Abstract

UMBACH, AMY THERESE. Bayesian Imputation Methods to Measure Quality of Life. (Under the direction of Dr. Sujit K. Ghosh)

The most widely used general health outcomes measure is the SF-36 Health Status questionnaire. The SF-36 is a 36 item general health survey which evaluates eight dimensions of health. This questionnaire is therapeutic non-specific. Often times, an analysis is done to determine if a subject's quality of life is better on one drug than another. This can be beneficial to the patient when selecting a drug and to the company when marketing a drug.

The SF-36 form is often used in clinical trials. One problem that is often encountered during a clinical trial is missing data. The industry standard for dealing with missing data of this type might not be the best. The industry standard of evaluating SF-36 data converts the data into eight score functions and treats the score functions as continuous data, even though they are discrete. We take a Bayesian perspective to obtain parameter estimates based on the posterior distribution of the model parameters. We employ Gibbs sampling to obtain simulation-based estimates. One of the practical advantages of our proposed method is that the MCMC method can be implemeted using `WinBUGS`. `WinBUGS` is a windows-based software package that is specialized for implementing MCMC-based analysis of full probability models. It allows the user to easily construct models and is available on the World Wide Web.

In this thesis, we begin by presenting background information for modeling SF-36 health survey data. We then develop the method of estimating missing responses in quality of life data, taking into account the ordering in the data. We present two simulation studies to validate our proposed method. This method is applied to data from a clinical trial conducted by GlaxoSmithKline Pharmaceutical company. The trial is an open-label, multinational, parallel group study to evaluate the impact of

oral Naratriptan 2.5mg on migraines.

It has been observed that people in different countries respond differently to the SF-36 questionnaire. In order to account for these differences, we conclude this thesis by fitting an ordinal response model with varying cut-points. One benefit of this type of model is that it allows one to compare treatments across countries.

Bayesian Imputation Methods to Measure Quality of Life

by

Amy T. Umbach

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

STATISTICS

in the

GRADUATE SCHOOL
at
NC STATE UNIVERSITY
2002

Dr. Sujit K. Ghosh, Chair	Dr. David A. Dickey
Dr. John F. Monahan	Dr. William H. Swallow

To my parents Nancy and Dale, my sister Erin,

and my special friend Ben Flynt,

without all of your love and support none of this would be possible.

Biography

Amy Umbach was born in Des Moines, Iowa in 1974 to Nancy and Dale Umbach. She moved to Muncie, Indiana in 1979. Amy earned her BS in Mathematics and Statistics in 1997 from Miami University in Oxford, Ohio. In the fall of 1997, she began graduate school at North Carolina State University in Statistics. She received her Master of Statistics in 1999 and her Ph.D. in 2003. While she was at NC State, she did an internship at GlaxoSmithKline Pharmaceutical Company for three years and taught an undergraduate statistics class for one year. She currently works for PharmaResearch in Wilmington, North Carolina.

Acknowledgements

All of the people in my life have made me who I am and have made it possible for me to be where I am today. I would like to thank and acknowledge a few of them.

First and foremost, I want to thank my parents. I feel very fortunate to have such wonderful parents. They have always encouraged me to go after my dreams and have done everything they can to help me achieve my goals. Having a dad as a statistician has been extremely helpful. He read my thesis and provided me with valuable comments. During my undergraduate and graduate course work, he was available to answer any statistical questions that I had. My mom really helped me through some of the tough times when I never thought I would finish.

I also want to thank my sister Erin. She has always been there for me. Having her in Charleston has been great. Going to visit her has helped me relax, take my mind off of school, and have fun.

I have made some great friends while in graduate school. Amy, Elizabeth, Kate, Kirsten, and Lisa have always reminded me that it is necessary to take some time away from studying. We have many great stories to tell. Without Elizabeth and Lisa in my study group, getting my masters would have been more difficult.

I would like to thank my Ph.D. study group, Palani, Kapil, and Prasheen. I couldn't imagine getting through inference and measure theory without them. I would also like to thank all the understanding students in the Hillsborough street office. Most of the students in the office tolerated my use of the PCs for my 8-12 hour simulation runs. I would especially like to thank Terry Byron for all of his computing help.

Special thanks to Kirsten for going to Europe with me. Even though our trip to Europe in May 2002 took me away from my research for three weeks, it was well

worth it. Not only did we have a great time, but I met a special friend, Ben, in Brussels. Ben has been very supportive during the short time that I have known him.

Four years ago, I found out that one of my childhood friends, Amy, was living one hour from Raleigh. We got together and it seemed like no time had passed since we had last seen each other (even though it had been eight years). Amy and I have had lots of good times together. Two years ago, I reunited with another friend, Stephanie. I have spent many weekends with her at her house on the lake in western North Carolina and at my apartment in Raleigh getting some much needed relaxation from school.

I would especially like to thank Dr. Cathy Barrows, my boss during most of my internship at Glaxo. She has been a very helpful and understanding mentor. During my internship, I learned about the pharmaceutical industry and decided that when I graduated I wanted to pursue a career as a biostatistician. Cathy also helped me get the data from Glaxo for my research.

There have been many teachers that have passed through my life. All of them have touched me in some way or another. Dr. Bolger was my favorite math professor at Miami. He is part of the reason that I majored in Math. Dr. See was my first statistics professor at Miami. She sparked my interest in statistics. All of my professors at NC State have not only been great teachers but they have also been very helpful. I would especially like to thank my committe members, Dr. Dickey, Dr. Monahan, and Dr. Swallow. They have been very helpful, supportive, and understanding. I definitely would not be where I am today without my advisor Dr. Ghosh. He provided much needed guidance and was very patient as I worked on my thesis.

Contents

List of Figures

List of Tables

Chapter 1

Quality of Life

1.1 Introduction

Quality of life data are often used in clinical trials. The data are usually collected through health surveys. Subjects are asked a number of questions about their health. The questionnaires are usually administered at least twice during a study, once at baseline and then at least one time while on drug. The data can then be analyzed to determine if the subject's quality of life is better on one drug than another. This can be beneficial when marketing a drug.

One problem that is often encountered in clinical trials is missing data. Another problem is that often the data are analyzed using statistical procedures that assume the response is continuous when in fact it is categorical or ordinal. This is the case with the SF-36 questionnaire (Ware et al., 1993). Thus, the underlying model assumptions could be incorrect and hence may produce biased results. Also, by treating the data as continuous, some information could be lost. For example, with many of the health survey questionnaires, the data are ordinal. In the case of the SF-36 analysis, the ordering information is not being used in the analysis.

The proposed research will involve developing a new method to estimate missing responses in health survey data. To test and develop our method, we will be focusing

on the SF-36 questionnaire. We then apply our method to data from a clinical trial conducted by GlaxoSmithKline. We will be analyzing the data from the clinical trial using statistical methods that treat the data as categorical. We will then compare our results with the results from the current method of SF-36 analysis.

In Section 1.2, we give background information about the SF-36 questionnaire. In Section 1.3, we review currently available literature and discuss the limitations of the current research. In Chapter 2, we present the categorical response models. We also give information about the data set from GlaxoSmithKline Pharmaceutical company. We discuss the results of our preliminary analysis of the GlaxoSmithKline data. This analysis was performed using data that contained no missing responses. In Chapter 3, we present the ordinal response models. We discuss the imputation process and present the imputation results for the Social Functioning subscale in Section 3.2. We compare the results from all the subscales to the results from the current method which uses two levels of imputation, the question level and the subscale level. The subscale imputation level uses Last Observation Carried Forward (LOCF) as the imputation method. At the end of Chapter 3 we present the results from two simulation studies to validate the imputation method. In Chapter 4, we present the results from the ordinal response model with varying cut-points. At the end of Chapter 4, we discuss future work.

The research that is presented in this thesis will be beneficial to health research. The imputation procedure is better than the current method and it is easy to implement. The use of varying cut-points allows one to compare treatment differences across populations. Also, we present a better way to analyze health survey data that takes into account the fact that the data is ordinal.

1.2 The SF-36 Health Survey Questionnaire

The SF-36 Health Status Questionnaire is a 36-item multipurpose health survey. The survey measures eight health concepts. These eight subscales are (Ware et al., 1993): Physical Functioning(PF), Role-Physical(RP), Bodily Pain(BP), General Health(GH), Vitality(VT), Social Functioning(SF), Role-Emotional(RE), and Mental Health(MH). The PF subscale consists of 10 questions and assesses limitations in physical activities because of health problems. The RP subscale consists of four questions and assesses limitations in usual role activities because of health problems. In the context of these questions, role activities apply to everyday responsibilities. Thus, the questions apply to people who work inside and outside the home. The BP subscale consists of two questions and assesses the extent to which pain interferes with everyday life, and the severity of the pain. The GH subscale consists of five questions and assesses physical health status. The VT subscale consists of four questions and assesses energy and fatigue. The SF subscale consists of two questions and assesses the limitations in social activities due to physical or emotional problems. The RE subscale consists of three questions and assesses limitations on usual role activities because of emotional problems. The MH subscale consists of five questions and assesses the major mental health dimensions of anxiety, depression, loss of emotional control, and psychological wellbeing. A sample of the SF-36 form can be found in Appendix A.

The SF-36 survey is one of the most widely used surveys to measure health outcomes. One of the reasons that this has become one of the most widely used health outcomes measure is because it is a generic measure. By a generic measure, we mean that it can be used across age, disease, or treatment groups. Generic measures assess health related quality of life outcomes. The SF-36 has been beneficial in comparing specific and general populations, comparing the relative burden of disease, screening patients, and differentiating the health benefits produced by a wide range of different

treatments. The SF-36 survey is a standardized and validated tool for monitoring the results of care and is used extensively in the USA and Europe. The eight subscales have been measured in the general non-patient population and in medical patients with a variety of conditions.

The questionnaire itself is public and can be found at `www.sf-36.com`. Also, a blank questionnaire can be found in Appendix A. Figure 1.1 shows which questions make up each subscale. The form is designed for use in a variety of settings. It can be used in surveys of general and specific populations, health policy evaluations, clinical trials, and clinical practice and research.

The form is designed to be administered in a variety of ways. It can be self-administered by people 14 years of age and older, or administered by trained interviewers in person or by telephone. This form is often used in the pharmaceutical industry. As more and more drugs become available, the results from health outcome measurements can assist a patient in deciding which drug to choose.

More than 1,000 articles documenting the use of the SF-36 have been published through 1998. These publications have been summarized in an annotated bibliography (Shiely, 1996). Among these references are a multitude of studies investigating different diseases and conditions, as well as different treatments undergone in various study designs. In the first SF-36 user manual, you can find information about the history and development of the SF-36, its psychometric evaluation, normative data, and studies of reliability and validity (Ware et al., 1993). McDowell and Newell (1996) offer one of the most complete independent accounts of SF-36 development.

The SF-36 has proven useful in measuring the health benefits produced by different treatments. The usefulness of the SF-36 is illustrated in articles describing more than 130 diseases and conditions (Shiely, 1996). However, when there are a number of missing observations, the current method of estimating missing responses might not be the best method. The current method consists of two levels of imputation. The first imputation level is described in the users manual (Ware et al., 1993). This method

consists of imputing values for individual question responses withing a subscale if 50 % or less of the questions in that subscale are missing for the subject. The imputed value is the mean of all the nonmissing question responses in that subscale for the subject. If more than 50% of the observations in the subscale are missing then the subscale score is set to missing. The second level of imputation consists of imputing missing subscale values. The current method of imputing missing subscales that is used in clinical trials is last observation carried forward.

The eight subscales that make up the SF-36 were selected from 40 concepts included in the Medical Outcomes Study (MOS) (Stewart, 1992). Most of the SF-36 items come from instruments that have been in use since the 1970s and 1980s (Stewart, 1992). Some of these instruments are the Health Perceptions Questionnaire and the General Psychological Well-Being Inventory (Ware, 2000). The MOS researchers selected and adapted questionnaire items from these and other sources. They also developed new measures. The resulting product was a 149-item Functioning and Well-Being Profile (FWBP) (Stewart, 1992). This is the source of the items on the SF-36. The wording and format for the questions was improved. The SF-36 form has been translated in more than 40 languages.

The questionnaire has 36 questions. The number of possible answers for each question ranges from two to six. The questions use ordinal scales. A patient is asked a question about his or her health and is asked to choose one and only one response from the two to six available. For some of the questions, a low value is given for a negative response. For other questions, a high value is given for a negative response. For example, Question 9e asks whether the person had a lot of energy in the past 4 weeks and assigns a value 1 if the answer is "Yes, all of the time". Question 9g asks whether the person felt worn out in the past 4 weeks and assigns a value of 1 if the answer is "Yes, all of the time". These responses cannot be compared directly. Therefore, we rescored Question 9e. If the person answered "Yes, all of the time", a value of 5 was assigned. Thus, the rescoring consisted of taking the number of

categories, adding one, and then subtracting the response value. The result is that for all of the questions, the higher the value, the less negative the response. Eleven questions were rescored in this manner. *The result is that for all questions, the higher the value, the better the quality of life.* The following questions were rescored: 1, 2, 6, 7, 8, 9a, 9d, 9e, 9h, 11b, and 11d.

1.3 Literature Review

Missing data in health research is a major problem. Mean or median imputation or last observation carried forward is frequently used. Multiple imputation is not yet used extensively. Proschan, McMahon, and Shih (2001) state that in clinical trials single imputation procedures are used due to their feasibility and simplicity. They also stated that much of the survey literature for multiple imputation involves complicated modeling using variables thought to be related to both the outcome and whether data are missing and that this complicated modeling seems invalid to non-statisticians. They discuss a method (WLP) proposed by Wittes et al. (1989) as a conservative approach. The letters WLP stand for Wittes, Lakatos, and Probstfield. The WLP method (opposite imputation) of dealing with missing binary data in clinical trials takes the observed proportion of events in the opposite arm as the imputation value for the proportion of events among missing data in one arm. The other WLP method (pooled imputation) takes the proportion of events in the pooled-observed data from both arms as the imputation value. They state that the WLP methods are the best estimates under conservative assumptions. Also, the WLP procedures are robust against deviations in the model. However, this is a single imputation method. Because of the good statistical properties of multiple imputation, more and more research is being done to discover ways to use multiple imputation techniques in health fields.

Zhou, Eckert, and Tierney (2001) used a cumulative logistic regression model to predict the missing values for a missing patient satisfaction question. They used a

cumulative logistic regression model since there were five ordered responses for each satisfaction question. There were many explanatory variables in the model. Due to the patterns of the missing data, multiple iterations of the procedure were necessary in order to impute all of the missing values. They used a multiple imputation method developed by Heitjan and Little (1991) and Rubin (1986). It is referred to as multiple imputation using predictive mean matching. They also compared the results to those using two single imputation methods and using only the complete data. All the imputation methods gave similar results but differed from the results using only the complete data. They attributed the lack in difference between single and multiple imputation methods to the fact that they were dealing with large numbers of observations. Thus, there is a need for better multiple imputation methods that can be applied to large data sets.

Gelman, King, and Liu (1998) use the Gibbs sampler to impute missing values. The data were obtained from a study of public opinion changes in the 1988 U.S. Presidential Election. The major purpose of the study was to examine changes in voter intentions over time for different subgroups of the population. However, not all of the questions were asked on all of the surveys. If imputations are available, then for the political scientists, the analyses would be simpler. One logistic model with all the variables could be fit to all the surveys. Without the imputations, not all of the variables could be included in the model. The Gibbs sampler was used to perform imputations at the single-survey level and parameters were estimated using information from all the surveys. They compared the available case analysis to the multiple-imputed analysis. They found that during the Republican convention, available polls did not ask questions about income or ideology. Thus, when analyzing the available case data, they had to skip these points. However, when the imputed data were analyzed, it was found that the different subgroups appeared to be moving together over time. The authors conclude that this makes sense politically. The results of this research show that the Gibbs sampler can be beneficial in imputing

missing responses in survey data when a large portion of the responses from post baseline surveys are missing.

Paulino and Pereira (1995) develop a Bayesian approach to the problem of incomplete categorical data under an informative general censoring pattern. The solution is based on Dirichlet priors for all the model parameters. The nonignorable model provides a general analysis of missing categorical data. However, it does not address one particular problem. Provided that the Dirichlet prior is not questioned, the method can be applied to any pattern of missing data without imposing any restrictions on the missing data mechanism. The method is applied to data that consist of the degree of sensitivity to dental caries, categorized into three risk levels: low, medium, and high. The implementation of the method resulted in the development of computational strategies for the evaluation of characteristics of interest as a main issue remaining open. Soares and Paulino (2001) is a follow up to the Paulino and Pereira (1995) paper. In this paper a Monte Carlo simulation approach based on an alternative parameterization is used to overcome the computational difficulties. This simulation method allows one to go beyond the computation of posterior expectations. This paper presents one way to overcome some of the computational difficulties. However, there is a need to do further research to develop ways to implement multiple imputation methods that are flexible and user friendly.

Due to the nice statistical properties of multiple imputation (Schafer, 1999) and the lack of multiple imputation methods in practice, there is also a need to create multiple imputation methods for missing data in surveys and health fields that are easy to implement. We will extend the methods of Soares and Paulino to the case of ordinal response models, where the cut-points are allowed to vary depending on covariates. To impute missing values, we will be using Bayesian imputation methods. In Chapters 3 and 4 we discuss the imputation process using Gibbs sampling. We use `WinBUGS` to do the Gibbs sampling because it is free and easy to use. Thus, the multiple imputation method discussed in Chapters 3 and 4 could be easily implemented

in health research.

1.3.1 Bayesian Imputation methods

We will be using Bayesian methods to impute missing values. We will first obtain the joint distribution of the parameters and the missing data, given the observed data and the modelling assumptions. Bayesian imputation of missing data is based on the following idea: Initial estimates for the missing data will be obtained using the complete data. By complete data, we mean subjects without any missing responses. These initial imputations will be used as starting points for Gibbs sampling. Gibbs sampling is a type of Markov Chain Monte Carlo (MCMC) simulation method. MCMC can be used as a method for evaluating expectations of the form $\mathrm{E}_\pi[f(\theta)] = \int f(\theta)\pi(\theta)d\theta$. In this expression, $f(\cdot)$ denotes a function of interest. In Monte Carlo integration, samples are drawn from $\pi(\cdot)$. These samples denoted by $(\theta^{(t)}, t = 1, \ldots, N)$ can be used to approximate the desired expectation. The expectation is approximated using

$$E[f(\theta)] \approx \frac{1}{N}\sum_{t=1}^{N} f(\theta^{(t)}). \tag{1.1}$$

Thus, the population mean of $f(\theta)$ is estimated using a sample mean. By the strong law of large numbers (SLLN), when the samples $\theta^{(t)}$ are independent and identically distributed, this approximation can be made fairly accurate by using a large sample size N. One thing to note is that N is not a fixed data sample size. It is the number of samples that you draw. Thus, it is controlled by the analyst. Unfortunately, many complicated models will not allow independent random draws. The main idea behind MCMC is to simulate realizations from a Markov chain having $\pi(\cdot)$ as its stationary distribution. This is one way of generating $\theta^{(t)}$. In this case the $\theta^{(t)}$ are not independent. The MCMC generates the $\theta^{(t)}$ by drawing samples throughout the support of $\pi(\cdot)$ in the correct proportions. There are a number of competing methods to simulate realizations from a Markov chain with $\pi(\cdot)$ as the stationary distribution. We will restrict our attention to a special case of single component Metropolis-Hastings,

the Gibbs sampler. The Gibbs sampler was given its name by Geman and Geman (1984) in the context of image restoration. To describe the algorithm defining the Gibbs sampler, we will first define the full conditional distribution. The full conditional distribution $\pi(\theta_i|\theta_{-i})$ is the distribution of the ith component of θ conditioning on all the remaining components. Note, θ has the distribution $\pi(\cdot)$. Thus, the full conditional distribution is:

$$\pi(\theta_i|\theta_{-i}) = \frac{\pi(\theta)}{\int \pi(\theta) d\theta_i} \propto \pi(\theta). \tag{1.2}$$

Given an arbitrary vector of starting values $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_k^{(0)})$, the first iteration of the Gibbs sampler proceeds by making random draws from the full conditional distribution as follows:

$$\begin{aligned}
\theta_1^{(1)} &\sim \pi(\theta_1|\theta_2^{(0)}, \ldots, \theta_k^{(0)}) \\
\theta_2^{(1)} &\sim \pi(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}) \\
&\vdots \\
\theta_j^{(1)} &\sim \pi(\theta_j|\theta_1^{(1)}, \ldots, \theta_{j-1}^{(1)}, \theta_{j+1}^{(0)}, \ldots, \theta_k^{(0)}) \\
&\vdots \\
\theta_k^{(1)} &\sim \pi(\theta_k|\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{k-1}^{(1)})
\end{aligned}$$

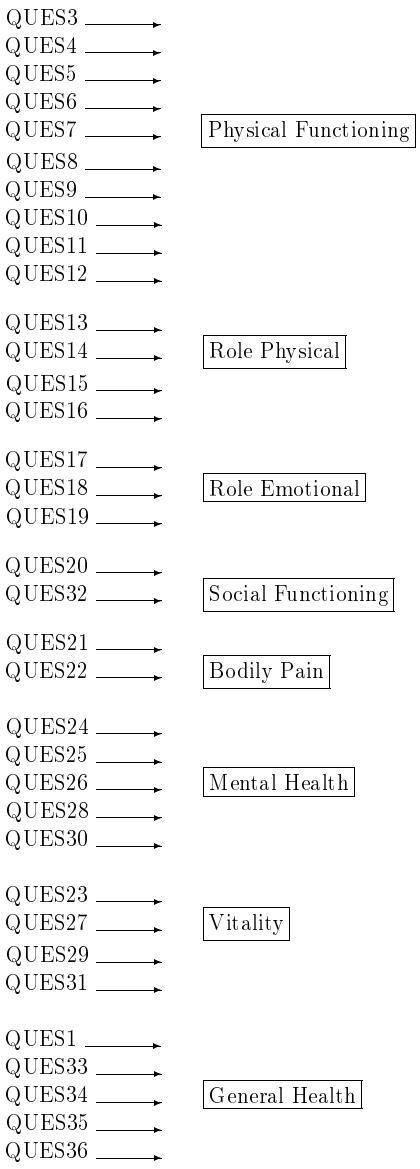
This completes a single cycle of the algorithm and defines a transition from $\theta^{(0)}$ to $\theta^{(1)} = (\theta_1^{(1)}, \ldots, \theta_k^{(1)})$. Thus, after t such iterations, we have $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_k^{(t)})$. The result will be a sequence $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots, \theta^{(t)}$ consisting of dependent draws. In describing the algorithm above, we updated the components of $\theta^{(t)}$ in a fixed order in which each component is updated in each iteration. While a fixed order is usual, it is not necessary. Often, the first portion of the simulated Markov chain is discarded in order to reduce the effect of the starting values. The portion that is discarded is referred to as the 'burn-in' iterations. Thus, if we let N denote the number of total iterations and we let M denote the number of 'burn-in' iterations, then $E[f(\theta)]$ can

be approximated using the ergodic average

$$\bar{f_N} = \frac{1}{N-M} \sum_{t=M+1}^{N} f(\theta^{(t)}). \tag{1.3}$$

The ergodic theorem ensures convergence of the ergodic average to the required expectation. The ergodic theorem states, if $\mathrm{E}_\pi[f(\theta)] < \infty$, and the Markov chain is ergodic, then $P[\bar{f}_n \to \mathrm{E}_\pi[f(\theta)]] = 1$, where $\mathrm{E}_\pi[f(\theta)] = \int f(\theta)\pi(\theta)d\theta$, the expectation of $f(\theta)$ is taken with respect to $\pi(\cdot)$. In defining the algorithm, we only focused on one chain. Raferty and Lewis (1992) created a method to diagnose convergence when using a single chain. Geweke (1992) also created a convergence diagnostic based on one chain. Running multiple chains is permissible. When multiple chains are run, the chains are run simultaneously with different starting values. When running multiple chains using `BUGS` (Spiegelhalter, 1996) or `WinBUGS` (Spiegelhalter, 1999), one should check the Gelman and Rubin (1992) convergence diagnostic. CODA (Best, 1995) can be used to obtain several graphical and numerical convergence diagnostics.

Figure 1.1: SF-36 plot



Chapter 2

Categorical Response Models

In this chapter, we present background information about modelling SF-36 health survey data. We analyze data from a clinical trial conducted by GlaxoSmithKline Pharmaceutical company using a multinomial model. We fit the model using the complete data. The complete data consists of subjects that answered all 36 questions for both baseline and Week 12.

2.1 Notations and assumptions

Before we present the models, we develop some notation that will be used throughout this chapter.

Let p_{ij} = Probability that a subject chooses answer j on question i, where $i = 1, \ldots, I_g$ and $j = 1, \ldots, J_g$.

Let N_{ij} = Number of subjects who answer j on question i, where $i = 1, \ldots, I_g$ and $j = 1, \ldots, J_g$.

For the SF-36 form:

g = 1, ..., 8.

g = 1 = PF. $I_1 = 10$ and $J_1 = 3$.

g = 2 = RP. $I_2 = 4$ and $J_2 = 2$.

g = 3 = RE. $I_3 = 3$ and $J_3 = 2$.

g = 4 = SF. $I_4 = 2$ and $J_4 = 5$.

g = 5 = BP. $I_5 = 2$ and $J_5 = 5, 6$.

g = 6 = MH. $I_6 = 5$ and $J_6 = 6$.

g = 7 = VT. $I_7 = 4$ and $J_7 = 6$.

g = 8 = GH. $I_8 = 5$ and $J_8 = 5$.

Ignoring the ordinal aspect of the response, a multinomial model can be used in general. As a patient chooses one of the J_g categories, it is reasonable to assume that

$$\mathbf{N_i} = (N_{i1}, \ldots, N_{iJ_g}) \sim \text{Multinomial}(N_i, p_{i1}, \ldots, p_{iJ_g}), \quad \text{where } N_i = \textstyle\sum_{j=1}^{J_g} N_{ij}, \quad i = 1, 2, \ldots, I_g, \text{ and } j = 1, 2, \ldots, J_g.$$

We reparameterize the p_{ij}'s as

$$p_{ij} = \frac{\mu_{ij}}{\sum_{j=1}^{J_g} \mu_{ij}} \quad \text{where } i = 1, 2, \ldots, I_g.$$

Where $\ln(\mu_{ij})$ = a linear model involving α_j and β_i, where α_j denotes the effect of response j and β_i denotes the effect of question i.

Equivalently, the above model can be expressed as a hierarchical Poisson model. We will assume that,

$$N_{ij} \sim \text{Poisson}(\mu_{ij}), \quad \text{where } i = 1, 2, \ldots, I_g \text{ and } j = 1, 2, \ldots, J_g.$$

Where $\ln(\mu_{ij})$ = a linear model involving α_j and β_i, where $i = 1, 2, \ldots, I_g$, and $j = 1, 2, \ldots, J_g$.

