INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6” x 9” black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700  800/521-0600
LONGITUDINAL ANALYSIS OF INCOMPLETE BINARY DATA

DISSERTATION

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

By

Theresa Marie Papa Stern, B.A., M.S.

*****

The Ohio State University

1997

Dissertation Committee:

Elizabeth A. Stasny, Adviser

Mark I. Irwin

Dennis K. Pearl

Approved by

Elizabeth A. Stasny

Adviser

Department of Statistics
ABSTRACT

The analysis of longitudinal binary data has recently been the focus of much interest. In addition to modeling the marginal responses, the correlation between the repeated observations must be accounted for. Missing and incomplete responses are a common occurrence in longitudinal studies. Techniques that allow for the analysis of complete and of drop-out data have been developed. However, it is sometimes desirable to have the ability to analyze data that has the form of “drop-in”, when the observation(s) are initially missing but all responses from some point onward are observed.

The motivation for this research is a 20-year longitudinal study, the Air Force Health Study (AFHS), designed to study the health of Air Force veterans occupationally exposed to herbicides in Vietnam. In particular, we study one variable, erythrocyte sedimentation rate. Sources of missing and incomplete data in the AFHS are also discussed.

Previous methodology for analyzing complete binary data, in addition to that for drop-out data, is extended to develop a general formulation for analyzing drop-in data. This formulation uses a likelihood-based regression model to relate the marginal response
probabilities to a set of covariates by a known link function. We assume that interest is based primarily on the regression parameters for the marginal expectations with the association between the responses largely considered a nuisance characteristic of the data. We consider logistic regression models for the drop-in missing-data mechanism. These models depend only on a set of known covariates and are, thus, missing-at-random (MAR).

We then further extend our drop-in model to data sets containing drop-in and drop-out cases, therefore allowing the analysis of a data set that contains complete, drop-out and drop-in data.

We present results of analyses conducted on the AFHS sedimentation rate data for drop-in data, and for drop-in and drop-out data. In addition, we include results of models fit to the AFHS sedimentation rate data which were based on the earlier methodologies. Some suggestions for future work are also given.
To my husband

Derek J. Stern
ACKNOWLEDGMENTS

My heartfelt thanks to

...my adviser, Dr. Elizabeth A. Stasny, for her guidance, patience and insight,

...the OSU Department of Statistics faculty, especially Dr. Douglas A. Wolfe who has been my second father,

...my dear friend, the late Dr. Craig A. Cooley, who provided me with endless help, encouragement and inspiration,

...my mother, Bette A. Papa, for her support, love and belief in me,

...my father, the late Richard A. Papa, who has been with me in spirit and whose guidance I have counted on, especially in the difficult times,

...Dr. Joel E. Michalek, Chief Principal Investigator of the Air Force Health Study, for being a wonderful mentor and for helping me acquire the data used in my model fitting,

...my father-in-law, Robert A. Stern, for help proofreading my dissertation,

...and my family and friends, who have provided me with love, support and an occasional study break.
VITA

September 19, 1969.......................... Born - Amsterdam, New York

May, 1991................................. B.A., Mathematics
   West Virginia University

December, 1991.......................... M.S., Statistics
   The Ohio State University

September, 1991-June 1992........... Graduate Teaching Associate
   The Ohio State University

   Graduate Fellow

September, 1995-June, 1997........... Graduate Research Associate
   The Ohio State University

PUBLICATIONS


**FIELDS OF STUDY**

Major Field: Statistics
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Abstract</th>
<th>ii</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xi</td>
</tr>
<tr>
<td>Chapters:</td>
<td></td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2. The Air Force Health Study</td>
<td>4</td>
</tr>
<tr>
<td>2.1 Study Design</td>
<td>5</td>
</tr>
<tr>
<td>2.2 Study Population</td>
<td>8</td>
</tr>
<tr>
<td>2.3 Study Participation</td>
<td>9</td>
</tr>
<tr>
<td>2.4 Statistical Methods</td>
<td>10</td>
</tr>
<tr>
<td>2.4.1 Preliminary Analyses</td>
<td>12</td>
</tr>
<tr>
<td>2.4.2 Intervening Variables</td>
<td>12</td>
</tr>
<tr>
<td>2.4.3 Paired versus Unpaired Analyses</td>
<td>13</td>
</tr>
<tr>
<td>2.4.4 Exposure Index Analyses</td>
<td>14</td>
</tr>
</tbody>
</table>

viii
2.4.5 Longitudinal Analyses ................................................................. 18

2.5 General Health Assessment in the AFHS ............................................. 20

2.6 Results of Sedimentation Rate Analyses in the AFHS ......................... 24
  2.6.1 General Assessment .................................................................. 25
  2.6.2 Herbicide Exposure Analyses ..................................................... 27
  2.6.3 Longitudinal Analyses ............................................................... 29

3. Analysis of Missing Data in the AFHS .................................................. 32
  3.1 Missing Data in the AFHS ............................................................... 33
    3.1.1 Missing Data Due to Noncompliance ......................................... 33
    3.1.2 Missing Data in Fully Compliant Subjects .................................. 35
    3.1.3 Missing Data Due to Study Design ............................................ 36
  3.2 Preliminary Analyses of Missing Data in the AFHS ......................... 39
    3.2.1 Analyses of Compliance Status ................................................. 39
    3.2.2 Analyses of Presence of Examination Data ............................... 46
    3.2.3 Analyses of Dioxin versus Compliance ..................................... 48

