>
<td>37</td>
<td>66</td>
<td>255</td>
<td>387</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>40</td>
<td>77</td>
<td>287</td>
<td>444</td>
</tr>
</tbody>
</table>

Table 5.1 Frequency of 24HR recalls and 7DDR recalls by recall occasions

Figure 5.1 presents the distribution of subject’s average observed saturated fat intake by group of 24HR recall occasions, and the distribution of subject’s standard
deviation of response error by group of 24HR recall occasions. Correspondingly, Figure 5.2 presents the cases of 7DDR recall.

Figure 5.1 Distributions of subject’s mean and standard deviation by 24HR recall occasions

Figure 5.2 Distributions of subject’s mean and standard deviation by 7DDR recall occasions

Figures 5.1 and 5.2 do not display a notable trend across different groups of recall occasions. We decide to include all of the 444 subjects in the further analysis regardless of recall occasions.

Referring to (3.16), for each subject, we assume that the mean of 24HR recalls is the subject’s true latent value $\mu_s$, and the response error variance of 24HR recalls is the
subject’s true response error variance $\sigma_{sr}^2$. Table 5.2 shows the values of parameters $\mu_s$ and $\sigma_{sr}^2$ of each subject for 24HR recall.

Table 5.2 Parameter values of each subject for 24HR and 7DDR, $N = 444$

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>24HR</th>
<th>7DDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s$</td>
<td>$\mu_s$</td>
<td>$\sigma_{sr}^2$</td>
</tr>
<tr>
<td>1</td>
<td>34.56</td>
<td>148.95</td>
</tr>
<tr>
<td>2</td>
<td>23.74</td>
<td>89.34</td>
</tr>
<tr>
<td>3</td>
<td>22.29</td>
<td>41.46</td>
</tr>
<tr>
<td>4</td>
<td>13.45</td>
<td>26.81</td>
</tr>
<tr>
<td>5</td>
<td>7.99</td>
<td>6.62</td>
</tr>
<tr>
<td>6</td>
<td>18.17</td>
<td>52.39</td>
</tr>
<tr>
<td>7</td>
<td>31.17</td>
<td>348.24</td>
</tr>
<tr>
<td>8</td>
<td>20.70</td>
<td>47.56</td>
</tr>
<tr>
<td>9</td>
<td>52.27</td>
<td>793.79</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>444</td>
<td>16.99</td>
<td>115.16</td>
</tr>
</tbody>
</table>

Corresponding to the 24HR recalls, we also obtain the mean and variance of 7DDR recalls for the same subject. Referring to (3.83), for each subject, we assume that the mean of 7DDR recalls is the sum of the parameters $\mu_s$ and $\tilde{\delta}_s$; the variance of 7DDR recalls is the sum of the variance parameters $\tilde{\sigma}_{sw}^2$ and $\frac{1}{7} \sigma_{sr}^2$ (in the simulation study we assume $W$ is large). Further, we can obtain $\tilde{\sigma}_{sw}^2$ by knowing $\tilde{\sigma}_{sw}^2 + \frac{1}{7} \sigma_{sr}^2$ and $\sigma_{sr}^2$. Table 5.2 shows the values of parameters $\mu_s + \tilde{\delta}_s$, $\tilde{\sigma}_{sw}^2 + \frac{1}{7} \sigma_{sr}^2$ and $\tilde{\sigma}_{sw}^2$ of each subject for 7DDR recall. For some subjects, such as $s = 7$, the observed variance of week bias is less than zero. For these subjects, we replace negative value of $\tilde{\sigma}_{sw}^2$ by zero and adjust the value of $\tilde{\sigma}_{sw}^2 + \frac{1}{7} \sigma_{sr}^2$ to fit the value of $\frac{1}{7} \sigma_{sr}^2$ from the 24HR recall. Among the 444 subjects, 92(20.7%) subjects’ values of $\tilde{\sigma}_{sw}^2$ are negative.
5.2 Selecting Subjects

We use simple random sampling without replacement method to select \( n \) subjects from a population of size \( N \). The \( n \) subjects comprise a sample \( c \) which we denote via the set \( \Omega_c \), such as \( \Omega_c = \{ \hat{\lambda}_s, \lambda_s, \hat{\lambda}_s \} \), \( c = 1, \ldots, C \) where \( C \) is the number of samples we obtain when we repeat the sampling process, we call \( C \) trials. We illustrate the process by using an example of generating \( C = 4 \) simple random samples of \( n = 3 \) from a population of \( N = 7 \). The 7 subjects are the first 7 subjects of 444 subjects. Table 5.3 presents parameter values of the 7 subjects, including subject id, subject’s true latent value and the true standard deviation of response error of saturated fat intake (g/day) for the 24HR recall.

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Subject’s latent value</th>
<th>Subject’s true StdDev of response error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>34.56</td>
<td>12.20</td>
</tr>
<tr>
<td>2</td>
<td>23.74</td>
<td>9.45</td>
</tr>
<tr>
<td>3</td>
<td>22.29</td>
<td>6.44</td>
</tr>
<tr>
<td>4</td>
<td>13.45</td>
<td>5.18</td>
</tr>
<tr>
<td>5</td>
<td>7.99</td>
<td>2.57</td>
</tr>
<tr>
<td>6</td>
<td>18.17</td>
<td>7.24</td>
</tr>
<tr>
<td>7</td>
<td>31.17</td>
<td>18.66</td>
</tr>
</tbody>
</table>

When we permute the subject IDs of the 7 subjects, there are 7! permutations. We randomly choose one permutation from the 7! permutations, and define the 3 subjects occupying the first 3 positions of the selected permutation as a sample of size 3. Table 5.4 displays the parameter values of the three selected subjects in four samples. For \( c = 1 \), \( \text{sid}1=5 \) indicates the first selected subject is subject 5 in the population, where \( \text{sm}1=7.99 \) and \( \text{ssd}1=2.57 \) correspond to the true latent value and the true standard deviation of response error of subject 5.
Table 5.4 Parameter values of 3 selected subjects in 4 samples \((N = 7, n = 3, C = 4)\)

<table>
<thead>
<tr>
<th>c</th>
<th>sid1</th>
<th>sid2</th>
<th>sid3</th>
<th>sm1</th>
<th>sm2</th>
<th>sm3</th>
<th>ssd1</th>
<th>ssd2</th>
<th>ssd3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>7.99</td>
<td>31.17</td>
<td>18.17</td>
<td>2.57</td>
<td>18.66</td>
<td>7.24</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>13.45</td>
<td>22.29</td>
<td>31.17</td>
<td>5.18</td>
<td>6.44</td>
<td>18.66</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>22.29</td>
<td>23.74</td>
<td>13.45</td>
<td>6.44</td>
<td>9.45</td>
<td>5.18</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>31.17</td>
<td>22.29</td>
<td>34.56</td>
<td>18.66</td>
<td>6.44</td>
<td>12.20</td>
</tr>
</tbody>
</table>

5.3 Generating 24HR Data

We generate the simulated 24HR recall data \(Y_{ij}^*\), where \(Y_{ij}^* = \sum_{s=1}^{N} U_{is} Y_{sj}^*\), \(i = 1, \ldots, n, \ j = 1, \ldots, m\). Let \(m = 12\), i.e. randomly select 12 days from \(D\) days for each subject \(i\). In studies of dietary intake assessment, it is most commonly accepted that variations in residues of dietary intake, such as saturated fat intake, are assumed to follow a lognormal distribution (Dodd et al. 2006). Hence, we generate \(Y_{sj}^*\) based on a lognormal distribution \((\mu_s^i, \sigma_{sr}^{i2})\), where \(\mu_s^i\) and \(\sigma_{sr}^{i2}\) are two parameters of the lognormal distribution. Next, we obtain the two distribution parameters \(\mu_s^i\) and \(\sigma_{sr}^{i2}\) for each subject \(s\) through known \(\mu_s\) and \(\sigma_{sr}^2\). From (3.16), we have \(E_R(Y_{sj}) = \mu_s\) and \(Var_R(Y_{sj}) = \sigma_{sr}^2\). Let \(Y_{sj}^\dagger = \log(Y_{sj}^*)\), then \(Y_{sj}^\dagger \sim N(\mu_s^i, \sigma_{sr}^{i2})\). When \(Y_{sj}^*\) is a variable of lognormal distribution, its expected value \(E_R(Y_{sj}^*) = e^{\mu_s^i + 0.5\sigma_{sr}^{i2}}\), variance \(Var_R(Y_{sj}^*) = e^{2(\mu_s^i + \sigma_{sr}^{i2})} - e^{2\mu_s^i + \sigma_{sr}^{i2}}\) (Casella and Berger 2002). Thus, \(\mu_s^i\) and \(\sigma_{sr}^{i2}\) can be obtained by solving

\[
\left\{ \begin{array}{l}
\mu_s = e^{\mu_s^i + 0.5\sigma_{sr}^{i2}} \\
\sigma_{sr}^2 = e^{2(\mu_s^i + \sigma_{sr}^{i2})} - e^{2\mu_s^i + \sigma_{sr}^{i2}} \end{array} \right.
\] (5.1)
We obtain

\[
\begin{align*}
\mu_i^* &= \ln(\mu_i) - 0.5\sigma_{iR}^{12} \\
\sigma_{iR}^{12} &= \ln\left(1 + \frac{\sigma_{iR}^2}{(\mu_i)^2}\right).
\end{align*}
\]  

(5.2)

Thus,

\[
Y_{ij}^* = \mu_i + z\sqrt{\sigma_{iR}^{12}}
\]  

(5.3)

where \(z\) is a standard normal variable. Therefore, we first generate \(Y_{ij}^*\), then get the simulated 24HR data \(Y_{ij}^{*'} = e^{Y_{ij}^*}\). This approach simulates responses that are greater than zero. It is appropriate to the data of total saturated fat intake which is the usual nutrient consumed daily by persons.

Table 5.5 illustrates 12 simulated 24HR responses for subject \(i\) in samples \(c = 1\) and \(c = 2\). For the same subject \(s\) in different sample, we obtain different simulated 24HR data.
Table 5.5 Simulated 12 24HR recalls of selected subject \( i \) (\( c = 1 \) and \( c = 2 \))

<table>
<thead>
<tr>
<th>c = 1</th>
<th>( \mu_i = 7.99, \sqrt{\sigma^2_{s\alpha}} = 2.57 )</th>
<th>c = 2</th>
<th>( \mu_i = 31.17, \sqrt{\sigma^2_{s\alpha}} = 18.66 )</th>
<th>c = 3</th>
<th>( \mu_i = 18.17, \sqrt{\sigma^2_{s\alpha}} = 7.24 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occasion</td>
<td>24HR</td>
<td>Occasion</td>
<td>24HR</td>
<td>Occasion</td>
<td>24HR</td>
</tr>
<tr>
<td>1</td>
<td>5.58</td>
<td>1</td>
<td>61.11</td>
<td>1</td>
<td>25.30</td>
</tr>
<tr>
<td>2</td>
<td>5.71</td>
<td>2</td>
<td>34.19</td>
<td>2</td>
<td>18.44</td>
</tr>
<tr>
<td>3</td>
<td>12.36</td>
<td>3</td>
<td>31.24</td>
<td>3</td>
<td>21.87</td>
</tr>
<tr>
<td>4</td>
<td>5.15</td>
<td>4</td>
<td>17.29</td>
<td>4</td>
<td>15.01</td>
</tr>
<tr>
<td>5</td>
<td>6.40</td>
<td>5</td>
<td>30.03</td>
<td>5</td>
<td>7.87</td>
</tr>
<tr>
<td>6</td>
<td>5.36</td>
<td>6</td>
<td>19.87</td>
<td>6</td>
<td>19.69</td>
</tr>
<tr>
<td>7</td>
<td>7.51</td>
<td>7</td>
<td>38.93</td>
<td>7</td>
<td>19.09</td>
</tr>
<tr>
<td>8</td>
<td>4.46</td>
<td>8</td>
<td>45.22</td>
<td>8</td>
<td>18.87</td>
</tr>
<tr>
<td>9</td>
<td>10.75</td>
<td>9</td>
<td>18.67</td>
<td>9</td>
<td>23.03</td>
</tr>
<tr>
<td>10</td>
<td>6.46</td>
<td>10</td>
<td>9.75</td>
<td>10</td>
<td>15.74</td>
</tr>
<tr>
<td>11</td>
<td>6.89</td>
<td>11</td>
<td>20.58</td>
<td>11</td>
<td>18.15</td>
</tr>
<tr>
<td>12</td>
<td>7.56</td>
<td>12</td>
<td>20.91</td>
<td>12</td>
<td>12.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c = 2</th>
<th>( \mu_i = 13.45, \sqrt{\sigma^2_{s\alpha}} = 5.18 )</th>
<th>c = 2</th>
<th>( \mu_i = 22.29, \sqrt{\sigma^2_{s\alpha}} = 6.44 )</th>
<th>c = 2</th>
<th>( \mu_i = 31.17, \sqrt{\sigma^2_{s\alpha}} = 18.66 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occasion</td>
<td>24HR</td>
<td>Occasion</td>
<td>24HR</td>
<td>Occasion</td>
<td>24HR</td>
</tr>
<tr>
<td>1</td>
<td>4.26</td>
<td>1</td>
<td>19.95</td>
<td>1</td>
<td>42.53</td>
</tr>
<tr>
<td>2</td>
<td>15.71</td>
<td>2</td>
<td>21.20</td>
<td>2</td>
<td>23.42</td>
</tr>
<tr>
<td>3</td>
<td>22.63</td>
<td>3</td>
<td>19.52</td>
<td>3</td>
<td>11.35</td>
</tr>
<tr>
<td>4</td>
<td>18.06</td>
<td>4</td>
<td>14.47</td>
<td>4</td>
<td>33.04</td>
</tr>
<tr>
<td>5</td>
<td>13.24</td>
<td>5</td>
<td>25.25</td>
<td>5</td>
<td>55.49</td>
</tr>
<tr>
<td>6</td>
<td>10.26</td>
<td>6</td>
<td>19.75</td>
<td>6</td>
<td>17.57</td>
</tr>
<tr>
<td>7</td>
<td>14.92</td>
<td>7</td>
<td>25.37</td>
<td>7</td>
<td>26.67</td>
</tr>
<tr>
<td>8</td>
<td>22.11</td>
<td>8</td>
<td>16.26</td>
<td>8</td>
<td>92.98</td>
</tr>
<tr>
<td>9</td>
<td>17.43</td>
<td>9</td>
<td>20.41</td>
<td>9</td>
<td>29.62</td>
</tr>
<tr>
<td>10</td>
<td>15.35</td>
<td>10</td>
<td>18.82</td>
<td>10</td>
<td>46.52</td>
</tr>
<tr>
<td>11</td>
<td>8.08</td>
<td>11</td>
<td>23.63</td>
<td>11</td>
<td>12.22</td>
</tr>
<tr>
<td>12</td>
<td>11.88</td>
<td>12</td>
<td>11.01</td>
<td>12</td>
<td>10.58</td>
</tr>
</tbody>
</table>

5.4 Generating 7DDR Data

We generate the simulated 7DDR recall data \( \tilde{Y}^*_u \), where \( \tilde{Y}^*_u = \sum_{s=1}^{N} U_{us} \tilde{Y}^*_s \),

\( i = 1, ..., n \), \( t = 1, ..., q \). Let \( q = 4 \), i.e. randomly select 4 weeks from \( W \) weeks for each subject \( i \). Similar to the 24HR recall, we generate \( \tilde{Y}^*_s \) for subject \( s \) based on a lognormal distribution (\( \tilde{\mu}_s, \tilde{\sigma}_{s\alpha}^2 \)), where \( \tilde{\mu}_s \) and \( \tilde{\sigma}_{s\alpha}^2 \) are two parameters of the lognormal distribution. From (3.83), we have \( E(\tilde{Y}_s) = \tilde{\mu}_s + \tilde{\delta}_s \) and
Var\(\left(\tilde{Y}_{st}\right) = \tilde{\sigma}_{sw}^2 + \frac{1}{7} \tilde{\sigma}_{sr}^2\). Let \(Y_{st}^{**} = \log(\tilde{Y}_{st})\), then \(Y_{st}^{**} \sim N\left(\tilde{\mu}_s, \tilde{\sigma}_{sr}^{12}\right)\). We have

\[
E\left(\tilde{Y}_{st}^{**}\right) = e^{\tilde{\mu}_s + 0.5\tilde{\sigma}_{sr}^{12}} \quad \text{and} \quad \text{Var}\left(\tilde{Y}_{st}^{**}\right) = e^{2(\tilde{\mu}_s + \tilde{\sigma}_{sr}^{12})} - e^{2\tilde{\mu}_s + \tilde{\sigma}_{sr}^{12}}.
\]

Similar to (5.1) and (5.2), we obtain

\[
\begin{aligned}
\tilde{\mu}_s &= \ln \left(\mu_s + \tilde{\delta}_s\right) - 0.5\tilde{\sigma}_{sr}^{12} \\
\tilde{\sigma}_{sr}^{12} &= \ln \left(1 + \frac{\tilde{\sigma}_{sw}^2 + \frac{1}{7} \tilde{\sigma}_{sr}^2}{(\mu_s + \tilde{\delta}_s)^2}\right).
\end{aligned}
\] (5.4)

Thus,

\[
Y_{st}^{**} = \tilde{\mu}_s + z\sqrt{\tilde{\sigma}_{sr}^{12}}
\] (5.5)

where \(z\) is a standard normal variable. Therefore, we first generate \(Y_{st}^{**}\), then get

\[
\tilde{Y}_{st}^{*} = e^{Y_{st}^{**}}.
\]

Table 5.6 illustrates 4 simulated 7DDR responses for subject \(i\) in samples \(c = 1\) and \(c = 2\).