The assumptions are being given for two different models. The first model is a multinomial model. The second model is a Poisson transformation model. It can be shown that the likelihoods for both models are equivalent. Thus, by likelihood principle, either model could be used for analysis. To show that the two likelihoods are equivalent, we will first consider the multinomial likelihood. For simplification

purposes, we will assume $i = 1, 2, \ldots, 36$ and $j = 1, 2, \ldots, 5$. The likelihood of the multinomial model is given by,

$$L(\underline{\alpha}, \underline{\beta}) = \prod_{i=1}^{36}(\prod_{j=1}^{5} \frac{p_{ij}^{N_{ij}}}{N_{ij}!})N_j!$$

Recall from the reparameterization that $p_{ij} = \frac{\mu_{ij}}{\sum_{j=1}^{5} \mu_{ij}}$. Note that $\mu_{i.} = \sum_{j=1}^{5} \mu_{ij}$.

$$\propto \prod_{i=1}^{36} \prod_{j=1}^{5} (\frac{\mu_{ij}}{\mu_{i.}})^{N_{ij}}.$$

Now we will use the fact that $\ln(\mu_{ij}) = \alpha_j + \beta_i$.

$$\propto \prod_{i=1}^{36} \frac{e^{\sum_{j=1}^{5} N_{ij}(\alpha_j+\beta_i)}}{(\sum_{j=1}^{5} e^{\alpha_j+\beta_i})^{N_i}}. \tag{2.1}$$

On the other hand, the Poisson transformation likelihood is given by:

$$\begin{aligned} L(\underline{\alpha}, \underline{\beta}) &= \prod_{i=1}^{36}(\prod_{j=1}^{5} \frac{\mu_{ij}^{N_{ij}} e^{-\mu_{ij}}}{N_{ij}!}) \\ &= \prod_{i=1}^{36} \prod_{j=1}^{5} \frac{(\mu_{i1} e^{\alpha_j+\beta_i})^{N_{ij}} e^{-\mu_{i1} e^{\alpha_j+\beta_i}}}{N_{ij}!} \\ &\propto \prod_{i=1}^{36} \frac{(\mu_{i1}^{N_i} e^{\sum_{j=1}^{5} N_{ij}(\alpha_j+\beta_i)} e^{-\mu_{i1} \sum_{j=1}^{5} e^{(\alpha_j+\beta_i)}})}{N_{i1}!N_{i2}!N_{i3}!N_{i4}!N_{i5}!}. \end{aligned}$$

Let μ_{i1} have independent Gamma(a,b) distributions with mean $\frac{a}{b}$ and variance $\frac{a}{b^2}$. Next we will write the joint distribution and integrate out μ_{i1}.

$$\propto \prod_{i=1}^{36} \frac{e^{\sum_{j=1}^{5} N_{ij}(\alpha_j+\beta_i)}(N_i + a)!}{N_{i1}!N_{i2}!N_{i3}!N_{i4}!N_{i5}!(b + \sum_{j=1}^{5} e^{(\alpha_j+\beta_i)})^{N_i+a}}$$

Now letting $\mathrm{a} \to 0, \mathrm{b} \to 0$, we get

$$\propto \prod_{i=1}^{36} \frac{e^{\sum_{j=1}^{5} N_{ij}(\alpha_j+\beta_i)}}{(\sum_{j=1}^{5} e^{\alpha_j+\beta_i})^{N_i}}. \tag{2.2}$$

Since right hand side of Equation 2.1 equals that of Equation 2.2, then with certain restrictions, the Multinomial Model and the Poisson Transformation model are equivalent. Thus either model can be used in the analysis. When we use the model to

analyze the data, we will expand the notation to include two other subscripts. The subscript that will correspond to treatment is k, and the subscript that will correspond to session is l. When considering the linear model, these are variables that can be included in the model. The linear model consists of the main effects

1. α_j = The effect of response j
2. β_i = The effect of question i
3. λ_k = The effect of treatment k
4. γ_l = The effect of session l

and interactions effects,

1. $(\alpha\beta)_{ij}$ = The effect of response j while in question i
2. $(\alpha\lambda)_{kj}$ = The effect of response j while in treatment k
3. $(\alpha\gamma)_{lj}$ = The effect of response j while in session l
4. $(\beta\lambda)_{ik}$ = The effect of question i while in treatment k
5. $(\beta\gamma)_{il}$ = The effect of question i while in session l
6. $(\lambda\gamma)_{kl}$ = The effect of treatment k while in session l

Appropriate parameterization is necessary to improve convergence and stability of the samples in MCMC methods and for identifiability of parameters. In order to do this, we will consider parameter restrictions. These assumptions are being used as part of the model in order to avoid identifiability issues. Since we have fixed effects and non-informative priors, we will be using corner constraints. According to the `BUGS` manual (1995), this is an appropriate and common parameterization for this situation. This is similar in spirit to the constraints used in SAS. For the main effects, we are assuming $\beta_1 = 0$, $\lambda_1 = 0$, and $\gamma_1 = 0$. For the interaction effects, we are assuming a

value of zero for all corner points. For example, $\alpha\beta_{1j} = 0$ for all values of j. Another possibility, at the cost of increased computation time, is the sum to zero constraint.

2.2 Simulation Study

We created a data set consisting of 600 patients in two treatment groups. There were 300 patients in each group.

In this data set, each question has five possible answers. We will assume the questions are independent. This data set is used for estimation and testing. We obtained the maximum likelihood estimates and the Bayes estimates for the data. We ran 500 simulations and obtained the estimates, Monte Carlo standard errors, and coverage probabilities for both the maximum likelihood estimates and the Bayes estimates. The estimates are equal to the mean of the estimates from the 500 runs of the simulation. The MC standard errors are equal to the the sample standard errors for the estimates. The coverage probability is the proportion of times the true value was captured in the confidence or credible interval. For the active treatment, the true values of the p_i's follow: $p_1 = .0016$, $p_2 = .0256$, $p_3 = .1536$, $p_4 = .4096$, and $p_5 = .4096$. For the placebo treatment, the true values of the p_i's follow: $p_1 = .4096$, $p_2 = .4096$, $p_3 = .1536$, $p_4 = .0256$, and $p_5 = .0016$.

2.2.1 Maximum Likelihood Estimator

We compute the likelihood function for each treatment group. For simplification purposes let $\sum_{i=1}^{36} N_{ij} = N_{.j}$ and $\sum_{i=1}^{36}\sum_{j=1}^{5} N_{ij} = N$. Assuming the questions are independent, the likelihood function is
$\mathrm{L}(\underline{p}) = \prod_{i=1}^{36}(\prod_{j=1}^{5} \frac{p_j^{N_{ij}}}{N_{ij}!})N_j!$ where $\underline{p} = (p_1 \dots p_5)$.
Taking into account the constraint that $\sum_{j=1}^{5} p_j = 1$, the log-likelihood of $\underline{p}$ is given by $\ln(\underline{p}) = \text{const.} + \sum_{j=1}^{5} N_{.j}\log\, p_j$.
The above function is maximized by $\hat{p_j} = \frac{N_{.j}}{N}$ with variance estimate as

$\hat{\text{Var}}(\hat{p_j}) = \frac{\hat{p_j}(1-\hat{p_j})}{N}$.

To obtain the 95% CI for p_j, we use the fact that $\text{logit}(\hat{p_j}) \sim \text{Normal}(\text{logit}(p_j), \frac{1}{Np_j(1-p_j)})$. This is used to make the normal approximation more accurate. Thus, the 95% CI for $\text{logit}(p_j)$ is $\text{logit}(\hat{p_j}) \pm 1.96\sqrt{\frac{1}{N\hat{p_j}(1-\hat{p_j})}}$. From this 95% CI, we obtain the 95% CI for p_j. For simplification purposes, let $C_L = \text{logit}(\hat{p_j}) - 1.96\sqrt{\frac{1}{N\hat{p_j}(1-\hat{p_j})}}$. Also, let $C_U = \text{logit}(\hat{p_j}) + 1.96\sqrt{\frac{1}{N\hat{p_j}(1-\hat{p_j})}}$. Thus, the 95% CI for p_j is $(\frac{\exp(C_L)}{1+\exp(C_L)}, \frac{\exp(C_U)}{1+\exp(C_U)})$. For the simulated data set, the mle's, standard error estimates, 95% CI's, coverage probability, and Monte Carlo standard errors are:

Table 2.1: Maximum Likelihood Estimates for Active Treatment

Parameter	Estimate	sd	95% CI lower limit	95% CI upper limit	Coverage Probability
p_1^1	.0016	.0004	.0010	.0026	.9440
MC Error	4.06E-04	4.90E-05	3.13E-04	4.96E-04	
p_2^1	.0256	.0015	.0228	.0288	.9480
MC Error	1.54E-03	4.44E-05	1.45E-03	1.62E-03	
p_3^1	.1537	.0035	.1471	.1607	.9580
MC Error	3.41E-03	3.15E-05	3.35E-03	3.47E-03	
p_4^1	.4094	.0047	.4002	.4187	.9540
MC Error	4.68E-03	8.28E-06	4.66E-03	4.69E-03	
p_5^1	.4096	.0047	.4004	.4189	.9340
MC Error	4.74E-03	8.37E-06	4.72E-03	4.75E-03	

Table 2.2: Maximum Likelihood Estimates for Placebo Treatment

Parameter	Estimate	sd	95% CI lower limit	95% CI upper limit	Coverage Probability
p_1^2	.4098	.0047	.4005	.4191	.9380
MC Error	5.00E-03	8.82E-06	4.98E-03	5.01E-03	
p_2^2	.4095	.0047	.4002	.4188	.9400
MC Error	5.02E-03	8.88E-06	5.00E-03	5.04E-03	
p_3^2	.1537	.0035	.1470	.1606	.9400
MC Error	3.76E-03	3.48E-05	3.69E-03	3.82E-03	
p_4^2	.0255	.0015	.0227	.0287	.9540
MC Error	1.44E-03	4.17E-05	1.36E-03	1.52E-03	
p_5^2	.0016	.0004	.0010	.0025	.9700
MC Error	3.73E-04	4.57E-05	2.87E-04	4.58E-04	

From Tables 2.1 and 2.2, we observe that the parameter estimates are very similar to the true values. One other important observation from Tables 2.1 and 2.2 is that the coverage probabilities are close to 95% for all the parameters.

2.2.2 Bayes Estimator

We now turn to Bayesian analysis of the previous model. We choose to use Jeffrey's noninformative prior for $p_1, p_2, p_3, p_4,$ and p_5. In general, Jeffrey's prior, J(p), is proportional to [*Expected Fisher information* for p]$^{\frac{1}{2}}$. For our example, J($\underline{p}$) $\propto \frac{N^2}{(p_1p_2p_3p_4p_5)^{\frac{1}{2}}} = N^2 p_1^{-\frac{1}{2}} p_2^{-\frac{1}{2}} p_3^{-\frac{1}{2}} p_4^{-\frac{1}{2}} p_5^{-\frac{1}{2}}$ which is the kernal of a Dirichlet($\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}$) distribution. Thus, we use a Dirichlet distribution with parameter ($\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}$) for our analysis. `BUGS` will be used to get the Bayes estimates for the parameters. The posterior distribution is $\propto \prod_{j=1}^{5} p_j^{N.j-\frac{1}{2}}$. The Bayes estimator under squared error loss is given by the posterior mean E[p_j |data] $= \frac{N_{.j}+\frac{1}{2}}{N+\frac{5}{2}}$ and the variance is given by Var(p_j |data) $= \frac{(N_{.j}+\frac{1}{2})(N+N_{.j}+2)}{(N+\frac{5}{2})^2(N+\frac{7}{2})}$.

For the simulated data set, the Bayes estimates, standard error estimates, 95% Credible Intervals, and Monte Carlo errors are:

Table 2.3: Bayes Estimates for Active Treatment

Parameter	Estimate	sd	95% CI lower limit	95% CI upper limit	Coverage Probability
p_1^1	.0016	.0004	.0010	.0025	.9260
MC Error	4.05E-04	4.84E-05	3.11E-04	5.03E-04	
p_2^1	.0257	.0015	.0228	.0287	.9600
MC Error	1.54E-03	4.42E-05	1.45E-03	1.62E-03	
p_3^1	.1537	.0035	.1470	.1606	.9600
MC Error	3.41E-03	3.53E-05	3.34E-03	3.48E-03	
p_4^1	.4093	.0048	.4000	.4186	.9540
MC Error	4.67E-03	1.50E-05	4.65E-03	4.69E-03	
p_5^1	.4095	.0047	.4004	.4189	.9340
MC Error	4.73E-03	1.40E-05	4.72E-03	4.75E-03	

Table 2.4: Bayes Estimates for Placebo Treatment

Parameter	Estimate	sd	95% CI lower limit	95% CI upper limit	Coverage Probability
p_1^2	.4096	.0047	.4004	.4189	.9380
MC Error	5.00E-03	1.35E-05	4.98E-03	5.01E-03	
p_2^2	.4094	.0047	.4001	.4187	.9440
MC Error	5.02E-03	1.49E-05	5.01E-03	5.04E-03	
p_3^2	.1536	.0035	.1468	.1607	.9460
MC Error	3.76E-03	3.52E-05	3.69E-03	3.82E-03	
p_4^2	.0256	.0015	.0227	.0286	.9580
MC Error	1.44E-03	4.08E-05	1.36E-03	1.51E-03	
p_5^2	.0016	.0004	.0010	.0025	.9600
MC Error	3.72E-04	4.47E-05	2.86E-04	4.60E-04	

From Tables 2.3 and 2.4, we observe that the parameter estimates are very similar to the true values. One other important observation from Tables 2.3 and 2.4 is that the coverage probabilities are close to 95% for all the parameters. A t-test revealed that there were no significant differences between the estimates based on mles and the Bayes estimates. The above analysis gets complicated in the presence of missing data. We presented the frequentist and Bayesian analysis for illustrative purposes

only.

2.3 Application to GlaxoSmithKline data set

Our first step in developing the estimation of missing responses is to analyze the data from a clinical trial by GlaxoSmithKline Pharmaceutical company. The trial is an open-label, multinational, parallel group study to evaluate the impact of oral Naratriptan 2.5mg on quality of life, indirect costs and patient satisfaction. The study aimed to prospectively measure health related quality of life, indirect costs, and patient satisfaction for patients treating their migraines with either Naratriptan 2.5mg oral or their own customary therapy. Data were collected in two parts. For the first part, treatment lasted a duration of 12 weeks with patients treating all attacks. The primary endpoint was to assess quality of life using non-migraine SF-36 Health Survey. Patients aged 18-65 years, diagnosed as currently experiencing a migraine, and who had last taken triptan 3 months prior to entry into the study and fulfilled other entry criteria, were eligible for the study. The study was an open-label, parallel group study conducted in six countries: Canada, Finland, Hungary, the Netherlands, New Zealand, and Spain. Patients were either randomized to Naratriptan or customary therapy to treat all attacks including mild, moderate, or severe. The first phase lasted over a 3-month period. The first phase was followed by an optional 3-month open-label treatment period in which all patients participating received only treatment with Naratriptan (Phase 2). For this section, we will only be using the data from Phase 1. During the study, the SF-36 quality of life questionnaire was completed at Weeks 1 (baseline) and 12 (end of Phase 1). Of the 966 patients who were randomized into the study, 480 were assigned to take their own customary treatment and 486 were assigned to take Naratriptan 2.5mg. Of these subjects, 478 in the customary treatment group and 484 in the Naratriptan treatment group filled out the SF-36 form at baseline. Treatment lasted a duration of 12 weeks with patients

treating all attacks. Of the 966 patients, 955 used study medication to treat at least one attack. We start by creating a data set that contains complete evaluable subjects. To be considered a complete evaluable subject, a subject must have answered all 36 questions for both Baseline and Week 12. In this data set, there are 391 subjects in Customary Therapy treatment group and 373 subjects in Naratriptan treatment group. This data was analyzed to determine if there were significant differences in the two treatment groups within each subscale.

2.4 Implementation of Bayes models

`WinBUGS` Version 1.3 was used to implement our Markov Chain Monte Carlo analysis. In order to submit a more efficient model in `WinBUGS`, the multinomial-Poisson transformation will be used. All unknown α's, β's, γ's, and λ's are initially given independent noninformative priors. The choice that we used for all our parameters was a normal flat prior. For each subscale, Gibbs sampling was used. In our Gibbs sampling, we ran three chains consisting of 255,000 iterates. The first 5000 iterates from each chain were discarded. These were considered a 'burn-in' period. Gelman-Rubin statistic plots were monitored to verify that a sufficient 'burn-in' period was used. Then, with the remaining iterates, every fifth value was used. Using every fifth value is not a requirement. However, to compute MC error of posterior estimates from `WinBUGS`, it is reasonable to use approximately independent samples from MCMC. Thus 50,000 iterates were used from each chain giving a sample size of 150,000 for analysis. For each model, it took an average of 200 seconds to run on a Pentium IV PC. Within each subscale, three chains were run for each model. We started out with a model consisting only of the main effects. Then insignificant terms were dropped and interactions were tested. The deviance was also monitored. The deviance is equal to -2log(likelihood). For a saturated model with all the main effects and all

the interactions, the deviance $= -2\sum_{ijkl} N_{ijkl} log(\frac{N_{ijkl}}{(\sum_{j=1}^{J_g} N_{ijkl})p_{ijkl}})$. A final model was selected for each subscale based on the lowest deviance. From the final model, the posterior mean, sd, median, 95% credible interval, and MC error are reported.

2.4.1 Physical Functioning

The final model for this subscale took 140 seconds to run. The best fitting log-linear model is: $\log(\mu_{jk}) = \alpha_j + \lambda_k + (\alpha\lambda)_{kj}$.

Since, λ_k and $(\alpha\lambda)_{kj}$ were found significant in the model, we can conclude that treatment has an effect on limitations in physical activities because of health problems. The results of this model are:

parameter	mean	sd	MC error	2.5%	median	97.5%
α_1	5.996	0.0500	3.821E-4	5.897	5.996	6.092
α_2	7.463	0.0239	6.563E-5	7.415	7.463	7.509
α_3	8.745	0.0126	3.368E-5	8.720	8.745	8.769
λ_2	-0.205	0.0746	8.130E-4	-0.351	-0.205	-0.060
$(\alpha\lambda)_{2,2}$	0.091	0.0824	8.376E-4	-0.070	0.090	0.252
$(\alpha\lambda)_{2,3}$	0.240	0.0767	8.329E-4	0.091	0.240	0.390
deviance	61.15	3.457	1.370E-2	56.4	60.5	69.64

2.4.2 Role Physical

The final model for this subscale took 140 seconds to run. The best fitting log-linear model is: $\log(\mu_{jl}) = \alpha_j + \gamma_l + (\alpha\gamma)_{lj}$.

Since no terms involving λ_k were found significant in the model, we can conclude that there is not sufficient evidence to say that treatment has an effect on limitations in usual role activities because of physical health problems. The results of this model are:

parameter	mean	sd	MC error	2.5%	median	97.5%
α_1	7.014	0.0300	9.854E-5	6.955	7.015	7.073
α_2	7.572	0.0227	5.995E-5	7.527	7.572	7.616
γ_2	-0.154	0.0442	1.666E-4	-0.241	-0.154	-0.068
$(\alpha\gamma)_{2,2}$	0.233	0.0542	2.077E-4	0.127	0.233	0.340
deviance	40.45	2.833	8.036E-3	36.93	39.81	47.6

2.4.3 Role Emotional

The final model for this subscale took 140 seconds to run. The best fitting log-linear model is: $\log(\mu_{jl}) = \alpha_j + \gamma_l + (\alpha\gamma)_{lj}$.

Since no terms involving λ_k were found significant in the model, we can conclude that there is not sufficient evidence to say that treatment has an effect on limitations in usual role activities because of emotional problems. The results of this model are:

parameter	mean	sd	MC error	2.5%	median	97.5%
α_1	6.438	0.0400	1.508E-4	6.359	6.438	6.516
α_2	7.418	0.0245	6.764E-5	7.370	7.418	7.466
γ_2	-0.330	0.0618	3.014E-4	-0.451	-0.329	-0.209
$(\alpha\gamma)_{2,2}$	0.430	0.0706	3.465E-4	0.291	0.430	0.569
deviance	38.83	2.823	8.199E-3	35.32	38.19	45.98

2.4.4 Social Functioning

The final model for this subscale took 140 seconds to run. The best fitting log-linear model is: $\log(\mu_j) = \alpha_j$.

Since no terms involving λ_k were found significant in the model, we can conclude that there is not sufficient evidence to say that treatment has an effect on limitations in social activities because of physical or emotional problems. The results of this model are:

parameter	mean	sd	MC error	2.5%	median	97.5%
α_1	3.597	0.1648	4.081E-4	3.262	3.602	3.907
α_2	5.179	0.0750	1.859E-4	5.029	5.180	5.323
α_3	6.404	0.0407	1.12E-4	6.324	6.405	6.483
α_4	6.966	0.0307	7.868E-5	6.905	6.966	7.025
α_5	7.069	0.0291	7.717E-5	7.012	7.070	7.126
deviance	43.42	3.158	7.774E-3	39.26	42.77	51.25

2.4.5 Mental Health

The final model for this subscale took 143 seconds to run. The best fitting log-linear model is: $\log(\mu_{jk}) = \alpha_j + \lambda_k + (\alpha\lambda)_{kj}$.

Since, λ_k and $(\alpha\lambda)_{kj}$ were found significant in the model, we can conclude that treatment has an effect on psychological distress and well-being. The results of this model are:

parameter	mean	sd	MC error	2.5%	median	97.5%
α_1	4.151	0.1255	1.480E-3	3.897	4.154	4.391
α_2	5.138	0.0765	2.062E-4	4.986	5.139	5.286
α_3	6.162	0.0459	1.221E-4	6.071	6.163	6.251
α_4	6.655	0.0360	9.546E-5	6.584	6.655	6.725
α_5	7.213	0.0271	7.163E-5	7.160	7.214	7.266
α_6	6.970	0.0306	7.761E-5	6.910	6.97	7.030
λ_2	-0.701	0.2165	4.354E-3	-1.128	-0.698	-0.284
$(\alpha\lambda)_{2,2}$	0.543	0.2441	4.392E-3	0.069	0.541	1.025
$(\alpha\lambda)_{2,3}$	0.513	0.2274	4.389E-3	0.074	0.511	0.963
$(\alpha\lambda)_{2,4}$	0.635	0.2226	4.381E-3	0.208	0.633	1.075
$(\alpha\lambda)_{2,5}$	0.719	0.2200	4.379E-3	0.295	0.716	1.154
$(\alpha\lambda)_{2,6}$	0.683	0.2206	4.372E-3	0.256	0.681	1.120
deviance	105.6	4.901	2.268E-2	97.97	104.9	116.9

2.4.6 Vitality

The final model for this subscale took 148 seconds to run. The best fitting log-linear model is: $\log(\mu_{jk}) = \alpha_j + \lambda_k + (\alpha\lambda)_{kj}$.

Since, λ_k and $(\alpha\lambda)_{kj}$ were found significant in the model, we can conclude that treatment has an effect on energy and fatigue. The results of this model are:

parameter	mean	sd	MC error	2.5%	median	97.5%
α_1	4.505	0.1054	9.577E-4	4.294	4.507	4.707
α_2	5.536	0.0628	1.63E-4	5.411	5.537	5.657
α_3	6.481	0.0392	1.007E-4	6.403	6.481	6.557
α_4	6.803	0.0333	8.833E-5	6.737	6.803	6.867
α_5	6.902	0.0317	8.534E-5	6.840	6.902	6.964
α_6	5.453	0.0656	1.768E-4	5.322	5.454	5.579
λ_2	-0.471	0.1709	2.431E-3	-0.810	-0.500	-0.140
$(\alpha\lambda)_{2,2}$	0.266	0.1948	2.482E-3	-0.114	0.265	0.650
$(\alpha\lambda)_{2,3}$	0.371	0.1800	2.459E-3	0.022	0.370	0.728
$(\alpha\lambda)_{2,4}$	0.506	0.1772	2.466E-3	0.163	0.505	0.858
$(\alpha\lambda)_{2,5}$	0.470	0.1765	2.456E-3	0.127	0.469	0.820
$(\alpha\lambda)_{2,6}$	0.329	0.1959	2.492E-3	-0.052	0.328	0.716
deviance	104.6	4.911	1.93E-2	96.99	103.9	116.0

2.4.7 General Health

The final model for this subscale took 274 seconds to run. The best fitting log-linear model is: $\log(\mu_{jk}) = \alpha_j + \lambda_k + \gamma_l + (\alpha\lambda)_{kj}$.
Since, λ_k and $(\alpha\lambda)_{kj}$ were found significant in the model, we can conclude that treatment has an effect on general health perceptions. The results of this model are:

parameter	mean	sd	MC error	2.5%	median	97.5%
α_1	4.114	0.0912	7.678E-4	3.931	4.115	4.288
α_2	5.290	0.0515	1.375E-4	5.188	5.29	5.389
α_3	6.123	0.0349	9.662E-5	6.054	6.123	6.191
α_4	6.522	0.0294	7.934E-5	6.464	6.522	6.579
α_5	6.325	0.0321	8.697E-5	6.261	6.325	6.387
λ_2	-0.312	0.1386	1.752E-3	-0.585	-0.312	-0.040
γ_2	0.000	0.0229	6.24E-5	-0.045	-0.000	0.045
$(\alpha\lambda)_{2,2}$	0.337	0.1556	1.788E-3	0.032	0.337	0.642
$(\alpha\lambda)_{2,3}$	0.305	0.1465	1.781E-3	0.018	0.305	0.594
$(\alpha\lambda)_{2,4}$	0.305	0.1438	1.776E-3	0.024	0.305	0.588
$(\alpha\lambda)_{2,5}$	0.176	0.1453	1.777E-3	-0.110	0.176	0.461
deviance	165.1	4.665	1.784E-2	157.9	164.4	175.9

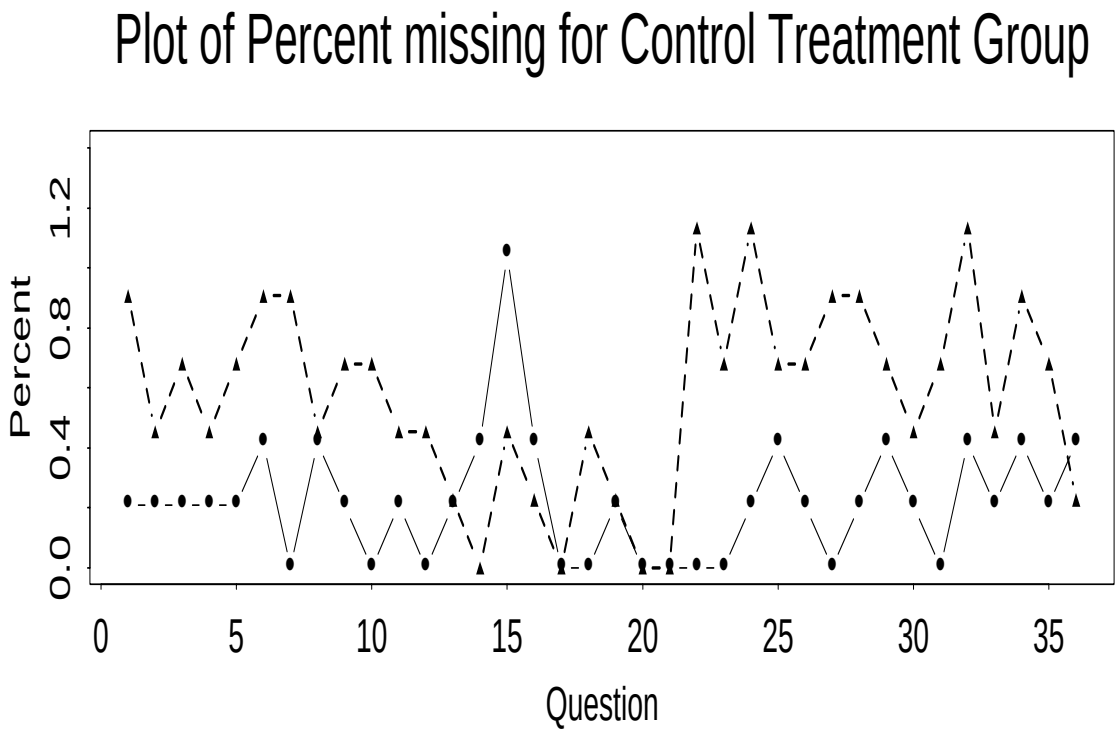
For all of the subscales the response effect was significant. Thus we can conclude that response has an effect on quality of life. However, we did not take into account the natural ordering of the responses. Thus, we need to incorporate this information. We can do this directly in the way we construct cumulative links in the ordinal response model. This model is discussed in Chapter 3. First we discuss what happens if missing values are ignored.