4. Literature Review ................................................................................ 55
  4.1 Regression Models for Longitudinal Binary Data ............................... 57
    4.1.1 Generalized Estimating Equations .......................................... 58
    4.1.2 Likelihood-Based Regression Models ....................................... 60
      4.1.2.1 Log-linear Representation ............................................... 61
      4.1.2.2 Parameterization via Marginal Moments .......................... 62
      4.1.2.3 Parameterization via Conditional Logits and...
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Models for Ranch Hand Compliance Status for Each Pair of Consecutive Study Cycles</td>
<td>52</td>
</tr>
<tr>
<td>3.2 Models for Comparison Compliance Status for Each Pair of Consecutive Study Cycles</td>
<td>53</td>
</tr>
<tr>
<td>3.3 Numbers of Ranch Hands and Comparisons with Available and Missing Physical Examination Data</td>
<td>54</td>
</tr>
<tr>
<td>3.4 Cycle 3 Dioxin versus Cycle 4 Compliance Status for Ranch Hands and Comparisons</td>
<td>54</td>
</tr>
<tr>
<td>5.1 Erythrocyte Sedimentation Rate by Group and Age at Cycle 1 for Subjects with Complete Data</td>
<td>101</td>
</tr>
<tr>
<td>5.2 Parameter Estimates, $G^2$ and $X^2$ Statistics and Degrees of Freedom for Sedimentation Rate Models Fit to Complete Data</td>
<td>103</td>
</tr>
<tr>
<td>5.3 Parameter Estimates, $G^2$ and $X^2$ Statistics and Degrees of Freedom for Different Covariate Structures Fit to Complete Data</td>
<td>105</td>
</tr>
<tr>
<td>5.4 Erythrocyte Sedimentation Rate by Group and Age at Cycle 1 for Subjects with Drop-Out Data</td>
<td>107</td>
</tr>
</tbody>
</table>
5.5 Parameter Estimates, $X^2$ Statistics and Degrees of Freedom for Logistic Regression Models fit to Drop-Out Mechanism .................................................109

5.6 Parameter Estimates, $G^2$ and $X^2$ Statistics and Degrees of Freedom for Sedimentation Rate Models fit to Drop-Out Data ..............................................111

6.1 Drop-In Patterns and Likelihood Contributions .................................................158

6.2 Erythrocyte Sedimentation Rate by Group and Age at Cycle 1 for Subjects with Drop-In Data ..............................................................................................159

6.3 Parameter Estimates, $X^2$ Statistics and Degrees of Freedom for Logistic Regression Models fit to Drop-In Mechanism .................................................161

6.4 Parameter Estimates, $G^2$ and $X^2$ Statistics and Degrees of Freedom for Sedimentation Rate Models fit to Drop-In Data .................................................163

6.5 Drop-In and Drop-Out Patterns and Likelihood Contributions ........................165

6.6 Parameter Estimates, $X^2$ Statistics and Degrees of Freedom for Logistic Regression Models fit to Drop-In Mechanism for Drop-In and Drop-Out Data ........................................................................................................166

6.7 Parameter Estimates, $G^2$ and $X^2$ Statistics and Degrees of Freedom for Full Model Fit to Drop-In and Drop-Out Sedimentation Rate Data ..........167
CHAPTER 1

INTRODUCTION

The research described in this dissertation is motivated by the need to analyze longitudinal binary response data from the Air Force Health Study (AFHS). The AFHS is a 20-year longitudinal study designed to study the health of Air Force veterans occupationally exposed to herbicides in Vietnam. One of the main goals of the AFHS is to investigate changes over time in some of the dependent variables measured. One such variable is the erythrocyte sedimentation rate, a nonspecific, though sensitive, indicator of overall general health.

In longitudinal studies, repeated observations are made over time on a response variable and a set of covariates. The response variables are usually correlated because repeated observations are made on the same subjects. Longitudinal regression models can be used to relate the expectation of the response variable to the covariates by a known link function. When the response variable is binary, the logit link function is commonly used. For longitudinal studies where the association between responses is the focus, interest is
on the regression parameters of the conditional expectations. When the relationship
between the marginal responses and the covariates is of primary interest, the regression
parameters for the marginal expectations are the focus. In this instance, the correlation
between responses is largely considered a nuisance characteristic of the data. Regardless
of the focus of a longitudinal study, the association between responses for a subject must
be accounted for when analyzing the data in order to draw valid conclusions.

Missing and incomplete data are common in longitudinal studies. Many times,
incomplete data takes the form of drop-out, when a subject is observed from the beginning
of the study until the time of drop-out, after which he is no longer observed. Responses
for a subject who drops out of a longitudinal study can be ordered such that if the $t^{th}$
response is missing, so is every response $t' > t$. Another form of incomplete data in a
longitudinal study is that of "drop-in". This occurs for a subject who enters the study after
it has begun and, once in the study, remains until the end. Subjects who drop in the study
are missing data on the first time period(s). Their responses can be ordered such that if
response $t$ is missing then every response $t' < t$ is missing. Although numerous papers
have discussed methodology to handle drop-out data, there has been very little, if any,
work on methodology to handle drop-in data.

The AFHS in particular has seen the occurrence of both drop-out and drop-in data.
Chapter 2 gives an overview of the study. Specifically, Section 2.6 summarizes results of
sedimentation rate analyses that have been conducted to date in the AFHS.

In Chapter 3, we describe how missing data occurs in the AFHS. We then give
results of preliminary analyses done on some of the missing data to discover factors
related to missingness in Section 3.2. We present a literature review in Chapter 4, with methods for analyzing complete longitudinal binary data given in Section 4.1 and for analyzing incomplete (drop-out) data in Section 4.2. We give results of fitting the models of Chapter 4 to the AFHS sedimentation rate data in Chapter 5. Section 5.1 discusses the models fit to subjects having complete sedimentation rate data, while Section 5.2 details models fit to subjects with drop-out sedimentation rate data. In Chapter 6, Sections 6.1 and 6.2, we develop methodology to deal with subjects having drop-in data. This methodology can be used for any number of observation time points $T$. In Section 6.3, we give results of using this drop-in methodology, with $T=4$, to analyze sedimentation rate data for subjects in the AFHS having drop-in data patterns. We further extend our drop-in methodology, in Section 6.4, to deal with subjects who have either drop-out or drop-in data. In Section 6.5, we apply our extended drop-in methodology, again with $T=4$, to the AFHS sedimentation rate data. Finally, in Chapter 7, we give our conclusions and topics of future work.
CHAPTER 2

THE AIR FORCE HEALTH STUDY

In January 1962, President John F. Kennedy approved the aerial distribution of herbicides in the Republic of Vietnam, for crop destruction and defoliation, to help with United States military actions (Lathrop et al., 1984). The military operation in charge of disseminating the herbicides, code named Operation Ranch Hand, ran from 1962 to 1971. During this period, it conducted over 6500 spray missions and sprayed over 19 million gallons of herbicides on between 10 and 20 percent of South Vietnam (Lathrop et al., 1984). Approximately 11 million gallons of Herbicide Orange, the main herbicide used in the operation, were sprayed. There was controversy over the Ranch Hand operation from the start. Although the focus was initially on the military, political and ecological consequences of the spray operations, since 1977 the focus has been on health issues (Grubbs et al., 1994). US military personnel from all branches of the service believe they were exposed to herbicides, particularly Agent Orange, during their duty in Vietnam.
Many believe that this exposure has led to negative health consequences for themselves and their children.