<table>
<thead>
<tr>
<th>Table 5.6 Simulated 4 7DDR recalls of selected subject (i) ((c = 1) and (c = 2))</th>
<th>(i = 1, s = 5)</th>
<th>(i = 2, s = 7)</th>
<th>(i = 3, s = 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c = 1)</td>
<td>(\mu_s = 7.99, \sqrt{\sigma_{sr}^2} = 2.57)</td>
<td>(\mu_s = 31.17, \sqrt{\sigma_{sr}^2} = 18.66)</td>
<td>(\mu_s = 18.17, \sqrt{\sigma_{sr}^2} = 7.24)</td>
</tr>
<tr>
<td>(\tilde{\delta}<em>s = 10.79, \tilde{\sigma}</em>{sw}^2 = 8.14)</td>
<td>(\tilde{\delta}<em>s = -1.69, \tilde{\sigma}</em>{sw}^2 = 0)</td>
<td>(\tilde{\delta}<em>s = 7.03, \tilde{\sigma}</em>{sw}^2 = 5.46)</td>
<td></td>
</tr>
<tr>
<td>Occasion</td>
<td>7DDR</td>
<td>Occasion</td>
<td>7DDR</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>11.90</td>
<td>1</td>
<td>27.34</td>
</tr>
<tr>
<td>2</td>
<td>12.19</td>
<td>2</td>
<td>24.66</td>
</tr>
<tr>
<td>3</td>
<td>17.58</td>
<td>3</td>
<td>32.68</td>
</tr>
<tr>
<td>4</td>
<td>18.21</td>
<td>4</td>
<td>29.15</td>
</tr>
<tr>
<td>(i = 1, s = 4)</td>
<td>(i = 2, s = 3)</td>
<td>(i = 3, s = 7)</td>
<td>(\mu_s = 13.45, \sqrt{\sigma_{sr}^2} = 5.18)</td>
</tr>
<tr>
<td>(\tilde{\delta}<em>s = 10.60, \tilde{\sigma}</em>{sw}^2 = 3.22)</td>
<td>(\tilde{\delta}<em>s = 8.18, \tilde{\sigma}</em>{sw}^2 = 4.31)</td>
<td>(\tilde{\delta}<em>s = -1.69, \tilde{\sigma}</em>{sw}^2 = 0)</td>
<td></td>
</tr>
<tr>
<td>Occasion</td>
<td>7DDR</td>
<td>Occasion</td>
<td>7DDR</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>28.87</td>
<td>1</td>
<td>28.06</td>
</tr>
<tr>
<td>2</td>
<td>21.99</td>
<td>2</td>
<td>38.74</td>
</tr>
<tr>
<td>3</td>
<td>20.21</td>
<td>3</td>
<td>32.68</td>
</tr>
<tr>
<td>4</td>
<td>26.42</td>
<td>4</td>
<td>33.19</td>
</tr>
</tbody>
</table>
In this way, we generate 12 24HR recalls and 4 7DDR recalls for each selected subject $i$. 
CHAPTER 6
COMPARING ESTIMATOR/PREDICTORS OF SUBJECT’S LATENT VALUE (SIMULATED 24HR DATA)

Based on the simulated 24HR data in Chapter 5, this chapter evaluates the performance of the estimator/predictor of subject’s latent value in two cases – known variance components and unknown variance components separately. In each case, we compare one estimator and two predictors in two situations – estimator/predictor of latent value for a randomly selected subject and estimator/predictor of latent value of a randomly selected subject, given the realized subject is subject $s$ separately. We evaluate the performance of these estimator/predictors by comparing the expected MSE of these estimator/predictors.

6.1 Known Variance Components

For the simulated 24HR data, when variance components are considered known, we compute one estimator and two predictors of subject’s latent value in two situations. Since any subject in the population could be in the $i^{th}$ position in a response vector, the estimator/predictor of the latent value for a randomly selected subject is an average of the estimator/predictor over all selected subjects, i.e. unconditional on the realized subject. We compute the average estimator/predictor for a randomly selected subject $i$ and compare their expected MSEs. This study uses indicator random variable to represent a specific subject in a sample. The estimator/predictor of the latent value of a randomly selected subject, given the realized subject is subject $s$ is an average of the
estimator/predictor over all samples that include subject $s$. We compute the estimator/predictor of a randomly selected subject, given the realized subject is subject $s$ and compare their expected MSEs.

### 6.1.1 Estimator/Predictor of the Latent Value for a Randomly Selected Subject

Let $\hat{P}_i^{(j)}$ denote an estimator/predictor of the latent value for the $i^{th}$ randomly selected subject, $\mu + B_i$ ($i \leq n$). We evaluate properties of the estimator $\hat{P}_i^{(j)}$ (ME) and two predictors $\hat{P}_2^{(j)}$ (MM) and $\hat{P}_3^{(j)}$ (FPMM).

Estimator $\hat{P}_1^{(j)}$ (ME) is defined by (3.17) after replacing $s$ by $i$, where

$$\hat{P}_1^{(j)} = \bar{Y}_i,$$

$$\bar{Y}_i = \frac{1}{m} \sum_{j=1}^{m} Y_{ij}.$$ Estimator $\hat{P}_1^{(j)}$ (ME) is an average of the observed responses for the $i^{th}$ randomly selected subject.

Predictor $\hat{P}_2^{(j)}$ (MM) is derived under the usual mixed model and defined by (3.21), where

$$\hat{P}_2^{(j)} = \hat{\mu}^s + k_i \left( \bar{Y}_i - \hat{\mu}^s \right),$$

$$\hat{\mu}^s = \sum_{i=1}^{n} w_i^s \bar{Y}_i, \quad w_i^s = \frac{1}{v_i^s}, \quad v_i^s = \sigma^2_s + \sigma^2_{B_i}/m, \quad k_i = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{B_i}/m}.$$ Predictor $\hat{P}_2^{(j)}$ (MM) is computed by using the weighted least squares estimate and the subject-specific response error variance.

Predictor $\hat{P}_3^{(j)}$ (FPMM) is derived under the FPMM and defined by (3.56), where
\[ \hat{P}_3^{(i)} = \bar{Y} + k (\bar{Y}_i - \bar{Y}) , \]
\[ \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} \bar{Y}_i , \quad k = \frac{\sigma^2}{\sigma^2_s + (\sigma^2_n + \sigma^2_E) / m} . \]

Predictor \( \hat{P}_3^{(i)} \) (FPMM) is computed by using the average response error variance over subjects in the population.

Although predictors \( \hat{P}_2^{(i)} \) (MM) and \( \hat{P}_3^{(i)} \) (FPMM) have similar expressions, they are based on different model assumptions. Theoretically, the sample under the usual mixed model is obtained from an infinite population. In fact, the predictor under the usual mixed model is also used when the population is finite. This study evaluates the performance of the three estimator/predictors in the common finite population setting (referring to Chapter 5 where we described how the simulated data were obtained).

Previous studies (Stanek and Singer 2004; San Martino et al. 2008) indicate that, with the common assumptions of known variance components and equal within-subject variance, the predictor under the FPMM has a smaller expected MSE than the predictor under the usual mixed model. In our study, the predictors \( \hat{P}_2^{(i)} \) (MM) and \( \hat{P}_3^{(i)} \) (FPMM) are slightly different from the predictors compared by San Martino et al. (2008). The difference between San Martino’s predictor under the usual mixed model and the predictor \( \hat{P}_2^{(i)} \) (MM) is that San Martino’s predictor used the average response error variance \( \sigma^2_R \). We use the subject-specific response error variance \( \sigma^2_{sr} \) in place of \( \sigma^2_R \) for the predictor \( \hat{P}_2^{(i)} \) (MM). The difference between San Martino’s predictor under the FPMM and the predictor \( \hat{P}_3^{(i)} \) (FPMM) in our study is that San Martino’s predictor
includes the unit (day) sampling fraction \( f = \frac{m}{D} \). In our simulation, \( f \) is considered to be zero (\( f = \frac{m}{D} \equiv 0 \)) for the predictor \( \hat{P}^{(i)} \) (FPMM).

This section compares the three different estimator/predictors to better understand the effects of different model specifications on estimator/predictor and to identify a best estimator/predictor for a randomly selected subject. The next section will evaluate how best to estimate the latent value of a randomly selected subject, given the realized subject is subject \( s \).

For illustration, based on the simulation data in Chapter 5, using the variable of individual saturated fat intake, Tables 6.1 to 6.3 present population data, the estimator/predictor and the expected MSE of the estimator/predictor based on the case \( N = 7, n = 3 \) where the number of repeated samples, i.e. trials \( C = 4 \) (\( m = 12 \), see Chapter 5 for more details about the simulated data). Table 6.1 shows each subject’s ID, subject’s latent value (true mean) and true standard deviation, as well as the known population parameters, \( \mu \), \( \sigma_s^2 \) and \( \sigma_R^2 \).

**Table 6.1 Characteristics of a population of \( N = 7 \)**

<table>
<thead>
<tr>
<th>Subject ID (s)</th>
<th>True mean (( \mu_s ))</th>
<th>True StdDev (( \sqrt{\sigma_{sr}^2} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>34.56</td>
<td>12.20</td>
</tr>
<tr>
<td>2</td>
<td>23.74</td>
<td>9.45</td>
</tr>
<tr>
<td>3</td>
<td>22.29</td>
<td>6.44</td>
</tr>
<tr>
<td>4</td>
<td>13.45</td>
<td>5.18</td>
</tr>
<tr>
<td>5</td>
<td>7.99</td>
<td>2.57</td>
</tr>
<tr>
<td>6</td>
<td>18.17</td>
<td>7.24</td>
</tr>
<tr>
<td>7</td>
<td>31.17</td>
<td>18.66</td>
</tr>
<tr>
<td>Average</td>
<td>( \mu = 21.62 )</td>
<td>( \sigma_s^2 = 87.98 )</td>
</tr>
<tr>
<td>Variance</td>
<td>( \sigma_s^2 = 101.97 )</td>
<td></td>
</tr>
</tbody>
</table>
Table 6.2 Estimates of the estimator/predictor and the squared errors of \( \hat{P}^{(i)} \) by \( c \) and \( i \) \((N = 7, n = 3, C = 4)\)

<table>
<thead>
<tr>
<th>( c )</th>
<th>( i )</th>
<th>( s )</th>
<th>( \mu_s^{(i)} )</th>
<th>( \hat{P}_1^{(i)} )</th>
<th>( \hat{P}_2^{(i)} )</th>
<th>( \hat{P}_3^{(i)} )</th>
<th>( se_1 )</th>
<th>( se_2 )</th>
<th>( se_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>7.99</td>
<td>7.01</td>
<td>7.08</td>
<td>7.98</td>
<td>0.95</td>
<td>0.83</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>7</td>
<td>31.17</td>
<td>28.98</td>
<td>26.01</td>
<td>28.01</td>
<td>4.76</td>
<td>26.55</td>
<td>9.93</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>6</td>
<td>18.17</td>
<td>17.98</td>
<td>17.93</td>
<td>17.98</td>
<td>0.04</td>
<td>0.06</td>
<td>0.04</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
<td>13.45</td>
<td>14.49</td>
<td>14.67</td>
<td>15.20</td>
<td>1.10</td>
<td>1.50</td>
<td>3.09</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>22.29</td>
<td>19.64</td>
<td>19.71</td>
<td>19.89</td>
<td>7.02</td>
<td>6.62</td>
<td>5.73</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>7</td>
<td>31.17</td>
<td>33.50</td>
<td>30.56</td>
<td>32.53</td>
<td>5.45</td>
<td>0.37</td>
<td>1.88</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>22.29</td>
<td>23.22</td>
<td>23.13</td>
<td>23.03</td>
<td>0.87</td>
<td>0.71</td>
<td>0.55</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>23.74</td>
<td>27.23</td>
<td>26.74</td>
<td>26.69</td>
<td>12.23</td>
<td>9.02</td>
<td>8.71</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td>13.45</td>
<td>12.64</td>
<td>12.85</td>
<td>13.38</td>
<td>0.64</td>
<td>0.36</td>
<td>0.004</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>7</td>
<td>31.17</td>
<td>34.22</td>
<td>33.84</td>
<td>34.12</td>
<td>9.31</td>
<td>7.16</td>
<td>8.72</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>22.29</td>
<td>22.86</td>
<td>23.23</td>
<td>23.76</td>
<td>0.33</td>
<td>0.90</td>
<td>2.18</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1</td>
<td>34.56</td>
<td>42.20</td>
<td>41.03</td>
<td>41.40</td>
<td>58.48</td>
<td>41.89</td>
<td>46.85</td>
</tr>
</tbody>
</table>

Table 6.2 shows estimates of the three estimator/predictors and the squared errors of the estimator/predictor by \( c \) and \( i \) \((N = 7, n = 3, C = 4)\). The squared error of the estimator/predictor \( \hat{P}^{(i)} \) is computed by \( se = (\hat{P}^{(i)} - \mu_s^{(i)})^2 \) where \( \mu_s^{(i)} = \sum_{s=1}^{N} u_s \mu_s \). We see that \( \hat{P}^{(i)} \) could refer to any subject in the population for \( i = 1 \) in a sample. For \( i = 2 \), \( \hat{P}^{(i)} \) has the same interpretation as \( i = 1 \), except that it refers to any subject in the population that could be in the \( i = 2 \) position in a sample. The expected MSE of each estimator/predictor is the mean of the squared differences between the estimator/predictor and the subject’s latent value. When we sort the data in Table 6.2 by \( i \) and compute the mean of the estimator/predictor and the MSE of the estimator/predictor for each \( i \), we obtain the results presented in Table 6.3.
Table 6.3 Average of the estimator/predictor \( \hat{P}^{(i)} \) and MSE of \( \hat{P}^{(i)} \) by \( i \) 
\((N = 7, n = 3, C = 4)\)

<table>
<thead>
<tr>
<th>( i )</th>
<th>( \mu_c^{(i)} )</th>
<th>( \overline{P}_1^{(i)} )</th>
<th>( \overline{P}_2^{(i)} )</th>
<th>( \overline{P}_3^{(i)} )</th>
<th>( m\text{se}_1^{(i)} )</th>
<th>( m\text{se}_2^{(i)} )</th>
<th>( m\text{se}_3^{(i)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18.72</td>
<td>19.74</td>
<td>19.68</td>
<td>20.08</td>
<td>3.06</td>
<td>2.55</td>
<td>3.09</td>
</tr>
<tr>
<td>2</td>
<td>24.87</td>
<td>24.68</td>
<td>23.92</td>
<td>24.59</td>
<td>6.09</td>
<td>10.77</td>
<td>6.64</td>
</tr>
</tbody>
</table>

In Table 6.3, an average of subject’s latent values of the four subjects who are selected in position \( i \) in four samples is given by \( \mu_c^{(i)} = \frac{1}{n} \sum_{i=1}^{n} u_i \mu_s \) where \( \Omega \) is the sample set. Correspondingly, \( \overline{P}^{(i)} \) is the average of the estimator/predictor \( \hat{P}^{(i)} \). The expected MSE of the \( \hat{P}^{(i)} \) is computed by \( m\text{se}^{(i)} = E_C \left( \hat{P}^{(i)} - \mu_s^{(i)} \right)^2 \) where the subscript \( C \) indicates the expectation with respect to the distribution over all possible samples. We see that \( \mu_c^{(i)} \) is not close to \( \mu \) due to the small number of simulation replications \( (C = 4) \).