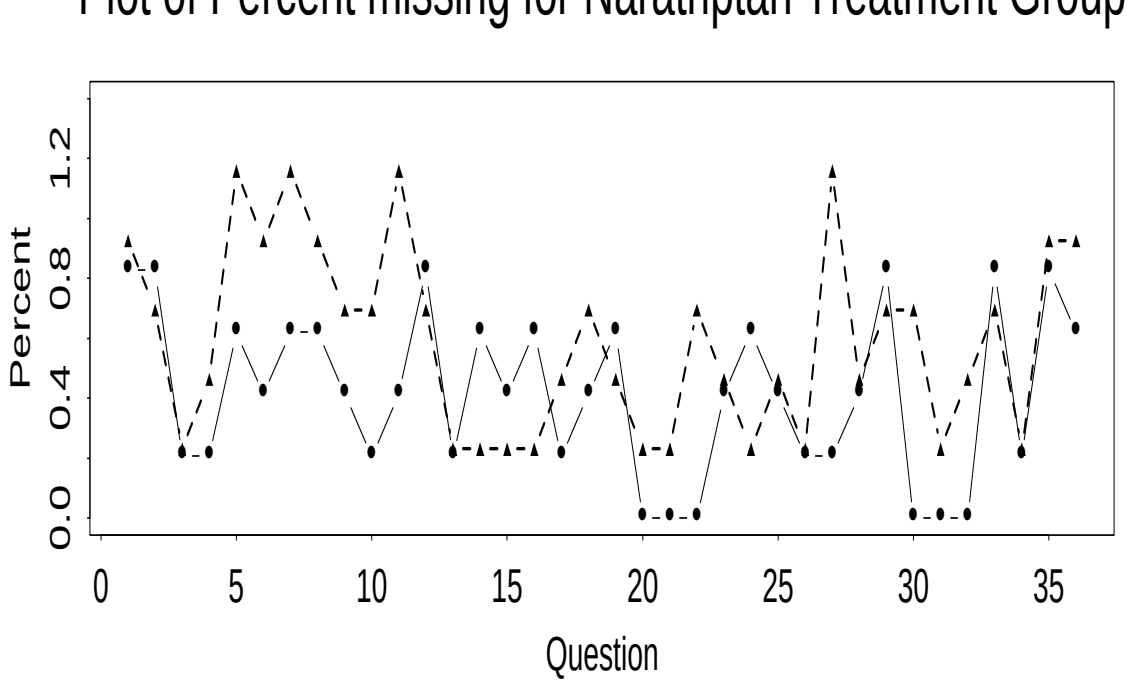
2.5 Missing Data

Schafer (1999) states that ignoring missing responses can be inefficient and may introduce bias if the data are not missing at random. He also states that a multiple imputation method can be highly efficient. Also one set of imputations may be used for many analyses. Before we discuss the imputation method we will be using, we

first need to investigate the missingness. As seen in Figure 2.1, the data seem to be missing at random. There does not seem to be any type of pattern in these plots. One thing to notice is that Week 12 has a greater percentage of missed values than Baseline for some of the questions. It turns out that the percentage of times that Week 12 is missed more than Baseline is about 58%.

Figure 2.1: Plots of percent missing where the dashed line represents week 12 and the solid line represents Baseline





Chapter 3

Ordinal Response Model

3.1 Introduction

In Chapter 2 we fit a multinomial model to the data. However, the responses have an underlying order. The higher the number, the better the quality of life. Due to this natural ordering in the responses, we decided to fit an ordinal model to the data. In this chapter we discuss the ordinal model. Also in Chapter 2, we discussed why missing responses should not be ignored. In this chapter we discuss the imputation process and present imputation results for missing responses. We also compare our results to the results from the current method of analyzing the data that uses last observation carried forward (LOCF) as the subscale imputation method. We conclude this chapter by presenting the results from two simulations to validate the imputation method.

First we develop some notation to be used throughout this chapter.

- Let Y_i denote an underlying latent continuous response variable having cdf $G_i(\cdot)$. Suppose $-\infty = a_0 < a_1 < \ldots < a_J = \infty$ are the "cut-points", such that R_i, the observed ith ordinal response in the data, satisfies $R_i = j$ if $a_{j-1} < Y_i \leq a_j$, where $j = 1, \ldots, J$.

- When the underlying response value Y_i falls in the jth interval $(a_{j-1}, a_j]$, then $P[R_i = j] = P[a_{j-1} < Y_i \leq a_j] = G_i(a_j) - G_i(a_{j-1})$.
  In the presence of covariates, we assume that $G_i(a_j) = G(a_j - \beta' x_i)$ for some link function $G(\cdot)$.

- The appropriate model for R_i applies link function $G(\cdot)$.
  Possible cumulative link functions:
  logit link, $G(t) = [1 + e^{-t}]^{-1}$,
  probit link, $G(t) = \Phi(t)$, where Φ denotes the cdf of N(0,1),
  log-log link, $G(t) = 1 - e^{-e^t}$.

- Note that $P[R_i = j] = G(a_j - \beta' x_i) - G(a_{j-1} - \beta' x_i)$,
  where $j = 1, \ldots, J$ such that $-\infty = a_0 < a_1 < \ldots < a_J = \infty$.

- In our model, we take $x_i' =$ (question[i], treatment[i], session[i])$'$ and $\beta' =$(QUES, TRT, SESS) such that
  $\beta' x_i =$QUES[question[i]] + TRT[treatment[i]] + SESS[session[i]].

The likelihood function of β and $(a_1, \ldots, a_{J-1})$ is given by
$L(\beta, \underline{a}) = \prod_{i=1}^{I} (G(a_{R_i} - \beta' x_i) - G(a_{R_i - 1} - \beta' x_i))$.
To choose between the link functions, we use the deviance = -2log$L(\beta, \underline{a})$.

3.2 Discussion

3.2.1 Model fitting ignoring missing values

The purpose of this section is to choose the link function to be used throughout this chapter and to show that results based on MLE and Bayes method are compatible. SAS and `BUGS` were used to fit the ordinal models to the complete data using both the logit link and the probit link. For `BUGS`, all unknown a's, SESS's, TRT's, and QUES's are initially given independent noninformative priors. The choice that

we used for all our parameters was an improper flat prior on the whole real line. Note that the posterior distribution is proper. For each health subscale, Gibbs sampling was used. In our Gibbs sampling, we ran three chains consisting of 10,000 iterates. The first 4000 iterates from each chain were discarded. These were considered a 'burn-in' period. Thus 6000 iterates were used from each chain giving a sample size of 18,000 for analysis. The Gelman Rubin diagnostic plots were examined to justify the use of 4000 iterates as the 'burn-in'. Figure 3.1 shows the estimate results with the logit link from SAS and `BUGS` on the same plot. Appendix B lists the estimates and their confidence intervals from SAS and `BUGS`. As seen in Figure 3.1, there is not a significant difference between the two devices. The estimates from the two softwares are on top of each other. The SAS results comparing the logit link and probit link are:

Group	Logit Link Log Likelihood	Probit Link Log Likelihood
1	-7603.0560	-7638.6766
2	-3878.6821	-3878.6616
3	-2450.9023	-2450.8747
4	-3882.9301	-3885.2324
6	-10494.1380	-10487.7647
7	-9200.1049	-9193.3474
8	-10446.5128	-10464.5585

For most of the groups, either the logit link gives a better model or both models are equivalent based on the log likelihood criteria. Thus the logit link was used for all of the models.

3.2.2 Bayesian Imputation Method

In this section, we extend the method to analyze data with missing responses. Gibbs sampling was used to impute values for the missing responses. If the logit link is used then the steps are:

1. Get initial estimates for the missing data and the parameters. For all parameters, except the a_j parameters, initial values of zero were chosen. For the a_j parameters, a combination of negative and positive values were used in increasing order. `WinBUGS` generated initial values for the missing responses.

2. Generate $Y_i^{(t)}$'s, where $Y_i^{(t)} \sim TL(x_i'\beta^{(t-1)}, 1)I(a_{R_i-1}^{(t-1)}, a_{R_i}^{(t-1)})$ and
   TL represents the truncated logistic distribution.

3. Update the $a_j^{(t)}$'s, where $a_j^{(t)} \sim U(L_j^{(t)}, U_j^{(t)})$, $L_j^{(t)} = max\{Y_i^{(t)}|R_i^{(t-1)} = j\}$, and $U_j^{(t)} = min\{Y_i^{(t)}|R_i^{(t-1)} = j-1\}$.

4. Impute the $R_i^{(t)}$'s as $R_i^{(t)} = \sum_{j=1}^{J} jI(a_{j-1}^{(t)} < Y_i^{(t)} \leq a_j^{(t)})$.

5. Sample $\beta^{(t)}$. This needs to be done using a Metropolis-Hastings step since the full conditional is not a standard distribution.

These steps using the logit link were carried out in `WinBUGS` using the code in Appendix C.

3.2.3 Social Functioning

This dataset consists of 3848 entries. There are 962 patients and each patient has a response for each question at each session. This subscale consists of 2 questions. There were 187 missing responses. Results from `WinBUGS`:

Table 3.1: Social Functioning Results

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1]	-4.312	0.1495	4.388E-3	-4.604	-4.309	-4.023
a[2]	-2.567	0.0849	3.532E-3	-2.724	-2.569	-2.395
a[3]	-1.008	0.0685	3.155E-3	-1.140	-1.009	-0.871
a[4]	0.457	0.0673	3.029E-3	0.325	0.455	0.591
QUES[1]	0.205	0.0614	1.560E-3	0.084	0.205	0.325
SESS[1]	-0.291	0.0619	1.730E-3	-0.411	-0.291	-0.169
TRT[1]	-0.050	0.0619	1.668E-3	-0.171	-0.049	0.072
resp[7]	4.102	0.9505	7.168E-3	2.0	4.0	5.0
resp[8]	4.020	0.9735	6.923E-3	2.0	4.0	5.0
resp[59]	4.115	0.9521	7.502E-3	2.0	4.0	5.0
resp[60]	4.009	0.9815	7.466E-3	2.0	4.0	5.0
resp[67]	4.151	0.9300	7.150E-3	2.0	4.0	5.0
resp[68]	4.032	0.9688	7.190E-3	2.0	4.0	5.0
resp[75]	4.135	0.9257	6.675E-3	2.0	4.0	5.0
resp[76]	4.032	0.9741	7.606E-3	2.0	4.0	5.0
resp[79]	4.133	0.9410	7.062E-3	2.0	4.0	5.0
resp[80]	4.033	0.9684	6.962E-3	2.0	4.0	5.0
response[95]	4.142	0.9419	7.525E-3	2.0	4.0	5.0
response[96]	4.045	0.9665	7.732E-3	2.0	4.0	5.0
response[147]	4.124	0.9436	7.041E-3	2.0	4.0	5.0
response[148]	4.028	0.973	6.683E-3	2.0	4.0	5.0
response[155]	4.118	0.9461	6.865E-3	2.0	4.0	5.0
response[156]	4.009	0.9788	7.240E-3	2.0	4.0	5.0
response[159]	4.107	0.9422	6.678E-3	2.0	4.0	5.0
response[160]	4.009	0.9673	7.281E-3	2.0	4.0	5.0
response[239]	4.141	0.9378	7.269E-3	2.0	4.0	5.0
response[240]	4.032	0.9634	7.298E-3	2.0	4.0	5.0
response[242]	3.853	1.0150	8.385E-3	2.0	4.0	5.0
response[295]	4.125	0.9383	6.556E-3	2.0	4.0	5.0
response[296]	4.006	0.9756	7.090E-3	2.0	4.0	5.0

The tables will look similar for all the subscales. The density plots for some of the missing parameters can be seen in Figure 3.2. The tables for the other subscales can be seen in Appendix D. The plots of the imputation values can be seen in Figure 3.3. This figure consists of a bar plot for each subscale. The possible response values for each subscale are listed on the x-axis. The height of each bar is the proportion of times that particular response value was chosen as the imputed value. One thing to note from Figure 3.3 is that for the Role Physical, the Role Emotional, and the Social Functioning subscales, only one value was chosen for the imputed value.

With the current method of analyzing the data, the first thing that is done is some rescoring of the questions. This can be somewhat complex depending on the subscale. For a detailed description of the rescoring that is done, refer to the SF-36 Health Survey Manual (Ware 1993). Next as the data are analyzed, two imputation levels of single imputation are carried out. The first level consists of imputing values for individual question responses within a subscale. If 50% or less of the questions in that subscale are missing then the missing responses are set to the mean of all the non-missing responses. After this imputation is done, subscale scores are calculated for each subject at each session. If more than 50% of the questions in that subscale are missing then the subscale score is set to missing. Otherwise, a subscale score is calculated by summing over all questions in that subscale and converting the sum onto a scale from 0-100. Then, the second imputation is carried out. For subjects with missing subscale scores, these are imputed using last observation carried forward (LOCF) as the imputation method. Next, a change from baseline value is computed for each subject for each subscale. The change from baseline value is the Week 12 value minus the Baseline value. Then, a mean score is calculated for each treatment group by averaging the scores for all the subjects in that treatment group. These scores are compared using a z-test. The results from the analysis carried out by GlaxoSmithKline had a significant difference between the Naratriptan and Customary

Therapy for one out of the seven subscales that we also analyzed. This subscale was the Vitality subscale. Our results differed slightly from the GlaxoSmithKline results. As seen in Table D.7, there was a difference between the two treatment groups for the General Health subscale. This conclusion is based on the fact that the 95% credible interval for the TRT parameter does not contain 0.

3.3 Simulation Study to Validate Imputation Method

We ran two simulations to test the imputation method. Both simulations were performed in the same manner with one difference. For the first simulation, the true value of TRT parameter was equal to 1. For the second simulation, the true value of the TRT parameter was equal to 5 which corresponds to small and large differences in treatment.

For the first simulation we created a data set consisting of 200 patients, 100 in each treatment group. In the data set, there were five responses, two questions, and one continuous covariate. The true values of the parameters are: BETA = 2, QUES = 1, TRT = 1, a_1 = -2, a_2 = -1, a_3 = 1, and a_4 = 2. In order to generate the data, the Y_i values are randomly drawn from a logistic distribution with location $\mu_i = x_i$*BETA + treatment*TRT + question*QUES. Where the x_i are generated from a standard normal distribution, treatment takes on a value of 1 or 0, and question takes on a value of 1 or 0. If the subject is in treatment 1 then the value of treatment is 1. If the subject is in treatment 2 then the value of treatment is 0. For question 1, the value of question is 1, and for question 2, the value of question is 0. Then, once Y_i has been generated, the observed R_i can be determined using the true a_j values as $R_i = \sum_{j=1}^{5}(j * I[a_{j-1} < Y_i \leq a_j])$.

Once the data were generated, PROC GENMOD was used to fit the ordinal model with logit link to the complete data. Next, a certain number of subjects were selected from each treatment group and their observed response R_i values were removed.

PROC GENMOD was used to fit the ordinal model with logit link to the data, ignoring the missing responses. Then, `WinBUGS` was used to fit the ordinal model with logit link to the data, imputing the missing responses. The simulation was done 100 times for 10, 35, and 50 subjects being removed from each treatment group. T-tests were done to compare the estimates to the true values of the parameters. The estimates were obtained by taking the mean value from the 100 runs of the simulation. The imputed results were compared to the true values from the complete data.

As seen in the tables(Appendices E and F), both devices gave good results. The parameter estimates were similar to the true values for most of the parameters. For example, refer to a_4 at 10% missing in Appendix E. The true value is 2. The value from SAS is equal to 1.99 and the value from `WinBUGS` is equal to 2.01. Both devices gave estimates that were .01 units away from the true value on the average. In the tables in Appendices E and F, the values in the "estimate" column were calculated by taking the mean of the 100 parameter estimates from the 100 runs of the simulation. The values in the "cov prob" column are equal to the proportion of times out of the 100 runs of the simulation, that the true value was captured in the confidence or credible interval. The "length" column refers to the length of the confidence or credible interval. The values in this column are computed by taking the mean of the 100 confidence or credible interval lengths from the 100 runs of the simulation. The MC Errors are equal to the the sample standard errors for the estimates. They are computed by taking the square root of the sum of squared errors divided by n-1.

For a_3, a_4, QUES, and TRT in the first simulation, the parameter estimates from `WinBUGS` and SAS were not significantly different from the true values for all levels of missing. The results from the second simulation were comparatively worse. The only parameter estimates that were not significantly different from the true values for all levels of missing from `WinBUGS` and SAS were the estimates for a_2. In addition for SAS only, there were no significant differences between the true value and the parameter estimates for a_1 at all levels of missing.

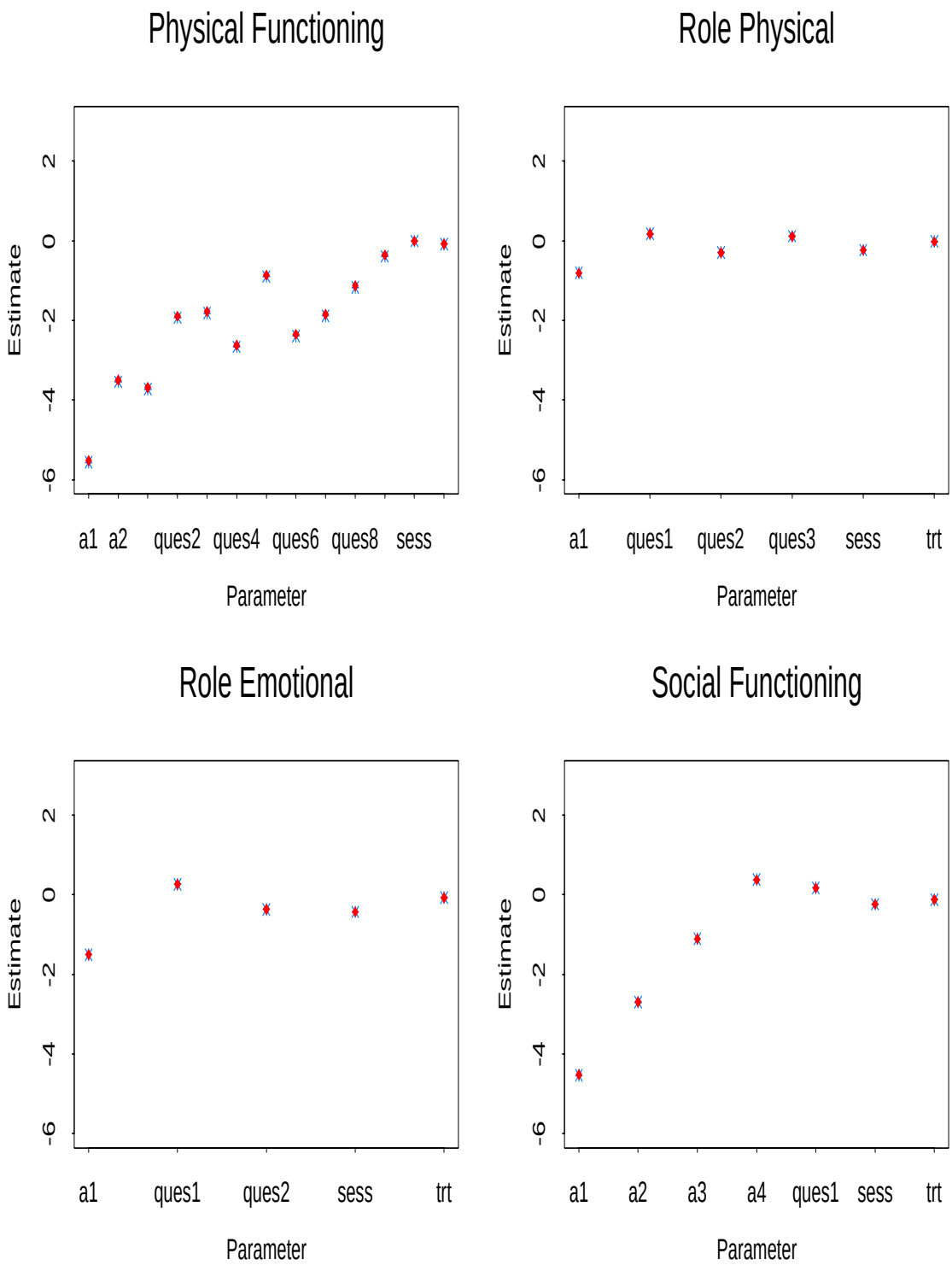
There was consistency between SAS and WinBUGS with regards to the coverage probability. For example, refer to TRT at 10% missing in Appendix E. The coverage probability for both SAS and WinBUGS was .94. For the coverage probabilities in both simulations, sometimes SAS and WinBUGS were equivalent, sometimes SAS was slightly higher, and sometimes WinBUGS was slightly higher. For the first simulation, the probabilities were all greater than .90. The majority of the time, the coverage probability was equal to .94, .95, or .96. For the second simulation, the majority of the time, the coverage probability was equal to .93, .94, .95, or .96. Occassionally the coverage probability dropped below .90. For example refer to a_4 for 35% missing in Appendix F. The coverage probability from WinBUGS is equal to .88. The parameter estimates from WinBUGS with imputation were more efficient than the parameter estimates from SAS. This gain in efficiency can be seen by comparing the average lengths of the confidence and credible intervals. The average length of the credible intervals were shorter than the average length of the confidence intervals for both simulations at all levels of missing. For example, refer to a_1 at 35% missing in Appendix E. The average length of the confidence interval from SAS was 1.102 wheras the average length of the credible interval from WinBUGS was .974.

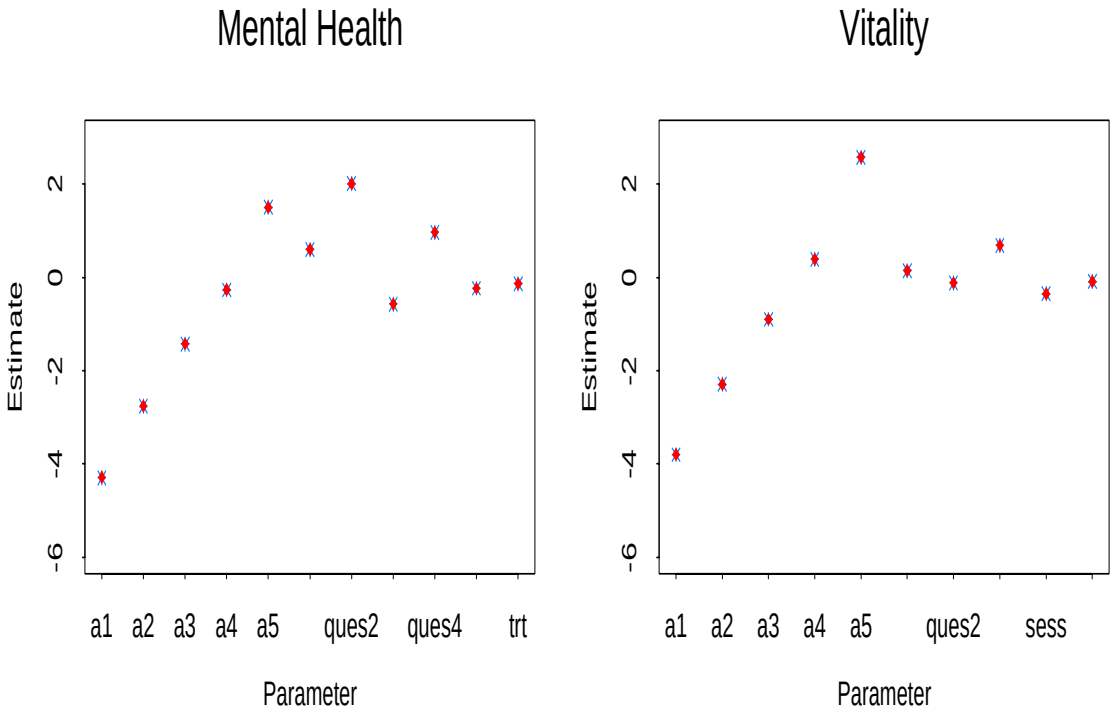
For the first simulation, for 10% missing, the imputed value was equal to the real value 50.95% of the time. For 35% missing, the imputed value was equal to the real value 50.11% of the time. For 50% missing, the imputed value was equal to the real value 50.79% of the time. Bar plots of the imputed values versus the real values can be seen in Figures 3.4 - 3.6. The height of the bar corresponds to the proportion of times that response value was chosen as the imputed value. In each figure, there is one bar plot for each real value. For the second simulation, for 10% missing, the imputed value was equal to the real value 68.25% of the time. For 35% missing, the imputed value was equal to the real value 68.74% of the time. For 50% missing, the imputed value was equal to the real value 68.54% of the time. Bar plots of the imputed values versus the real values can be seen in Figures 3.7 - 3.9. Figures 3.4 - 3.9 show that for

real values of 1, 3, and 5, the imputed value that is chosen is equal to the real value more often than not. The bars for 1,3, and 5 are the highest.

The results from the two simulations showed that both SAS and `WinBUGS` perform better when there are small treatment differences than when there are large treatment differences. They also showed that there is not a significant difference between the two devices in regards to the parameter estimates. However, using `WinBUGS` with imputation will result in more efficient parameter estimates. In addition, it will result in a complete data set that can be used for other analyses.