In October 1978, the US Air Force Deputy Surgeon General assured both Congress and the White House that the health of Ranch Hand veterans would be thoroughly examined. The Ranch Hands were those Air Force veterans in charge of handling and distributing nearly all of the herbicides disseminated in Vietnam and, thus, are the subgroup of Vietnam veterans who had the most contact with the herbicides. Although as a group the Ranch Hands were the most-exposed subgroup of veterans, the exposure in individual Ranch Hand veterans ranged from none to extreme. No records were kept, during the time the Ranch Hand unit was active, as to who was exposed and, if so, to what extent. Therefore, each of the Ranch Hand veteran’s true exposure is unknown. An epidemiologic study design was prepared by the US Air Force School of Aerospace Medicine and, following an extensive review, a final study protocol was published and the epidemiologic study, called the Air Force Health Study (AFHS), was begun (Lathrop et al., 1984).

2.1 Study Design

The AFHS is designed to determine whether adverse health effects exist and can be attributed to occupational exposure to the herbicide or its contaminant 2,3,7,8-tetrachlorodibenzo-p-dioxin, also called TCDD or dioxin (Lathrop et al., 1984). The study uses a matched cohort design in a noncurrent prospective setting, using traditional
epidemiologic measurements, including self-reports, structured interviews, medical record data, physical examinations, biopsy results, laboratory tests, morbidity indices, and mortality outcomes (Lathrop et al., 1984). The study is noncurrent in that Ranch Hand veterans were exposed over time between 1962 and 1971.

Historical data records were used to identify Ranch Hand veterans who served in the Republic of Vietnam during the period when the spray operations were conducted. Comparison subjects are from the population of Air Force veterans assigned to Air Force units operating C-130 cargo aircraft in Southeast Asia during the same time as the Ranch Hand operation. Cargo mission aircrew and support personnel were chosen to be the Comparisons because their population is of a satisfactory size, and their military training, background, and mentalities were comparable to that of the Ranch Hand group (Lathrop et al., 1984). The Comparisons were not members of Operation Ranch Hand and, so, were not occupationally exposed to herbicides during the war. The Comparison population was identified through the same historical data sources that were used to identify the Ranch Hand population.

Up to ten Comparison individuals were matched to each Ranch Hand veteran by military rank, military occupation, race and age to the nearest month of birth. From each of these matched sets, five Comparisons were randomly selected for the 1982 Baseline mortality study. The mortality segment of the AFHS looks at mortality since the time of Vietnam duty. In addition, each living Ranch Hand veteran and the first living member of his Comparison set (the Original Comparison) were invited to the Baseline morbidity segment, which was also held in 1982. The Baseline morbidity study constructed each
participant’s medical history by reviewing past medical records, and investigated current mental and physical health, through the use of questionnaires and physical examinations (Lathrop et al., 1987).

The follow-up portion of the AFHS also includes mortality and morbidity segments. The Ranch Hands and Comparisons are to be followed in the mortality segment, through annual updates, until the end of the AFHS in 2002. These annual updates will investigate possible consequences of herbicide exposure, such as specific mortality patterns or disease clusters. The Baseline mortality study and the first four updates used the 1:5 study design; after 1987, however, the mortality study was expanded to include all of the veterans in the Comparison population. In addition, eligible participants in the morbidity study were invited to take part in follow-up study cycles that were conducted in 1985, 1987 and 1992; a follow-up study cycle is taking place in 1997 and a future follow-up is planned for 2002. Since the purpose of the morbidity follow-ups is to cover the latency period of diseases that could stem from occupational exposure to herbicides in Vietnam, they focus on examining possible long-term health effects following Ranch Hand herbicide exposure (Lathrop et al., 1987). The follow-ups allow for confirmation or negation of results from previous study cycles, and for exploration of possible longitudinal changes.
2.2 Study Population

The exposed population, "Ranch Hand", is defined in the Study Protocol as those individuals formally assigned to the USAF organizations that aerially disseminated herbicides and insecticides in the Republic of Vietnam from 1962 through 1971 (Lathrop et al., 1984). The initial search of historical data sources, just prior to the Baseline study, found 1,264 Ranch Hand veterans; since the Baseline study, 5 additional Ranch Hand members have been identified. The Comparison population is defined as those assigned to operations that flew cargo missions throughout Southeast Asia during the same time as the Ranch Hand operation.

The Study Protocol called for matching Ranch Hands and Comparisons on the variables military rank, military occupation, race, and age. The protocol also required that all morbidity study Comparisons be alive at the start of the morbidity segment of the study. A total of 1,208 living Ranch Hand veterans and 1,238 Comparisons agreed to take part in the Baseline morbidity segment of the study.

For the morbidity segment of the AFHS, all living Ranch Hand veterans and their Original Comparisons were asked to complete a comprehensive personal and family health questionnaire given by a civilian contractor experienced in survey research. After completing the questionnaire, subjects were asked to undergo a physical examination. These two-day physical examinations were all at a single location, with the contractor personnel performing the examinations blind to each participant's exposure group.
2.3 Study Participation

Study participation at each of the study cycles is classified as fully compliant, when the questionnaire and physical examination are both completed by the subject; partially compliant, when the questionnaire is completed, but the physical examination is not; noncompliant, when neither the questionnaire nor the physical examination is completed; or unlocateable, when the subject can not be located.