When we increase the number of simulation replications, Table 6.4 shows the results of \( N = 7, n = 3, C = 20,000 \).

Table 6.4 Average of the estimator/predictor \( \hat{P}^{(i)} \) and MSE of \( \hat{P}^{(i)} \) 
\((N = 7, n = 3, C = 20,000)\)

<table>
<thead>
<tr>
<th>( i )</th>
<th>( \mu_c^{(i)} )</th>
<th>( \overline{P}_1^{(i)} )</th>
<th>( \overline{P}_2^{(i)} )</th>
<th>( \overline{P}_3^{(i)} )</th>
<th>( m\text{se}_1^{(i)} )</th>
<th>( m\text{se}_2^{(i)} )</th>
<th>( m\text{se}_3^{(i)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21.66</td>
<td>21.63</td>
<td>21.18</td>
<td>21.64</td>
<td>8.60</td>
<td>7.77</td>
<td>8.10</td>
</tr>
</tbody>
</table>

* 95% confidence interval.

In Table 6.4, \( \mu_c^{(i)} \) is close to the population mean \( \mu = 21.62 \). For each estimator/predictor, the point estimates of the average of the estimator/predictor \( \overline{P}^{(i)} \) are different from \( \mu_c^{(i)} \). We calculate the interval estimate of the \( \overline{P}^{(i)} \) to justify that the
difference is due to the random variability in the simulation process or other reasons. The 95% confidence interval of $\hat{P}(i)$ is calculated based on the 20,000 estimates of $\hat{P}(i)$ ($\hat{P}(i) \pm 1.96$ std. error). We see that 95% CIs of $\hat{P}_1(i)$ (ME) and $\hat{P}_3(i)$ (FPMM) contain $\mu_i$ except for $\hat{P}_2(i)$ (MM). It indicates that $\hat{P}_2(i)$ (MM) is biased for the population mean $\mu$. Since the predictor $\hat{P}_2(i)$ (MM) is computed by using the weighted least squares estimator $\hat{\mu}^*$, should the observed $\hat{P}_2(i)$ (MM) be unbiased for the weighted population mean $\mu^*$? $\mu^*$ is computed by using the population data (see Table 6.5). We define

$$\mu^* = \sum_{s=1}^{N} \tilde{w}_s \mu_s$$

(6.1)

where $\tilde{w}_s$ is the population weighting factor, $\tilde{w}_s = \frac{1}{v_s}$ and $v_s = \sigma_s^2 + \frac{\sigma_{sr}^2}{m}$. In general, the weighted population mean $\mu^*$ is different from the population mean $\mu$. In Table 6.5, we also see $\mu \neq \mu^*$.

<table>
<thead>
<tr>
<th>Subject ID $s$</th>
<th>$\mu_s$</th>
<th>$\tilde{w}_s$</th>
<th>$\tilde{w}_s \mu_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>34.56</td>
<td>0.136</td>
<td>4.700</td>
</tr>
<tr>
<td>2</td>
<td>23.74</td>
<td>0.143</td>
<td>3.395</td>
</tr>
<tr>
<td>3</td>
<td>22.29</td>
<td>0.150</td>
<td>3.344</td>
</tr>
<tr>
<td>4</td>
<td>13.45</td>
<td>0.152</td>
<td>2.044</td>
</tr>
<tr>
<td>5</td>
<td>7.99</td>
<td>0.154</td>
<td>1.230</td>
</tr>
<tr>
<td>6</td>
<td>18.17</td>
<td>0.148</td>
<td>2.689</td>
</tr>
<tr>
<td>7</td>
<td>31.17</td>
<td>0.117</td>
<td>3.647</td>
</tr>
</tbody>
</table>

Ave. $\mu = 21.620$ Sum $\mu^* = 21.044$

To address the relation of the predictor $\hat{P}_2(i)$ (MM) and the weighted population mean $\mu^*$, next we scrutinize the predictor $\hat{P}_2(i)$ (MM).
6.1.1.1 Relation of the Predictor $\hat{P}^{(i)}_2$ (MM) and the Weighted Population Mean $\mu^*$

In this example, using subject ID $s$, we can identify each specific subject for a randomly selected sample $c$. It helps us to understand the relationship between the average of the predictor $\hat{P}^{(i)}_2$ (MM) and the weighted population mean $\mu^*$. Since the observed $\hat{P}^{(i)}_2$ (MM) is computed from the sample sets when $N = 7$, $n = 3$, and $C = 20,000$, correspondingly, we compute a parameter $\mu^*$ — a weighted mean based on the possible sample sets when $N = 7$ and $n = 3$.

We define the parameter

$$\mu^* = \sum_{s=1}^{N} \widetilde{w}_s^* \mu_s$$

(6.2)

where $\widetilde{w}_s^* = \frac{\widetilde{w}_s^*}{\sum_{s=1}^{N} \widetilde{w}_s^*}$, $\widetilde{w}_s^*$ is an average of the weighting factor $w_i^*$ of subject $s$. The coefficient $\widetilde{w}_s^*$ is computed from the possible sample sets. In order to evaluate the coefficient, we define $\delta_{sc} = \begin{cases} 1 & \text{if } s \in \Omega_c \\ 0 & \text{otherwise} \end{cases}$ and $c_s = \sum_{c=1}^{C} \delta_{sc}$. Then $\widetilde{w}_s^* = \frac{\sum_{c=1}^{C} \delta_{sc} w_i^*}{c_s}$ where

$$w_i^* = \frac{1}{v_i^*}, \quad \sum_{i=1}^{n} 1/v_i^*$$

$\sum_{c=1}^{C} \delta_{sc} w_i^*$ is the sum of all $w_i^*$ when subject $s$ is selected in a sample, $c_s$ is the number of samples that include subject $s$. Thus, when $N = 7$ and $n = 3$, there are

$$\binom{N}{n} = \binom{7}{3} = \frac{7 \times 6 \times 5}{3 \times 2 \times 1} = 35$$

possible sample sets. Subject “$s$” is in $\binom{N-1}{n-1} = \frac{6 \times 5}{2} = 15$ of these
We calculate \( w^*_i \) for the 15 sets for a subject \( s \in \Omega_c \), and hence calculate \( \tilde{w}^*_s \). When subject \( s = 1 \), the results are shown in Table 6.6, where
\[
\tilde{w}^*_s = \frac{1}{\sum \frac{1}{v^*_i}}.
\]

Table 6.6 Example of \( \tilde{w}^*_s \) computation based on 15 possible sample sets when subject \( s = 1 \) (\( m = 12 \))

<table>
<thead>
<tr>
<th>Set ( i )</th>
<th>( \sigma^2_s )</th>
<th>( \sigma^2_i )</th>
<th>( v^*_i )</th>
<th>( 1/v^*_i )</th>
<th>( \sum (1/v^*_i) )</th>
<th>( W^*_s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>148.95</td>
<td>89.34</td>
<td>41.46</td>
<td>87.98</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>148.95</td>
<td>89.34</td>
<td>26.81</td>
<td>87.98</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>6</td>
<td>148.95</td>
<td>89.34</td>
<td>52.39</td>
<td>87.98</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>7</td>
<td>148.95</td>
<td>89.34</td>
<td>348.24</td>
<td>87.98</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
<td>148.95</td>
<td>41.46</td>
<td>26.81</td>
<td>87.98</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>148.95</td>
<td>41.46</td>
<td>6.62</td>
<td>87.98</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>6</td>
<td>148.95</td>
<td>41.46</td>
<td>52.39</td>
<td>87.98</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>7</td>
<td>148.95</td>
<td>41.46</td>
<td>348.24</td>
<td>87.98</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>5</td>
<td>148.95</td>
<td>26.81</td>
<td>6.62</td>
<td>87.98</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>6</td>
<td>148.95</td>
<td>26.81</td>
<td>52.39</td>
<td>87.98</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>7</td>
<td>148.95</td>
<td>26.81</td>
<td>348.24</td>
<td>87.98</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>6</td>
<td>148.95</td>
<td>6.62</td>
<td>52.39</td>
<td>87.98</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>7</td>
<td>148.95</td>
<td>6.62</td>
<td>348.24</td>
<td>87.98</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>7</td>
<td>148.95</td>
<td>52.39</td>
<td>348.24</td>
<td>87.98</td>
</tr>
</tbody>
</table>

Ave.

\( \bar{w}^*_s = \frac{1}{m} \sum_{i=1}^{m} W^*_s = 0.32159 \)

Note: This table is for the convenience of illustration, subject \( s = 1 \) could be in any position in each set.

Table 6.7 shows the results of \( \bar{w}^*_s \) for all subjects in the population. This is obtained by a similar process as illustrated in Table 6.6 for \( \bar{w}^*_s \) (\( s = 1 \)). Based on (6.2), we obtain the parameter \( \bar{\mu}^* = 21.164 \).
Table 6.7 Parameter $\hat{\mu}^*$ — weighted mean based on the possible sample sets when $N = 7$ and $n = 3$ ($m = 12$)

<table>
<thead>
<tr>
<th>Subject ID $s$</th>
<th>$\mu_s$</th>
<th>$\hat{w}_s^*$</th>
<th>$\hat{w}_s^{**}$</th>
<th>$\hat{w}_s^{**} \mu_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>34.56</td>
<td>0.322</td>
<td>0.138</td>
<td>4.763</td>
</tr>
<tr>
<td>2</td>
<td>23.74</td>
<td>0.335</td>
<td>0.143</td>
<td>3.405</td>
</tr>
<tr>
<td>3</td>
<td>22.29</td>
<td>0.346</td>
<td>0.148</td>
<td>3.033</td>
</tr>
<tr>
<td>4</td>
<td>13.45</td>
<td>0.350</td>
<td>0.150</td>
<td>2.014</td>
</tr>
<tr>
<td>5</td>
<td>7.99</td>
<td>0.355</td>
<td>0.152</td>
<td>1.214</td>
</tr>
<tr>
<td>6</td>
<td>18.17</td>
<td>0.343</td>
<td>0.147</td>
<td>2.671</td>
</tr>
<tr>
<td>7</td>
<td>31.17</td>
<td>0.284</td>
<td>0.122</td>
<td>3.795</td>
</tr>
<tr>
<td>Sum</td>
<td>2.335</td>
<td>1.000</td>
<td></td>
<td>$\hat{\mu}^* = 21.164$</td>
</tr>
</tbody>
</table>

Compared with the results in Table 6.5, we see that the $\hat{\mu}^*$ is different from the population mean $\mu = 21.620$ and the weighted population mean $\hat{\mu} = 21.044$. The reason for the difference between $\mu^*$ and $\hat{\mu}$ is that $\hat{\mu}$ is computed using the weighting factor $w^*_i$ which is conditional on the samples. When we increase the sample size to $n = N = 7$, $\hat{\mu}^* = \mu^*$. From this perspective, the weighted population mean $\mu^*$ is a special case of the parameter $\hat{\mu}^*$ when $n = N$.

Further, compared with the simulation result of $\hat{P}_2^{(i)}$ (MM) in Table 6.4 when $N = 7, n = 3, C = 20,000$, we see $\hat{P}_2^{(i)} = 21.18$, the 95% CI is (21.06~21.30). The confidence interval contains $\mu^*$, but not include the population mean $\mu$. When $n = N$, we obtain $\hat{P}_2^{(i)} = 21.086$, the 95% CI is (20.97~21.20), it contains $\mu^*$ and $\mu^*$, but not $\mu$. It implies the $\hat{P}_2^{(i)}$ (MM) is a biased predictor of the population mean $\mu$ when the weighted least squares estimator is conditional on samples with heterogeneous within-subject variance. And the $\hat{P}_2^{(i)}$ (MM) is a unbiased predictor of the weighted population mean $\mu^*$.
When \( N = 7, n = 3, C = 20,000 \), Table 6.4 shows the expected MSE of the predictor \( \hat{P}_2^{(i)} \) (MM), \( mse_2^{(i)} \) (MM) is the smallest. A simulation study conducted by San Martino et al. (2008) suggested that different cluster sampling fraction (\( F = n / N \)), unit sampling fraction (\( f = m / D \)), cluster intra-class correlation coefficient and unit intra-class correlation coefficient have effects on predictor when comparing the performance of different predictors in the finite population setting. Therefore, in our study, we change the cluster sampling fraction \( F \) (\( N \) and \( n \)) to compare the performance of different estimator/predictor. In Table 6.8 (\( N = 50, n = 3, C = 20,000 \)), compared with the results in Table 6.4, we found that the \( mse_3^{(i)} \) (FPMM) is the smallest rather than the \( mse_2^{(i)} \) (MM), indicating that the predictor \( \hat{P}_2^{(i)} \) (MM) is not uniformly better than the others.

Table 6.8 Average of the estimator/predictor \( \hat{P}^{(i)} \) and MSE of \( \hat{P}^{(i)} \)
(\( N = 50, n = 3, C = 20,000, \mu = 27.45 \))

<table>
<thead>
<tr>
<th>( i )</th>
<th>( \mu_c^{(i)} )</th>
<th>Average of estimator/predictor ( \hat{P}^{(i)} )</th>
<th>MSE of ( \hat{P}^{(i)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( \hat{P}_1^{(i)} ) (ME)</td>
<td>( \hat{P}_2^{(i)} ) (MM)</td>
</tr>
<tr>
<td>1</td>
<td>27.43</td>
<td>27.37</td>
<td>26.95</td>
</tr>
</tbody>
</table>

Table 6.9 shows the results of comparisons of \( mse^{(i)} \) s for different sampling settings when \( C = 20,000 \). Since sometimes the difference of the expected MSEs of two predictors (or the estimator and the predictor) is slight, and to a certain extent variability exists in a simulation study, as in the previous study (San Martino et al. 2008), we use relative percent increase (RPI) to identify whether two predictors (or the estimator and the predictor) are considered to be equivalent,

\[
RPI = \frac{mse_{estimator/predictor} - mse_{best estimator/predictor}}{mse_{best estimator/predictor}} \times 100\% \tag{6.3}
\]
where \( mse_{\text{best estimator/predictor}} \) is the expected MSE of the best estimator/predictor we identify with the minimum expected MSE in each setting. We use the same criterion as used by San Martino et al. (2008), and classify the two predictors (or the estimator and the predictor) are equivalent when RPI < 5%. In Table 6.9, although the absolute value of \( mse_2^{(i)} \) (MM) is the smallest in one setting \( (N = 7) \), the RPI of the \( mse_3^{(i)} \) (FPMM) relative to the \( mse_2^{(i)} \) (MM) is less than 5% (4.6%), indicating the two predictors are equivalent. Comparing the \( mse^{(i)} \)s and the RPIs in different sampling settings, we see that the predictor \( \hat{P}_3^{(i)} \) (FPMM) is always better than the estimator \( \hat{P}_1^{(i)} \) (ME) and the predictor \( \hat{P}_2^{(i)} \) (MM).

Table 6.9 Comparisons of the \( mse^{(i)} \) and the RPI for different sampling settings \( (C = 20,000) \)

<table>
<thead>
<tr>
<th>Different sampling setting</th>
<th>Comparisons of ( mse^{(i)} ) and RPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N = 7, n = 3, \mu = 21.62 (F = 0.43) )</td>
<td>( mse_1^{(i)} &gt; mse_3^{(i)} &gt; mse_2^{(i)} )  ( \text{RPI: (11% \ (4.6%)} )</td>
</tr>
<tr>
<td>( N = 50, n = 3, \mu = 27.45 (F = 0.06) )</td>
<td>( mse_1^{(i)} &gt; mse_2^{(i)} &gt; mse_3^{(i)} )  ( \text{RPI: (8% \ (4.6%)} )</td>
</tr>
<tr>
<td>( N = 50, n = 10, \mu = 27.45 (F = 0.20) )</td>
<td>( mse_1^{(i)} &gt; mse_2^{(i)} &gt; mse_3^{(i)} )  ( \text{RPI: (11% \ (10%)} )</td>
</tr>
<tr>
<td>( N = 200, n = 50, \mu = 25.88 (F = 0.25) )</td>
<td>( mse_2^{(i)} &gt; mse_1^{(i)} &gt; mse_3^{(i)} )  ( \text{RPI: (23% \ (10%)} )</td>
</tr>
<tr>
<td>( N = 444*, n = 200, \mu = 25.30 (F = 0.45) )</td>
<td>( mse_2^{(i)} &gt; mse_1^{(i)} &gt; mse_3^{(i)} )  ( \text{RPI: (20% \ (11%)} )</td>
</tr>
</tbody>
</table>

*Note: * \( N = 444 \) is the population shown in Table 5.2.