Figure 3.1: Plot of SAS vs BUG estimates where the BUG estimate is a diamond and the SAS estimate is a star




Mental Health
Estimate
2
0
-2
-4
-6
a1
a2
a3
a4
a5
ques2
ques4
trt
Parameter

Vitality
Estimate
2
0
-2
-4
-6
a1
a2
a3
a4
a5
ques2
sess
Parameter

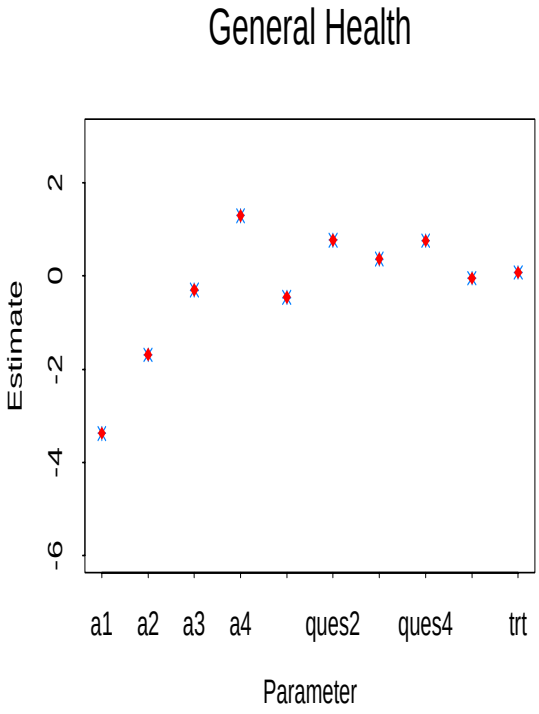

General Health
Estimate
2
0
-2
-4
-6
a1
a2
a3
a4
ques2
ques4
trt
Parameter

Figure 3.2: Density Plots

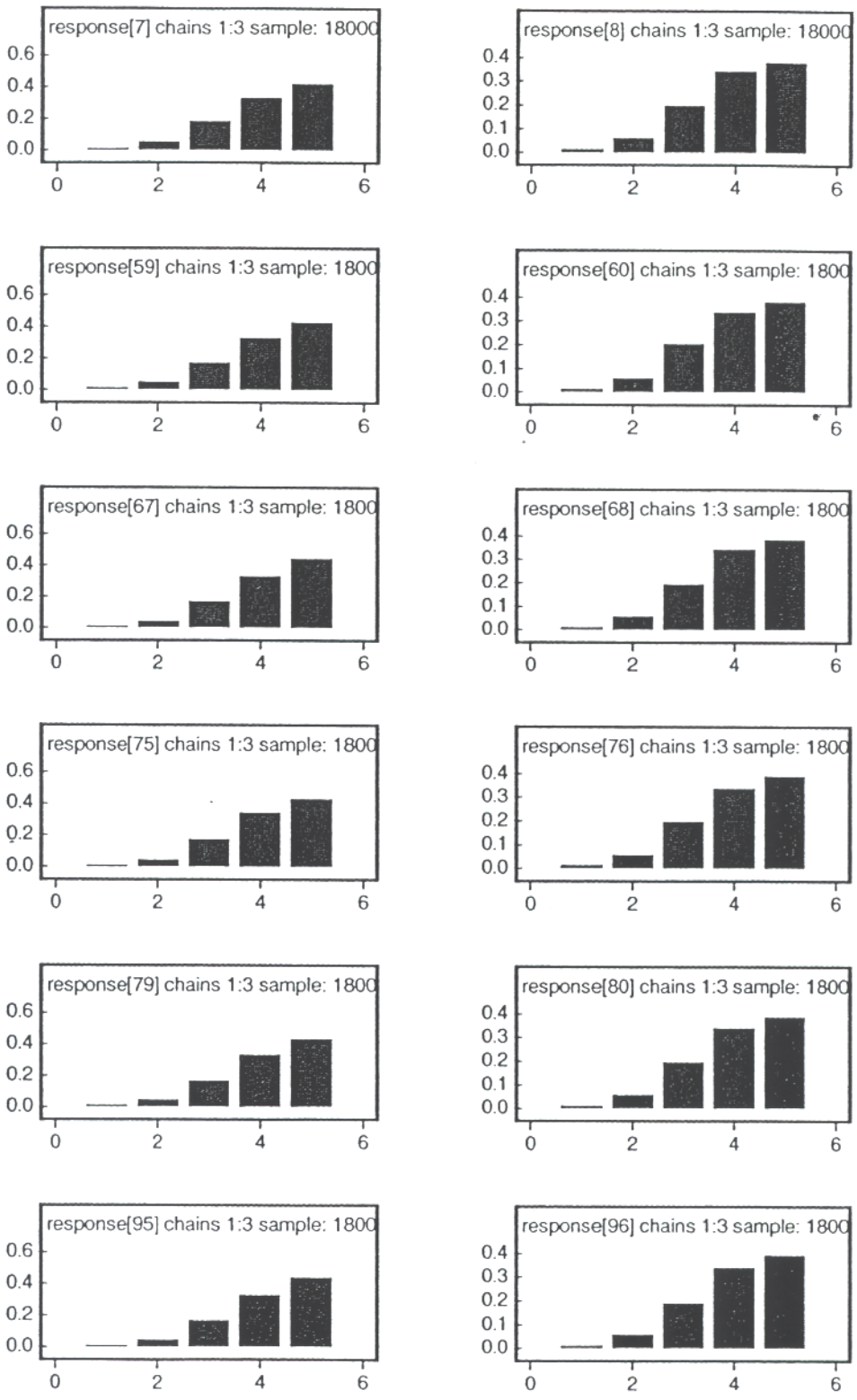
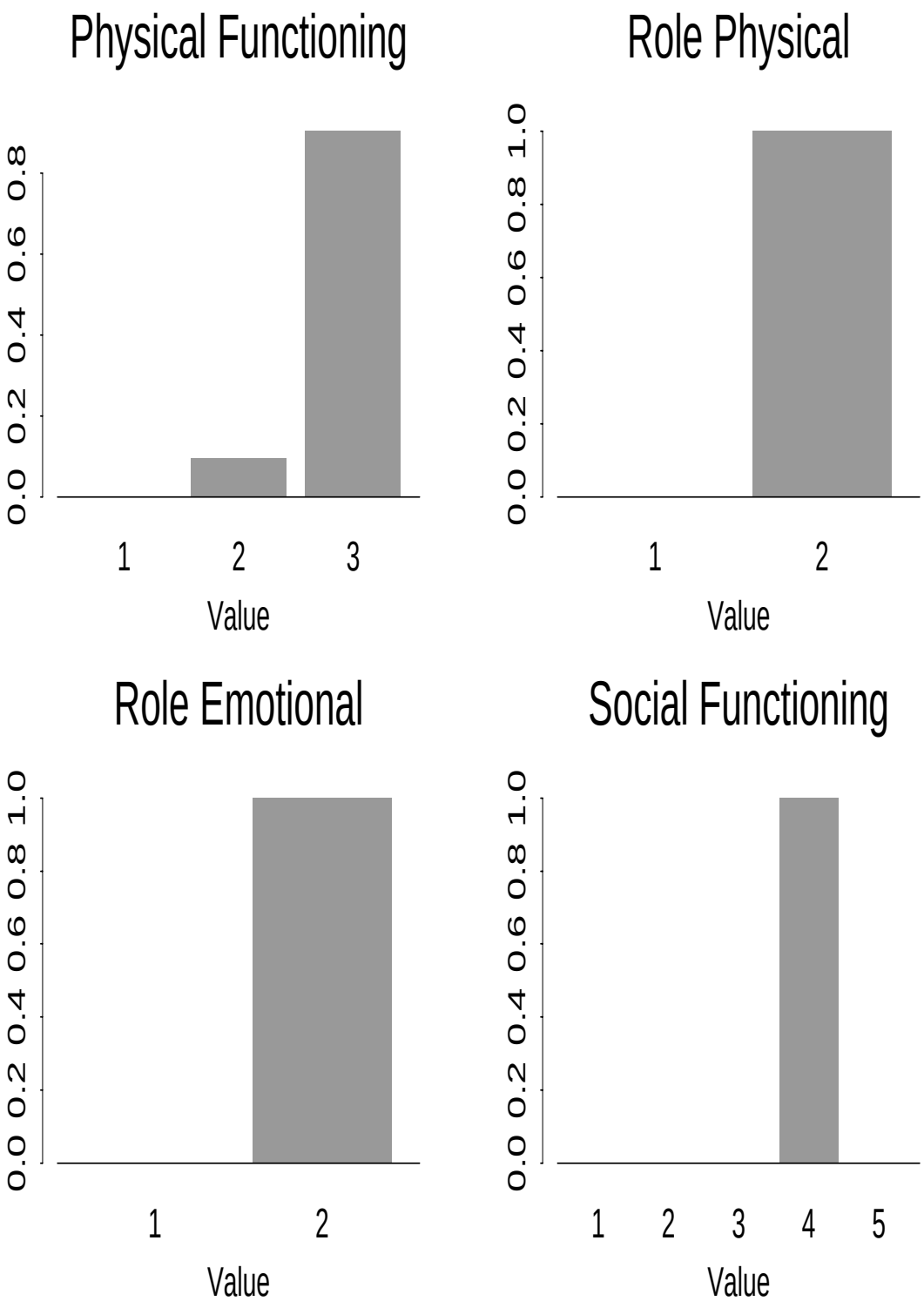
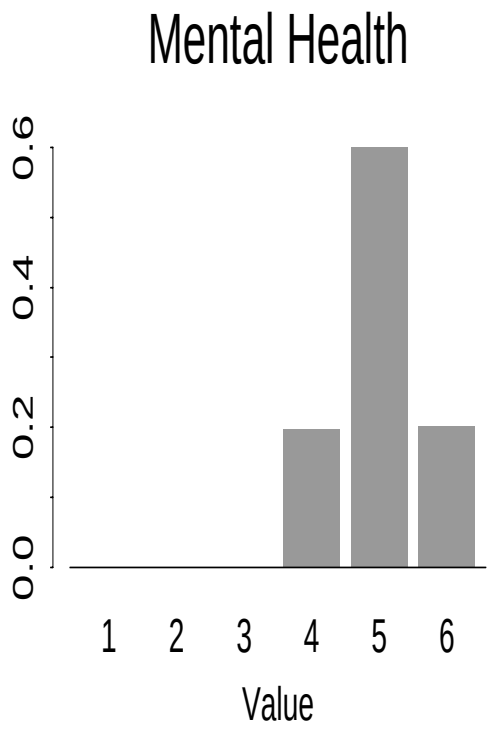
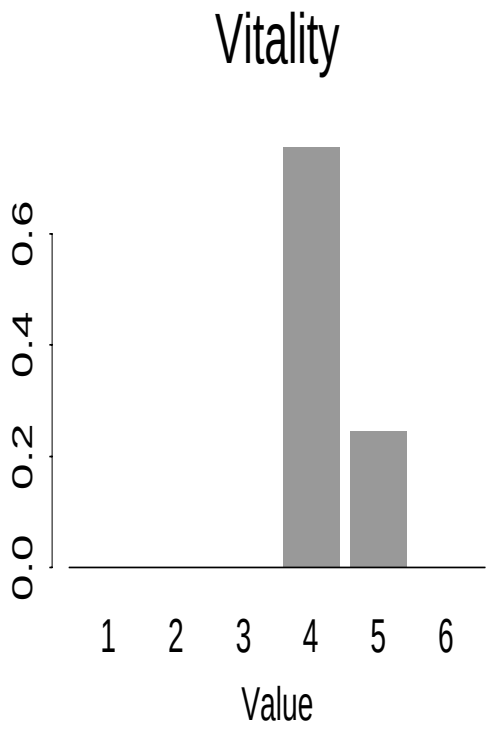