The authors of the Study Protocol believed the participation rates for the Baseline questionnaire and Baseline physical examination would be quite low, but that more Ranch Hands than Comparisons would participate. The main reason for expecting low participation rates was that compliance to the AFHS was discouraged by the employers of those veterans working in flying-related jobs. The Ranch Hand Association gave strong support to Ranch Hands taking part in the AFHS, but the Comparisons did not have such support backing their participation. Because loss to follow-up in the Comparison group was anticipated by the authors of the Study Protocol, it was mandated that if the Original Comparison refused to participate in the study, the next randomly ordered Comparison for the Ranch Hand with the same self-perception of health as the refusing Comparison would be invited to participate in the study. The protocol called for Original and Replacement Comparisons to be matched on self-perception of health to minimize potential bias from differential compliance to the examination.

For the first follow-up examination in 1985, all Ranch Hands and Comparisons invited to the Baseline study, including those who had refused or who had only partially
complied in 1982, were again invited to take part. Ranch Hand veterans who were newly locatable (had been unlocateable in 1982 but were now locatable) or who were just identified as being members of the Ranch Hand group, and their respective Comparisons, were also invited to the study.

Ranch Hands and Comparisons invited to the Baseline and/or the 1985 follow-up were again invited to the 1987 follow-up study. As in the 1985 follow-up, newly located and newly identified Ranch Hands and their Comparisons were invited. Likewise, for the 1992 follow-up, all participants asked to the Baseline, 1985 and/or 1987 studies were invited to attend, along with newly located Ranch Hands and their Comparisons.

There was a total of 2,269 fully compliant participants at the 1982 Baseline study, with 1,045 Ranch Hands and 1,224 Comparisons. The population for the 1985 follow-up included 1,016 Ranch Hands and 1,293 Comparisons, for a total of 2,309 subjects. There were 995 Ranch Hands and 1,299 Comparisons fully compliant to the 1987 follow-up, with a total of 2,294 veterans fully compliant. The 1992 follow-up had 952 fully compliant Ranch Hands and 1,281 fully compliant Comparisons, giving a study population of 2,233 subjects.

2.4 Statistical Methods

The main objective of the AFHS is to compare the health of the Ranch Hand and Comparison groups. It is unknown if herbicide exposure in humans has significant long-term negative health effects. Physical chemistry data, animal toxicity data, human
exposure case reports, and epidemiologic studies have not found clear, specific and objective medical problems following chronic exposure to herbicides (Grubbs et al., 1995). The symptoms, illnesses and syndromes thought to be related to herbicide exposure are confounded by other variables, and symptoms and signs that may exist could be common to other diseases (Grubbs et al., 1995). Without any specific, objective measures of herbicide exposure, the AFHS uses an extensive examination that focuses on organ systems known or believed to be affected by herbicides and dioxin, including dermatology, neurology, psychiatry, endocrinology, reproduction, and immunology (Grubbs et al., 1995).

During each study cycle, significant amounts of data are collected on each participant. In the Baseline report alone, more than 190 dependent variables were analyzed (Lathrop et al., 1984). Consistent with many epidemiologic studies, some of the variables in the AFHS are highly correlated. To help with this, the AFHS groups variables most associated with one another into clinical categories and, within categories, uses summary indices when possible. One of the main uses of the summary indices in the AFHS is to protect against falsely inferring a herbicide effect. Since a falsely-inferred herbicide effect may be in contrast to what is expected based on results from previous study cycles or other clinical categories, patterns of statistically significant results, between clinical categories and from cycle to cycle, are followed (Grubbs et al., 1995).
2.4.1 Preliminary Analyses

Preliminary analyses in the AFHS entail computing basic descriptive statistics, such as frequency distribution, histogram, mean, median, standard deviation, and range, for dependent and independent variables for both Ranch Hands and Comparisons. The preliminary analyses serve two functions. First, the analyses summarize each variable, give the relationship of the Ranch Hand group to the Comparison group, suggest normal/abnormal cutpoints, and help determine transformations of asymmetrically distributed continuous dependent variables (Thomas et al., 1990). Secondly, the preliminary analyses assess relationships between covariates and dependent variables, and among covariates. The statistical significance of these relationships is given by the p-values corresponding to tests of their strength; in the AFHS, relationships with p-values ≤ 0.05 are termed “significant” and those with p-values between 0.05 and 0.10 are “marginally significant” (Thomas et al., 1990).

2.4.2 Intervening Variables

The existence or nonexistence of statistically significant differences between the Ranch Hand and Comparison groups in the preliminary analyses neither confirms nor refutes a cause-effect relationship between herbicide exposure and disease. Hence, group associations, regardless of their statistical significance, are further analyzed. More rigorous analyses adjust for covariates that may explain or mask a herbicide effect.
Even though it is impossible to identify and include every possible covariate, identified potentially important covariates are included in the analyses. Adjustment is also made for confounding variables, to give less biased estimates of relationships between exposure and health effects. In addition, interactions are included in adjusted models, to investigate whether any subpopulations have increased or decreased risks.

2.4.3 Paired versus Unpaired Analyses

In the AFHS, the Comparisons are matched to their respective Ranch Hands by age (to the nearest month of birth), race (Black, non-Black), and military rank and occupation (officer-pilot, officer-navigator, officer-nonflyer, enlisted flyer, enlisted groundcrew). Although paired analyses that include the matching variables increase both group comparability and statistical power, most analyses in the AFHS are unpaired analyses that do not take advantage of the paired design of the study (Lathrop et al., 1987). Unpaired analyses on paired data lead to a loss of power and less chance of detecting the possibility of a herbicide effect, but can also be more powerful than paired analyses if noncompliance leads to large numbers of broken pairs (Lathrop et al., 1987). The comparability of the groups in unpaired analyses is still increased by the matching, and also by adjustment through stratification and covariates. Relationships between the matching variables and the explanatory variables are investigated to ensure the matching variables and explanatory variables do not interact to affect dependent variables.
2.4.4 Exposure Index Analyses

Although each of the Ranch Hand veteran's true exposure to herbicide is unknown, the AFHS has tried, throughout the study, to accurately quantify this true, but unknown, exposure. Thus far, the AFHS has used two different exposure indices as surrogates for herbicide exposure.