In conclusion, when variance components are known, we suggest using the predictor \( \hat{P}_3^{(i)} \) (FPMM) derived under the FPMM to predict the latent value for a
randomly selected subject. Emphasis should be paid that our suggestion is based on the simulation results in this study, evaluations from different studies are valuable for comparison.

6.1.2 Estimator/Predictor of the Latent Value of a Randomly Selected Subject,

Given the Realized Subject is Subject \( s \)

In Table 6.2, a sample of three subjects is obtained through simple random sampling without replacement when \( N = 7 \), we see that one subject could be selected in different samples. When we sort the data in Table 6.2 by subject ID \( s \), the results are shown in Table 6.10.

Table 6.10 Estimates of the estimator/predictor \( \hat{P}^{(i)} \) and the squared error of \( \hat{P}^{(i)} \) sorted by \( s \) \((N = 7, n = 3, C = 4)\)

<table>
<thead>
<tr>
<th>( s )</th>
<th>( c )</th>
<th>( i )</th>
<th>( \mu_{s}^{(i)} )</th>
<th>( \hat{P}_{1}^{(i)} ) (ME)</th>
<th>( \hat{P}_{2}^{(i)} ) (MM)</th>
<th>( \hat{P}_{3}^{(i)} ) (FPMM)</th>
<th>( s_{e_{1}} ) ( \hat{P}^{(i)} ) (ME)</th>
<th>( s_{e_{2}} ) ( \hat{P}^{(i)} ) (MM)</th>
<th>( s_{e_{3}} ) ( \hat{P}^{(i)} ) (FPMM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>34.56</td>
<td>42.20</td>
<td>41.03</td>
<td>41.40</td>
<td>58.48</td>
<td>41.89</td>
<td>46.85</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
<td>23.74</td>
<td>27.23</td>
<td>26.74</td>
<td>26.69</td>
<td>12.23</td>
<td>9.02</td>
<td>8.71</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>22.29</td>
<td>19.64</td>
<td>19.71</td>
<td>19.89</td>
<td>7.02</td>
<td>6.62</td>
<td>5.73</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>22.29</td>
<td>23.22</td>
<td>23.13</td>
<td>23.03</td>
<td>0.87</td>
<td>0.71</td>
<td>0.55</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>2</td>
<td>22.29</td>
<td>22.86</td>
<td>23.23</td>
<td>23.76</td>
<td>0.33</td>
<td>0.90</td>
<td>2.18</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1</td>
<td>13.45</td>
<td>14.49</td>
<td>14.67</td>
<td>15.20</td>
<td>1.10</td>
<td>1.50</td>
<td>3.09</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>13.45</td>
<td>12.64</td>
<td>12.85</td>
<td>13.38</td>
<td>0.64</td>
<td>0.36</td>
<td>0.04</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>7.99</td>
<td>7.01</td>
<td>7.08</td>
<td>7.98</td>
<td>0.95</td>
<td>0.83</td>
<td>0.0001</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>3</td>
<td>18.17</td>
<td>17.98</td>
<td>17.93</td>
<td>17.98</td>
<td>0.04</td>
<td>0.06</td>
<td>0.04</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>2</td>
<td>31.17</td>
<td>28.98</td>
<td>26.01</td>
<td>28.01</td>
<td>4.76</td>
<td>26.55</td>
<td>9.93</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>3</td>
<td>31.17</td>
<td>33.50</td>
<td>30.56</td>
<td>32.53</td>
<td>5.45</td>
<td>0.37</td>
<td>1.88</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>1</td>
<td>31.17</td>
<td>34.22</td>
<td>33.84</td>
<td>34.12</td>
<td>9.31</td>
<td>7.16</td>
<td>8.72</td>
</tr>
</tbody>
</table>

When \( C \) is large, we compute the average of the estimator/predictor \( \hat{P}^{(i)} \) by subject \( s \), \( \hat{P}^{(s)} \) and the expected MSE, \( mse^{(s)} \) for each subject \( s \), we obtain the results presented in Table 6.11 \((N = 7, n = 3, C = 20,000)\).
Table 6.11 Estimates of the estimator/predictor $\hat{P}^{(s)}$ and MSE of $\hat{P}^{(s)}$ by $s$
($N = 7, n = 3, C = 20,000$)

<table>
<thead>
<tr>
<th>$s$</th>
<th>$c_s$</th>
<th>$\mu_s$</th>
<th>$\hat{P}_1^{(s)}$ (ME)</th>
<th>$\hat{P}_2^{(s)}$ (MM)</th>
<th>$\hat{P}_3^{(s)}$ (FPMM)</th>
<th>$mse_1^{(s)}$ (ME)</th>
<th>$mse_2^{(s)}$ (MM)</th>
<th>$mse_3^{(s)}$ (FPMM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>8533</td>
<td>7.99</td>
<td>8.00</td>
<td>8.06</td>
<td>8.94</td>
<td>0.54</td>
<td>0.55</td>
<td>1.48</td>
</tr>
<tr>
<td>4</td>
<td>8431</td>
<td>13.45</td>
<td>13.47</td>
<td>13.62</td>
<td>14.03</td>
<td>2.28</td>
<td>2.25</td>
<td>2.49</td>
</tr>
<tr>
<td>6</td>
<td>8686</td>
<td>18.17</td>
<td>18.14</td>
<td>18.26</td>
<td>18.39</td>
<td>4.34</td>
<td>4.11</td>
<td>4.02</td>
</tr>
<tr>
<td>3</td>
<td>8482</td>
<td>22.29</td>
<td>22.27</td>
<td>22.24</td>
<td>22.23</td>
<td>3.48</td>
<td>3.34</td>
<td>3.23</td>
</tr>
<tr>
<td>2</td>
<td>8519</td>
<td>23.74</td>
<td>23.69</td>
<td>23.54</td>
<td>23.55</td>
<td>7.39</td>
<td>6.78</td>
<td>6.71</td>
</tr>
<tr>
<td>7</td>
<td>8694</td>
<td>31.17</td>
<td>31.23</td>
<td>29.19</td>
<td>30.57</td>
<td>29.82</td>
<td>24.86</td>
<td>26.82</td>
</tr>
<tr>
<td>1</td>
<td>8655</td>
<td>34.56</td>
<td>34.56</td>
<td>33.27</td>
<td>33.68</td>
<td>12.61</td>
<td>12.39</td>
<td>12.00</td>
</tr>
</tbody>
</table>

Table 6.11 shows the results for subject $s$ in increasing order of $\mu_s$. $c_s$ is the number of samples among 20,000 simulation replications which selected subject $s$. The estimator/predictor $\hat{P}^{(s)}$ is an average of the estimator/predictor $\hat{P}^{(i)}$ for subject $s$ over $c_s$. The expected MSE of $\hat{P}^{(s)}$ is computed by $mse^{(s)} = E_{c_s} \left( \hat{P}^{(i)} - \mu_s^{(i)} \right)^2$, where the subscript $c_s$ denotes the expectation with respect to the distribution over all possible samples where $s$ is in the sample, i.e. $s \in \Omega_c$.

For each subject $s$, from the expression of each predictor, we know that only the predictor $\hat{P}_1^{(s)}$ (ME) is unbiased, predictors $\hat{P}_2^{(s)}$ (MM) and $\hat{P}_3^{(s)}$ (FPMM) are biased due to the shrinkage factors. The results in Table 6.11 show the bias.

6.1.2.1 Bias of the Predictor $\hat{P}_3^{(s)}$ (FPMM)

We define the bias of the predictor $\hat{P}_3^{(s)}$ (FPMM) as $B_3^{(s)}$, it is conditional on a specific subject $s$. $B_3^{(s)} = E \left( \hat{P}_3^{(s)} \right) - \mu_s$, the 95% CI of $B_3^{(s)}$ is computed using $B_3^{(s)} \pm 1.96$ * std. error (there are more than 8,000 observed values for each subject). We compare
$B_{3(s)}$ with a bias of the FPMM-BLUP $\hat{P}_{li}$ investigated by Stanek and Singer (2011) which is conditional on a given sample sequence, i.e. $\tilde{h}_i = 1$, where $\tilde{h}$ denotes the possible sequences of $n$ subjects and $\tilde{h} = 1, ..., \tilde{H}$, $\tilde{H} = \frac{N!}{(N-n)!}$; $\tilde{I}_h$ is an indicator random variable, $\tilde{I}_h = 1$ when sample sequence $\tilde{h}$ is selected, and $\tilde{I}_h = 0$ otherwise. For a given sample sequence $\tilde{I}_h = 1$, the FPMM-BLUP $\hat{P}_{li}$ is a predictor for the subject $s$ on the $i$th position in a sample sequence $\tilde{h}$, and it is biased. We define the bias of the FPMM-BLUP $\hat{P}_{li}$ as $\tilde{B}_{s(l)}$. Stanek and Singer (2011) derived the bias $\tilde{B}_{s(l)} = -(1-k)(\mu_s - \mu)$, where $k$ is the shrinkage factor. When $N = 7, n = 3, C = 20,000$, we have $k = 0.9119$, $\mu = 21.62$. From Table 6.12, we see that $B_{3(s)}$ and $\tilde{B}_{s(l)}$ are different. $B_{3(s)}$ is smaller than $\tilde{B}_{s(l)}$. After drawing a plot of $B_{3(s)}$ vs. $\tilde{B}_{s(l)}$ and fitting a linear regression equation (Figure 6.0), we obtain the association of the $B_{3(s)}$ and $\tilde{B}_{s(l)}$ follows

$$B_{3(s)} = 0.773 \ast \tilde{B}_{s(l)}.$$

However, we have not yet interpreted the association. It is an interesting query for us to explore further.

Table 6.12 Comparing two biases, $B_{3(s)}$ and $\tilde{B}_{s(l)}$

<table>
<thead>
<tr>
<th>$s$</th>
<th>$\mu_s$</th>
<th>$\hat{P}_{3(s)}$ (FPMM)</th>
<th>$B_{3(s)}$</th>
<th>95% CI of $B_{3(s)}$</th>
<th>$\tilde{B}_{s(l)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>7.99</td>
<td>8.94</td>
<td>0.951</td>
<td>(0.935~0.967)</td>
<td>1.200</td>
</tr>
<tr>
<td>4</td>
<td>13.45</td>
<td>14.03</td>
<td>0.589</td>
<td>(0.558~0.621)</td>
<td>0.720</td>
</tr>
<tr>
<td>6</td>
<td>18.17</td>
<td>18.39</td>
<td>0.219</td>
<td>(0.177~0.261)</td>
<td>0.304</td>
</tr>
<tr>
<td>3</td>
<td>22.29</td>
<td>22.23</td>
<td>-0.054</td>
<td>(-0.092~0.016)</td>
<td>-0.059</td>
</tr>
<tr>
<td>2</td>
<td>23.74</td>
<td>23.55</td>
<td>-0.183</td>
<td>(-0.238~0.128)</td>
<td>-0.187</td>
</tr>
<tr>
<td>7</td>
<td>31.17</td>
<td>30.57</td>
<td>-0.595</td>
<td>(-0.703~0.487)</td>
<td>-0.841</td>
</tr>
<tr>
<td>1</td>
<td>34.56</td>
<td>33.68</td>
<td>-0.874</td>
<td>(-0.945~-0.803)</td>
<td>-1.140</td>
</tr>
</tbody>
</table>
When comparing the MSE of $\hat{P}^{(s)}$, we see that the smallest $mse^{(s)}$ changes for different subjects. Since the population of $N = 7$ is small, we increase population size to further analyze the expected MSE of $\hat{P}^{(s)}$.

Next, we show the results based on the cases $N = 200, n = 50, C = 20,000$ and $N = 444, n = 200, C = 20,000$. When $N$ is large, we use scatter plot to demonstrate the $\hat{P}^{(s)}$ with the smallest $mse^{(s)}$ for each subject $s$. Since the distribution of $\hat{P}^{(s)}$ shows a certain cluster pattern with different subject’s latent value and variance, we explore the cluster patterns by plotting the scatter plots of the subject’s true standard deviation $(\sqrt{\sigma_{tr}^2})$ vs. the subject’s latent value ($\mu_t$), $\hat{P}^{(s)}_1$ (ME), $\hat{P}^{(s)}_2$ (MM) and $\hat{P}^{(s)}_3$ (FPMM) by the
\( \hat{P}^{(s)} \) with the smallest \( mse^{(s)} \), respectively. Once we reveal the cluster pattern of \( \hat{P}^{(s)} \) with the smallest \( mse^{(s)} \), it is valuable for us to decide which estimator/predictor should be used for a given subject \( s \). For example, the cluster profile can help the practitioners in a clinic to estimate the latent value for a subject with observed mean and standard deviation after they repeatedly measured the saturated fat intake for the subject. Therefore, in the following, we focus on analyzing the cluster profile of \( \hat{P}^{(s)} \) with the smallest \( mse^{(s)} \) so that we can classify a randomly selected subject, given the realized subject is subject \( s \), into a cluster of \( \hat{P}^{(s)} \) with the smallest \( mse^{(s)} \).

First, we look at Figure 6.1a, the difference in MSE between the estimator \( \hat{P}_1^{(s)} \) (ME) and the predictor \( \hat{P}_2^{(s)} \) (MM) vs. the subject’s latent value by the \( \hat{P}^{(s)} \) with the smaller \( mse^{(s)} \).

![Figure 6.1a Plot of mse difference (mse2-mse1) vs. subject’s latent value (N=200, n=50, C=20000, 24HR, mse1 and mse2)](image)
When $N = 200, n = 50, C = 20,000$, Figure 6.1a compares $mse_1^{(s)}$ (ME) and $mse_2^{(s)}$ (MM) for the $N = 200$ subjects. Since there is one subject with very large difference of $mse_2^{(s)} - mse_1^{(s)}$, we cannot clearly see the distribution of the other dots. We investigated the dot with very large MSE difference and found it is the subject with ID = 60, $\mu = 88.38 \text{ g/day}$ and $\sqrt{\sigma_{sr}^2} = 32.16 \text{ g/day}$. The relatively large $\sqrt{\sigma_{sr}^2}$ causes the very large MSE difference. It is reasonable in reality. We include the subject in the further analysis.

Figure 6.1b shows the case when $N = 444, n = 200, C = 20,000$ for the $N = 444$ subjects, we see that there are more subjects with large MSE difference.
Figure 6.1c enlarges Figure 6.1a without including the dot with the large MSE difference. We see that the $mse_{1}^{(s)}$ (ME) is smaller than the $mse_{2}^{(s)}$ (MM) when a subject’s latent value is greater than 40g/day, or a subject’s latent value is very small, such as $\mu_{s} < 7$ g/day, the $mse_{1}^{(s)}$ (ME) is smaller. The results are consistent with the results shown in Table 6.11. Further, we consider the plot of the subject’s true standard deviation $(\sqrt{\sigma_{IR}^{2}})$ vs. the subject’s latent value ($\mu_{s}$) by the $\hat{P}^{(s)}$ with smaller $mse^{(s)}$ (Figure 6.1d).
From Figure 6.1d, we reach the same conclusion as the one from Figure 6.1c, indicating the subject’s true standard deviation has no effect on the distribution of $\hat{P}^{(s)}$ with smaller $mse^{(s)}$. 
Figures 6.2a and 6.2b compare the $mse_1^{(s)}$ (ME) and the $mse_3^{(s)}$ (FPMM). Figure 6.2a is a scatter plot of the $mse^{(s)}$ difference ($mse_3^{(s)} - mse_1^{(s)}$) vs. subject’s latent value, dots indicate the $mse_1^{(s)}$ (ME) is smaller, stars indicate the $mse_3^{(s)}$ (FPMM) is smaller. We see that the $mse_3^{(s)}$ (FPMM) is smaller when a subject’s latent value is between 20-40 g/day. The areas with smaller $mse_1^{(s)}$ (ME) and smaller $mse_3^{(s)}$ (FPMM) cross each other between 15-20g/day and between 40-55g/day. Further, Figure 6.2b provides subject’s true standard deviation information. We see that the distribution of the smaller $mse^{(s)}$ is also associated with the subject’s true standard deviation and presents a certain cluster pattern.
Figures 6.3a and 6.3b compare the $mse_2^{(s)}$ (MM) and the $mse_3^{(s)}$ (FPMM). We see that the $mse_2^{(s)}$ (MM) is smaller when $\mu_s < 20$ g/day; the $mse_3^{(s)}$ (FPMM) is smaller when $\mu_s > 40$ g/day; between $20 < \mu_s < 40$ g/day, the $mse_3^{(s)}$ (FPMM) is smaller when a
subject’s true standard deviation is relatively small, otherwise, the $mse_2^{(s)}$ (MM) is smaller.