Figure 3.3: Plots of Imputed Values




Mental Health
0.0
0.2
0.4
0.6
1
2
3
4
5
6
Value


Vitality
0.0
0.2
0.4
0.6
1
2
3
4
5
6
Value

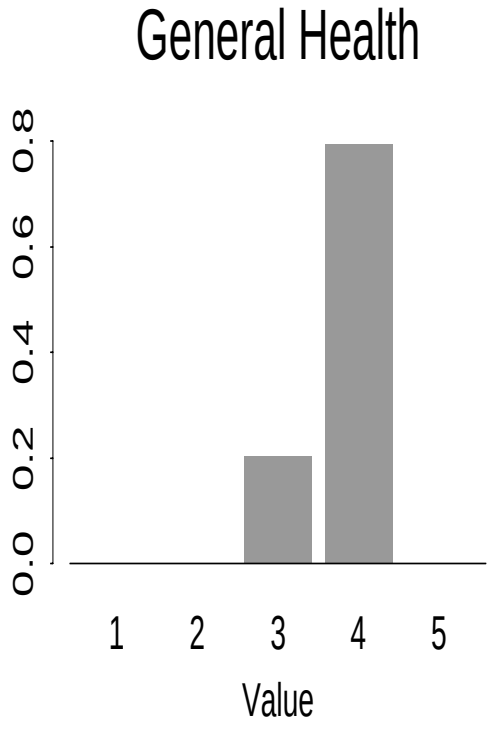

General Health
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
Value

Figure 3.4: Plot of Imputed Values for 10% Missing Where TRT = 1

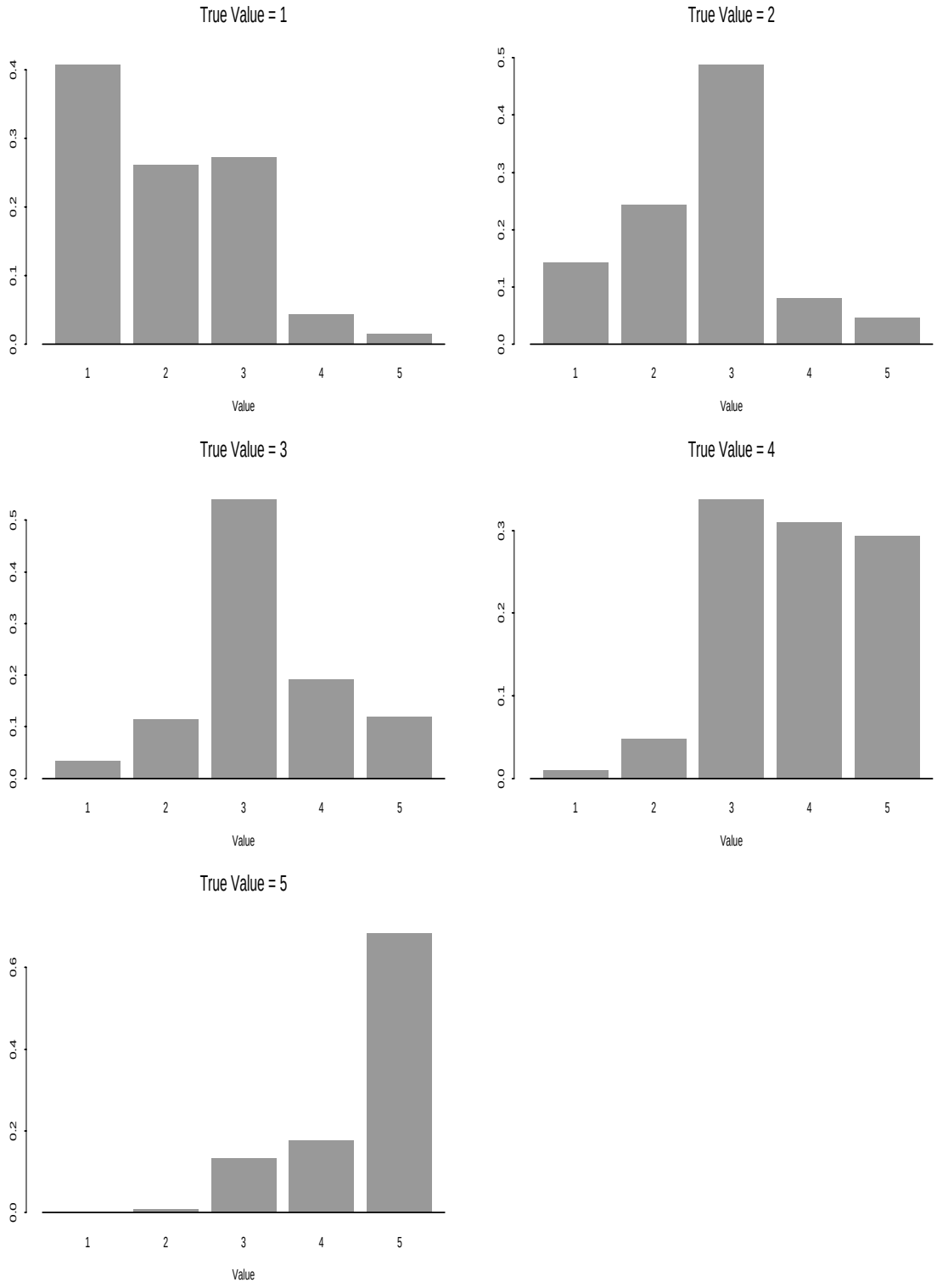


Figure 3.5: Plot of Imputed Values for 35% Missing Where TRT = 1

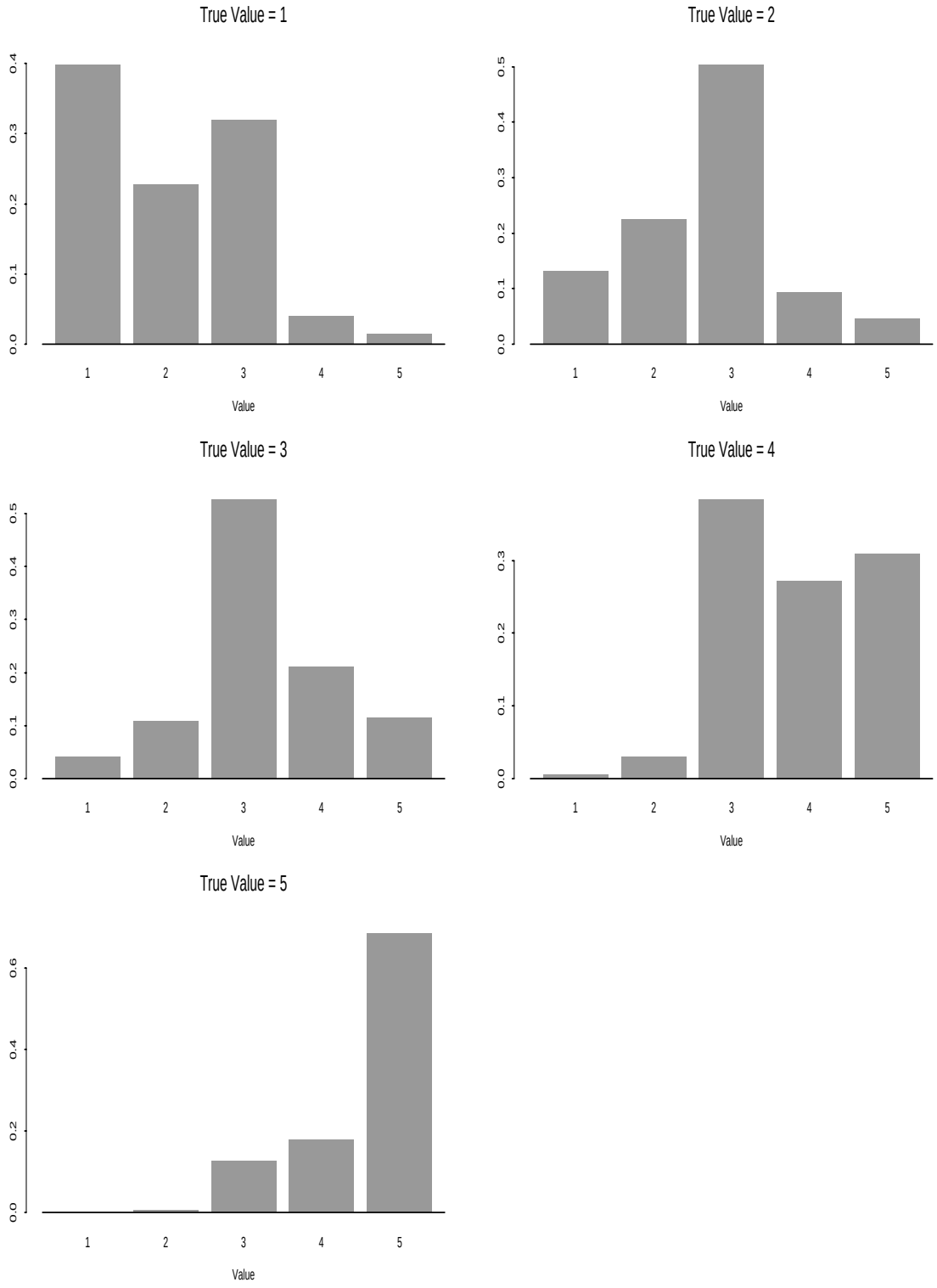


Figure 3.6: Plot of Imputed Values for 50% Missing Where TRT = 1

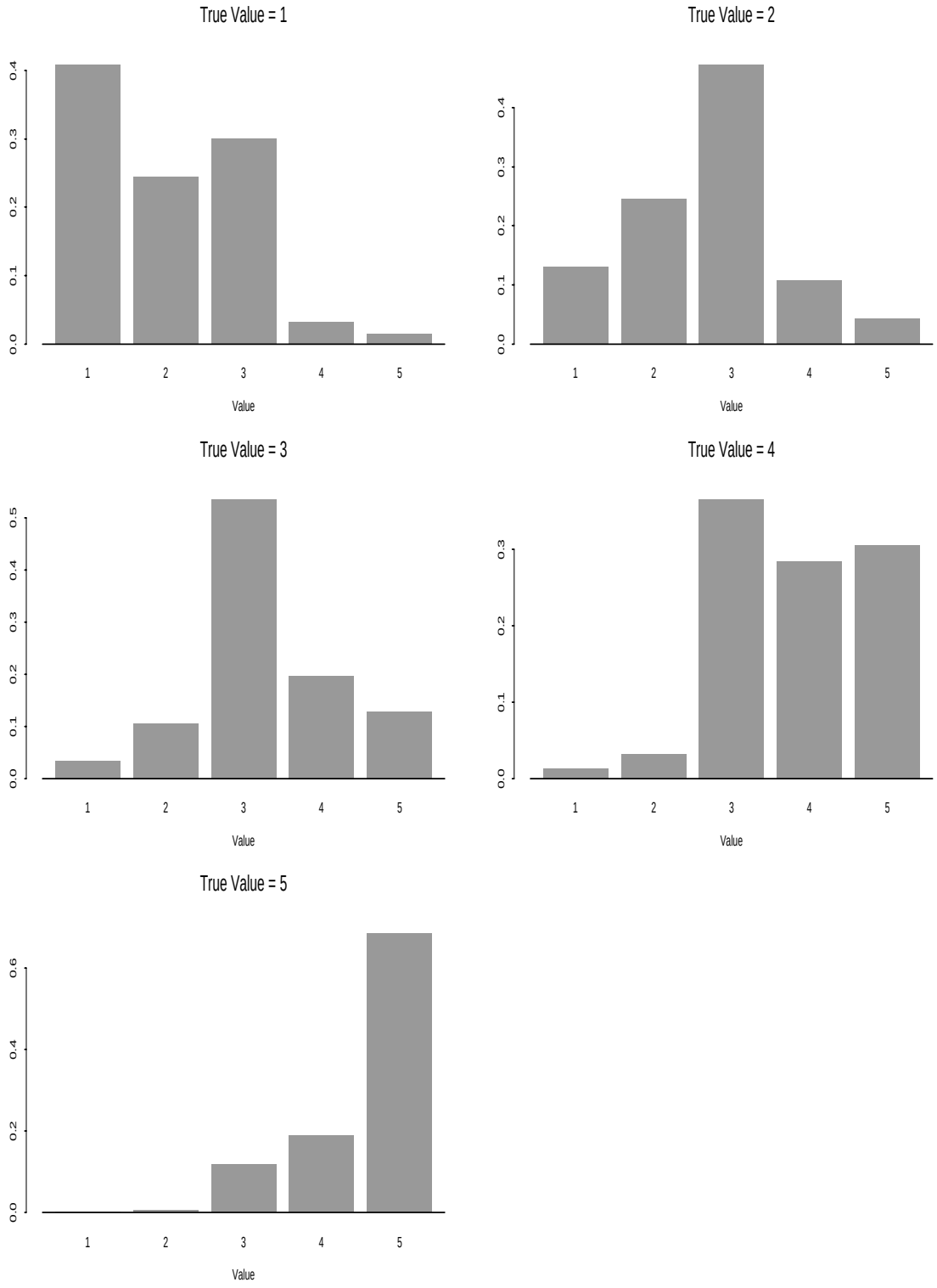


Figure 3.7: Plot of Imputed Values for 10% Missing Where TRT = 5

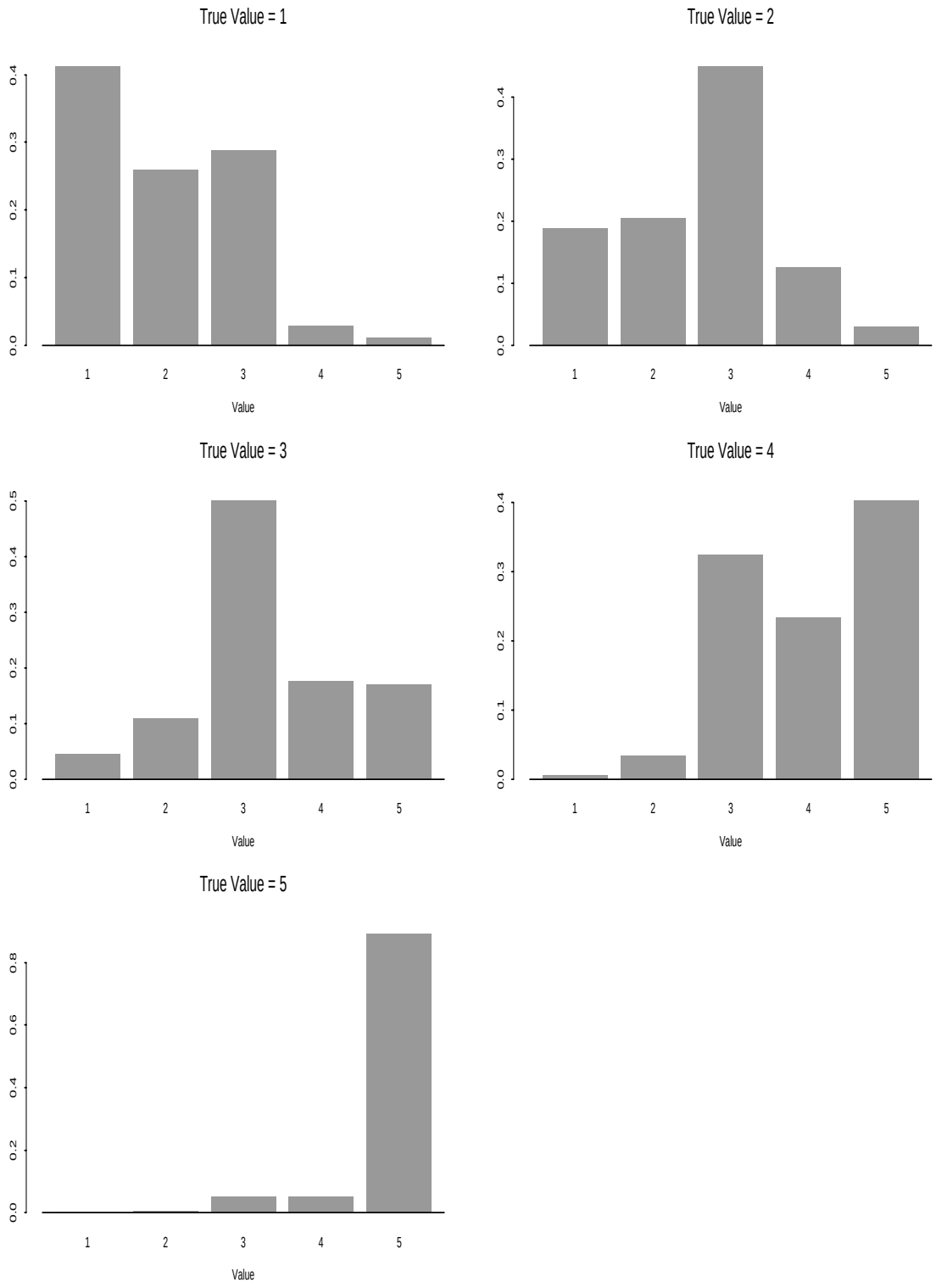


Figure 3.8: Plot of Imputed Values for 35% Missing Where TRT = 5

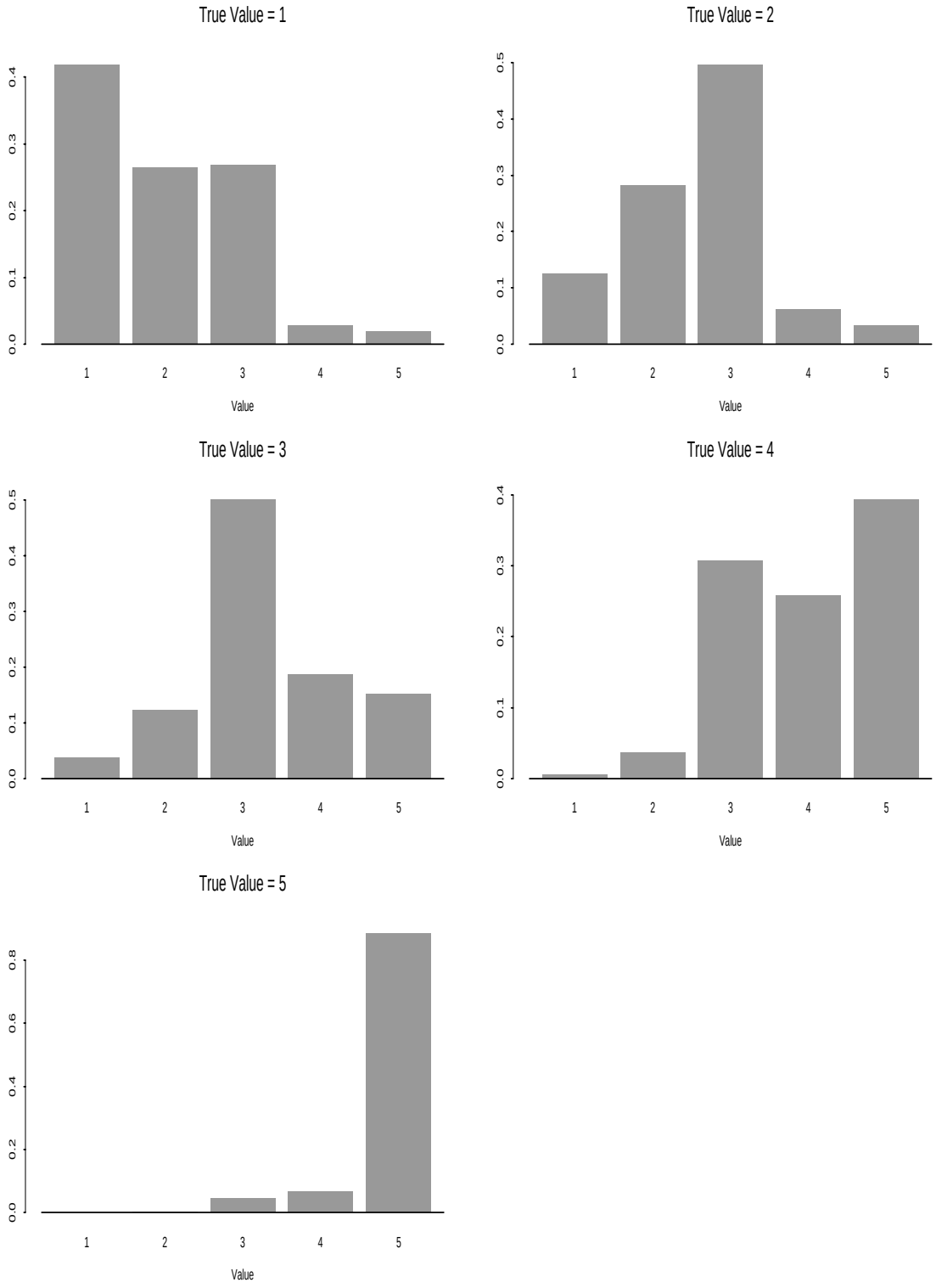
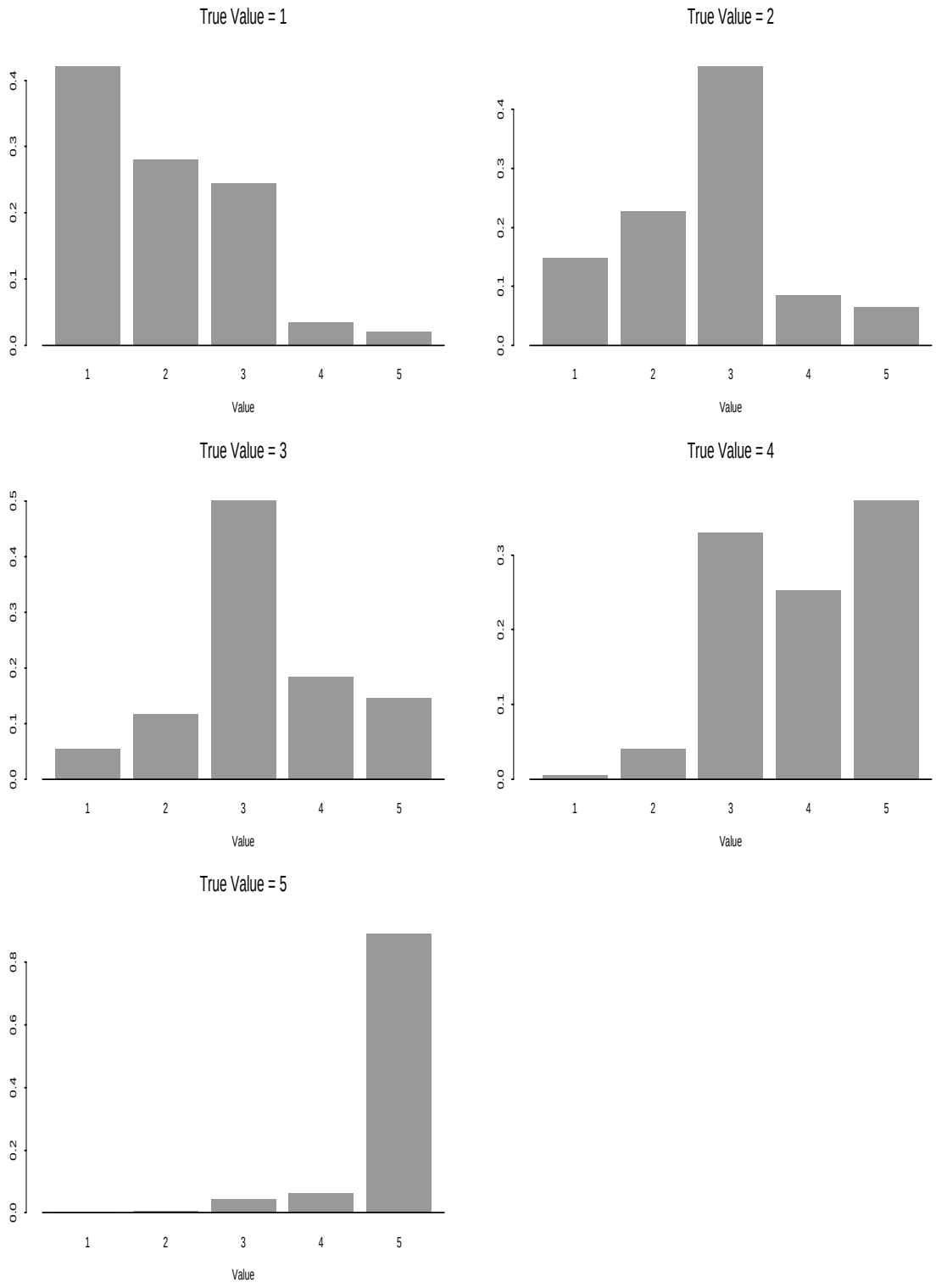


Figure 3.9: Plot of Imputed Values for 50% Missing Where TRT = 5



Chapter 4

Ordinal Response Model with Varying Cut-Points

4.1 Introduction

In Chapter 3 we fit an ordinal response model to the data. However, we did not take into account the population differences. Due to the diversity among populations, the probability of an individual responding in any given response category is different across populations. Adjustments are needed to make survey results comparable across populations. When categorical variables are involved, analyses must account for differences in response category cut-points (a_j). Cut-points are also likely to vary within a cultural or sociodemographic group.

The World Health Organization (WHO) used a sample from the WHO Multi-Country Study it conducted to demonstrate the assessment of the predictive validity of the hierarchical ordered probit model (HOPIT) (WHO 2001). This model is a modification of the standard ordinal response model with probit link. The modification incorporates components that refer to estimation of cut-points using responses to vignettes. A vignette is a question that the respondents are asked. The question asks the respondents to evaluate their level of ability on a given domain. In the WHO

study, the vignette assesses the respondents level of ability on the mobility domain. The vignette has the same number of responses as the questions on the survey. The purpose of the vignette is as a means of correction of the responses from the survey which is a self-report survey. The correction makes the responses comparable across populations. The vignette fixes the level of ability such that variations in categorical responses are attributable to variations in response category cut-points. This introduction of information in the form of responses to vignettes allows one to identify the effects of a set of sociodemographic covariates (such as age, sex, country of residence, etc.) on both the level of the underlying latent variable that is being estimated as well as on the cut-points. In the WHO Multi-Country Study, there are six vignettes for the domain of mobility, each designed to capture a different level of ability on this domain. The sample showed that the HOPIT model seemed to perform well in terms of predictive validity. We will be modifying our standard ordinal response model with logit link in a similar manner.

It has been shown that people in different countries respond differently to the SF-36 form. In order to investigate the differences between countries, we fit a separate model for each country. The analysis was performed in `WinBUGS`. Plots of the cut-points by country for each group can be seen in Figure 4.1. This figure shows that the cut-points changed depending on the country. It also shows that the cut-points are not significantly different for all countries. Thus, a separate cut point for each country may not be needed. For example, refer to Figure 4.1 for Role Physical. It appears that the cut-points for New Zealand, Hungary, Canada, and The Netherlands seem to be similar. It also appears that the cut-points for Spain and Finland are slightly different, with the cut-point for Spain being significantly different then the cut-points for Hungary, Canada, and Finland. The conclusion of significant difference is drawn from the fact that the credible intervals do not overlap. Tables 4.1 - 4.7 summarize the findings. An X indicates that the two cut-points are significantly different for those countries (in other words, their credible intervals do not overlap). For our example

above, notice that in Table 4.2, the cut-point for Spain is significantly different than the cut-points for Hungary, Canada, and Finland.

One important benefit of allowing the cut-points to vary depending on country is that the treatment difference can be compared across countries. Another benefit of using the varying cut-points is that it allows the information from similar countries to be pooled together. This will be beneficial if not all responses are being selected on the survey. For example, refer to Figure 4.1 for Mental Health. Notice that the a_1 estimate for The Netherlands is not on this plot. This is because the response value of 1 was never selected by people in The Netherlands for the Mental Health questions. Thus the parameter estimate is extremely negative.

Here we extend our model in Chapter 3 to include the effect of varying cut-points.

- $L_c = w$, where $w = 1, \ldots W$ and c, the country number, takes on values of 1, ..., 6. W is equal to the number of cut-points needed.

- Let Y_i denote an underlying latent continuous response variable. Suppose $-\infty = a_{w0} < a_{w1} < \ldots < a_{wJ} = \infty$ are the "cut-points", such that R_i, the observed ith ordinal response in the data, satisfies $R_i = j$ if $a_{w(j-1)} < Y_i \leq a_{wj} \mid L_{c_i} = w$, where w = 1, ..., W, $-\infty = a_{w0} < a_{w1} < \ldots < a_{wJ} = \infty$, and c_i is the country number for the ith ordinal response in the data.

- When the underlying response value Y_i falls in the jth interval $(a_{w(j-1)}, a_{wj}] \mid L_{c_i} = w$, then $P[R_i = j \mid L_{c_i} = w] = P[a_{w(j-1)} < Y_i \leq a_{wj}]P(L_{c_i} = w) = [G_i(a_{wj}) - G_i(a_{w(j-1)})]\nu_w$. Thus, $P[R_i = j] = \sum_{w=1}^{W}[G_i(a_{wj}) - G_i(a_{w(j-1)})]\nu_w$. In the presence of covariates, we assume that $G_i(a_{wj}) = G(a_{wj} - \beta' x_i)$.

- In our model, we take x_i' = (question[i], treatment[i], session[i])$'$ and β' =(QUES, TRT, SESS) such that $\beta' x_i$ =QUES[question[i]] + TRT[treatment[i]] + SESS[session[i]].

The likelihood function of β and $\underline{a}' = (a_{11}, \ldots, a_{1(J-1)}, a_{21}, \ldots, a_{2(J-1)}, \ldots, a_{w1}, \ldots, a_{(w)J-1})'$ is given by

$L(\beta, \underline{a}) = \prod_{i=1}^{m} \sum_{w=1}^{W} (G(a_{(L_w)(R_i)} - \beta' x_i) - G(a_{(L_w)(R_i-1)} - \beta' x_i))\nu_w$.

4.1.1 Imputation

In this section, we extend the method to analyze data with missing responses. Gibbs sampling was used to impute values for the missing responses using the logit link. The steps are:

1. Get initial estimates for the missing data and the parameters. For the L_c parameters initial values were chosen between 1 and W. For all other parameters, except the a_{wj} parameters, initial values of zero were chosen. For the a_{wj} parameters, a combination of negative and positive values were used in increasing order within each w. `WinBUGS` generated initial values for the missing responses.

2. Generate $Y_i^{(t)}$'s where $Y_i^{(t)} \sim TL(x_i'\beta^{(t-1)}, 1)I(a^{(t-1)}_{(L_{c_i})(R_{(i-1)})}, a^{(t-1)}_{(L_{c_i})(R_i)})$ and TL represents the truncated logistic distribution.

3. Update the $a_{wj}^{(t)}$'s, where $a_{wj}^{(t)} \sim U(L_{wj}^{(t)}, U_{wj}^{(t)})$,
   where $L_{wj}^{(t)} = max\{Y_i^{(t)} | R_i^{(t-1)} = j, L_{c_i}^{(t-1)} = w\}$, and
   $U_{wj}^{(t)} = min\{Y_i^{(t)} | R_i^{(t-1)} = j-1, L_{c_i}^{(t-1)} = w\}$.

4. Update the $\nu_w^{(t)}$'s, where $(\nu_1^{(t)}, \ldots, \nu_W^{(t)})$ ~Dirichlet$(1+\sum_{c=1}^{6} I(L_c^{(t-1)} = 1), \ldots, 1+\sum_{c=1}^{6} I(L_c^{(t-1)} = W)$.

5. Update the $L_c^{(t)}$'s as
   $L_c^{(t)} = \sum_{w=1}^{W} \nu_{wi}^{*(t)} I[L_{c_i}^{(t-1)} = w]$ where
   $\nu_{wi}^{*(t)} = \frac{[G_i(a_{wj})^{(t)} - G_i(a^{(t)}_{w(j-1)})]\nu_w^{(t)}}{\sum_{d=1}^{W}[G_i(a_{wj}^{(t)} - G_i(a^{(t)}_{w(j-1)})]\nu_d^{(t)}}$.

6. Impute the $R_i^{(t)}$'s as
   $R_i^{(t)} = \sum_{j=1}^{J} \sum_{w=1}^{W} jI(a^{(t)}_{w(j-1)} < Y_i^{(t)} \leq a_{wj}^{(t)}) I(L_{c_i}^{(t)} = w)$.

7. Sample $\beta^{(t)}$. This needs to be done using a Metropolis-Hastings step since the full conditional is not a standard distribution.

Before these steps using the logit link were carried out in WinBUGS using the code in Appendix G, the following was done. First, Figure 4.1 was examined to determine how many cut-points were needed for each subscale. For example, for the Role Physical subscale, it appears that three cut-points are needed. New Zealand, Hungary, Canada, and The Netherlands seem to be grouped together. Spain and Finland appear to need their own cut-points. Next we fit an ordinal model to the data with varying cut-point variables using the code in Appendix G. For this step we used one more cut-point than we thought was necessary. In other words if we thought three cut-points were needed, we fit the model with W = 4. Then, the model was examined to determine if the correct number of cut-points was being used. By monitoring the L_c variable in WinBUGS, one can determine how many independent values are being chosen. This is how many cut-points are needed. As seen in Table H.2, L_c has three distinct values (1,2,3). Thus, three cut-points are needed for the Physical Functioning subscale. Upon examining the L_c values, if it was determined that the W being used was not correct, the model was re-run using one less cut-point. This was continued until the correct number of cut-points was being used.

Once the correct number of cut-points was identified, the final models were fit. The imputation using the new model with logit link was carried out in WinBUGS using the code in Appendix G. We fit one chain consisting of 10,000 iterates. We used a 'burn-in' of 5000. Thus, a sample size of 5000 was used for analysis. The tables of the results can be seen in Appendix H. By comparing the table parameter estimates in Appendix H to the values in Figure 4.1, one will see that the parameter estimates are similar. For example, refer to Table H.2 and Figure 4.1 Role Physical. In Figure 4.1 Role Physical, the estimate for New Zealand is approximately -.6, the estimate for Hungary is approximately -.9, the estimate for Canada is approximately -.7, and the estimate for The Netherlands is approximately -.5. All of these are similar to the parameter estimate in Table H.2 of -.596. Thus, we can conclude that we are able to get correct cut-point parameter estimates without having to fit a separate model for

each country. Therefore, we only need to fit one model per subscale which allows us to compare treatments across populations.

Recall that one benefit of fitting only one model per subscale is that is allows the information from similar countries to be pooled together. As seen in Table H.5 for the model with varying cut-points, the parameter estimate for a_1 for people in The Netherlands is -5.008. This is a more reasonable estimate than the estimate from fitting a separate model for The Netherlands. The estimate from the separate model for the Netherlands was extremely negative.

The plots of the imputation values can be seen in Figure 4.2. When comparing the imputation plots (Figure 4.2) to the imputation plots from the model in Chapter 3 (Figure 3.3) there are a few major differences. For the Role Physical subscale, values of 1 and 2 are selected in the new model. In Chapter 3, only the value of 2 was selected. For the Social Functioning subscale, values of 4 and 5 are selected in the new model. In Chapter 3, only the value of 4 was selected. Thus, one can conclude that allowing the cut-points to vary, will result in better imputed values. The imputed values are better because they take into account the differences in the countries. For example, in Chapter 3, for the Social Functioning subscale, all imputed values were 4. During the exploratoy analysis, when a separate model was fit for each country, the imputed values were 4 for all countries except Finland. For Finland, the majority of the values were 5. When the model with varying cut-points was fit to the data, for the Social Functioning subscale, values of 4 and 5 were selected. The times that a 5 was selected were for subjects from Finland.

Recall that the results from the analysis carried out by GlaxoSmithKline had a significant difference between the two treatment groups for the Vitality Subscale. Our results differed slightly from the GlaxoSmithKline results. As seen in Table H.7, there was a difference between the two treatment groups for the General Health subscale. This conclusion is based on the fact that the credible interval does not contain 0.

4.2 Benefits of Using WinBUGS

At the end of Chapter 3, we presented the results from two validation simulations. Both showed that SAS and `WinBUGS` gave similar parameter estimates. The validation simulations also showed that using `WinBUGS` with imputation will result in more efficient parameter estimates. Another benefit to using `WinBUGS` is that you end up with a complete set of data. Due to the nature of the drug approval process, it may be necessary to continue to analyze the data using the current industry standard. By using the imputation methods presented in this paper, you will have a complete set of data with good imputation values from a multiple imputation method that you can use for the analysis.

One result that comes from using `WinBUGS` to analyze the data is density plots for the imputed values. An example of these plots can be seen in Figure 3.2. These plots can be examined to determine how probable the imputed value is.

Another benefit to using `WinBUGS` is that it can be used to fit the model with varying cut-points that was discussed in Section 4.1. SAS cannot be used to fit this model in the same way. As you fit the model using `WinBUGS`, the program will determine how many cut-points are needed and which countries belong to each group. If you wanted to fit the model in SAS, you would have to examine the by country plots yourself and determine how many cut-points are needed and which countries belong to each group. This becomes more difficult as the number of cut-points and the number of countries increase. There exist many advantages to using the model with varying cut-points. Fitting this model will allow you to compare treatment differences across populations. It also results in reasonable parameter estimates if some response choices are not present in the data, as was seen with response 1 for the Mental Health questions for people from The Netherlands.

4.3 Future Work

The methods presented in this paper are for health surveys with the same number of responses for all questions in a subscale. The SF-36 health survey, has one subscale that does not follow this pattern. The subscale has two questions in it, one of the questions has 5 possible choices, the other one has 6. This subscale was not included in the work presented here. This next phase of this research will need to develop a way to handle this problem. Also, the methods develped in this paper were tested on the SF-36 health survey. There are many therapeutic specific health surveys currently being used in clinical trials. The methods presented here need to be tested on some of the other health surveys.

We assumed that the data was missing at random. The next phase of this research will need to consider that the data is not missing at random. This will involve joint modeling of the data and the pattern of missing.

Table 4.1: Physical Functioning Cut-point Summary

Par	Country	NZ	SP	HG	CD	FN	HL
a[1]	NZ					X	
	SP					X	
	HG					X	
	CD					X	
	FN						X
	HL						
a[2]	NZ						
	SP						
	HG						
	CD					X	
	FN						X
	HL						

Table 4.2: Role Physical Cut-point Summary

Par	Country	NZ	SP	HG	CD	FN	HL
a[1]	NZ					X	
	SP			X	X	X	
	HG						
	CD						X
	FN						
	HL						

Table 4.3: Role Emotional Cut-point Summary

Par	Country	NZ	SP	HG	CD	FN	HL
a[1]	NZ						
	SP						
	HG				X	X	X
	CD						
	FN						
	HL						

Table 4.4: Social Functioning Cut-point Summary

Par	Country	NZ	SP	HG	CD	FN	HL
a[1]	NZ						
	SP						
	HG					X	
	CD					X	
	FN						
	HL						
a[2]	NZ					X	
	SP						
	HG					X	
	CD						
	FN						
	HL						
a[3]	NZ					X	
	SP					X	
	HG					X	
	CD					X	
	FN						X
	HL						
a[4]	NZ					X	
	SP					X	
	HG						X
	CD						
	FN	X					X
	HL						

Table 4.5: Mental Health Cut-point Summary

Par	Country	NZ	SP	HG	CD	FN	HL
a[1]	NZ			X			X
	SP						X
	HG				X	X	X
	CD						X
	FN						X
	HL						
a[2]	NZ		X	X			
	SP				X		X
	HG				X	X	X
	CD						
	FN						
	HL						
a[3]	NZ		X	X			
	SP				X	X	X
	HG				X	X	X
	CD						
	FN						
	HL						
a[4]	NZ		X				
	SP				X	X	X
	HG			X	X	X	
	CD						
	FN						
	HL						
a[5]	NZ						
	SP			X	X		X
	HG						
	CD						
	FN						X
	HL						