In the 1982, 1985 and 1987 study cycles, the AFHS quantified herbicide exposure in a way consistent with many other epidemiologic studies, by using a work-history based exposure index. This work history-based index, the Military Records Index (MRI), estimated potential exposure to any of the four TCDD-containing herbicides from fixed-wing spray missions (Michalek, 1989). The MRI was calculated for each Ranch Hand, from military records of the veterans and of the herbicides sprayed.

The MRI was defined as:

\[ \text{MRI} = \left( \sum_{j=1}^{m} C_j \times G_j / N_j \right) / 10^4 \]

where \( m \) is the number of months of the veteran's Vietnam tour and, for month \( j \), \( C_j \) is the weighting factor for the concentration of the herbicide sprayed, \( G_j \) is the number of gallons of herbicide sprayed, and \( N_j \) is the number of veterans sharing the \( j^{th} \) veteran's duties. The TCDD concentration of the herbicide used, \( C_j \), is calculated on the basis of two facts: different herbicides were sprayed at different times during the Ranch Hand operation, and concentrations of these different herbicides changed. Prior to January 1,
1965, a mixture of Herbicides Purple, Pink and Green was sprayed, while after January 1, 1965, only Herbicide Orange was sprayed (Buckingham, 1982). The estimated mean TCDD concentrations of the four herbicides were 2 parts per million (ppm) for Herbicide Orange, 33 ppm for Herbicide Purple, and 66 ppm for Herbicides Pink and Green (Buckingham, 1982). Military records of the herbicide gallons sprayed indicate that the mean TCDD concentration of the herbicides before January 1, 1965 was approximately 48 ppm and that after January 1, 1965, the mean TCDD concentration was 2 ppm (Buckingham, 1982). Thus, based on the ratio of the herbicide concentrations being 48:2 or 24:1, \( C_j \) is defined to be 24.0 for months before January 1, 1965 and to be 1.0 for months after January 1, 1965. For \( G_j \), the gallons of herbicide sprayed during month \( j \), the gallons of TCDD-containing herbicide sprayed each month of the Ranch Hand operation was determined and then gallons of Herbicides Purple, Pink, and Green were converted to Herbicide Orange equivalent gallons. Finally, the number of veterans sharing duties during month \( j \), \( N_j \), is calculated from military records that give each Ranch Hand's occupational category and the date(s) of his tour(s).

The purpose of the exposure analyses based on the MRI was the investigation of a possible dose-response relationship between the MRI and each dependent variable (Lathrop et al., 1987). Because the MRI is not specific to each individual, it was believed to be inaccurate, especially in underestimating exposure for Ranch Hands who had daily contact with the herbicides (Thomas et al., 1990). The MRI also failed to uncover dose-response relationships or medically plausible results (Grubbs et al., 1995). Therefore, the
Air Force continued to develop new surrogates to estimate herbicide exposure. One such
development was the measurement of serum dioxin levels.

Serum dioxin is the serum concentration of 2,3,7,8-tetrachlorodibenzo-p-dioxin.
Over the last 10 years, an assay that can detect minute amounts of dioxin in human serum
has been developed, allowing for accurate measurement of dioxin body burdens and for
investigation of the relationship between serum dioxin and health.

Current body burden dioxin levels were measured by the Center for Disease
Control (CDC) through serum samples taken from Ranch Hands and Comparisons,
beginning with the 1987 physical examination. It is hoped that estimates of the half-life of
dioxin, along with current dioxin body burdens, can give a more objective estimate of past
exposure. Based on a pilot study of serum levels in 36 AFHS subjects (Wolfe et al.,
1988), it was believed at the 1987 and 1992 follow-ups that the half-life of dioxin is 7.1
years. However, continuing research by the principal investigators of the AFHS and the
CDC since the 1992 follow-up has led to the estimate of dioxin half-life being revised.
The latest research gives 8.7 years as the estimated half-life of dioxin (Michalek et al.,
1996).

The accuracy of the assay in determining herbicide exposure is reinforced by
results indicating that, on average, Ranch Hands have higher serum dioxin levels than
Comparisons; that dioxin levels for the three Ranch Hand occupational cohorts are
consistent with what is known about the opportunity for herbicide exposure for each of
the occupations; and that, relative to other groups exposed to dioxin, the Ranch Hand
dioxin measurements are at an appropriate level (Grubbs et al., 1995).
Because the assay is a direct measure of TCDD exposure and the MRI an indirect measure, and because the assay is individually specific and the MRI is not, the assay is preferred over the MRI. When the relationship between the assay results and the MRI was analyzed, results indicated that classifying veterans using the MRI, instead of dioxin body burden, can lead to substantial misclassification (Roegner et al., 1991).

In the 1987 and 1992 follow-ups of the AFHS, models that included dioxin levels as the surrogate for herbicide exposure were used. These models were based on assumptions from the Ranch Hand dioxin pilot study and half-life substudy, used a dioxin half-life of 7.1 years, and classified veterans with serum dioxin levels at or below 10 ppt as having “background” levels (Wolfe et al., 1988; Pirkle et al., 1989).

For the 1987 follow-up, Ranch Hands were analyzed with two different models, to decrease the possibility that inaccurate assumptions about dioxin elimination would give inaccurate results (Roegner et al., 1991). The first model used an extrapolated initial dioxin dose as the surrogate of herbicide exposure, and the second model used current dioxin body burden as the surrogate. Initial dioxin values can be extrapolated for only those Ranch Hands with current dioxin values above background. It will probably never be known whom, if any, of the Ranch Hands with background current dioxin levels at the 1987 follow-up had an initial dose that was above background levels.

Another model based on dioxin body burdens was also used at the 1987 examination. This third model investigated the relationships between health and dioxin body burdens in both Ranch Hands and Comparisons, by categorizing into 4 groups. Comparisons with current dioxin greater than 10 ppt were excluded from the analyses.
Ranch Hands with current dioxin between 10 and 15 ppt were also excluded, to decrease misclassification into the unknown (up to 10 ppt) and low (between 15 and 33.3 ppt) categories due to possible variability in the dioxin assay (Roegner et al., 1991).