When $N = 444$, $n = 200$, $C = 20,000$, we plot the same figures as Figures 6.1, 6.2 and 6.3 for the $N = 444$ subjects, the figures show quite similar cluster patterns. To demonstrate more reliable results, the subsequent analyses are based on the case $N = 444$, $n = 200$, $C = 20,000$ for the $N = 444$ subjects when comparing the $mse^{(s)}$ of the estimator/predictor in one figure.

![Figure 6.4a Plot of subject’s true StdDev vs. subject’s latent value by estimator/predictor with smallest mse(N=444,n=200,C=20000,24HR)](image)

(subject's true StdDev vs. subject's latent value by estimator/predictor with smallest mse(N=444,n=200,C=20000,24HR))

106
Figure 6.4b Plot of subject's true StdDev vs. estimator P1 by estimator/predictor with smallest mse (N=444, n=200, C=20000, 24HR)

Figure 6.4c Plot of subject's true StdDev vs. predictor P2 by estimator/predictor with smallest mse (N=444, n=200, C=20000, 24HR)
Figures 6.4a-6.4d show the plots of the subject’s true standard deviation vs. the subject’s latent value ($\mu_s$), $\hat{P}_1^{(s)}$ (ME), $\hat{P}_2^{(s)}$ (MM) and $\hat{P}_3^{(s)}$ (FPMM) by the $\hat{P}^{(s)}$ with the smallest $mse^{(s)}$, respectively, where dots indicate the $mse_1^{(s)}$ (ME) is smallest, circles indicate the $mse_2^{(s)}$ (MM) is smallest, stars indicate the $mse_3^{(s)}$ (FPMM) is smallest. In Figure 6.4a, the smallest $mse^{(s)}$ locates in different regions associated with subject’s latent value and subject’s true standard deviation. The distribution of the smallest $mse^{(s)}$ shows a certain cluster pattern. Figures 6.4a-6.4d show the similar cluster patterns but with different range of the estimator/predictor. Further, Figures 6.4a and 6.4b show the quite similar cluster patterns since the $\hat{P}_1^{(s)}$ (ME) is an unbiased estimator of the subject’s latent value. Moreover, in reality we don’t know the subject’s latent value and in practice the $\hat{P}_1^{(s)}$ (ME) can be computed easily as an empirical estimate of a subject’s latent value. Hence, we choose Figure 6.4b for further analysis. In Figure 6.4b, for each estimator/predictor with the smallest MSE, the subjects are roughly gathered into two
clusters: subjects with smaller estimate of the $\hat{P}_1^{(s)}$ (ME) and smaller subject’s true standard deviation; subjects with larger estimate of the $\hat{P}_1^{(s)}$ (ME) and larger subject’s true standard deviation. Figure 6.4b shows the 444 subjects are classified into several clusters.

We calculate the RPI in $mse^{(s)}$ for the estimator $\hat{P}_1^{(s)}$ (ME) and the predictors $\hat{P}_2^{(s)}$ (MM) and $\hat{P}_3^{(s)}$ (FPMM) relative to the best estimator/predictor for each subject, we consider the two predictors (or the estimator and the predictor) equivalent when RPI < 5%. Based on Figure 6.4b, we obtain Figure 6.5 with the distribution of the best or equivalent to the best estimator/predictor, such as dots indicate only the estimator $\hat{P}_1^{(s)}$ (ME) is the best, triangles indicate the estimator $\hat{P}_1^{(s)}$ (ME) and the predictor $\hat{P}_2^{(s)}$ (MM) are the best equivalently. Here we introduce two concepts of cluster analysis in data mining: strict partitioning clustering and overlapping clustering. Strict partitioning clustering denotes a subject belongs to exactly one cluster. Overlapping clustering denotes a subject may belong to more than one cluster (Cluster Analysis, Wikipedia). In Figure 6.5, the subjects within the circles belong to the strict partitioning clustering, and those subjects outside the circles belong to the overlapping clustering.
Corresponding to Figures 6.4b and 6.5, more generally, we develop Figures 6.6 and 6.7 using the percentile of the subject’s true standard deviation vs. the percentile of the estimator $\hat{P}_1^{(s)} (ME)$.

Based on the clusters in Figure 6.6 and combining the information of the overlapping clustering in Figure 6.7, we divide the subjects into six clusters. Misclassification may occur around the cut-off points, or when the percentile of the $\hat{P}_1^{(s)} (ME)$ is within $[89, 96]$ and the percentile of the subject’s true standard deviation is small, or when the percentile of the $\hat{P}_1^{(s)} (ME)$ is within $(96, 99]$ and the percentile of the subject’s true standard deviation is very large.
Figure 6.6 Plot of percentile of the subject’s true StdDev vs. percentile of the estimator P1 by estimator/predictor with smallest mse (N=444, n=200, C=20000, 24HR)

Figure 6.7 Plot of percentile of the subject’s true StdDev vs. percentile of the estimator P1 by the best or equivalent to the best estimator/predictor (N=444, n=200, C=20000, 24HR)

(Note: 1=ME 2=MM 3=FPMM)
Based on the classification of the clusters in Figure 6.6, Table 6.13 lists the
descriptive statistics of the best estimator/predictor for the \( N = 444 \) subjects. \( mse^{(s)} \) is
the average \( mse^{(s)} \). The \( mse^{(s)} \) of the best estimator/predictor is smallest within the
cluster.

Table 6.13 Descriptive statistics of cluster profile based on the \( N = 444 \) subjects

<table>
<thead>
<tr>
<th>Percentile of the ( \hat{P}_1^{(s)} ) ( X )</th>
<th>Percentile of the subject’s true StdDev ( Y )</th>
<th>( n )</th>
<th>( mse_1^{(s)} ) (ME)</th>
<th>( mse_2^{(s)} ) (MM)</th>
<th>( mse_3^{(s)} ) (FPMM)</th>
<th>Best estimator/predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X \in [0,2) )</td>
<td>any</td>
<td>8</td>
<td>2.19</td>
<td>2.20</td>
<td>5.03</td>
<td>( \hat{P}_1^{(s)} ) (ME)</td>
</tr>
<tr>
<td>( X \in [2,30) )</td>
<td>any</td>
<td>125</td>
<td>4.90</td>
<td>4.56</td>
<td>5.22</td>
<td>( \hat{P}_2^{(s)} ) (MM)</td>
</tr>
<tr>
<td>( X \in [30,89) )</td>
<td>( Y \in [0,60) )</td>
<td>135</td>
<td>8.10</td>
<td>7.16</td>
<td>6.81</td>
<td>( \hat{P}_3^{(s)} ) (FPMM)</td>
</tr>
<tr>
<td>( X \in [30,89) )</td>
<td>( Y \in [60,99] )</td>
<td>128</td>
<td>21.91</td>
<td>16.58</td>
<td>18.15</td>
<td>( \hat{P}_2^{(s)} ) (MM)</td>
</tr>
<tr>
<td>( X \in [89,96] )</td>
<td>any</td>
<td>35</td>
<td>27.20</td>
<td>30.61</td>
<td>25.44</td>
<td>( \hat{P}_3^{(s)} ) (FPMM)</td>
</tr>
<tr>
<td>( X \in (96,99] )</td>
<td>any</td>
<td>13</td>
<td>40.60</td>
<td>136.35</td>
<td>44.89</td>
<td>( \hat{P}_1^{(s)} ) (ME)</td>
</tr>
</tbody>
</table>

In conclusion, when variance components are known, we want to predict the
latent value of a randomly selected subject, given the realized subject is subject \( s \), we
suggest referring to Figures 6.4b and 6.5 if knowing the \( \hat{P}_1^{(s)} \) and the subject’s true
standard deviation, or referring to Figure 6.6 and Table 6.13 if knowing the percentile of
the \( \hat{P}_1^{(s)} \) and the percentile of the subject’s true standard deviation to choose the best
estimator/predictor. However, usually, we do not know the subject’s true standard
deviation. In next section we will compute one estimator and two empirical predictors
based on estimated variance components and make comparisons with the results of
known variance components.

112
The conclusion is based on subject’s saturated fat intake, further work is needed to claim the cluster patterns are specific or can be generalized, such as to analyze the subject’s total fat intake or other variables with similar characteristics.

6.2 Unknown Variance Components

For the simulated 24HR data, when variance components are considered unknown, we compare the estimator \( \hat{P}_1^{(i)} \) (ME) and the predictors \( \hat{P}_{2e}^{(i)} \) (MM) and \( \hat{P}_{3e}^{(i)} \) (FPMM), where \( \hat{P}_{2e}^{(i)} \) (MM) and \( \hat{P}_{3e}^{(i)} \) (FPMM) are the empirical predictors under the usual mixed model and under the FPMM. Based on (3.23) and (3.60), we need to estimate the variance components \( \sigma_S^2 \), \( \sigma_R^2 \) and \( \sigma_{ir}^2 \) in order to compute the empirical predictors \( \hat{P}_{2e}^{(i)} \) (MM) and \( \hat{P}_{3e}^{(i)} \) (FPMM). Similar to Chapter 4, we use method of moments to obtain the estimates of \( \hat{\sigma}_S^2 \), \( \hat{\sigma}_R^2 \) and \( \hat{\sigma}_{ir}^2 \) by equating the observed mean squares to the expected mean squares from a one-way ANOVA table. From (4.8), (4.9) and (4.10),

\[
\hat{\sigma}_R^2 = MSR, \\
\hat{\sigma}_S^2 = \max\left(0, \frac{MSB - MSR}{m}\right),
\]

and

\[
\hat{\sigma}_{ir}^2 = SS_i^2.
\]

The observed MSR, observed MSB and \( SS_i^2 \) are obtained based on (4.1), (4.2) and (4.3). Thus, corresponding to (3.21) and (3.56), we replace \( \sigma_S^2 \), \( \sigma_R^2 \), \( \sigma_{ir}^2 \) by \( \hat{\sigma}_S^2 \), \( \hat{\sigma}_R^2 \),\( \hat{\sigma}_{ir}^2 \).
and obtain the empirical predictor under the usual mixed model as (3.23) and the empirical predictor under the FPMM as (3.60).

### 6.2.1 Estimator/Empirical Predictor of the Latent Value for a Randomly Selected Subject

Based on (3.23) and (3.60), we compute the empirical predictors $\hat{P}_{2e}^{(i)}$ (MM) and $\hat{P}_{3e}^{(i)}$ (FPMM). Similar to Table 6.4, we compute the $\tilde{P}_{2e}^{(i)}$ (MM) and $\tilde{P}_{3e}^{(i)}$ (FPMM), and the expected empirical MSE ($mse_e^{(i)}$) for position $i$ ($N = 7, n = 3, C = 20,000$), the results show $mse_{2e}^{(i)}$ (MM) is smallest (Table 6.14).

Table 6.14 Average of the estimator/empirical predictor $\hat{P}^{(i)}$ and MSE/empirical MSE of $\hat{P}^{(i)}$ ($N = 7, n = 3, C = 20,000$)

<table>
<thead>
<tr>
<th>$i$</th>
<th>$\mu_r^{(i)}$</th>
<th>Average of estimator/empirical predictor $\hat{P}^{(i)}$</th>
<th>MSE/empirical MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\hat{P}_{1}^{(i)}$</td>
<td>$\hat{P}_{2e}^{(i)}$</td>
</tr>
<tr>
<td>1</td>
<td>21.66</td>
<td>21.63</td>
<td>21.15</td>
</tr>
</tbody>
</table>

Table 6.15 presents the results of comparisons of the expected MSE/empirical MSE and the RPI in the expected MSE/empirical MSE for different sampling settings when $C = 20,000$. When the $mse_{2e}^{(i)}$ (MM) is smallest, we see that the RPIs of the $mse_{3e}^{(i)}$ (FPMM) relative to the $mse_{2e}^{(i)}$ (MM) are less than 5%, indicating the empirical predictors $\hat{P}_{2e}^{(i)}$ (MM) and $\hat{P}_{3e}^{(i)}$ (FPMM) are equivalent. When the $mse_{3e}^{(i)}$ (FPMM) is smallest, the RPIs of the $mse_{1}^{(i)}$ (ME) and the $mse_{2e}^{(i)}$ (MM) relative to the $mse_{3e}^{(i)}$ (FPMM) are greater than 5%, indicating the empirical predictor $\hat{P}_{3e}^{(i)}$ (FPMM) is the best.
Table 6.15 Comparisons of the $mse^{(i)} / mse_e^{(i)}$ and the RPI for different sampling settings ($C = 20,000$)

<table>
<thead>
<tr>
<th>Different sampling setting</th>
<th>Comparisons of $mse^{(i)} / mse_e^{(i)}$ and RPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N = 7, n = 3, \mu = 21.62 (F = 0.43)$</td>
<td>$mse_1^{(i)} &gt; mse_3^{(i)} &gt; mse_2^{(i)}$</td>
</tr>
<tr>
<td></td>
<td>RPI: (4.6%) (4.4%)</td>
</tr>
<tr>
<td>$N = 50, n = 3, \mu = 27.45 (F = 0.06)$</td>
<td>$mse_1^{(i)} &gt; mse_3^{(i)} &gt; mse_2^{(i)}$</td>
</tr>
<tr>
<td></td>
<td>RPI: (1.9%) (1.8%)</td>
</tr>
<tr>
<td>$N = 50, n = 10, \mu = 27.45 (F = 0.20)$</td>
<td>$mse_1^{(i)} &gt; mse_3^{(i)} &gt; mse_2^{(i)}$</td>
</tr>
<tr>
<td></td>
<td>RPI: (8.2%) (0.4%)</td>
</tr>
<tr>
<td>$N = 200, n = 50, \mu = 25.88 (F = 0.25)$</td>
<td>$mse_1^{(i)} &gt; mse_3^{(i)} &gt; mse_2^{(i)}$</td>
</tr>
<tr>
<td></td>
<td>RPI: (9.9%) (7.5%)</td>
</tr>
<tr>
<td>$N = 444, n = 200, \mu = 25.30 (F = 0.45)$</td>
<td>$mse_1^{(i)} &gt; mse_3^{(i)} &gt; mse_2^{(i)}$</td>
</tr>
<tr>
<td></td>
<td>RPI: (10.9%) (10.3%)</td>
</tr>
</tbody>
</table>

In conclusion, when variance components are unknown, similar to the case of known variance components, we suggest using the empirical predictor $\hat{P}_3^{(i)}$ (FPMM) derived under the FPMM to predict the latent value for a randomly selected subject.

6.2.2 Estimator/Empirical Predictor of the Latent Value of a Randomly Selected Subject, Given the Realized Subject is Subject $s$

Table 6.16 shows the estimates of the empirical predictors $\hat{P}_2^{(s)}$ (MM) and $\hat{P}_3^{(s)}$ (FPMM), and the expected empirical MSE $mse_e^{(s)}$ for each subject $s$ ($N = 7, n = 3, C = 20,000$). Still, the estimator $\hat{P}_1^{(s)}$ (ME) is unbiased, the empirical predictors $\hat{P}_2^{(s)}$ (MM) and $\hat{P}_3^{(s)}$ (FPMM) are biased. The smallest $mse_e^{(s)} / mse_e^{(s)}$ changes
for different subjects. We further analyze the estimator/empirical predictor and the MSE/empirical MSE of \( \hat{P}^{(s)} \) based on a larger population of \( N = 444 \).