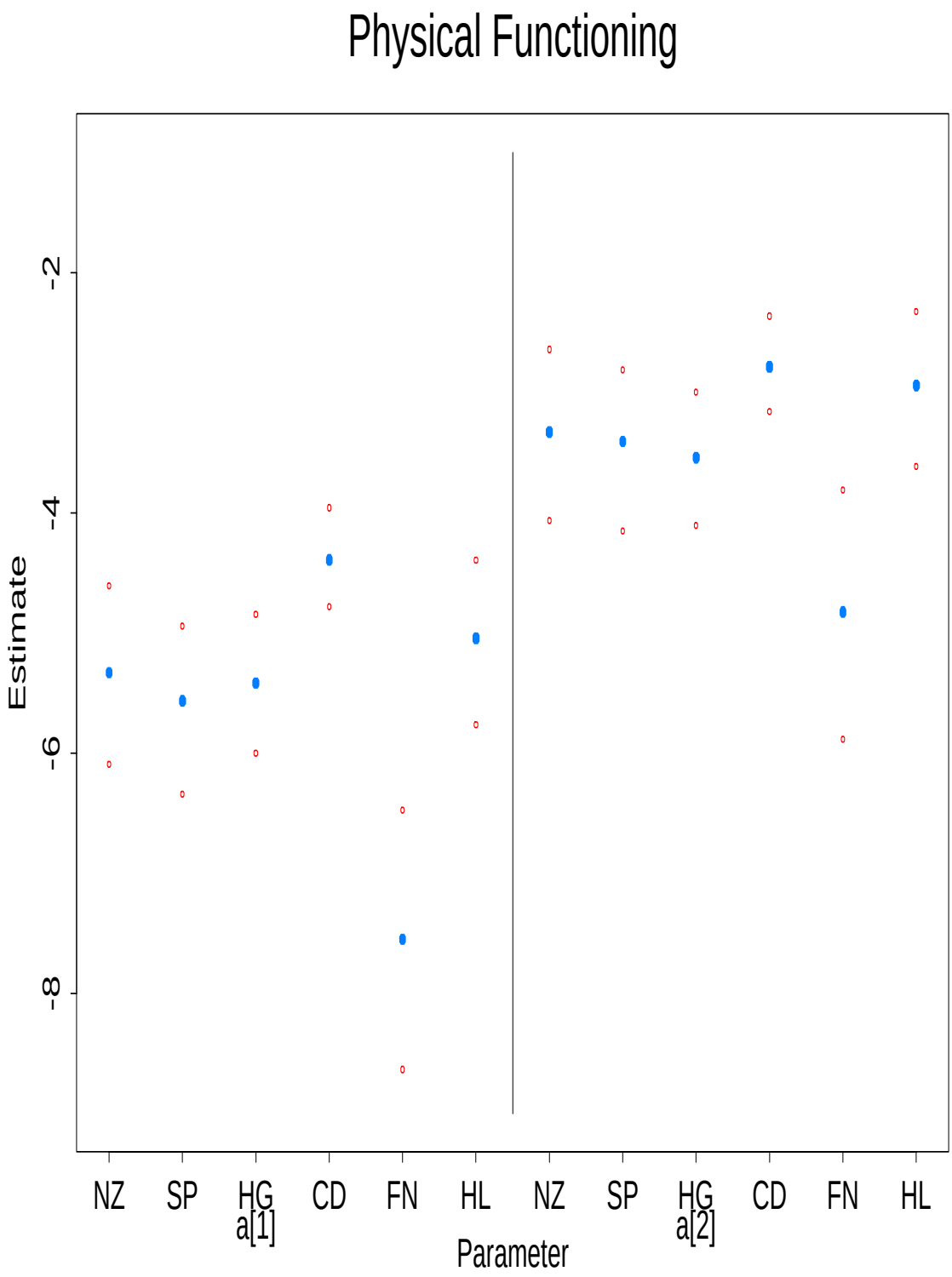
Table 4.6: Vitality Cut-point Summary

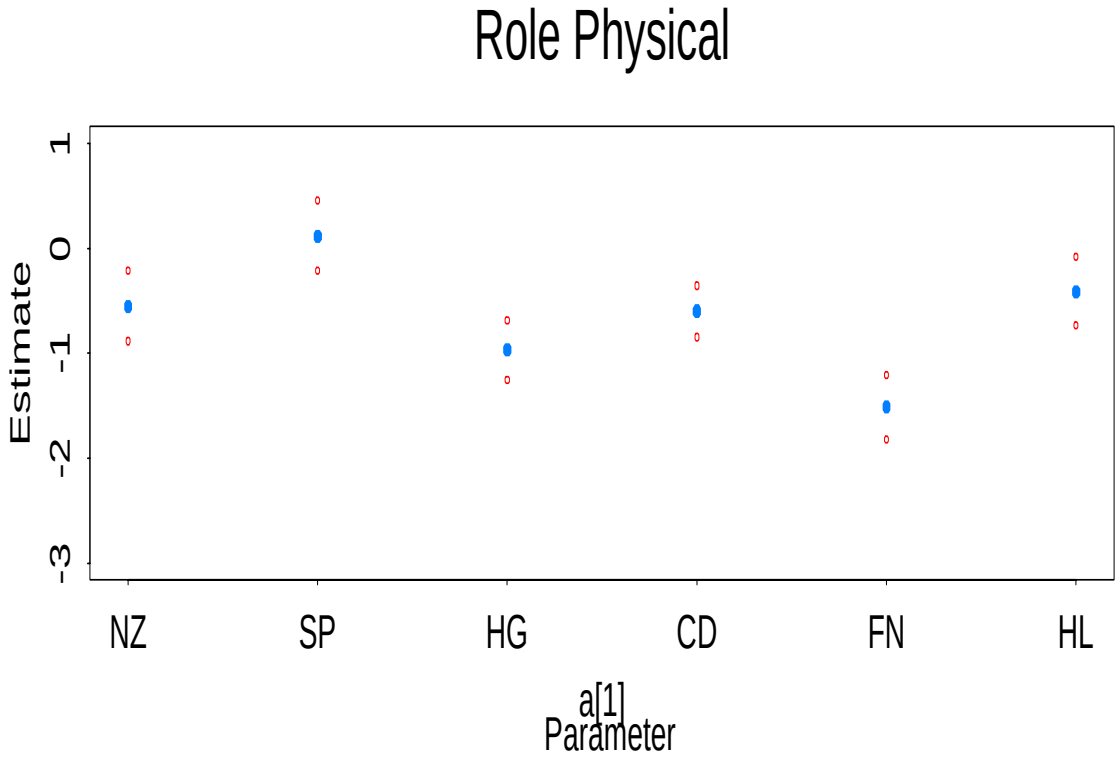
Par	Country	NZ	SP	HG	CD	FN	HL
a[1]	NZ			X			
	SP			X			
	HG				X	X	X
	CD						
	FN						
	HL						
a[2]	NZ						
	SP						
	HG					X	
	CD					X	
	FN						X
	HL						
a[3]	NZ					X	
	SP					X	
	HG					X	
	CD					X	
	FN						X
	HL						
a[4]	NZ					X	
	SP					X	
	HG					X	
	CD					X	
	FN						X
	HL						
a[5]	NZ			X			
	SP			X			
	HG				X	X	X
	CD						
	FN						
	HL						

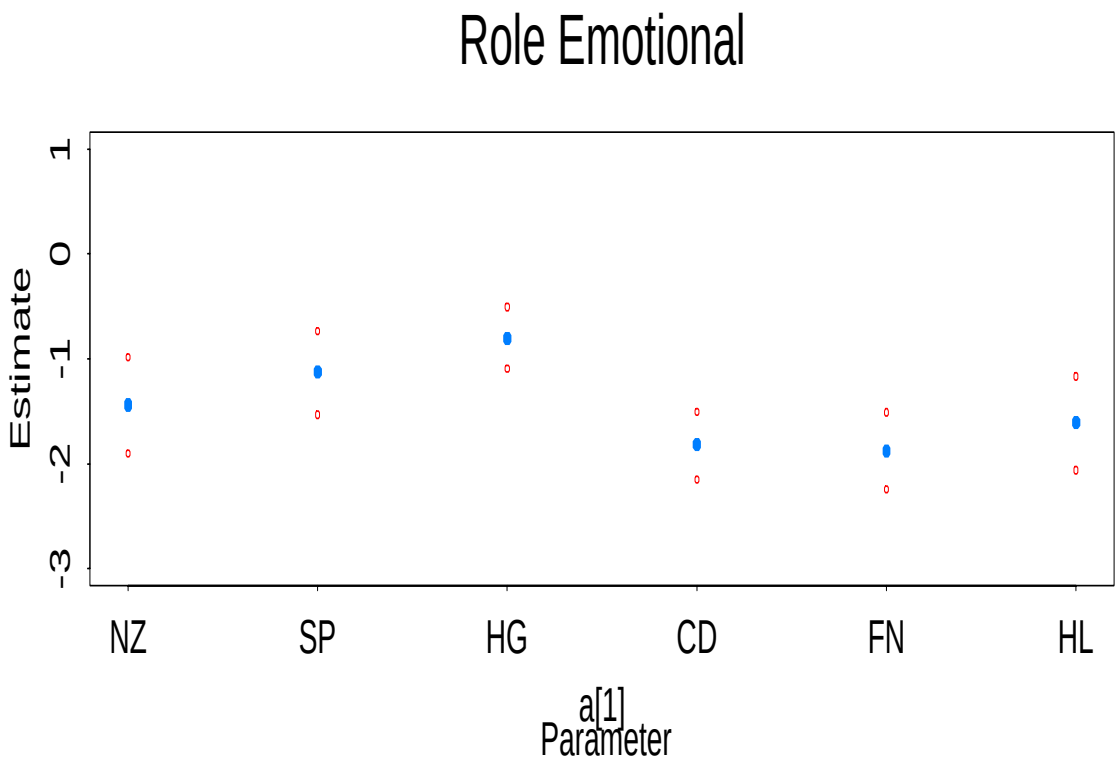
Table 4.7: General Health Cut-point Summary

Par	Country	NZ	SP	HG	CD	FN	HL
a[1]	NZ				X	X	X
	SP			X			
	HG				X	X	X
	CD						
	FN						
	HL						
a[2]	NZ			X			
	SP				X	X	
	HG				X	X	X
	CD						
	FN						
	HL						
a[3]	NZ		X	X			
	SP				X	X	X
	HG				X	X	
	CD						X
	FN						
	HL						
a[4]	NZ		X	X			
	SP				X	X	X
	HG				X	X	X
	CD						
	FN						
	HL						

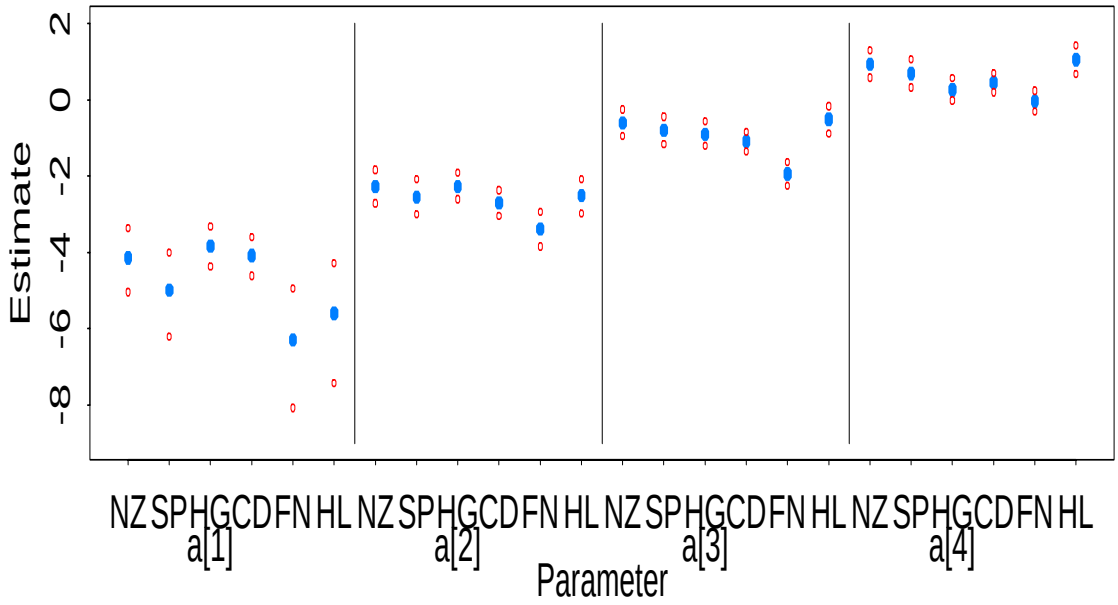
Figure 4.1: Plots of Cut-Points



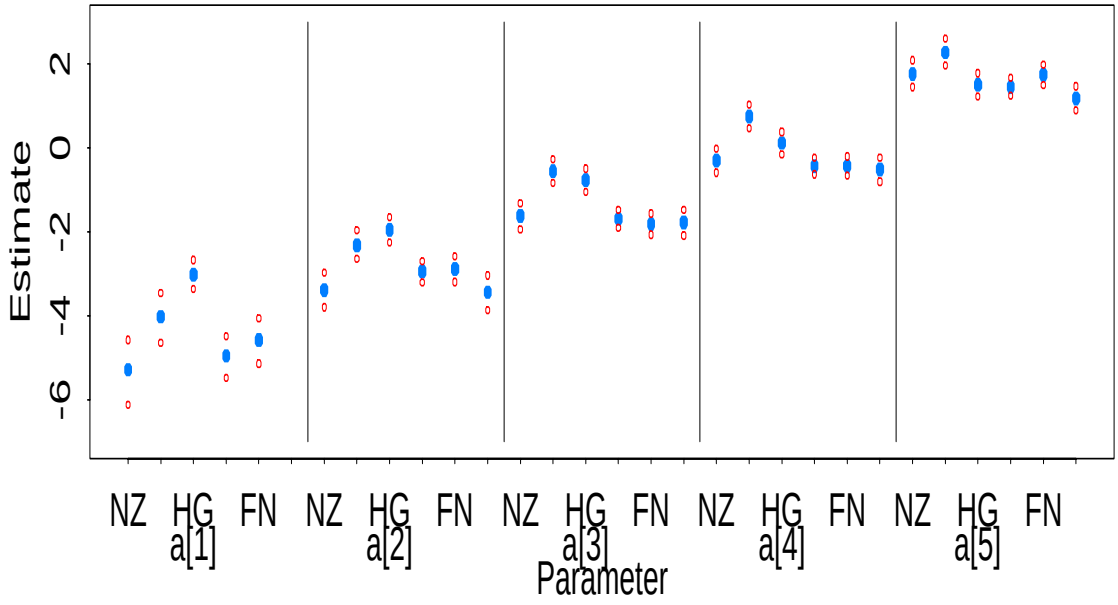

Role Physical
Estimate
1
0
-1
-2
-3
NZ
SP
HG
CD
FN
HL
a[1]
Parameter


Role Emotional
Estimate
1
0
-1
-2
-3
NZ
SP
HG
CD
FN
HL
a[1]
Parameter

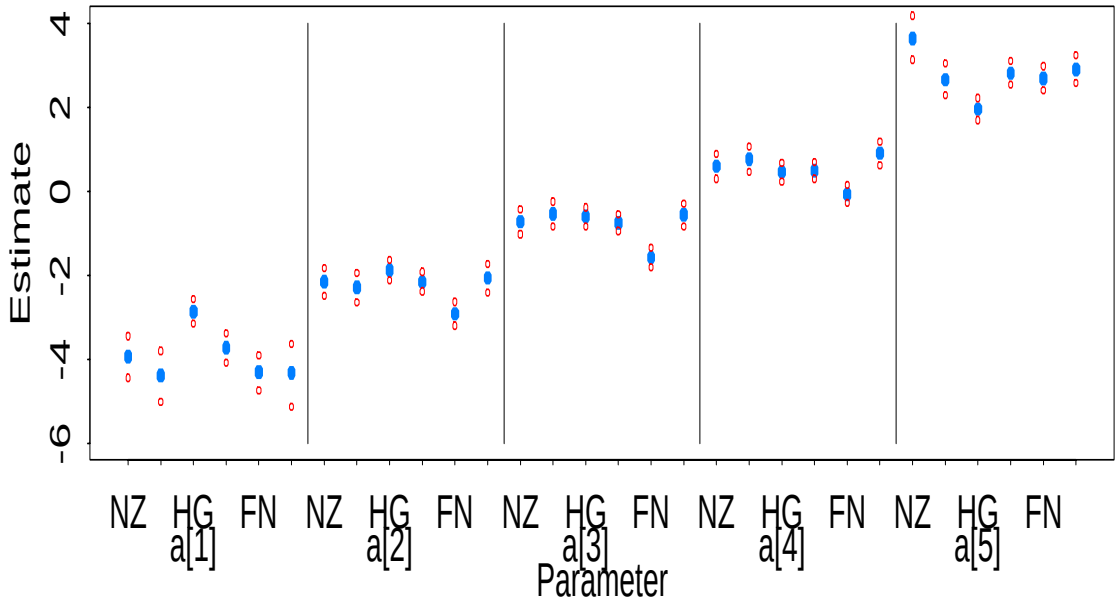
Social Funtioning



Mental Health



Vitality



General Health

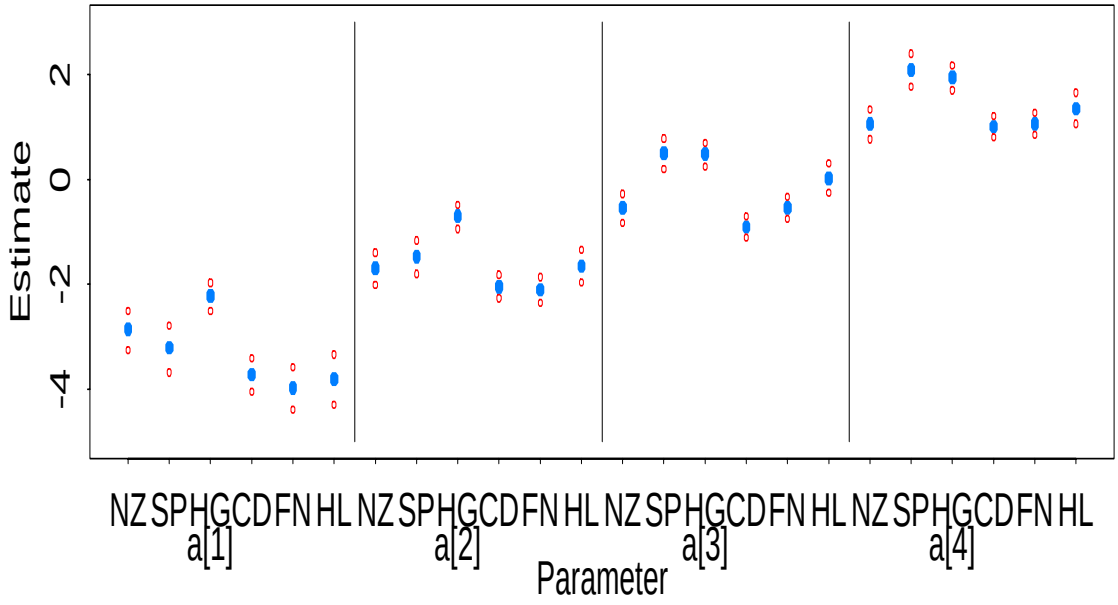
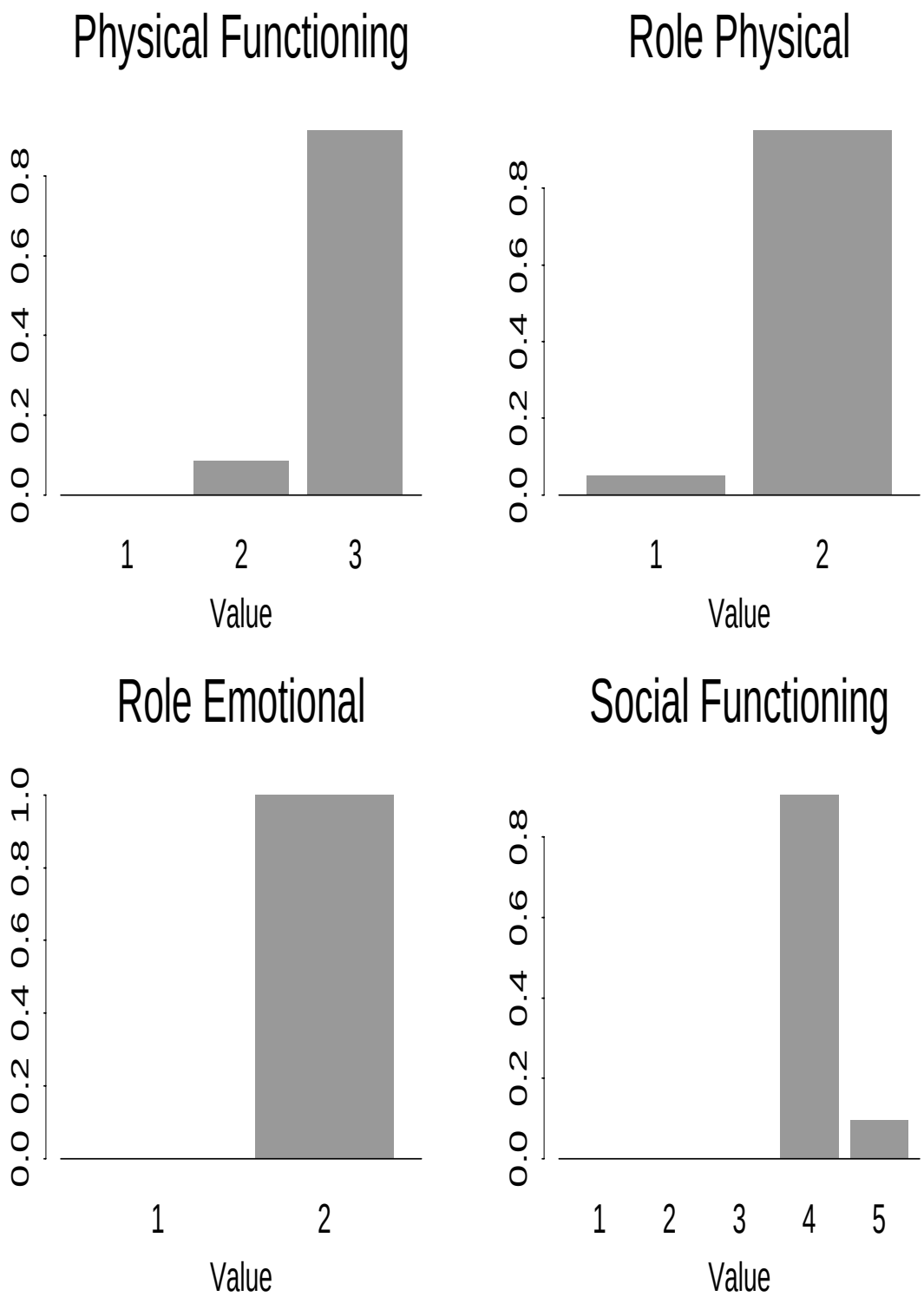
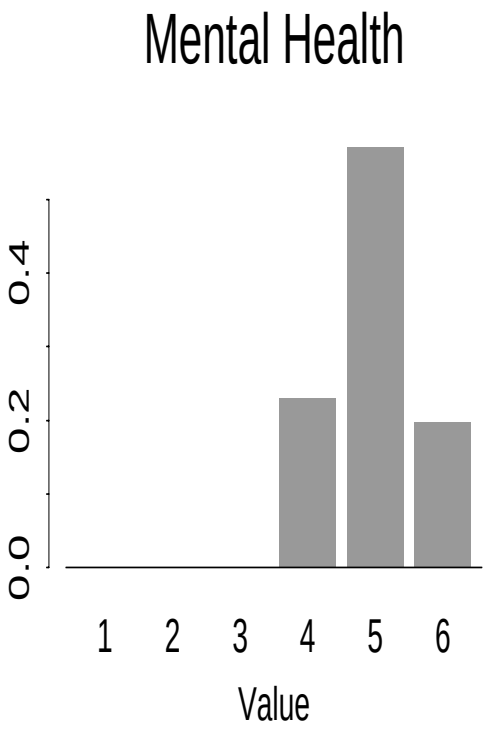
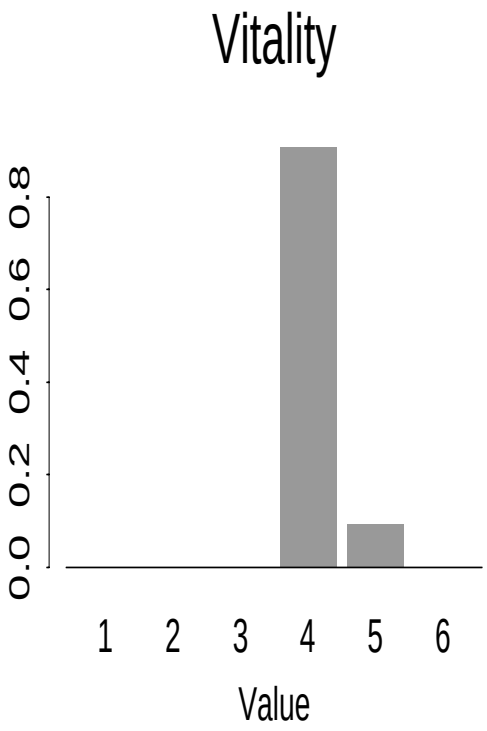
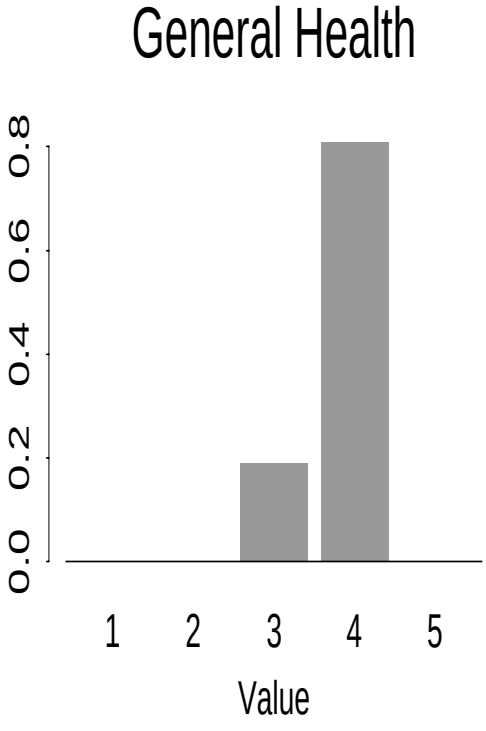


Figure 4.2: Plots of Imputed Values




Mental Health
0.0
0.2
0.4
1
2
3
4
5
6
Value


Vitality
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
6
Value


General Health
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
Value

Bibliography

[1] Best, NG, Cowles, MK, and Vines, SK (1995), CODA Convergence Diagnosis and Output Analysis software for Gibbs Sampler output: Version 0.3.

[2] Casella, G, Berger, RL, (1990), *Statistical Inference*, Wadsworth and Brooks/Cole, Belmont.

[3] Evidence and Information for Policy Cluster World Health Organization (2001), Statistical Methods for Enhancing Cross-Population Comparibility of Survey Results, draft document.

[4] Gelman, A, King, G, Liu, C (1998), Not Asked and Not Answered: Multiple Imputation for Multiple Surveys, *Journal of the American Statistical Association*, **93**, 846-857.

[5] Gelman, A and Rubin, DB (1992), Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**, 457-72.

[6] Geweke, J (1992), Evaluating the accuracy of sampling-based approaches to calculating posterior moments, *Bayesian Statistics*, Fourth edition, Clarendon Press, Oxford, UK.

[7] Gilks, WR, Richardson, S, Spiegelhalter, DJ, (1996), *Markov Chain Monte Carlo in Practice*, Chapman and Hill, Boca Raton.

[8] Heitjan, DF, Little, RJA (1991), Multiple imputation for the fatal accident reporting system, *Applied Statistics*, **40**, 13-29.

[9] McDowell, I, Newell, C, (1996), *Measuring Health: A Guide to Rating Scales and Questionnaries*, second ed., Oxford University Press, New York.

[10] Paulino, C, Pereira, C (1995), Bayesian Methods for Categorical Data Under Informative General Censoring, *Biometrika*, **82**, 439-446.

[11] Proschan, MA, McMahon, RP, Shih, JH, Hunsberger, SA, Geller, NL, Knatterud, G, Wittes, J (2001), Sensitivity analysis using an imputation method for missing binary data in clinical trials, *Journal of Statistical Planning and Inference*, **96**, 155-165.

[12] Raferty, AL and Lewis, S (1992), How many iterations in the Gibbs sampler?, *Bayesian Statistics*, Fourth edition, Oxford University Press, 763-74.

[13] Rubin, DB (1986), Statistical matching using file concatenation with adjusted weights and multiple imputations, *Journal of Business And Economic Statistics*, **4**, 87-94.

[14] Schafer, JL, (1999), "Multiple Imputation: Fabricate your Data Well," Classroom Day presentation, Joint Statistical Meetings, Baltimore, MD, `www.stat.psu.edu/~jls`.

[15] Schafer, JL and Olsen, MK (1997), Multiple Imputation for Multivariate Missing Data Problems: A Data Analyst's Perspective, `www.stat.psu.edu/~jls`.

[16] Shiely, J-C, Bayliss, MS, Keller, SD, et al., (1996), *SF-36 Health Survey Annotated Bibliography: The First Edition (1988-1995)*, The Health Institute New England Medical Center, Boston.

[17] Soares, P, Paulino, C (2001), Incomplete Categorial Data Analysis: A Bayesian Perspective, *Journal of Statistical Computing and Simulation*, **69**, 157-170.

[18] Spiegelhalter, DJ,Thomas, A, and Best, NG (1999), WinBUGS Version 1.2 User Manual MRC Biostatistics Unit.

[19] Spiegelhalter, DJ,Thomas, A, Best, NG, and Gilks, WR (1996), BUGS: Bayesian inference Using Gibbs Sampling, Version 0.5.

[20] Steel, RG, Torrie, JH, Dickey, DA, (1997), *Principles and Procedures of Statistics A Biomedical Approach*, Third edition, McGraw-Hill, New York.

[21] Stewart, AL, Ware, JE, (1992), *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach*, Duke University Press, Durham.

[22] Ware, JE, (2000) SF-36 Health Survey Update, *SPINE*, **Vol.25, No.24**, 3130-3139.

[23] Ware, JE, Snow, KK, Kosinski, M, et al., (1993), *SF-36 Health Survey Manual and Interpretation Guide*, The Health Institute New England Medical Center, Boston.