The analyses of the 1992 follow-up centered on six different models, each one using a different surrogate for exposure. The first model used group and military occupation as the estimate of exposure and was the only model from the 1992 study cycle that did not include a serum dioxin measurement as a surrogate of herbicide exposure. The second model used initial dioxin in Ranch Hands as the surrogate, while the third model categorized Ranch Hands and Comparisons based on either initial or current dioxin. Since the scientific community has yet to agree on the “appropriate” current dioxin measurement, the final three models used at the 1992 follow-up used different forms of the “current” dioxin variable (Grubbs et al., 1995). The fourth model used lipid-weight current dioxin as the surrogate; the fifth model used whole-weight current dioxin; and the last model used whole-weight current dioxin with total lipid weight of the sample as an independent variable that was not removed during stepwise procedures. Section 2.6 gives results of exposure index analyses conducted to date on AFHS data.

2.4.5 Longitudinal Analyses

Another objective of the AFHS is to investigate changes in the levels of variables or in the occurrence of abnormal findings or disease between the Ranch Hand and Comparison groups (Thomas et al., 1990). Changes in health status between the 1982
Baseline examination and each follow-up examination as a function of dioxin exposure are investigated. In general, longitudinal analyses in the AFHS use data from two timepoints: the Baseline examination and the current follow-up examination (e.g. in 1987, the data used was the Baseline data and the 1987 data).

In the longitudinal analyses of discrete variables, only participants with "normal" health in 1982 are included in the analysis of participants' health at the follow-up examination. Participants who were "abnormal" in 1982 are not included in the AFHS longitudinal analyses because the focus is on the temporal effects of dioxin exposure between 1982 and the follow-up (Grubbs et al., 1995). Since those subjects "abnormal" in 1982 were already abnormal before this period, only those who were "normal" in 1982 are considered to be at risk when the effects of dioxin over time are investigated. The rate of abnormalities in these longitudinal analyses is then the approximate cumulative incidence rate between 1982 and the follow-up (Grubbs et al., 1995).

The dependent variable for the longitudinal analysis is the health of the participant at the follow-up examination who was of "normal" health in 1982. Age is a risk factor for almost all clinical areas, and even though Ranch Hands and Comparisons are matched on age, the serum dioxin levels are not (Grubbs et al., 1995). Therefore, both age in 1982 and the surrogate for dioxin exposure are the independent variables in the longitudinal analyses. Models that include dioxin are adjusted for percent body fat during the Vietnam tour and for change in percent body fat from the Vietnam tour to the blood draw for dioxin. Finally, the only dioxin models longitudinally analyzed at the 1992 follow-up were those that used initial dioxin, since current dioxin, the estimate of exposure in the last
three models, changes over time and is not available for all participants in 1982 or 1992.

Section 2.6 gives results of longitudinal analyses conducted to date on AFHS data.

2.5 General Health Assessment in the AFHS

Although several organ systems are known to be affected by heavy, acute exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin, the effects of low-dose chronic exposure to TCDD are not known (Grubbs et al., 1995). In laboratory animals, dioxin retention and toxicity is species- and strain-specific and seems to be correlated with a protein receptor found in the cytosol of certain organs. Even though these protein receptors have been found in some human tissues, it is unknown if they play any role in human toxicity to dioxin. Regardless, epidemiologic studies focus on those biologic endpoints affected in animals, including immunotoxicity, carcinogeneity, genetic/reproductive outcomes, hepatotoxicity, and neurotoxicity (Grubbs et al., 1995). There is also the possibility that humans exposed to chronic, low doses of dioxin experience subtle, interrelated effects in their overall general health, rather than experiencing problems with specific organs (Grubbs et al., 1995). Therefore, several “general health” variables are investigated in the AFHS, as well as the more specific biological endpoints found in animal studies.

The AFHS’s investigation of general health looks at variables frequently used by clinicians in outpatient practice. The general health variables chosen for the AFHS are both sensitive and indirect measures of overall health rather than being specific to an organ.
Of the five variables included in the general health assessment of the AFHS, there are three subjective and two objective measures that are based on data collected from the questionnaire, physical examination, and laboratory examination. No participants are excluded for medical reasons for analysis of general health variables.

The questionnaire variable that is part of the general health assessment is self-perception of health. During the health interview, each study participant is asked, "Compared to other people your age, would you say your health is excellent, good, fair or poor?" Although self-perception of health is subjective and susceptible to varying degrees of conscious and subconscious bias, it is included as an indication of each subject’s overall health. For statistical analyses in the AFHS, self-reported health is dichotomized as "excellent or good" and "fair or poor".

The three variables from the physical examination in the general health investigation are appearance of illness or distress, relative age, and percent body fat. For the appearance of illness or distress, the examining physician states whether or not illness or distress is indicated by the subject's facial appearance. This assessment is made before the start of the physical examination, when the physician first observes the participant, and is answered as either "yes" or "no". Relative age, how well the veteran's apparent age and his true chronological age agree, is also assessed before the start of the physical examination. The physician indicates whether the subject appears younger than, older than, or the same as his stated age. Relative age is dichotomized as "older than" and "same as or younger than" for statistical analyses. To the extent the examining physicians remain blind to the group membership of each participant, these two variables will be less
biased than self-perception of health. The third variable from the physical examination, percent body fat, measures the relative body mass of an individual based on his height (in meters) and weight (in kilograms) at the physical examination. Body fat is calculated from the formula (Knapik et al., 1983):

\[
\text{Body fat (in percent)} = \left[ \frac{\text{wt (kg)}}{\text{ht}^2 \ (\text{m})} \right] \times 1.264 - 13.305
\]

In the AFHS, body fat is analyzed in both discrete and continuous forms. A natural logarithmic transformation is used in the continuous analyses, for normalizing purposes, while discrete analyses dichotomize into “lean or normal” (≤ 25 percent) and “obese” (> 25 percent).