Table 6.16 Estimates of the estimator/empirical predictor \( \hat{P}^{(s)} \) and MSE/empirical MSE of \( \hat{P}^{(s)} \) by \( s \) (\( N = 7, n = 3, C = 20,000 \))

<table>
<thead>
<tr>
<th>( s )</th>
<th>( c_s )</th>
<th>( \mu_s )</th>
<th>( \hat{P}_1^{(s)} ) (ME)</th>
<th>( \hat{P}_2^{(s)} ) (MM)</th>
<th>( \hat{P}_3^{(s)} ) (FPMM)</th>
<th>( mse_1^{(s)} ) (ME)</th>
<th>( mse_2^{(s)} ) (MM)</th>
<th>( mse_3^{(s)} ) (FPMM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>8533</td>
<td>7.99</td>
<td>8.00</td>
<td>8.06</td>
<td>8.59</td>
<td>0.54</td>
<td>0.58</td>
<td>1.08</td>
</tr>
<tr>
<td>4</td>
<td>8431</td>
<td>13.45</td>
<td>13.47</td>
<td>13.68</td>
<td>14.11</td>
<td>2.28</td>
<td>2.62</td>
<td>3.19</td>
</tr>
<tr>
<td>6</td>
<td>8686</td>
<td>18.17</td>
<td>18.14</td>
<td>18.31</td>
<td>18.66</td>
<td>4.34</td>
<td>4.55</td>
<td>5.05</td>
</tr>
<tr>
<td>3</td>
<td>8482</td>
<td>22.29</td>
<td>22.27</td>
<td>22.19</td>
<td>22.45</td>
<td>3.48</td>
<td>3.48</td>
<td>3.91</td>
</tr>
<tr>
<td>2</td>
<td>8519</td>
<td>23.74</td>
<td>23.69</td>
<td>23.36</td>
<td>23.69</td>
<td>7.39</td>
<td>6.96</td>
<td>7.23</td>
</tr>
<tr>
<td>7</td>
<td>8694</td>
<td>31.17</td>
<td>31.23</td>
<td>29.06</td>
<td>30.18</td>
<td>29.82</td>
<td>26.30</td>
<td>26.74</td>
</tr>
<tr>
<td>1</td>
<td>8655</td>
<td>34.56</td>
<td>34.56</td>
<td>33.35</td>
<td>33.73</td>
<td>12.61</td>
<td>13.33</td>
<td>13.16</td>
</tr>
</tbody>
</table>

When \( N = 444, n = 200, C = 20,000 \), we compare \( mse_1^{(s)} \) (ME), \( mse_2^{(s)} \) (MM) and \( mse_3^{(s)} \) (FPMM) of the three estimator/empirical predictors in one figure. Figures 6.8a-6.8d show the plots of the subject’s estimated standard deviation vs. the subject’s latent value (\( \mu_s \)), \( \hat{P}_1^{(s)} \) (ME), \( \hat{P}_2^{(s)} \) (MM) and \( \hat{P}_3^{(s)} \) (FPMM) by the \( \hat{P}^{(s)} \) with the smallest \( mse^{(s)}/mse_e^{(s)} \), respectively, where dots indicate the \( mse_1^{(s)} \) (ME) is smallest, circles indicate the \( mse_2^{(s)} \) (MM) is smallest, stars indicate the \( mse_3^{(s)} \) (FPMM) is smallest.
Figure 6.8a Plot of subject's estimated StdDev vs. latent value by estimator/empirical predictor with smallest mse (N=444, n=200, C=20000, 24HR, unknown variance)

Figure 6.8b Plot of subject's estimated StdDev vs. estimator P1 by estimator/empirical predictor with smallest mse (N=444, n=200, C=20000, 24HR, unknown variance)
Figures 6.8a-6.8d show the similar cluster patterns, especially for Figures 6.8a and 6.8b. We choose Figure 6.8b for further analysis with the same reason as we choose Figure 6.4b. In Figure 6.8b, we see there are five clusters of the estimator/empirical predictor with the smallest MSE/empirical MSE.
We calculate the RPI in $\frac{mse(n)}{mse(e)}$ for the estimator $\hat{P}_1^{(i)}$ (ME) and the empirical predictors $\hat{P}_2^{(i)}$ (MM) and $\hat{P}_3^{(i)}$ (FPMM) relative to the best estimator/empirical predictor for each subject, we still consider the two empirical predictors (or the estimator and the empirical predictor) equivalent when RPI < 5%. Figure 6.9 presents the distribution of the estimator $\hat{P}_1^{(i)}$ (ME) and the empirical predictors $\hat{P}_2^{(i)}$ (MM) and $\hat{P}_3^{(i)}$ (FPMM), such as dots indicate only the estimator $\hat{P}_1^{(i)}$ (ME) is the best, triangles indicate the estimator $\hat{P}_1^{(i)}$ (ME) and the empirical predictor $\hat{P}_2^{(i)}$ (MM) are the best equivalently. The subjects within the circles belong to the strict partitioning clustering, and those subjects outside the circles belong to the overlapping clustering.

![Figure 6.9 Plot of subject’s estimated StdDev vs. estimator P1 by the best or equivalent to the best estimator/empirical predictor (N=444, n=200, C=20000, 24HR, unknown variance)](image)

Corresponding to Figures 6.8b and 6.9, we develop Figures 6.10 and 6.11 using the percentile of the subject’s estimated standard deviation vs. the percentile of the $\hat{P}_1^{(i)}$. 

119
We see the scatter plot roughly has the “football” shape. Referring to Figure 6.11, we
divide the 444 subjects into five clusters in Figure 6.10. The estimator \( \hat{P}_1^{(s)} \) (ME) with the
smallest \( mse^{(s)} \) locates two ends of the football. The empirical predictors \( \hat{P}_{2e}^{(i)} \) (MM) and
\( \hat{P}_{3e}^{(i)} \) (FPMM) locate in the middle with a trend that the \( \hat{P}_{3e}^{(i)} \) (FPMM) with the smallest
\( mse^{(i)} \) locates the area of relatively smaller percentile of the \( \hat{P}_1^{(s)} \) and smaller percentile of
the subject’s estimated standard deviation, and the \( \hat{P}_{2e}^{(i)} \) (MM) with the smallest \( mse^{(s)} \)
locates the area of relatively larger percentile of the \( \hat{P}_1^{(s)} \) and larger percentile of the
subject’s estimated standard deviation.

**Figure 6.10 Plot of percentile of the subject’s estimated StdDev vs. percentile
of the estimator P1 by estimator/empirical predictor with smallest mse
(N=444,n=200,C=20000,24HR,unknown variance)**
In conclusion, when variance components are unknown, we want to predict the latent value of a randomly selected subject, given the realized subject is subject \( s \), we suggest referring to Figures 6.8b and 6.9 if knowing the \( \hat{P}_1^{(s)} \) and the subject’s estimated standard deviation, or referring to Figures 6.10 and 6.11 if knowing the percentile of the \( \hat{P}_1^{(s)} \) and the percentile of the subject’s estimated standard deviation to choose the best estimator/empirical predictor.

We compare Figures 6.6 and 6.7 of known variance components and Figures 6.10 and 6.11 of unknown variance components. From Figures 6.7 and 6.10, we found, in the middle of the football, the clusters of the circled areas of the strict partitioning clustering in Figure 6.7 are consistent with the clusters of the circled areas in Figure 6.10. The clusters outside of the circles in the middle could be the predictor \( \hat{P}_2^{(s)} \) (MM).
\( \hat{P}_3^{(s)} \) (FPMM) with the smallest \( mse^{(s)} \), or the empirical predictor \( \hat{P}_2^{(s)} \) (MM) / \( \hat{P}_3^{(s)} \) (FPMM) with the smallest \( mse_\epsilon^{(s)} \). The triangle area with smaller percentile of the \( \hat{P}_1^{(s)} \) could be the \( \hat{P}_1^{(s)} \) (ME) / \( \hat{P}_2^{(s)} \) (MM) with the smallest \( mse^{(s)} \) when variance components are known, however, the cluster change to the \( \hat{P}_1^{(s)} \) (ME) with the smallest \( mse^{(s)} \) when variance components are unknown.

![Figure 6.7 Plot of percentile of the subject's true StdDev vs. percentile of the estimator P1 by the best or equivalent to the best estimator/predictor (N=444,n=200,C=20000,24HR)](image)

(Note: 1=ME 2=MM 3=FPMM)
Figure 6.10 Plot of percentile of the subject’s estimated StdDev vs. percentile of the estimator P1 by estimator/empirical predictor with smallest mse (N=444,n=200,C=20000,24HR,unknown variance)
CHAPTER 7

COMPARING PREDICTORS OF SUBJECT’S LATENT VALUE

(SIMULATED 24HR AND 7DDR DATA)

This chapter aims to compare predictors of subject’s latent value from the 24HR measures only and from the 24HR and 7DDR combined measures based on the simulated 24HR and 7DDR data in Chapter 5. We evaluate the predictors in two cases – known variance components and unknown variance components separately. In each case, similar to Chapter 6, we compute the predictor of subject’s latent value in two situations – predictor of the latent value for a randomly selected subject and predictor of the latent value of a randomly selected subject, given the realized subject is subject \( s \). The performance of these predictors is evaluated by comparing the expected MSEs of these predictors.

7.1 Known Variance Components

For the simulated 24HR and 7DDR data, when variance components are considered known, we compare the two predictors of subject’s latent value of different sources of data under the usual mixed model and under the FPMM separately.

7.1.1 Predictor of the Latent Value for a Randomly Selected Subject

For the 24HR data, the predictor \( \hat{P}_2^{(i)} \) under the usual mixed model is defined by (3.21), where

\[
\hat{P}_2^{(i)} = \hat{\mu} + k_i (\bar{Y}_i - \hat{\mu}^*) .
\]
The predictor $\hat{P}^{(i)}_3$ under the FPMM model is defined by (3.56), where

$$\hat{P}^{(i)}_3 = \bar{Y} + k \left( \bar{Y} - \bar{Y} \right).$$

For the 24HR and 7DDR combined data, the predictor $\hat{P}^{(i)}_2$ under the usual mixed model is defined by (3.121), where

$$\hat{P}^{(i)}_2 = \bar{Y} + k_1,$$

$\bar{Y}$ and $k_1$ are obtained from (3.119) and (3.120). The predictor $\hat{P}^{(i)}_3$ under the FPMM model is defined by (3.130), where

$$\hat{P}^{(i)}_3 = \bar{Y} + k^*_1,$$

$\bar{Y}$ and $k^*_1$ are obtained from (3.128) and (3.129).

Similar to Table 6.2, we compute the estimates of the predictors and the squared errors of the predictors for each position $i$ in each sample $c$, then sort the data by $i$ and compute the average of the predictors and the expected MSE of the predictors for the position $i$, we obtain the results presented in Table 7.1 when $N = 7, n = 3$, and $C = 20,000$.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$\mu^{(i)}_c$</th>
<th>Average of predictor $\hat{P}^{(i)}$</th>
<th>MSE of $\hat{P}^{(i)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MM</td>
<td>FPMM</td>
</tr>
<tr>
<td></td>
<td>$\hat{P}^{(i)}_2$</td>
<td></td>
<td>$\hat{P}^{(i)}_3$</td>
</tr>
<tr>
<td>(24)</td>
<td>(24)</td>
<td>(24)</td>
<td>(24)</td>
</tr>
<tr>
<td></td>
<td>$mse^{(i)}_c$</td>
<td>$mse^{(i)}_c$</td>
<td>$mse^{(i)}_c$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>21.66</td>
<td>21.18</td>
<td>20.79</td>
</tr>
</tbody>
</table>

Since the population size $N = 7$ is small, we develop Table 7.2 to see the results of different sampling settings. Similarly, we make comparisons by using the MSE over all positions to obtain more stable results. Still, when RPI < 5%, we consider the two
predictors $\hat{P}_2(i)$ (MM, 24HR) and $\hat{P}_{c2}(i)$ (MM, 24HR+7DDR), or $\hat{P}_3(i)$ (FPMM, 24HR) and $\hat{P}_{c3}(i)$ (FPMM, 24HR+7DDR) to be equivalent. For both the usual mixed model and the FPMM model, the predictors based on the combined data with smaller MSE are better than the ones based on the 24HR data only except for the case of smaller population size of $N = 7$.

Table 7.2 Comparisons of the $mse(i)$ and the RPI for different sampling settings ($C = 20,000$)

<table>
<thead>
<tr>
<th>Different sampling setting</th>
<th>Comparisons of $mse(i)$ and RPI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MM</td>
</tr>
<tr>
<td></td>
<td>$mse_i(24)$</td>
</tr>
<tr>
<td></td>
<td>$mse_i(24+7)$</td>
</tr>
<tr>
<td>$N = 7, n = 3, \mu = 21.62 (F = 0.43)$</td>
<td>7.81 &lt; 8.79</td>
</tr>
<tr>
<td>$N = 50, n = 3, \mu = 27.45 (F = 0.06)$</td>
<td>13.41 &gt; 11.88</td>
</tr>
<tr>
<td>$N = 50, n = 10, \mu = 27.45 (F = 0.20)$</td>
<td>13.75 &gt; 11.70</td>
</tr>
<tr>
<td>$N = 200, n = 50, \mu = 25.88 (F = 0.25)$</td>
<td>15.79 &gt; 13.71</td>
</tr>
<tr>
<td>$N = 444, n = 200, \mu = 25.30 (F = 0.45)$</td>
<td>14.67 &gt; 13.05</td>
</tr>
</tbody>
</table>

Note: * indicates $C = 1,000$ for the purpose of demonstration.

In conclusion, when variance components are known, to predict the latent value for a randomly selected subject, the predictors $\hat{P}_{c2}(i)$ (MM) and $\hat{P}_{c3}(i)$ (FPMM) based on the combined data have better performance.

7.1.2 Predictor of the Latent Value of a Randomly Selected Subject, Given the Realized Subject is Subject $s$

Similar to Table 6.11, when $C$ is large, we compute the average of the predictor $\hat{P}(i)$ by subject $s$, $\hat{P}(s)$ and the expected MSE, $mse(s)$ for each subject $s$, we obtain the
results presented in Table 7.3 \((N = 7, n = 3, C = 20,000)\). We see that the \(mse^{(s)}\) based on the combined data is not always smaller for different subjects.