[24] Zhou, Xiou-Hua, Eckert, George, Tierney, William (2001), Multiple imputation in public health research, *Statistics in Medicine*, **20**, 1541-1549.

Appendix A

SF-36 Questionnaire

SF-36 Health Survey Scoring Demonstration

This survey asks for your views about your health. This information will help you keep track of how you feel and how well you are able to do your usual activities. Answer every question by selecting the answer as indicated. If you are unsure about how to answer a question, please give the best answer you can.

1. In general, would you say your health is:

Fair	Poor
○	○

2. Compared to one year ago, how would you rate your health in general now?

Much better now	Somewhat better now	About the same	Somewhat worse now	Much worse now
○	○	○	○	○

3. The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

	Yes, limited a lot	Yes, limited a little	No, not limited at all
a. Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports	○	○	○
b. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	○	○	○
c. Lifting or carrying groceries	○	○	○
d. Climbing several flights of stairs	○	○	○
e. Climbing one flight of stairs	○	○	○
f. Bending, kneeling, or stooping	○	○	○
g. Walking more than a mile	○	○	○
h. Walking several blocks	○	○	○
i. Walking one block	○	○	○
j. Bathing or dressing yourself	○	○	○

4. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health?

	Yes	No
a. Cut down on the amount of time you spent on work or other activities	○	○
b. Accomplished less than you would like	○	○
c. Were limited in the kind of work or other activities	○	○
d. Had difficulty performing the work or other activities (for example, it took extra effort)	○	○

5. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?

	Yes	No
a. Cut down on the amount of time you spent on work or other activities	○	○
b. Accomplished less than you would like	○	○
c. Didn't do work or other activities as carefully as usual	○	○

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?

Not at all	Slightly	Moderately	Quite a bit	Extremely
○	○	○	○	○

7. How much bodily pain have you had during the past 4 weeks?

None	Very mild	Mild	Moderate	Severe	Very severe
○	○	○	○	○	○

8. During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?

Not at all	A little bit	Moderately	Quite a bit	Extremely
○	○	○	○	○

9. These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past 4 weeks

	All of the time	Most of the time	A good bit of the time	Some of the time	A little of the time	None of the time
a. Did you feel full of pep?	○	○	○	○	○	○
b. Have you been a very nervous person?	○	○	○	○	○	○
c. Have you felt so down in the dumps that nothing could cheer you up?	○	○	○	○	○	○
d. Have you felt calm and peaceful?	○	○	○	○	○	○
e. Did you have a lot of energy?	○	○	○	○	○	○
f. Have you felt downhearted and blue?	○	○	○	○	○	○
g. Did you feel worn out?	○	○	○	○	○	○
h. Have you been a happy person?	○	○	○	○	○	○
i. Did you feel tired?	○	○	○	○	○	○

10. During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting friends, relatives, etc.)?

All of the time	Most of the time	Some of the time	A little of the time	None of the time
○	○	○	○	○

11. How TRUE or FALSE is each of the following statements for you?

	Definitely true	Mostly true	Don't know	Mostly false	Definitely false
a. I seem to get sick a little easier than other people	○	○	○	○	○
b. I am as healthy as anybody I know	○	○	○	○	○
c. I expect my health to get worse	○	○	○	○	○
d. My health is excellent	○	○	○	○	○

Appendix B

SAS vs BUGS Parameter Estimates

Table B.1: SAS vs BUGS Parameter Estimates

Group	Software	Parameter	Estimate	lower limit	upper limit
1	BUG	a1	-5.5620	-5.9120	-5.2710
1	SAS	a1	-5.5192	-5.8282	-5.2102
1	BUG	a2	-3.5360	-3.8720	-3.2580
1	SAS	a2	-3.4950	-3.7923	-3.1977
1	BUG	trt	-0.0807	-0.1682	0.0058
1	SAS	trt	-0.0797	-0.1670	0.0075
1	BUG	sess	-0.0067	-0.0930	0.0803
1	SAS	sess	-0.0056	-0.0928	0.0815
1	BUG	ques1	-3.7260	-4.0690	-3.4360
1	SAS	ques1	-3.6844	-3.9903	-3.3784
1	BUG	ques2	-1.9320	-2.2800	-1.6280
1	SAS	ques2	-1.8934	-2.2124	-1.5745
1	BUG	ques3	-1.8170	-2.1710	-1.5120

1	SAS	ques3	-1.7772	-2.0985	-1.4560
1	BUG	ques4	-2.6610	-3.0110	-2.3660
1	SAS	ques4	-2.6223	-2.9321	-2.3126
1	BUG	ques5	-0.8964	-1.2740	-0.5552
1	SAS	ques5	-0.8622	-1.2128	-0.5116
1	BUG	ques6	-2.3910	-2.7360	-2.0900
1	SAS	ques6	-2.3512	-2.6635	-2.0388
1	BUG	ques7	-1.8840	-2.2350	-1.5810
1	SAS	ques7	-1.8462	-2.1664	-1.5261
1	BUG	ques8	-1.1710	-1.5320	-0.8440
1	SAS	ques8	-1.1338	-1.4734	-0.7942
1	BUG	ques9	-0.3887	-0.7858	-0.0123
1	SAS	ques9	-0.3554	-0.7364	0.0255
2	BUG	a1	-0.8072	-0.9347	-0.6770
2	SAS	a1	-0.8098	-0.9416	-0.6781
2	BUG	trt	-0.0193	-0.1248	0.0865
2	SAS	trt	-0.0211	-0.1277	0.0854
2	BUG	sess	-0.2336	-0.3384	-0.1279
2	SAS	sess	-0.2346	-0.3412	-0.1279
2	BUG	ques1	0.1739	0.0244	0.3270
2	SAS	ques1	0.1727	0.0200	0.3255
2	BUG	ques2	-0.2947	-0.4404	-0.1520
2	SAS	ques2	-0.2961	-0.4437	-0.1486
2	BUG	ques3	0.1134	-0.0358	0.2610
2	SAS	ques3	0.1109	-0.0409	0.2627
3	BUG	a1	-1.5070	-1.6700	-1.3490
3	SAS	a1	-1.5064	-1.6671	-1.3457

3	BUG	trt	-0.0750	-0.2104	0.0592
3	SAS	trt	-0.0761	-0.2142	0.0621
3	BUG	sess	-0.4345	-0.5725	-0.2956
3	SAS	sess	-0.4359	-0.5749	-0.2970
3	BUG	ques1	0.2605	0.0838	0.4395
3	SAS	ques1	0.2605	0.0832	0.4379
3	BUG	ques2	-0.3731	-0.5357	-0.2099
3	SAS	ques2	-0.3721	-0.5356	-0.2086
4	BUG	a1	-4.5360	-4.9100	-4.2090
4	SAS	a1	-4.5170	-4.8614	-4.1726
4	BUG	a2	-2.7000	-2.8830	-2.5220
4	SAS	a2	-2.6969	-2.8774	-2.5165
4	BUG	a3	-1.1110	-1.2440	-0.9764
4	SAS	a3	-1.1109	-1.2505	-0.9713
4	BUG	a4	0.3749	0.2442	0.5077
4	SAS	a4	0.3727	0.2382	0.5072
4	BUG	trt	-0.1238	-0.2541	0.0076
4	SAS	trt	-0.1241	-0.2543	0.0060
4	BUG	sess	-0.2437	-0.3728	-0.1142
4	SAS	sess	-0.2449	-0.3751	-0.1146
4	BUG	ques1	0.1690	0.0407	0.2995
4	SAS	ques1	0.1679	0.0377	0.2981
6	BUG	a1	-4.2960	-4.5220	-4.0750
6	SAS	a1	-4.2839	-4.5094	-4.0585
6	BUG	a2	-2.7630	-2.9040	-2.6260
6	SAS	a2	-2.7571	-2.8986	-2.6156
6	BUG	a3	-1.4260	-1.5420	-1.3100

6	SAS	a3	-1.4212	-1.5376	-1.3048
6	BUG	a4	-0.2678	-0.3743	-0.1581
6	SAS	a4	-0.2650	-0.3746	-0.1553
6	BUG	a5	1.5050	1.3900	1.6200
6	SAS	a5	1.5054	1.3895	1.6213
6	BUG	trt	-0.1343	-0.2157	-0.0504
6	SAS	trt	-0.1330	-0.2158	-0.0503
6	BUG	sess	-0.2357	-0.3192	-0.1529
6	SAS	sess	-0.2340	-0.3168	-0.1511
6	BUG	ques1	0.5978	0.4682	0.7280
6	SAS	ques1	0.5983	0.4696	0.7271
6	BUG	ques2	2.0090	1.8680	2.1520
6	SAS	ques2	2.0082	1.8676	2.1488
6	BUG	ques3	-0.5731	-0.6993	-0.4470
6	SAS	ques3	-0.5716	-0.6988	-0.4443
6	BUG	ques4	0.9686	0.8388	1.1010
6	SAS	ques4	0.9688	0.8382	1.0995
7	BUG	a1	-3.8010	-3.9980	-3.6010
7	SAS	a1	-3.7945	-3.9872	-3.6018
7	BUG	a2	-2.2920	-2.4290	-2.1540
7	SAS	a2	-2.2901	-2.4222	-2.1581
7	BUG	a3	-0.8964	-1.0160	-0.7759
7	SAS	a3	-0.8952	-1.0092	-0.7811
7	BUG	a4	0.3877	0.2700	0.5061
7	SAS	a4	0.3882	0.2764	0.5001
7	BUG	a5	2.5740	2.4280	2.7210
7	SAS	a5	2.5731	2.4330	2.7132

7	BUG	trt	-0.0906	-0.1811	0.0026
7	SAS	trt	-0.0902	-0.1806	0.0002
7	BUG	sess	-0.3499	-0.4432	-0.2590
7	SAS	sess	-0.3496	-0.4404	-0.2589
7	BUG	ques1	0.1442	0.0136	0.2711
7	SAS	ques1	0.1447	0.0183	0.2711
7	BUG	ques2	-0.1159	-0.2444	0.0145
7	SAS	ques2	-0.1157	-0.2427	0.0112
7	BUG	ques3	0.6872	0.5568	0.8207
7	SAS	ques3	0.6873	0.5590	0.8156
8	BUG	a1	-3.3800	-3.5520	-3.2060
8	SAS	a1	-3.3700	-3.5379	-3.2021
8	BUG	a2	-1.6940	-1.8140	-1.5710
8	SAS	a2	-1.6884	-1.8076	-1.5691
8	BUG	a3	-0.3065	-0.4148	-0.1927
8	SAS	a3	-0.3024	-0.4109	-0.1939
8	BUG	a4	1.2890	1.1760	1.4050
8	SAS	a4	1.2908	1.1790	1.4026
8	BUG	trt	0.0726	-0.0076	0.1557
8	SAS	trt	0.0746	-0.0068	0.1561
8	BUG	sess	-0.0517	-0.1331	0.0299
8	SAS	sess	-0.0504	-0.1318	0.0310
8	BUG	ques1	-0.4631	-0.5872	-0.3378
8	SAS	ques1	-0.4605	-0.5839	-0.3370
8	BUG	ques2	0.7638	0.6304	0.8975
8	SAS	ques2	0.7648	0.6332	0.8965
8	BUG	ques3	0.3607	0.2332	0.4895

8	SAS	ques3	0.3625	0.2350	0.4900
8	BUG	ques4	0.7530	0.6213	0.8872
8	SAS	ques4	0.7541	0.6224	0.8858

Appendix C

WinBUGS

C.1 Implementation Using WinBUGS

For implementing Gibbs sampling we will be using `WinBUGS`. `WinBUGS` is a windows based software package that is specialized for implementing MCMC-based analysis of full probability models. `WinBUGS` provides a graphical interface to the `BUGS` language. It allows the user to easily construct models and is available on the World Wide Web. To implement a graphical model using `WinBUGS`, the model must be constructed and a distribution must be assigned to each node. The model is then checked by `WinBUGS` to see that it is syntactically correct. The observed data is then read in. Next, the model is compiled. When `BUGS` compiles the model, it is building data structures that are needed to carry out the Gibbs sampling. The model is also checked for consistency with the data. Initial values for the nodes must be loaded, then the MCMC simulation can be run. The simulated values will need to be checked for convergence. When compiling the model in `WinBUGS`, the user is asked the number of chains. An integer is entered and this denotes how many chains you want to run simultaneously. The *Sample Monitor Tool* has two open slots labelled "beg" and "end" which can be used to specify which iterations should be included in the calculation of any summary statistics. By specifying the "beg" equal to $m + 1$, you are denoting the first m iterations as the 'burn-in'. There is also an open slot labelled "thin". By putting a number greater than 1, you are not using all the updated components in each iteration. There is also an open slot labelled "thin" in the *Model Update Tool*. By making this value greater than 1, you are not updating all the components in each iteration.

C.2 WinBUGS Code

```
#
```

```
 Model {
#
    for (m in 1:M) {
         for (j in 1 : J-1) {
#
# Cumulative probability of better response than j
#
                logit(Q[m, j]) <- (a[j] - mu[m] )
            }
#
# Probability of response = j
#
            p[m, 1] <-  Q[m, 1]
            for (j in 2 : (J-1))
   { p[m,  j] <- Q[m,  j ] - Q[m,  j-1] }
            p[m,  J] <- 1 - Q[m, (J-1)]

#           response[m] ~ dcat(p[m,  ])
#
# Fixed effects
#
# logistic mean for group i in period t
   mu[m] <- QUES [question[m]] + TRT [treatment[m]]
            + SESS [session[m]]
    }
#priors
#
#code from Dr. Ghosh for ranked alpha
for (j in 1 : (J-1)){
   alphastar[j] ~ dflat()
   a[j] <- ranked(alphastar[], j) }

QUES[I] <- 0
for (i in 2 : (I-1)) {
  QUES[i] ~ dflat() }

    TRT[K] <- 0
for (k in 2 : (K-1)) {
  TRT[k] ~ dflat() }

SESS[L] <- 0
for (l in 2 : (L-1)) {
```

```
      SESS[l] ~ dflat() }

    }
################
#data for group 1
################
list( M=19240,J=3,I=10,K=2,L=2)

#inits
list(alphastar=c(-1.5,1.5), QUES=c(0,0,0,0,0,0,0,0,0,NA), TRT=c(0,NA),
SESS=c(0,NA))

list(alphastar=c(-2.5,2.5), QUES=c(-3,-3,-3,-3,-3,-3,-3,-3,-3,NA),
TRT=c(-3,NA), SESS=c(-3,NA))

list(alphastar=c(-3.5,3.5), QUES=c(3,3,3,3,3,3,3,3,3,NA), TRT=c(3,NA),
SESS=c(3,NA))

################
#data for group 2
################
list( M=7696,J=2,I=4,K=2,L=2)

#inits
list(alphastar=c(1.5), QUES=c(0,0,0,NA), TRT=c(0,NA), SESS=c(0,NA))

list(alphastar=c(2.5), QUES=c(-3,-3,-3,NA), TRT=c(-3,NA),
SESS=c(-3,NA))

list(alphastar=c(3.5), QUES=c(3,3,3,NA), TRT=c(3,NA), SESS=c(3,NA))

################
#data for group 3
################
list( M=5772,J=2,I=3,K=2,L=2)

#inits
list(alphastar=c(1.5), QUES=c(0,0,NA), TRT=c(0,NA), SESS=c(0,NA))

list(alphastar=c(2.5), QUES=c(-3,-3,NA), TRT=c(-3,NA), SESS=c(-3,NA))

list(alphastar=c(3.5), QUES=c(3,3,NA), TRT=c(3,NA), SESS=c(3,NA))
```

```
################
#data for group 4
################
list( M=3848,J=5,I=2,K=2,L=2)

#inits
list(alphastar=c(-1.5,-.5,.5,1.5), QUES=c(0,NA), TRT=c(0,NA),
SESS=c(0,NA))

list(alphastar=c(-2.5,-1,1,2.5), QUES=c(-3,NA), TRT=c(-3,NA),
SESS=c(-3,NA))

list(alphastar=c(-3.5,-1.5,1.5,3.5), QUES=c(3,NA), TRT=c(3,NA),
SESS=c(3,NA))

################
#data for group 6
################
list( M=9620,J=6,I=5,K=2,L=2)

#inits
list(alphastar=c(-1.5,-1,-.5,1,2), QUES=c(0,0,0,0,NA), TRT=c(0,NA),
SESS=c(0,NA))

list(alphastar=c(-2.5,-1.5,-.5,1.5,3), QUES=c(-3,-3,-3,-3,NA),
TRT=c(-3,NA),
SESS=c(-3,NA))

list(alphastar=c(-3.5,-2,-.5,2,4), QUES=c(3,3,3,3,NA), TRT=c(3,NA),
SESS=c(3,NA))

################
#data for group 7
################
list( M=7696,J=6,I=4,K=2,L=2)

#inits
list(alphastar=c(-1.5,-1,-.5,1,2), QUES=c(0,0,0,NA), TRT=c(0,NA),
SESS=c(0,NA))

list(alphastar=c(-2.5,-1.5,-.5,1.5,3), QUES=c(-3,-3,-3,NA),
```

```
TRT=c(-3,NA),
SESS=c(-3,NA))

list(alphastar=c(-3.5,-2,-.5,2,4), QUES=c(3,3,3,NA), TRT=c(3,NA),
SESS=c(3,NA))

################
#data for group 8
################
list( M=9620,J=5,I=5,K=2,L=2)

#inits
list(alphastar=c(-1.5,-.5,.5,1.5), QUES=c(0,0,0,0,NA), TRT=c(0,NA),
SESS=c(0,NA))

list(alphastar=c(-2.5,-1,1,2.5), QUES=c(-3,-3,-3,-3,NA),
TRT=c(-3,NA), SESS=c(-3,NA))

list(alphastar=c(-3.5,-1.5,1.5,3.5), QUES=c(3,3,3,3,NA), TRT=c(3,NA),
SESS=c(3,NA))
```

Appendix D

Imputation Results

Table D.1: Physical Functioning

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1]	-5.206	0.1305	7.883E-3	-5.450	-5.206	-4.940
a[2]	-3.289	0.1262	7.798E-3	-3.519	-3.289	-3.029
QUES[1]	-3.484	0.1295	7.438E-3	-3.729	-3.484	-3.220
QUES[2]	-1.810	0.1362	7.488E-3	-2.067	-1.811	-1.534
QUES[3]	-1.678	0.1378	7.512E-3	-1.940	-1.678	-1.396
QUES[4]	-2.452	0.1326	7.498E-3	-2.702	-2.453	-2.181
QUES[5]	-0.851	0.1499	7.562E-3	-1.136	-0.853	-0.544
QUES[6]	-2.187	0.1333	7.459E-3	-2.440	-2.187	-1.915
QUES[7]	-1.739	0.1369	7.470E-3	-2.002	-1.739	-1.463
QUES[8]	-1.121	0.1440	7.456E-3	-1.394	-1.121	-0.828
QUES[9]	-0.320	0.1616	7.473E-3	-0.633	-0.322	0.011
SESS[1]	-0.102	0.0395	6.318E-4	-0.180	-0.101	-0.024
TRT[1]	-0.022	0.0389	5.977E-4	-0.098	-0.022	0.054
deviance	19550	5.217	1.173E-1	19540	19550	19560

Table D.2: Role Physical

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1]	-0.702	0.0609	2.398E-3	-0.825	-0.702	-0.584
QUES[1]	0.188	0.0700	1.688E-3	0.050	0.188	0.325
QUES[2]	-0.305	0.0681	1.684E-3	-0.437	-0.304	-0.171
QUES[3]	0.123	0.0696	1.679E-3	-0.013	0.122	0.260
SESS[1]	-0.276	0.0488	9.958E-4	-0.371	-0.276	-0.180
TRT[1]	0.065	0.0483	9.058E-4	-0.029	0.064	0.159
deviance	9453	3.4440	5.645E-2	9448	9452	9461

Table D.3: Role Emotional

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1]	-1.433	0.0753	2.813E-3	-1.583	-1.433	-1.291
QUES[1]	0.241	0.0808	1.760E-3	0.084	0.241	0.400
QUES[2]	-0.372	0.0757	1.763E-3	-0.521	-0.372	-0.223
SESS[1]	-0.447	0.0638	1.370E-3	-0.574	-0.446	-0.324
TRT[1]	-0.042	0.0635	1.237E-3	-0.166	-0.043	0.083
deviance	6050	3.16	0.04801	6045	6049	6057

Table D.4: Social Functioning

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1]	-4.312	0.1495	4.388E-3	-4.604	-4.309	-4.023
a[2]	-2.567	0.0849	3.532E-3	-2.724	-2.569	-2.395
a[3]	-1.008	0.0685	3.155E-3	-1.140	-1.009	-0.871
a[4]	0.456	0.0673	3.029E-3	0.325	0.455	0.591
QUES[1]	0.205	0.0614	1.560E-3	0.084	0.205	0.325
SESS[1]	-0.291	0.0619	1.730E-3	-0.412	-0.291	-0.169
TRT[1]	-0.049	0.0619	1.668E-3	-0.171	-0.049	0.072
deviance	9466	3.767	7.684E-2	9461	9466	9475

Table D.5: Mental Health

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1]	-4.192	0.1033	3.615E-3	-4.394	-4.189	-3.990
a[2]	-2.666	0.0669	3.108E-3	-2.799	-2.666	-2.537
a[3]	-1.334	0.0565	2.908E-3	-1.444	-1.335	-1.224
a[4]	-0.187	0.0535	2.770E-3	-0.289	-0.188	-0.080
a[5]	1.538	0.0560	2.730E-3	1.429	1.537	1.650
QUES[1]	0.565	0.0609	1.972E-3	0.445	0.564	0.685
QUES[2]	1.945	0.0651	1.931E-3	1.819	1.945	2.074
QUES[3]	-0.566	0.0599	2.014E-3	-0.682	-0.567	-0.449
QUES[4]	0.936	0.0616	1.973E-3	0.815	0.936	1.058
SESS[1]	-0.296	0.0388	9.541E-4	-0.373	-0.295	-0.221
TRT[1]	-0.058	0.0388	9.014E-4	-0.134	-0.058	0.017
deviance	25590	4.731	9.440E-2	25580	25590	25600

Table D.6: Vitality

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1]	-3.660	0.0840	2.933E-3	-3.824	-3.660	-3.495
a[2]	-2.193	0.0586	2.658E-3	-2.310	-2.192	-2.081
a[3]	-0.799	0.0497	2.430E-3	-0.898	-0.798	-0.702
a[4]	0.463	0.0498	2.382E-3	0.366	0.463	0.561
a[5]	2.567	0.0629	2.455E-3	2.444	2.567	2.692
QUES[1]	0.103	0.0579	1.718E-3	-0.012	0.103	0.215
QUES[2]	-0.131	0.0582	1.707E-3	-0.244	-0.131	-0.018
QUES[3]	0.657	0.0585	1.660E-3	0.542	0.656	0.770
SESS[1]	-0.373	0.0419	9.679E-4	-0.457	-0.373	-0.291
TRT[1]	-0.005	0.0414	8.487E-4	-0.086	-0.004	0.076
deviance	22320	4.469	7.920E-2	22320	22320	22330

Table D.7: General Health

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1]	-3.191	0.0743	2.935E-3	-3.340	-3.190	-3.044
a[2]	-1.594	0.0564	2.816E-3	-1.707	-1.593	-1.485
a[3]	-0.243	0.0525	2.736E-3	-0.351	-0.242	-0.142
a[4]	1.324	0.0540	2.706E-3	1.216	1.325	1.428
QUES[1]	-0.416	0.0585	2.027E-3	-0.530	-0.416	-0.302
QUES[2]	0.753	0.0615	1.990E-3	0.631	0.753	0.872
QUES[3]	0.376	0.0603	1.995E-3	0.256	0.377	0.492
QUES[4]	0.772	0.0620	1.984E-3	0.650	0.772	0.892
SESS[1]	-0.084	0.0382	8.971E-4	-0.160	-0.083	-0.009
TRT[1]	0.107	0.0384	8.787E-4	0.032	0.107	0.182
deviance	25280	4.47	9.135E-2	25280	25280	25290

Appendix E

Validation Simultation Results Where TRT = 1

Parameter	device	miss	estimate	sd	cov prob	length	t	p-val
a_1	TRUE		-2					
a_1	SAS	0	-2.0455	0.2249	0.94	0.882	-2.024	0.0430
Error	SAS	0	0.2118	0.0090				
a_1	SAS	10	-2.0518	0.2375	0.99	0.931	-2.179	0.0294
Error	SAS	10	0.2233	0.0102				
a_1	BUG	10	-2.0792	0.2300	0.94	0.906	-3.438	0.0006
Error	BUG	10	0.2204	0.0134				
a_1	SAS	35	-2.0646	0.2811	0.96	1.102	-2.296	0.0217
Error	SAS	35	0.2681	0.0139				
a_1	BUG	35	-2.0893	0.2498	0.94	0.974	-3.567	0.0004
Error	BUG	35	0.2533	0.0158				
a_1	SAS	50	-2.0243	0.3180	0.96	1.247	-0.764	0.4449
Error	SAS	50	0.3242	0.0191				
a_1	BUG	50	-2.0727	0.2606	0.95	1.017	-2.784	0.0054

Error	BUG	50	0.2620	0.0164				
a_2	TRUE		-1					
a_2	SAS	0	-1.0375	0.1985	0.93	0.778	-1.890	0.0588
Error	SAS	0	0.2116	0.0067				
a_2	SAS	10	-1.0444	0.2095	0.92	0.821	-2.117	0.0342
Error	SAS	10	0.2198	0.0074				
a_2	BUG	10	-1.0495	0.2028	0.91	0.795	-2.436	0.0148
Error	BUG	10	0.2178	0.0105				
a_2	SAS	35	-1.0333	0.2476	0.93	0.971	-1.345	0.1786
Error	SAS	35	0.2723	0.0102				
a_2	BUG	35	-1.0447	0.2199	0.94	0.862	-2.030	0.0423
Error	BUG	35	0.2439	0.0130				
a_2	SAS	50	-1.0434	0.2814	0.94	1.103	-1.542	0.1232
Error	SAS	50	0.3005	0.0129				
a_2	BUG	50	-1.0565	0.2304	0.94	0.905	-2.449	0.0143
Error	BUG	50	0.2508	0.0148				
a_3	TRUE		1					
a_3	SAS	0	0.9836	0.1965	0.94	0.770	-0.833	0.4050
Error	SAS	0	0.1956	0.0055				
a_3	SAS	10	0.9701	0.2070	0.96	0.811	-1.446	0.1482
Error	SAS	10	0.1959	0.0062				
a_3	BUG	10	0.9893	0.2018	0.95	0.789	-0.528	0.5972
Error	BUG	10	0.1963	0.0135				
a_3	SAS	35	0.9958	0.2450	0.93	0.961	-0.170	0.8649
Error	SAS	35	0.2737	0.0097				
a_3	BUG	35	0.9933	0.2171	0.94	0.850	-0.306	0.7596
Error	BUG	35	0.2289	0.0151				

a_3	SAS	50	0.9791	0.2782	0.97	1.091	-0.752	0.4520
Error	SAS	50	0.2570	0.0110				
a_3	BUG	50	0.9934	0.2275	0.98	0.892	-0.290	0.7718
Error	BUG	50	0.2142	0.0146				
a_4	TRUE		2					
a_4	SAS	0	1.9885	0.2139	0.97	0.838	-0.538	0.5903
Error	SAS	0	0.2159	0.0070				
a_4	SAS	10	1.9743	0.2252	0.97	0.883	-1.141	0.2538
Error	SAS	10	0.2181	0.0077				
a_4	BUG	10	2.0100	0.2201	0.97	0.862	0.454	0.6500
Error	BUG	10	0.2220	0.0143				
a_4	SAS	35	1.9927	0.2665	0.94	1.044	-0.274	0.7844
Error	SAS	35	0.2938	0.0124				
a_4	BUG	35	2.0155	0.2369	0.93	0.926	0.654	0.5130
Error	BUG	35	0.2485	0.0171				
a_4	SAS	50	1.9851	0.3031	0.98	1.188	-0.492	0.6225
Error	SAS	50	0.2876	0.0134				
a_4	BUG	50	2.0190	0.2479	0.97	0.966	0.763	0.4452
Error	BUG	50	0.2352	0.0153				
QUES	TRUE		1					
QUES	SAS	0	1.0066	0.2061	0.96	0.808	0.320	0.7491
Error	SAS	0	0.2127	0.0044				
QUES	SAS	10	1.0076	0.2175	0.96	0.853	0.347	0.7284
Error	SAS	10	0.2117	0.0050				
QUES	BUG	10	1.0207	0.2135	0.96	0.837	0.970	0.3320
Error	BUG	10	0.2130	0.0080				
QUES	SAS	35	1.0095	0.2565	0.92	1.006	0.371	0.7110