The final general health variable, from the laboratory examination, is the erythrocyte sedimentation rate. The erythrocyte sedimentation rate is a nonspecific, though sensitive, indicator of general health, with unusually high elevated sedimentation rates being associated with some serious ongoing disease process (Grubbs et al., 1995). Sedimentation rate is also analyzed in the AFHS in both discrete and continuous forms. A natural logarithmic transformation is used in continuous analyses, with 0.1 added to the measurement before transforming, because of the presence of zeros. For discrete analyses, subjects are dichotomized into “normal” and “abnormal” categories, with cutpoints for the two categories changing over study cycles because of laboratory differences, more sensitive detection, and the effect of age on what are considered “normal” values. The 1982 Baseline examination had a cutpoint of 12 mm/hr for all subjects; those subjects with a sedimentation rate greater than 12 mm/hr were “abnormal”, and those with a sedimentation rate less than or equal to 12 mm/hr were
“normal”. The cutpoint was 20 mm/hr for both the 1985 and 1987 examinations. The 1992 follow-up had age-dependent cutpoints: the cutpoint was 15 mm/hr for subjects 50 years of age or younger, and 20 mm/hr for those at least 50 years old.

The five variables used in the general health assessment of the AFHS, self-perception of health, appearance of illness or distress, relative age, percent body fat, and erythrocyte sedimentation rate, are indirect indicators of overall health. A negative self-perception of health, an unhealthy appearance, an abnormal body weight and a high sedimentation rate are all traditional indicators of illness (Grubbs et al., 1995).

Overall, so far throughout the course of the AFHS, the general health assessment has found classical associations between the clinical measures of ill health, such as sedimentation rate, obesity/leanness, age, self-perception of health and the appearance of distress (Grubbs et al., 1995). The erythrocyte sedimentation rate variable has been the only general health variable consistently found to be significantly related to herbicide exposure (Grubbs et al., 1995). However, some of the results from analyses on the erythrocyte sedimentation rate have been contradictory. In addition, the longitudinal analyses have not used all the available data and have also been minimal. Therefore, we will expand on the longitudinal analyses conducted on the AFHS sedimentation rate data. In Chapter 5, we give results of longitudinal analyses on subjects who have data from all four study cycles. These analyses are based on methods developed by Fitzmaurice and Laird (1993). Chapter 5 also contains results of longitudinal analyses conducted on subjects who have data from all four study cycles or who have dropped out of the study. The “drop-out” analyses are based on the methods of Fitzmaurice, Laird and Zahner.
Finally, in Chapter 6, we develop, and then use, methodology to deal with subjects who have entered the study after it began in 1982 and methodology to deal with subjects who have complete data, who have dropped out of the study, or who have entered the study after 1982.

2.6 Results of Sedimentation Rate Analyses in the AFHS

The results of analyses conducted on sedimentation rate data from the four study cycles have been published by the AFHS after each cycle. Following is an overview of the results, from the general assessments, the herbicide exposure analyses, and the longitudinal analyses, for the four study cycles. The summarized results from the first study cycles, in 1982, are from Chapter 9 of *Air Force Health Study. An Epidemiologic Investigation of Health Effects in Air Force Personnel Following Exposure to Herbicides. Baseline Morbidity Study Results.* (Lathrop et al., 1984); results from the 1985, second study cycle are summarized from Chapter 9 of *Air Force Health Study. An Epidemiologic Investigation of Health Effects in Air Force Personnel Following Exposure to Herbicides. First Follow-up Examination Results.* (Lathrop et al., 1987); general assessments and MRI results from the third study cycle, in 1987, are from Chapter 9 of *Air Force Health Study. An Epidemiologic Investigation of Health Effects in Air Force Personnel Following Exposure to Herbicides. 1987 Follow-up Examination Results.* (Thomas et al., 1990) and the 1987 serum dioxin results are from Chapter 6 of *Air Force
The covariates age, race, military rank and occupation, and personality type have been included in AFHS analyses of sedimentation rate in pairwise associations and in adjusted analyses. The matching variables age, race, military rank and military occupation are included in the analyses of all dependent variables. Personality type is determined from the Jenkins Activity Survey, given during the physical examination, with Type A personalities indicated by positive Jenkins scores and Type B personalities by negative scores (Grubbs et al., 1995). Personality type was used in its continuous form for the analyses at the 1985 and 1987 examinations, and was dichotomized as Type A or Type B for analyses at the 1992 examination. Because the Jenkins Activity Survey was not given at the 1987 follow-up, subjects who attended the 1987 examination, but not the 1985 examination, were missing the personality type variable.

2.6.1 General Assessment

Baseline analyses of sedimentation rate by group membership, age ($\leq 40$, $> 40$) and the appearance of illness or distress determined that Ranch Hands have significantly
fewer sedimentation rate abnormalities than Comparisons, among those who were 40 years of age or less. There was no group difference among those over the age of forty. Sedimentation rate was also significantly positively associated with the appearance of illness or distress and with percent body fat. This is consistent with the fact that a high sedimentation rate, an abnormal body weight and an ill appearance are all traditional indicators of illness.

For the 1985 study cycle, continuous analyses found a significantly higher mean sedimentation rate in older subjects, but no difference between Ranch Hands and Comparisons. Discrete analyses found a significantly higher percentage of sedimentation rate abnormalities in Ranch Hands than in Comparisons, a significantly higher percentage of abnormalities in older subjects, and a somewhat higher percentage of abnormalities in enlisted flyers than in officers or enlisted groundcrew. This finding from 1985 was in direct opposition to the Baseline examination, where Ranch Hands 40 years old or less had significantly fewer sedimentation rate abnormalities than Comparisons.

The 1987 analyses did not uncover any significant group difference in mean sedimentation rates, but did find a significantly higher percentage of sedimentation rate abnormalities in the Ranch Hand group. This finding was similar to that of the 1985 follow-up examination. The percentage of sedimentation rate abnormalities increased significantly with increasing age, in addition to there being a significant relationship between occupation and sedimentation rate, with enlisted flyers having the highest mean sedimentation rates and the highest percent abnormal, and officers the lowest mean and the lowest percent abnormal.
In 1992, both continuous and discrete analyses failed to find any significant differences between Ranch Hands and Comparisons, but did find significant differences based on military occupation, with enlisted groundcrew having higher mean sedimentation rates and percentages of abnormalities. There was also a significantly higher incidence of abnormal sedimentation rates in older subjects than in younger subjects. These analyses confirmed results of the previous two examination cycles, with older veterans and enlisted subjects having more abnormal sedimentation rates than younger subjects and officers, respectively.