Table 7.3 Estimates of the predictor \(\hat{P}^{(s)}\) and MSE of \(\hat{P}^{(s)}\) by \(s\) 
\((N = 7, n = 3, C = 20,000)\)

<table>
<thead>
<tr>
<th>(s)</th>
<th>(c_s)</th>
<th>(\mu_s)</th>
<th>Predictor (\hat{P}^{(s)})</th>
<th>MSE of (\hat{P}^{(s)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>8533</td>
<td>7.99</td>
<td>(\hat{P}^{(s)}) (\mu_s) (24)</td>
<td>(mse^{(s)}) (24)</td>
</tr>
<tr>
<td>4</td>
<td>8431</td>
<td>13.45</td>
<td>13.62</td>
<td>14.03</td>
</tr>
<tr>
<td>6</td>
<td>8686</td>
<td>18.17</td>
<td>18.26</td>
<td>18.39</td>
</tr>
<tr>
<td>3</td>
<td>8482</td>
<td>22.29</td>
<td>22.24</td>
<td>22.23</td>
</tr>
<tr>
<td>2</td>
<td>8519</td>
<td>23.74</td>
<td>23.54</td>
<td>23.55</td>
</tr>
<tr>
<td>7</td>
<td>8694</td>
<td>31.17</td>
<td>29.19</td>
<td>30.57</td>
</tr>
<tr>
<td>1</td>
<td>8655</td>
<td>34.56</td>
<td>33.27</td>
<td>33.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(s)</th>
<th>(c_s)</th>
<th>(\mu_s)</th>
<th>Predictor (\hat{P}^{(s)})</th>
<th>MSE of (\hat{P}^{(s)})</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>(\hat{P}^{(s)}) (24+7)</td>
<td>(mse^{(s)}) (24+7)</td>
</tr>
<tr>
<td>5</td>
<td>8533</td>
<td>7.99</td>
<td>8.06</td>
<td>8.94</td>
</tr>
<tr>
<td>4</td>
<td>8431</td>
<td>13.45</td>
<td>13.62</td>
<td>14.03</td>
</tr>
<tr>
<td>6</td>
<td>8686</td>
<td>18.17</td>
<td>18.26</td>
<td>18.39</td>
</tr>
<tr>
<td>3</td>
<td>8482</td>
<td>22.29</td>
<td>22.24</td>
<td>22.23</td>
</tr>
<tr>
<td>2</td>
<td>8519</td>
<td>23.74</td>
<td>23.54</td>
<td>23.55</td>
</tr>
<tr>
<td>7</td>
<td>8694</td>
<td>31.17</td>
<td>29.19</td>
<td>30.57</td>
</tr>
<tr>
<td>1</td>
<td>8655</td>
<td>34.56</td>
<td>33.27</td>
<td>33.68</td>
</tr>
</tbody>
</table>

When \(N = 444, n = 200, C = 20,000\), for the usual mixed model, based on the formulae (3.118) to (3.121), we know \(\hat{P}^{(s)}\) (MM, 24HR+7DDR) is associated with the combined observed data, subject’s true variance, subject’s week bias variance. We explore their associations with the predictor \(\hat{P}^{(s)}\) (MM, 24HR) and \(\hat{P}^{(s)}\) (MM, 24HR+7DDR) and illustrate some results. Figures 7.1 to 7.4 show plots of subject’s true standard deviation, subject’s week bias (due to the 7DDR bias from week to week) standard deviation, the observed 7DDR recall verse subject’s latent value, respectively by predictor with smaller MSE (Figures 7.1 to 7.3) or by the better or equivalent to the better predictor (Figure 7.4). In Figures 7.1 and 7.2, the two kinds of dots distribute randomly, indicating the better predictor is not associated with subject’s true variance and subject’s week bias variance.
Figure 7.3 shows an interesting result. There are 16.2% subjects with smaller MSE of predictor $\hat{P}_2^{(s)}$ (MM, 24HR). These subjects are distributed into two clusters away from the diagonal line with the cut point around 30g/day of subject’s latent value. It indicates 7DDR recall is over-reported when subject’s latent value is less than 30g/day and is under-reported when subject’s latent value is greater than 30g/day. For these
subjects, adding the 7DDR data is bad when the 7DDR bias is large. In Figure 7.4, considering that the two predictors are equivalent when RPI < 5% (labelled “both” with star sign), the cluster pattern is similar to Figure 7.3.
We define a subject’s average response ratio, where

\[
\text{subject's average response ratio} = \frac{\text{subject's average response}_{24\text{HR}} - \text{subject's average response}_{7\text{DDR}}}{\text{subject's average response}_{24\text{HR}}}.
\]

We calculate subject’s average response ratio and plot Figures 7.5 and 7.6. We see that the predictor \( \hat{P}_2^{(s)} \) (MM, 24HR) is better when the subject’s average response ratio is greater than 0.7 and subject’s latent value is less than 30g/day, or when the subject’s average response ratio is less than -0.15 and subject’s latent value is greater than 30g/day.
In practice, we do not know the subject’s latent value, for the same reason as in Chapter 6, when we use the subject’s latent value and the estimator $\hat{P}_1^{(s)}$ (the subject’s average response of 24HR), the plots show the quite similar cluster patterns. We develop Figures 7.7 and 7.8 by using the estimator $\hat{P}_1^{(s)}$ (24HR) to replace the subject’s latent value. The plots show similar patterns to Figures 7.5 and 7.6. The advantage of using Figures 7.7 and 7.8 is that $\hat{P}_1^{(s)}$ (24HR) can be computed easily from the data.
Corresponding to Figures 7.7 and 7.8, more generally, we develop Figures 7.9 and 7.10 by using the percentile of the estimator $\hat{P}_1^{(s)} (24HR)$. The cut point of the percentile of the estimator $\hat{P}_1^{(s)} (24HR)$ is around 70 percentile. To know whether the cut points of 0.7, -0.15 and 70 percentile can be generalized, a study based on other population or other variables should be conducted for future work.
Next, we analyze the better predictor of the latent value of a randomly selected subject, given the realized subject is subject $s$ for the FPMM model. Since the repeated trials $C = 20,000$ is very large, and the matrices become larger for the combined data, the program runs pretty slow. For the purpose of demonstration, we perform the subsequent analyses using replication $C = 1,000$. When $N = 444, n = 200, C = 1,000$, for the FPMM model, there are 21.6% subjects with smaller MSE of predictor $\hat{P}_3^{(s)}$ (FPMM, 24HR).

Figures 7.11 and 7.12 show similar patterns to Figures 7.3 and 7.4.
From Figures 7.13 and 7.14, we see that the predictor $\hat{P}_3^{(c)}$ (FPMM, 24HR) is better when the subject’s average response ratio is greater than 0.5 and subject’s latent value is less than 30g/day, or when the subject’s average response ratio is less than -0.15 and subject’s latent value is greater than 30g/day.
Similarly, we plot Figures 7.15 and 7.16 using the estimator \( \hat{P}_{1}^{(s)} \) (24HR), plot Figures 7.17 and 7.18 using the percentile of the estimator \( \hat{P}_{1}^{(s)} \) (24HR). The cut point of percentile is still around 70.
Figure 7.16 Plot of subject's average response ratio vs. estimator P1 by the better or equivalent to the better predictor (N=444, n=200, C=1000, 24HR+7DDR, FPMM model)

Figure 7.17 Plot of subject's average response ratio vs. percentile of the estimator P1 by the better predictor (N=444, n=200, C=1000, 24HR+7DDR, FPMM model)
In conclusion, when variance components are known, based on the observed 24HR and 7DDR data, we suggest using the predictor based on the combined data to predict the latent value of a randomly selected subject, given the realized subject is subject $s$, except that the predictor based on the 24HR data should be used when the subject’s average response ratio is greater than 0.7 (mixed model) or 0.5 (FPMM model) and the percentile of the $\hat{P}_1^{s}(24HR)$ is less than 70, or when the subject’s average response ratio is less than -0.15 and the percentile of the $\hat{P}_1^{s}(24HR)$ is greater than 70.

In future work we would investigate whether these cut points are specific or could be generalized, and what values of these cut points would be for other variables.

### 7.2 Unknown Variance Components

When variance components are considered unknown, we compare two empirical predictors $\hat{P}_{2r}^{s}(24HR)$ and $\hat{P}_{c2r}^{s}(24HR+7DDR)$ under the usual mixed model, and two
empirical predictors $\hat{P}^{(i)}_{3e}(24\text{HR})$ and $\hat{P}^{(i)}_{c3e}(24\text{HR}+7\text{DDR})$ under the FPMM model. The empirical predictor $\hat{P}^{(i)}_{2e}(\text{MM, 24HR})$ is defined by (3.23), where
\[
\hat{P}^{(i)}_{2e} = \hat{\mu} + \hat{k} \left( \bar{Y} - \hat{\mu} \right),
\]
The empirical predictor $\hat{P}^{(i)}_{c2e}(\text{MM, 24HR})$ is defined by (3.124), where
\[
\hat{P}^{(i)}_{c2e} = \hat{\mu} + \hat{B}_i,
\]
$\hat{\mu}$ and $\hat{B}_i$ are obtained from (3.122) and (3.123). The empirical predictor $\hat{P}^{(i)}_{3e}(\text{FPMM, 24HR})$ is defined by (3.60), where
\[
\hat{P}^{(i)}_{3e} = \bar{Y} + \hat{k} \left( \bar{Y} - \bar{Y} \right).
\]
The empirical predictor $\hat{P}^{(i)}_{c3e}(\text{FPMM, 24HR+7DDR})$ is defined by (3.133), where
\[
\hat{P}^{(i)}_{c3e} = \hat{\mu} + \hat{B}_i^*,
\]
$\hat{\mu}$ and $\hat{B}_i^*$ are obtained from (3.131) and (3.132).

Similar to Chapter 4, for the 24HR data, we need to estimate $\sigma^2_S$, $\sigma^2_R$ and $\sigma^2_{IR}$; for the 24HR+7DDR data, we need to estimate $\sigma^2_S$, $\sigma^2_R$, $\sigma^2_S$, $\sigma^2_W$ and $\sigma^2_{IR}$. For the 24HR data, we already obtained $\hat{\sigma}^2_S$, $\hat{\sigma}^2_R$ and $\hat{\sigma}^2_{IR}$ in Section 6.2. For the combined 24HR+7DDR data, the method is similar to the method in Section 4.2. When $N = 7, n = 3, C = 1,000$, after we simulate 16 repeated measures (12 24HR recalls and 4 7DDR recalls) for each selected subject in a sample, we create a variance-covariance matrix for each sample. Following the structure of parametric variance and covariance matrix in Table 4.4, we use method of moments to obtain the estimated variance components. Instead of solving the simultaneous equations, we use SAS proc reg
procedure to obtain $\hat{\sigma}^2_S$, $\hat{\sigma}^2_R$, $\hat{\sigma}^2_S$ and $\hat{\sigma}^2_W$. The results of both methods are the same.

When negative value appears, we set the negative component to be zero. From (4.2), we see $\hat{\sigma}^2_R$ is obtained based on the sample. Therefore, when variance components are unknown, the $\hat{\sigma}^2_R$ changes for different sample. However, when variance components are known, $\sigma^2_R$ (3.59) is the same for different sample. We use the same method as in Section 4.2 to estimate $\hat{\sigma}^2_W$. We have $\hat{\sigma}^2_{iw} = \max \left( 0, SS_i^{*2} - \frac{1}{q} \hat{\sigma}^2_{ir} \right)$ where

$$SS_i^{*2} = \frac{1}{q-1} \sum_{t=1}^{q} (Y_{it} - \bar{Y}_i)^2, q = 4$$ based on the 7DDR data, and $\hat{\sigma}^2_{ir}$ is obtained using 24HR data.

When we use the estimated variance components to compute the empirical predictor $\hat{P}^{(i)}_{e3e}$ (FPMM, 24HR+7DDR), in some cases, the matrix $\hat{V}_c$ is singular. To solve this problem, we replace $\hat{\sigma}^2_R = 0$ by $\hat{\sigma}^2_R = 1$. We compare the cases of $\hat{\sigma}^2_R = 1$ and $\hat{\sigma}^2_R = 0.1$ to see the sensitivity of assuming different values of $\hat{\sigma}^2_R$. When

$N = 7, n = 3, C = 1,000$, there are 4 trials (0.4%) with $\hat{\sigma}^2_R = 0$, and using $\hat{\sigma}^2_R = 1$ or $\hat{\sigma}^2_R = 0.1$, the values of $\hat{P}^{(i)}_{e3e}$ (FPMM, 24HR+7DDR) change very slightly. Among the 12 values of $\hat{P}^{(i)}_{e3e}$ (FPMM, 24HR+7DDR), the largest relative difference is 0.036%. We conclude from this example that the values of $\hat{P}^{(i)}_{e3e}$ (FPMM, 24HR+7DDR) is not sensitive to the change of $\hat{\sigma}^2_R$. When $N = 50, n = 3, C = 1,000$, there are 22 trials (2.2%) with $\hat{\sigma}^2_R = 0$; when $N = 50, n = 10, C = 1,000$, there are 12 trials (1.2%) with $\hat{\sigma}^2_R = 0$. 

139
There are no trials where $\hat{\sigma}_r^2 = 0$ when $N = 200, n = 50, C = 1,000$ and $N = 444, n = 200, C = 1,000$.

### 7.2.1 Empirical Predictor of the Latent Value for a Randomly Selected Subject

Similar to Table 6.1, when $N = 7, n = 3, C = 1,000$, we compute the $\tilde{P}_{2e}^{(i)} (24\text{HR})$ and $\tilde{P}_{e2e}^{(i)} (24\text{HR}+7\text{DDR})$ under the usual mixed model, the $\tilde{P}_{3e}^{(i)} (24\text{HR})$ and $\tilde{P}_{e3e}^{(i)} (24\text{HR}+7\text{DDR})$ under the FPMM model for position $i$, separately. Correspondingly, we compute their expected empirical MSE ($mse_e^{(i)}$). In Table 7.4, we see the empirical MSEs of the empirical predictors based on the 24HR data are smaller both under the usual mixed model and under the FPMM model.

**Table 7.4 Average of the empirical predictor $\hat{P}^{(i)}$ and the empirical MSE of $\hat{P}^{(i)}$ ($N = 7, n = 3, C = 1,000$)**

<table>
<thead>
<tr>
<th>$i$</th>
<th>$\mu^{(i)}$</th>
<th>Average of empirical predictor</th>
<th>Empirical MSE of $\hat{P}^{(i)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\hat{P}_{2e}^{(i)}$ (24)</td>
<td>$mse_{2e}^{(i)}$ (24)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{P}_{e2e}^{(i)}$ (24+7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{P}_{3e}^{(i)}$ (24)</td>
<td>$mse_{3e}^{(i)}$ (24)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{P}_{e3e}^{(i)}$ (24+7)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>21.43</td>
<td>21.03</td>
<td>8.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19.84</td>
<td>15.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>21.51</td>
<td>9.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>21.42</td>
<td>11.14</td>
</tr>
</tbody>
</table>

Table 7.5 shows the results of different sampling settings by using the MSE over all positions. Still, when RPI < 5%, we consider the two empirical predictors to be equivalent. For both the usual mixed model and the FPMM model, compared with the combined 24HR+7DDR data, the empirical predictor based on the 24HR data has smaller or equivalent to smaller MSE.
Table 7.5 Comparisons of the empirical $mse^{(i)}_e$ and the RPI for different sampling settings ($C = 1,000$)

<table>
<thead>
<tr>
<th>Different sampling setting</th>
<th>Comparisons of empirical $mse^{(i)}_e$ and RPI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MM</td>
</tr>
<tr>
<td></td>
<td>$mse^{(2i)}<em>{e2e}$ $mse^{(2i)}</em>{e2e}$ (24) $mse^{(2i)}_{e2e}$ (24)</td>
</tr>
<tr>
<td>$N = 7, n = 3, \mu = 21.62 (F = 0.43)$</td>
<td>15.36 $&gt;$ 8.37 RPI: (84%)</td>
</tr>
<tr>
<td>$N = 50, n = 3, \mu = 27.45 (F = 0.06)$</td>
<td>16.53 $&gt;$ 13.75 RPI: (20%)</td>
</tr>
<tr>
<td>$N = 50, n = 10, \mu = 27.45 (F = 0.20)$</td>
<td>14.59 $&gt;$ 13.01 RPI: (12%)</td>
</tr>
<tr>
<td>$N = 200, n = 50, \mu = 25.88 (F = 0.25)$</td>
<td>15.57 $&gt;$ 13.91 RPI: (12%)</td>
</tr>
<tr>
<td>$N = 444, n = 200, \mu = 25.30 (F = 0.45)$</td>
<td>13.43 $\approx$ 13.43 RPI: (0%)</td>
</tr>
</tbody>
</table>

In conclusion, when variance components are unknown, to predict the latent value for a randomly selected subject, the empirical predictors $\hat{P}^{(i)}_{c2e}$ (MM) and $\hat{P}^{(i)}_{e3e}$ (FPMM) based on the 24HR data have better performance. This conclusion is different from the case that the predictors $\hat{P}^{(i)}_{c2e}$ (MM) and $\hat{P}^{(i)}_{c3e}$ (FPMM) based on the combined data have better performance when variance components are known in Section 7.1.1.