Error	SAS	35	0.2771	0.0076				
QUES	BUG	35	1.0139	0.2284	0.96	0.895	0.607	0.5440
Error	BUG	35	0.2410	0.0084				
QUES	SAS	50	1.0060	0.2923	0.94	1.146	0.204	0.8386
Error	SAS	50	0.3116	0.0099				
QUES	BUG	50	1.0220	0.2399	0.97	0.941	0.915	0.3601
Error	BUG	50	0.2497	0.0086				
TRT	TRUE		1					
TRT	SAS	0	0.9827	0.2062	0.93	0.808	-0.838	0.4021
Error	SAS	0	0.2101	0.0044				
TRT	SAS	10	0.9668	0.2176	0.94	0.853	-1.526	0.1269
Error	SAS	10	0.2220	0.0051				
TRT	BUG	10	0.9859	0.2124	0.94	0.832	-0.662	0.5081
Error	BUG	10	0.2190	0.0067				
TRT	SAS	35	0.9840	0.2570	0.94	1.007	-0.622	0.5340
Error	SAS	35	0.2575	0.0066				
TRT	BUG	35	1.0074	0.2285	0.93	0.896	0.322	0.7472
Error	BUG	35	0.2324	0.0085				
TRT	SAS	50	0.9682	0.2930	0.97	1.149	-1.083	0.2788
Error	SAS	50	0.2878	0.0097				
TRT	BUG	50	1.0004	0.2405	0.94	0.941	0.018	0.9853
Error	BUG	50	0.2417	0.0078				
BETA	TRUE		2					
BETA	SAS	0	2.0278	0.1500	0.96	0.588	1.851	0.0641
Error	SAS	0	0.1472	0.0079				
BETA	SAS	10	2.0312	0.1585	0.95	0.621	1.967	0.0492
Error	SAS	10	0.1585	0.0091				

BETA	BUG	10	2.0562	0.1554	0.95	0.610	3.609	0.0003
Error	BUG	10	0.1587	0.0090				
BETA	SAS	35	2.0308	0.1868	0.96	0.732	1.645	0.0999
Error	SAS	35	0.1829	0.0123				
BETA	BUG	35	2.0617	0.1671	0.98	0.655	3.687	0.0002
Error	BUG	35	0.1652	0.0103				
BETA	SAS	50	2.0201	0.2137	0.96	0.838	0.939	0.3476
Error	SAS	50	0.2128	0.0171				
BETA	BUG	50	2.0499	0.1744	0.92	0.684	2.857	0.0043
Error	BUG	50	0.1691	0.0104				

Appendix F

Validation Simultation Results Where TRT = 5

Parameter	device	miss	mean	sd	cov prop	length	t	p-val
a_1	TRUE		-2					
a_1	SAS	0	-2.0160	0.2625	0.92	1.029	-0.608	0.5435
Error	SAS	0	0.2814	0.0158				
a_1	SAS	10	-2.0168	0.2771	0.94	1.086	-0.605	0.5453
Error	SAS	10	0.3070	0.0173				
a_1	BUG	10	-2.0572	0.2684	0.93	1.055	-2.126	0.0335
Error	BUG	10	0.2969	0.0195				
a_1	SAS	35	-2.0614	0.3304	0.93	1.295	-1.853	0.0639
Error	SAS	35	0.3658	0.0270				
a_1	BUG	35	-2.0907	0.2932	0.91	1.150	-3.085	0.0020
Error	BUG	35	0.3082	0.0218				
a_1	SAS	50	-2.0348	0.3775	0.96	1.480	-0.918	0.3585
Error	SAS	50	0.3786	0.0311				
a_1	BUG	50	-2.0826	0.3089	0.96	1.214	-2.665	0.0077

Error	BUG	50	0.3218	0.0233				
a_2	TRUE		-1					
a_2	SAS	0	-0.9933	0.2264	0.94	0.887	0.296	0.7669
Error	SAS	0	0.2374	0.0098				
a_2	SAS	10	-0.9889	0.2389	0.98	0.936	0.464	0.6425
Error	SAS	10	0.2469	0.0105				
a_2	BUG	10	-1.0043	0.2306	0.94	0.904	-0.186	0.8523
Error	BUG	10	0.2455	0.0134				
a_2	SAS	35	-0.9982	0.2832	0.92	1.110	0.064	0.9491
Error	SAS	35	0.3236	0.0171				
a_2	BUG	35	-1.0047	0.2504	0.93	0.982	-0.189	0.8502
Error	BUG	35	0.2776	0.0178				
a_2	SAS	50	-0.9963	0.3248	0.96	1.273	0.113	0.9101
Error	SAS	50	0.3196	0.0189				
a_2	BUG	50	-1.0043	0.2655	0.95	1.038	-0.160	0.8730
Error	BUG	50	0.2679	0.0176				
a_3	TRUE		1					
a_3	SAS	0	1.0514	0.2240	0.93	0.878	2.294	0.0218
Error	SAS	0	0.2587	0.0098				
a_3	SAS	10	1.0593	0.2368	0.92	0.928	2.503	0.0123
Error	SAS	10	0.2684	0.0108				
a_3	BUG	10	1.0666	0.2308	0.90	0.905	2.879	0.0040
Error	BUG	10	0.2740	0.0144				
a_3	SAS	35	1.0402	0.2794	0.95	1.095	1.436	0.1510
Error	SAS	35	0.3181	0.0158				
a_3	BUG	35	1.0684	0.2466	0.89	0.966	2.767	0.0057
Error	BUG	35	0.2885	0.0182				

a_3	SAS	50	1.1099	0.3230	0.94	1.266	3.397	0.0007
Error	SAS	50	0.3549	0.0189				
a_3	BUG	50	1.0937	0.2625	0.94	1.028	3.562	0.0004
Error	BUG	50	0.2866	0.0189				
a_4	TRUE		2					
a_4	SAS	0	2.0499	0.2474	0.93	0.970	2.016	0.0438
Error	SAS	0	0.2678	0.0119				
a_4	SAS	10	2.0640	0.2617	0.95	1.026	2.441	0.0146
Error	SAS	10	0.2660	0.0129				
a_4	BUG	10	2.0954	0.2567	0.93	1.006	3.707	0.0002
Error	BUG	10	0.2794	0.0170				
a_4	SAS	35	2.0455	0.3085	0.92	1.209	1.472	0.1411
Error	SAS	35	0.3411	0.0193				
a_4	BUG	35	2.0948	0.2736	0.88	1.072	3.456	0.0005
Error	BUG	35	0.3106	0.0210				
a_4	SAS	50	2.0904	0.3558	0.92	1.395	2.536	0.0112
Error	SAS	50	0.3788	0.0236				
a_4	BUG	50	2.1181	0.2911	0.94	1.144	4.047	$<.0001$
Error	BUG	50	0.2865	0.0205				
QUES	TRUE		1					
QUES	SAS	0	1.0597	0.2570	0.94	1.007	2.323	0.0202
Error	SAS	0	0.2700	0.0098				
QUES	SAS	10	1.0734	0.2716	0.91	1.064	2.700	0.0069
Error	SAS	10	0.3009	0.0114				
QUES	BUG	10	1.0888	0.2663	0.90	1.043	3.330	0.0009
Error	BUG	10	0.2903	0.0123				
QUES	SAS	35	1.0578	0.3200	0.95	1.255	1.803	0.0713

Error	SAS	35	0.3432	0.0150				
QUES	BUG	35	1.0834	0.2848	0.93	1.118	2.925	0.0034
Error	BUG	35	0.3074	0.0154				
QUES	SAS	50	1.0961	0.3681	0.89	1.443	2.607	0.0091
Error	SAS	50	0.4287	0.0209				
QUES	BUG	50	1.1009	0.3016	0.95	1.183	3.341	0.0008
Error	BUG	50	0.3258	0.0163				
TRT	TRUE		5					
TRT	SAS	0	5.0900	0.4134	0.96	1.621	2.171	0.0299
Error	SAS	0	0.3982	0.0322				
TRT	SAS	10	5.1118	0.4382	0.96	1.718	2.544	0.0110
Error	SAS	10	0.4291	0.0371				
TRT	BUG	10	5.1981	0.4313	0.95	1.695	4.578	<.0001
Error	BUG	10	0.4231	0.0368				
TRT	SAS	35	5.0821	0.5146	0.98	2.017	1.589	0.1121
Error	SAS	35	0.4708	0.0480				
TRT	BUG	35	5.2025	0.4627	0.95	1.812	4.359	<.0001
Error	BUG	35	0.4678	0.0427				
TRT	SAS	50	5.2000	0.6002	0.93	2.353	3.308	0.0009
Error	SAS	50	0.6141	0.0727				
TRT	BUG	50	5.2734	0.4932	0.91	1.934	5.514	<.0001
Error	BUG	50	0.4813	0.0515				
BETA	TRUE		2					
BETA	SAS	0	2.0398	0.1844	0.96	0.723	2.151	0.0315
Error	SAS	0	0.1897	0.0136				
BETA	SAS	10	2.0478	0.1950	0.96	0.765	2.442	0.0146
Error	SAS	10	0.2053	0.0154				

BETA	BUG	10	2.0841	0.1913	0.93	0.749	4.383	<.0001
Error	BUG	10	0.2012	0.0148				
BETA	SAS	35	2.0678	0.2313	0.93	0.907	2.922	0.0035
Error	SAS	35	0.2436	0.0206				
BETA	BUG	35	2.0836	0.2059	0.94	0.808	4.048	<.0001
Error	BUG	35	0.2176	0.0170				
BETA	SAS	50	2.0708	0.2673	1.00	1.048	2.635	0.0084
Error	SAS	50	0.2259	0.0256				
BETA	BUG	50	2.1114	0.2195	0.94	0.860	5.056	<.0001
Error	BUG	50	0.2132	0.0191				

Appendix G

WinBUGS Code

```
#
 Model {
#
    for (m in 1:M) {
        for (c in 1:C) {
         for (j in 1 : J-1) {

#
# Cumulative probability of better response than j
#
 logit(Q[m,c, j]) <- (a[latent[c],j] - mu[m] )

                }




#
# Probability of response = j
#
   p[m,c,1] <-  Q[m, c,1]
   for (j in 2 : (J-1))
       { p[m, c, j] <- Q[m, c, j ] - Q[m,c, j-1] }
   p[m,c, J] <- 1 - Q[m,c, (J-1)]}
```

```
#
# Fixed effects
#

# logistic mean for group i in period t
  mu[m] <- QUES [question[m]] + TRT [treatment[m]]
           + SESS [session[m]]




response[m] ~ dcat(p[m,country[m], ])

}

  for (c in 1:C) {
  latent[c] ~ dcat(prob[]) }


#priors

for (w in 1:W){
probstar[w] ~ dflat()
prob[w] <- ranked(probstar[], w)}


for (w in 1:W){
for (j in 1 : (J-1)){
   alphastar[w,j] ~ dflat()
   a[w,j] <- ranked(alphastar[w,],  j) }  }

QUES[I] <- 0
for (i in 1 : (I-1)) {
  QUES[i] ~ dflat() }

TRT[K] <- 0
for (k in 1 : (K-1)) {
  TRT[k] ~ dflat() }

SESS[L] <- 0
for (l in 1 : (L-1)) {
```

```
     SESS[l] ~ dflat() }

   }

################
#data for group 1
################
list( M=19240,J=3,I=10,K=2,L=2,C=6,W=3)

#inits
list(alphastar=structure(.Data=c(-1.5,1.5,
                        -1.5,1.5,
                    -1.5,1.5), .Dim=c(3,2)),
     QUES=c(0,0,0,0,0,0,0,0,0,NA), TRT=c(0,NA), SESS=c(0,NA),
     latent=c(1,1,1,2,3,1), probstar=c(.15,.35,.50))


################
#data for group 2
################
list( M=7696,J=2,I=4,K=2,L=2,C=6,W=3)


#inits
list(alphastar=structure(.Data=c(1.5,1.5,1.5), .Dim=c(3,1)),
     QUES=c(0,0,0,NA), TRT=c(0,NA), SESS=c(0,NA),
     latent=c(1,2,1,1,3,1), probstar=c(.15,.35,.5))

################
#data for group 3
################
list( M=5772,J=2,I=3,K=2,L=2,C=6,W=2)


#inits
list(alphastar=structure(.Data=c(1.5,1.5), .Dim=c(2,1)),
     QUES=c(0,0,NA), TRT=c(0,NA), SESS=c(0,NA),
     latent=c(1,1,1,2,2,2), probstar=c(.25,.75))

################
#data for group 4
################
```

```
list( M=3848,J=5,I=2,K=2,L=2,C=6,W=2)

#inits
list(alphastar=structure(.Data=c(-1.5,-.5,.5,1.5,
                                 -1.5,-.5,.5,1.5), .Dim=c(2,4)),
     QUES=c(0,NA), TRT=c(0,NA), SESS=c(0,NA),
     latent=c(1,1,1,1,2,1), probstar=c(.25,.75))

################
#data for group 6
################
list( M=9620,J=6,I=5,K=2,L=2,C=6,W=3)

#inits
list(alphastar=structure(.Data=c(-1.5,-1,-.5,1,2,
                                 -1.5,-1,-.5,1,2,
                                 -1.5,-1,-.5,1,2), .Dim=c(3,5)),
     QUES=c(0,0,0,0,NA), TRT=c(0,NA), SESS=c(0,NA),
     latent=c(1,2,3,1,1,1), probstar=c(.15,.35,.5))

#################
# data for group 7
################

list( M=7696,J=6,I=4,K=2,L=2,C=6,W=2)

#inits
list(alphastar=structure(.Data=c(-1.5,-1,-.5,1,2,
                                 -1.5,-1,-.5,1,2), .Dim=c(2,5)),
     QUES=c(0,0,0,NA), TRT=c(0,NA), SESS=c(0,NA),
     latent=c(1,1,1,1,2,2), probstar=c(.3,.7))


#################
# data for group 8
################

list( M=9620,J=5,I=5,K=2,L=2,C=6,W=2)


#inits
list(alphastar=structure(.Data=c(-1.5,-.5,.5,1.5,
```

```
                                -1.5,-.5,.5,1.5), .Dim=c(2,4)),
QUES=c(0,0,0,0,NA), TRT=c(0,NA), SESS=c(0,NA),
latent=c(1,1,2,1,1,1),probstar=c(.4,.6))
```

Appendix H

Final Model Imputation Results

Table H.1: Physical Functioning

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1,1]	-6.648	0.2070	1.698E-2	-7.064	-6.653	-6.245
a[1,2]	-4.004	0.1472	1.570E-2	-4.263	-4.007	-3.715
a[2,1]	-5.461	0.1662	1.672E-2	-5.773	-5.465	-5.134
a[2,2]	-3.360	0.1460	1.612E-2	-3.613	-3.370	-3.067
a[3,1]	-4.896	0.1457	1.606E-2	-5.136	-4.903	-4.608
a[3,2]	-3.104	0.1414	1.589E-2	-3.337	-3.114	-2.827
QUES[1]	-3.559	0.1442	1.536E-2	-3.817	-3.564	-3.275
QUES[2]	-1.843	0.1505	1.552E-2	-2.118	-1.845	-1.544
QUES[3]	-1.710	0.1519	1.547E-2	-1.985	-1.716	-1.410
QUES[4]	-2.496	0.1464	1.531E-2	-2.760	-2.499	-2.207
QUES[5]	-0.875	0.1603	1.543E-2	-1.172	-0.879	-0.557
QUES[6]	-2.226	0.1466	1.524E-2	-2.489	-2.228	-1.942
QUES[7]	-1.770	0.1500	1.532E-2	-2.039	-1.774	-1.471
QUES[8]	-1.148	0.1568	1.524E-2	-1.435	-1.152	-0.833

QUES[9]	-0.341	0.1728	1.530E-2	-0.662	-0.347	0.006
SESS[1]	-0.098	0.0394	1.053E-3	-0.176	-0.099	-0.023
TRT[1]	-0.028	0.0397	1.204E-3	-0.106	-0.029	0.048
L[1]	3	0	1.414E-12	3	3	3
L[2]	2	0	1.414E-12	2	2	2
L[3]	3	0	1.414E-12	3	3	3
L[4]	3	0	1.414E-12	3	3	3
L[5]	1	0	1.414E-12	1	1	1
L[6]	2	0	1.414E-12	2	2	2
deviance	19210	5.821	0.2227	19200	19200	19220

Table H.2: Role Physical

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1,1]	-0.026	0.0913	4.863E-3	-0.211	-0.026	0.143
a[2,1]	-1.572	0.0850	4.526E-3	-1.743	-1.566	-1.406
a[3,1]	-0.596	0.0650	4.394E-3	-0.727	-0.597	-0.468
QUES[1]	0.198	0.0724	3.275E-3	0.054	0.199	0.340
QUES[2]	-0.319	0.0691	3.122E-3	-0.456	-0.321	-0.184
QUES[3]	0.129	0.0726	3.372E-3	-0.015	0.130	0.267
SESS[1]	-0.274	0.0509	1.697E-3	-0.372	-0.274	-0.176
TRT[1]	0.077	0.0500	1.608E-3	-0.021	0.078	0.174
L[1]	2.990	0.1438	3813E-3	3	3	3
L[2]	1	0	1.414E-12	1	1	1
L[3]	3	0	1.414E-12	3	3	3
L[4]	3	0	1.414E-12	3	3	3
L[5]	2	0	1.414E-12	2	2	2
L[6]	3	0	1.414E-12	3	3	3
deviance	9132	3.998	0.1076	9127	9132	9142

Table H.3: Role Emotional

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1,1]	-0.854	0.0889	3.866E-3	-1.025	-0.855	-0.683
a[2,1]	-1.591	0.0715	4.205E-3	-1.728	-1.590	-1.449
QUES[1]	0.247	0.0788	2.615E-3	0.089	0.247	0.398
QUES[2]	-0.377	0.0739	2.787E-3	-0.525	-0.377	-0.231
SESS[1]	-0.463	0.0640	2.046E-3	-0.585	-0.461	-0.338
TRT[1]	-0.039	0.0632	1.964E-3	-0.163	-0.040	0.083
L[1]	2	0	1.414E-12	2	2	2
L[2]	2	0	1.414E-12	2	2	2
L[3]	1	0	1.414E-12	1	1	1
L[4]	2	0	1.414E-12	2	2	2
L[5]	2	0	1.414E-12	2	2	2
L[6]	2	0	1.414E-12	2	2	2
deviance	5960	3.34	0.09996	5955	5959	5968

Table H.4: Social Functioning

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1,1]	-4.107	0.1485	7.218E-3	-4.399	-4.105	-3.819
a[1,2]	-2.419	0.0846	5.363E-3	-2.582	-2.416	-2.258
a[1,3]	-0.813	0.0662	4.742E-3	-0.947	-0.810	-0.687
a[1,4]	0.611	0.0653	4.463E-3	0.483	0.614	0.7338
a[2,1]	-6.359	0.8101	2.896E-2	-8.362	-6.303	-5.046
a[2,2]	-3.438	0.2026	8.914E-3	-3.852	-3.435	-3.041
a[2,3]	-1.997	0.1210	6.295E-3	-2.225	-1.998	-1.755
a[2,4]	-0.087	0.0910	4.792E-3	-0.257	-0.085	0.0957
QUES[1]	0.202	0.0605	2.290E-3	0.085	0.202	0.323
SESS[1]	-0.291	0.0603	2.637E-3	-0.410	-0.290	-0.177
TRT[1]	-0.046	0.0603	2.460E-3	-0.168	-0.045	0.068
L[1]	1	0	1.291E-12	1	1	1
L[2]	1	0	1.291E-12	1	1	1
L[3]	1	0	1.291E-12	1	1	1
L[4]	1	0	1.291E-12	1	1	1
L[5]	2	0	1.291E-12	2	2	2
L[6]	1	0	1.291E-12	1	1	1
deviance	9325	4.765	0.1479	9318	9324	9336

Table H.5: Mental Health

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1,1]	-4.857	0.2557	1.076E-2	-5.394	-4.841	-4.352
a[1,2]	-3.179	0.1293	7.780E-3	-3.435	-3.177	-2.918
a[1,3]	-2.125	0.0920	6.552E-3	-2.317	-2.126	-1.949
a[1,4]	-0.799	0.0715	5.369E-3	-0.941	-0.801	-0.658
a[1,5]	1.232	0.0708	4.980E-3	1.101	1.233	1.379
a[2,1]	-3.362	0.1160	6.054E-3	-3.588	-3.356	-3.139
a[2,2]	-2.164	0.0769	5.232E-3	-2.313	-2.166	-2.016
a[2,3]	-0.773	0.0598	4.793E-3	-0.895	-0.773	-0.652
a[2,4]	0.285	0.0574	4.686E-3	0.169	0.284	0.395
a[2,5]	1.726	0.0638	4.587E-3	1.600	1.727	1.854
a[3,1]	-5.008	0.1902	8.861E-3	-5.381	-4.994	-4.647
a[3,2]	-2.866	0.0833	5.789E-3	-3.036	-2.866	-2.701
a[3,3]	-1.434	0.0616	5.323E-3	-1.555	-1.434	-1.307
a[3,4]	-0.198	0.0543	4.624E-3	-0.303	-0.198	-0.085
a[3,5]	1.642	0.0569	4.310E-3	1.537	1.637	1.758
QUES[1]	0.609	0.0568	3.096E-3	0.497	0.609	0.720
QUES[2]	1.994	0.0628	3.137E-3	1.870	1.994	2.118
QUES[3]	-0.587	0.0581	3.274E-3	-0.699	-0.588	-0.473
QUES[4]	0.978	0.0581	3.155E-3	0.866	0.978	1.092
SESS[1]	-0.297	0.0392	1.768E-3	-0.373	-0.297	-0.218
TRT[1]	-0.058	0.0385	1.429E-3	-0.131	-0.058	0.019
L[1]	3	0	1.414E-12	3	3	3
L[2]	2	0	1.414E-12	2	2	2
L[3]	2	0	1.414E-12	2	2	2
L[4]	3	0	1.414E-12	3	3	3

L[5]	1	0	1.414E-12	1	1	1
L[6]	3	0	1.414E-12	3	3	3
deviance	25200	6.24	0.2155	25190	25200	25210

Table H.6: Vitality

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1,1]	-3.741	0.1178	6.932E-3	-3.967	-3.746	-3.499
a[1,2]	-2.041	0.0816	7.389E-3	-2.214	-2.038	-1.872
a[1,3]	-0.562	0.0810	8.272E-3	-0.778	-0.557	-0.423
a[1,4]	0.706	0.0751	7.379E-3	0.525	0.713	0.847
a[1,5]	3.006	0.0937	5.848E-3	2.839	3.004	3.207
a[2,1]	-3.587	0.1203	9.107E-3	-3.815	-3.588	-3.340
a[2,2]	-2.350	0.0835	7.608E-3	-2.499	-2.352	-2.163
a[2,3]	-1.035	0.0727	7.335E-3	-1.153	-1.042	-0.839
a[2,4]	0.250	0.0607	5.772E-3	0.144	0.245	0.376
a[2,5]	2.261	0.1028	1.009E-2	1.969	2.269	2.436
QUES[1]	0.097	0.0586	3.443E-3	-0.012	0.096	0.218
QUES[2]	-0.136	0.0585	3.385E-3	-0.246	-0.136	-0.017
QUES[3]	0.660	0.0601	3.524E-3	0.543	0.659	0.782
SESS[1]	-0.369	0.0420	2.073E-3	-0.453	-0.369	-0.286
TRT[1]	-0.004	0.0425	2.181E-3	-0.085	-0.005	0.081
L[1]	1	0	1.414E-12	1	1	1
L[2]	1	0	1.414E-12	1	1	1
L[3]	2	0	1.414E-12	2	2	2
L[4]	1	0	1.414E-12	1	1	1
L[5]	1.942	0.2337	2.749E-2	1	2	2
L[6]	2	0	1.414E-12	2	2	2
deviance	22170	9.779	0.9708	22160	22170	22200

Table H.7: General Health

parameter	mean	sd	MC error	2.5%	median	97.5%
a[1,1]	-3.719	0.1203	5.682E-3	-3.944	-3.715	-3.476
a[1,2]	-1.827	0.0675	4.870E-3	-1.958	-1.826	-1.702
a[1,3]	-0.185	0.0554	4.626E-3	-0.298	-0.184	-0.076
a[1,4]	1.281	0.0568	4.603E-3	1.159	1.283	1.390
a[2,1]	-2.864	0.0804	5.220E-3	-3.025	-2.860	-2.715
a[2,2]	-1.406	0.0605	4.986E-3	-1.531	-1.405	-1.292
a[2,3]	-0.285	0.0541	4.632E-3	-0.391	-0.283	-0.179
a[2,4]	1.377	0.0568	4.750E-3	1.266	1.380	1.484
QUES[1]	-0.419	0.0579	3.618E-3	-0.531	-0.420	-0.303
QUES[2]	0.756	0.0617	3.533E-3	0.632	0.757	0.877
QUES[3]	0.378	0.0599	3.503E-3	0.260	0.378	0.496
QUES[4]	0.775	0.0620	3.539E-3	0.656	0.775	0.896
SESS[1]	-0.083	0.0378	1.530E-3	-0.158	-0.083	-0.009
TRT[1]	0.113	0.0371	1.383E-3	0.039	0.113	0.186
L[1]	2	0	1.414E-12	2	2	2
L[2]	1	0	1.414E-12	1	1	1
L[3]	2	0	1.414E-12	2	2	2
L[4]	2	0	1.414E-12	2	2	2
L[5]	1	0	1.414E-12	1	1	1
L[6]	1	0	1.414E-12	1	1	1
deviance	25170	5.161	0.1598	25160	25170	25180