7.2.2 Empirical Predictor of the Latent Value of a Randomly Selected Subject, Given the Realized Subject is Subject $s$

Similar to Table 6.16, we compute the average of the empirical predictor $\hat{P}^{(i)}$ by subject $s$, $\hat{P}^{(i)}$ and the empirical expected MSE, $mse^{(i)}$ for each subject $s$, we obtain the results presented in Table 7.6 ($N = 7, n = 3, C = 1,000$). We see that, for most subjects except for subject $s = 7$, the $mse^{(s)}$ based on the combined data is larger.
Table 7.6 Estimates of the empirical predictor $\hat{P}^{(s)}$ and the empirical MSE of $\hat{P}^{(s)}$ by $s$ $(N = 7, n = 3, C = 1,000)$

<table>
<thead>
<tr>
<th>$s$</th>
<th>$c$</th>
<th>$\mu$</th>
<th>Empirical predictor $\hat{P}^{(s)}$</th>
<th>Empirical MSE of $\hat{P}^{(s)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>MM</td>
<td>FPMM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\hat{p}_{2e}^{(s)}$</td>
<td>$\hat{p}_{c2e}^{(s)}$</td>
</tr>
<tr>
<td>5</td>
<td>436</td>
<td>7.99</td>
<td>8.04</td>
<td>8.20</td>
</tr>
<tr>
<td>6</td>
<td>447</td>
<td>18.17</td>
<td>18.33</td>
<td>17.77</td>
</tr>
<tr>
<td>3</td>
<td>409</td>
<td>22.29</td>
<td>22.11</td>
<td>21.97</td>
</tr>
<tr>
<td>2</td>
<td>437</td>
<td>23.74</td>
<td>23.29</td>
<td>22.65</td>
</tr>
<tr>
<td>7</td>
<td>422</td>
<td>31.17</td>
<td>28.94</td>
<td>24.60</td>
</tr>
<tr>
<td>1</td>
<td>438</td>
<td>34.56</td>
<td>33.29</td>
<td>30.78</td>
</tr>
</tbody>
</table>

When we increase the population size, $N = 444, n = 200, C = 1,000$, for the usual mixed model, similar to Figures 7.1 and 7.2, we plot subject’s estimated standard deviation and subject’s estimated week bias standard deviation vs. subject’s latent value, respectively by the empirical predictor with smaller MSE (Figures 7.19 and 7.20). The two kinds of dots distribute randomly, indicating the better empirical predictor is not associated with subject’s estimated variance and subject’s estimated week bias variance.

Figure 7.19 Plot of subject’s estimated StdDev vs. subject’s latent value by the better empirical predictor $(N=444, n=200, C=1000, 24HR+7DDR, \text{mixed model})$
There are 30.0% subjects with smaller MSE of empirical predictor $\hat{P}_{2e}^{(s)}$ (MM, 24HR). Similar to Figures 7.3 to 7.6, we plot Figures 7.21 to 7.24. Figures 7.21 and 7.22 show similar cluster patterns to Figures 7.3 and 7.4 with around 30g/day of the cut point of subject’s latent value (except for a few subjects out of the cut point). Figures 7.23 and 7.24 show the empirical predictor $\hat{P}_{2e}^{(s)}$ (MM, 24HR) is better when the subject’s average response ratio is greater than 0.5 and subject’s latent value is less than 30g/day, or when the subject’s average response ratio is less than 0 and subject’s latent value is greater than 30g/day.
Figure 7.21 Plot of average daily saturated fat intake (7DDR) vs. subject's latent value by the better empirical predictor (N=444, n=200, C=1000, 24HR + 7DDR, mixed model)

Figure 7.22 Plot of average daily saturated fat intake (7DDR) vs. subject's latent value by the better or equivalent to the better empirical predictor (N=444, n=200, C=1000, 24HR + 7DDR, mixed model)
Using the estimator \( \hat{P}_1^{(s)} \) (24HR) to replace the subject’s latent value, Figures 7.25 and 7.26 show similar patterns to Figures 7.23 and 7.24. Using the percentile of the estimator \( \hat{P}_1^{(s)} \) (Figures 7.27 and 7.28), when the subject’s average response ratio is greater than 0.5 and the percentile of the estimator \( \hat{P}_1^{(s)} \) is less than 70, or when the subject’s average response ratio is less than 0 and the percentile of the estimator \( \hat{P}_1^{(s)} \) is
greater than 70, the empirical predictor $\hat{P}_{2e}^{(s)}$ (MM, 24HR) is better. A few subjects with the subject’s average response ratio greater than 0.5 and the percentile greater than 70, and with the subject’s average response ratio less than 0 and the percentile smaller than 70 show the empirical predictor $\hat{P}_{2e}^{(s)}$ (MM, 24HR) is better.
For the FPMM model, there are 30.9% subjects with smaller MSE of empirical predictor $\hat{P}_{3e}^{(s)}$ (FPMM, 24HR). Figures 7.29 to 7.32 show similar patterns to Figures 7.21 to 7.24. Figures 7.31 and 7.32 show the empirical predictor $\hat{P}_{3e}^{(s)}$ (FPMM, 24HR) is better when the subject’s average response ratio is greater than 0.4 and subject’s latent value is
less than 30g/day, or when the subject’s average response ratio is less than -0.15 and subject’s latent value is greater than 30g/day.
Using the estimator $\hat{P}_1^{(s)}$ (24HR), Figures 7.33 and 7.34 show similar patterns to Figures 7.31 and 7.32. Using the percentile of the estimator $\hat{P}_1^{(s)}$ (Figures 7.35 and 7.36), when the subject’s average response ratio is greater than 0.4 and the percentile is less than 70, or when the subject’s average response ratio is less than -0.15 and the percentile is greater than 70, the empirical predictor $\hat{P}_3^{(s)}$ (FPMM, 24HR) is better. Still, a few
subjects with the subject’s average response ratio greater than 0.4 and the percentile
greater than 70, and with the subject’s average response ratio less than -0.15 and the
percentile smaller than 70 show the empirical predictor $\hat{P}^{(c)}_{3c}$ (FPMM, 24HR) is better.
In conclusion, when variance components are unknown, we have similar suggestion as the case of known variance components. Based on the observed 24HR and 7DDR data, we suggest using the empirical predictor based on the combined data to predict the latent value of a randomly selected subject, given the realized subject is subject $s$, except that the empirical predictor based on the 24HR data should be used
when the subject’s average response ratio is greater than 0.5 (mixed model) or 0.4 (FPMM model) and the percentile of the $\hat{P}_1^{(s)}$ (24HR) is less than 70, or when the subject’s average response ratio is less than 0 (mixed model) or -0.15 (FPMM model) and the percentile of the $\hat{P}_1^{(s)}$ (24HR) is greater than 70.

7.3 Summary

In practice, such as for a specific subject in a clinic, with repeatedly observed saturated fat intake using 24HR and 7DDR instruments, respectively. How to predict the subject’s usual saturated fat intake? We refer to Figures 7.25, 7.33 and 6.8b to make decision. Based on the observed 24HR and 7DDR responses, the estimator $\hat{P}_1^{(s)}$ (24HR), subject’s average response ratio, and subject’s estimated standard deviation can be computed easily. Using the cut points of the estimator $\hat{P}_1^{(s)}$ (24HR) and the subject’s average response ratio, we first decide to use 24HR data or the combined data; then, when using the 24HR data, we further refer to Figure 6.8b to decide which predictor is best.

In this way, we obtain the best estimate for each specific subject. Furthermore, for the population of these subjects, the average of the best estimate for each subject is the best estimate for the population with the smallest MSE.
CHAPTER 8
CONCLUSION AND DISCUSSION

8.1 Conclusion

This study proposed alternative method of predicting a subject’s latent value of usual dietary intake using finite population mixed model with heterogeneous within-subject variances. In some settings, the FPMM predictor is better than the usual mixed model predictor, this study provided a guidance for their use in practice. We developed a straightforward linear model approach for biased panel data. After accounting for the bias in the 7DDR data, we compared the performance of predictors based on 24HR data and 24HR+7DDR data. We have mainly two conclusions based on this study that help to guide estimation of specific subject latent value. These conclusions apply to a specific setting where response is saturated fat intake among adults in Worcester, Massachusetts.

1) Based on the observed 24HR recall of individual saturated fat intake, we suggest using the predictor \( \hat{P}_3^{(i)} \) or the empirical predictor \( \hat{P}_{3e}^{(i)} \) derived under the FPMM to predict the latent value for a randomly selected subject. In practice, for the case to predict the latent value of a randomly selected subject, given the realized subject is subject \( s \), we suggest referring to Figures 6.8b and 6.9 or Figures 6.10 and 6.11 to decide the best estimator or empirical predictor. For a subject, after computing the percentile of the estimator \( \hat{P}_1^{(i)} \) and the percentile of the subject’s estimated standard deviation, the location of the subject in the figure can be decided, and then the corresponding best estimator or empirical predictor is indicated in Figure 6.10.
2) When both 24HR data and 7DDR data of individual saturated fat intake are available, to predict the latent value for a randomly selected subject, we suggest using the predictor under either the usual mixed model or the FPMM model based on the combined data when variance components are known. When variance components are unknown, we suggest using the 24HR data for the empirical predictor under the usual mixed model; and using either the 24HR data or the combined data for the empirical predictor under the FPMM model.

In practice, when predicting the latent value of a randomly selected subject, given the realized subject is subject $s$, after computing the percentile of the estimator $\hat{P}_1^{(i)}(24HR)$ and the subject’s average response ratio, we recommend using the empirical predictor based on the combined data except for the cases of the percentile of the estimator $\hat{P}_1^{(i)}(24HR)$ less than 70 and the subject’s average response ratio greater than 0.5 (MM) or 0.4 (FPMM), and the cases of the percentile of the estimator $\hat{P}_1^{(i)}(24HR)$ greater than 70 and the subject’s average response ratio less than 0 (MM) or -0.15 (FPMM).

3) In addition to these conclusions, this study proved that the expected value of the MM-BLUP, $\bar{P}_2^{(i)}$, is biased for the population mean when the weighted least squares estimator is conditional on samples and heterogeneous within-subject variance occurs. Table 8.1 summarizes the conclusions when variance components are unknown in practice. It can serve as a reference for practitioner.
Table 8.1 Summary of conclusions when variance components are unknown

- **For a randomly selected subject**
  1. Use the 24HR data
  2. FPMM empirical predictor \( \hat{P}_{3e}^{(i)} \) is the best

- **For a randomly selected subject \( | s \)**
  Step 1: Refer to Figures 7.25 or 7.33 to decide on using 24HR data or 24HR+7DDR data;
  Step 2: When using the 24HR data, refer to Figures 6.8b and 6.9 (or Figures 6.10 and 6.11) for choosing a more accurate estimator/predictor.

In summary, the contributions of this study are that: 1) it provided guidance for predicting subject’s latent value of usual saturated fat intake using two source data (24HR, 7DDR) via a mixed model framework; 2) it examined the performance of predictors conditional on sampled subject; 3) it showed that WLS estimator is a biased estimator of the average latent value of the population when heterogeneous within-subject variance occurs.

### 8.2 Discussion

The FPMM model does not require the assumption of parametric distribution for observed response. For this reason, it is an appealing model that is appropriate for skewed individual dietary intake data. The FPMM model avoids transforming and back-transforming the data. This study evaluated the predictor of saturated fat intake at the individual level. The use of subject’s label is important to interpret what is meant by a random effect and to track a subject’s heterogeneous measurement error. The FPMM can account for the two-stage sampling design and measurement with error. This avoids lack of fit. In the simulation study, we assume the number of days, \( D \) and the number of
weeks, $W$ are very large so that the sampling fractions for days and weeks are zero. We simulated 12 24HR recalls and 4 7DDR recalls for each selected subject in each sample. Thus, this study did not account for the more complex situations of unbalanced data in the 24HR or the 7DDR, or issues that may arise from non-ignorable missing data.

This study compared the performance of predictors of subject’s latent value based on the combined data and based on the 24HR data. Based on the simulation study, around 80% (known variance) or 70% (unknown variance) of the subjects $(N = 400, n = 200, c = 1000)$ have a more accurate predictor based on the combined data. For these subjects, the combination of two instruments provides more accurate prediction of subject’s latent value than repeated measures from one instrument. This may be not only due to more data used but also due to the 7DDR data catching more information of day-to-day variability in individual dietary intake, especially from occasionally consumed food. Since the computation of the predictor uses the observed 7DDR recall, the bias in the 7DDR data has an effect on the predictor. The merit of the combined data is offset by large bias of over-reported or under-reported dietary intake from 7DDR recall. For subjects with large bias, such as the two clusters indicated in Figure 7.25, the predictor based on 24HR data is better.

This study developed a simultaneous model for the combined data. It provided a simple approach for analyzing the multiple source data where bias may occur. The combination is to fuse 24HR response and 7DDR response into one response variable using a mixed model framework. This is different from the combination that the 7DDR serves as a covariate in a model. The combination method maintains the subject link that
is in the multiple sources of data. The combined data have more observations for each subject, and result in a parsimonious covariance structure.

The 7DDR model (3.70) has a different format from the model (2.3) used in previous studies related to adjust for bias of dietary measurement (Freedman 2011), but includes the same bias components (additive bias, multiplicative bias and subject-specific bias). In the model (2.3), \( T_i \) corresponds to the latent value, \( \mu_i \), where \( \mu_i = \mu + \beta_i \). After replacing \( T_i \) using \( \mu + \beta_i \), we have

\[
R_{ij} = \beta_0 + \beta_i T_i + r_i + e_{ij}
\]

\[
= \beta_0 + \beta_i (\mu + \beta_i) + r_i + e_{ij}
\]

\[
= \beta_0 + \beta_i \mu + \beta_i^2 + r_i + e_{ij}
\]

\[
= \mu + [\beta_0 + (\beta_i - 1) \mu] + \beta_i^2 + [r_i - 1\beta_i + r_i] + e_{ij}
\]

Compared with the model (3.70) \( \tilde{Y}_{swk}^* = \mu_s + \tilde{\delta}_s + \omega_{sw} + \tilde{\delta}_{sw} + F_{swk}^* \), we have \( \tilde{\delta}_s \) corresponding to \( \beta_0 + (\beta_i - 1) \mu \) and \( \tilde{\delta}_{sw} \) corresponding to \( (\beta_i - 1) \beta_i + r_i \). The system bias \( \beta_i \) is associated with subject’s latent value through the term \( \beta_i T_i \), correspondingly, we use the subject-specific deviation in 7DDR measure, \( \tilde{\delta}_s \). The term \( (\beta_i - 1) \beta_i + r_i \) contains random subject-specific bias, correspondingly, we use the week bias \( \tilde{\delta}_{sw} \) which is nested in \( \tilde{\delta}_s \).

One area investigated in this study is the use of a weighted estimator of the population mean. A weighted estimator is often used to estimate the population mean and is more accurate than the sample mean. Buonaccorsi (2006) discussed the conflict in the usual mixed model assumption that assumes fixed heterogeneous within-subject variance, but allowing subject’s effect is random. It is difficult to justify the better performance of
the weighted estimator with the theory when accounting for identifiable subjects. Stanek and Singer (2011) pointed out the conflict in the assumption brings in the artificial sample space and the partial use of subject’s label in the model. In this study, tracking subject’s effect and subject’s response error variance through the subject’s label, we follow the whole physical process from randomly selecting a subject to observing each sample and obtaining a predictor for a randomly selected subject. The weighting factor \( w^*_i \) is conditional on the sample. We proved that, using the predictor \( \hat{P}^{(i)}_2 \) derived under the usual mixed model to predict the latent value for a randomly selected subject, the \( \hat{P}^{(i)}_2 \) is a biased predictor of the population mean \( \mu \); and it is a unbiased predictor of the weighted population mean \( \mu^* \).

This study use the observed 24HR and 7DDR data of individual saturated fat intake which is a commonly consumed nutrient every day. Both 24HR and 7DDR are short-term dietary assessment instruments. The conclusions might not be appropriate for the usual intake of other episodically consumed nutrient. Future research is needed to examine other nutrients and evaluate optimal predictor of subject’s latent value. It is possible that inclusion of auxiliary variable into future analysis will improve the prediction of subject’s latent value.

Since there are only a few dietary biomarkers suitable for the epidemiological study (Hedrick et al. 2012), usually the 24HR is used as a reference instrument in many studies. We assumed the 24HR recall is unbiased and serves as a reference when adjusting for the bias of 7DDR recall. In fact, bias may exist in the 24HR recall based on some studies using the biomarker as the unbiased reference (Kipnis et al. 2003; Freedman et al. 2004). This study also assumed that the random measurement errors are not
correlated between different subjects and between different measures within subject. This assumption may not be reasonable in practice.

Finally, the results obtained in this study are based on a simulation study in the context of the Seasons study. This study applies the theories of finite population mixed model and mixed model to practical problem, it provided further understanding and interpretation when using these models for prediction. The conclusions of this study provide guidance on methods and model selection for estimating subject’s latent value in practice. Caution should be taken when using the conclusions with the understanding of assumptions and limitations.
BIBLIOGRAHY


Ockene, I. S. (1993), Grant Proposal of Seasons Study (NHLBI, R01-HL52745), Seasons study website: http://www.umass.edu/seasons/pdf/files/seasons-grant-proposal.PDF.