2.6.2 Herbicide Exposure Analyses

In the 1982 Baseline study, the military records occupation index (MRI) was used to investigate the possible existence of a dose-response effect in the sedimentation rate. Although it is now known that this exposure index is inaccurate in its attempt to quantify true herbicide exposure, the results from the herbicide exposure analyses are given here, so that they may be compared to other study results. The MRI was categorized into low, medium, and high for each of the occupational strata, and each stratum was analyzed separately. No statistically significant relationships were detected in the Ranch Hand group, for any of the dose-response analyses of the sedimentation rate, at the Baseline study.

Again, in 1985, the MRI was used to investigate possible dose-response relationships, and was again categorized as low, medium and high, with separate analyses
performed within each occupational stratum. For the enlisted groundcrew, there was a marginally significant increase in mean sedimentation rate with an increase in the MRI level; however, upon further adjustment for age, this mean difference became less significant. Discrete MRI analyses found a significantly higher percentage of abnormalities in enlisted flyers.

For the 1987 follow-up examination, sedimentation rate was analyzed in relation to both the MRI and serum dioxin measurements. Of the continuous and discrete MRI analyses, the only significant dose-response relationship was seen in the continuous adjusted analyses of enlisted flyers, where older subjects had significantly higher mean sedimentation rates.

In the 1987 analysis of the association between sedimentation rate and serum dioxin, significant positive relationships were detected between sedimentation rate and initial dioxin and between sedimentation rate and current dioxin for subjects whose time since Vietnam tour was longer than 18.6 years. There was a significantly higher percentage of sedimentation rate abnormalities in the Ranch Hands when the contrasts of categorized current dioxin were analyzed.

The 1987 herbicide exposure results were consistent with those from the 1985 follow-up. A significantly higher percentage of sedimentation rate abnormalities was seen in the Ranch Hands, in a dose-response pattern. In addition, there was a significantly higher percentage of abnormalities in Ranch Hand veterans with a longer time between their tour of duty in Vietnam and the follow-up.
In 1992, both discrete and continuous sedimentation rate analyses found a significant association between sedimentation rate and categorized dioxin, with higher mean sedimentation rates and higher incidences of abnormalities seen in Ranch Hands. There was also a marginally significant relationship between current lipid-adjusted dioxin and both continuous and discrete sedimentation rate. The model that included both whole-weight dioxin and total lipid weight also uncovered a significant relationship between whole-weight dioxin and sedimentation rate.

The two examinations directly prior to 1992 found significantly higher percentages of sedimentation rate abnormalities in Ranch Hands than in Comparisons, in a dose-response pattern. For the 1992 follow-up study cycle, however, neither continuous nor discrete analyses found significant differences between Ranch Hands and Comparisons. Although the enlisted groundcrew, the Ranch Hand occupation at greatest risk for exposure-related problems, did have a marginally higher mean sedimentation rate than did the Comparisons, the difference was not clinically significant. A few analyses of models using “current” serum dioxin as the exposure estimate detected results in a dose-response pattern, but the differences were small and the biological meaning unknown.

2.6.3 Longitudinal Analyses

In 1985, a longitudinal analysis of discrete sedimentation rate was run to investigate longitudinal differences in the 1982 Baseline examination and the 1985 follow-up examination. The off-diagonal data (normal to abnormal, abnormal to normal) from the
2 examinations were compared by group membership. This longitudinal analysis found highly significant group differences in the change in sedimentation rate abnormalities, because of a reversal of findings between the 1982 and 1985 examinations. While the Baseline examination had a significantly higher percentage of sedimentation rate abnormalities in younger Comparisons, the 1985 follow-up had a significantly higher percentage of abnormalities in Ranch Hands. It needs to be noted that the sedimentation rate laboratory test procedure used at the 1985 follow-up was more sensitive than that used at the Baseline examination.

Sedimentation rate was again longitudinally analyzed at the 1987 study cycle, based on data from the 1982 Baseline and the 1987 follow-up examinations. Although a lower percentage of Ranch Hands were abnormal at the Baseline, a higher percentage were abnormal at the 1987 follow-up; this is the same result as was found in the longitudinal analysis of the 1985 follow-up. There was a highly significant group difference in the change from Baseline to 1987 which can be explained by the fact that four times as many Ranch Hands changed from normal at Baseline to abnormal at the 1987 follow-up than vice versa, but nearly the same numbers changed in each direction in the Comparison group. A nonsignificant positive association between initial dioxin and the percentage of Ranch Hands who were abnormal at the 1987 examination, and a nonsignificant interaction between current dioxin and time since Vietnam tour were also uncovered. A dose-response pattern was seen in the longitudinal analyses of categorized dioxin, with those in “higher” categories of dioxin having significantly higher percentages of sedimentation rate abnormalities.
As in the previous two follow-up cycles, longitudinal analyses of sedimentation rate examined the difference between the 1982 and 1992 examinations, with subjects “abnormal” in 1982 excluded from the longitudinal analyses. These longitudinal analyses found nonsignificant results between Ranch Hands and Comparisons, in the differences between initial dioxin and percent abnormal in 1992, and in the contrasts between categories of dioxin.

Some of the longitudinal results uncovered in 1992 are in direct contrast to previous examination results. Both 1985 and 1987 longitudinal analyses detected the Ranch Hands as having significantly higher percentages of abnormal sedimentation rates, giving support for the possibility of a subtle occult disease process following dioxin exposure. However, the 1992 longitudinal analyses found neither significant group differences nor biologic differences in the means across dioxin categories. Because of the interesting findings related to sedimentation rate, we will use the sedimentation rate data in our longitudinal analyses in Chapters 3, 5 and 6.
CHAPTER 3

ANALYSIS OF MISSING DATA IN THE AFHS

One of the main goals of the AFHS is to investigate changes in laboratory variables and to determine whether rates of abnormalities differ over time between Ranch Hands and Comparisons. To help accomplish this goal, the AFHS conducts longitudinal analyses on some of the dependent variables. One variable that has been analyzed longitudinally at each follow-up examination is the erythrocyte sedimentation rate, since the effects of chronic low-dose dioxin exposure in humans is unknown and since the erythrocyte sedimentation rate is a nonspecific, though sensitive, indicator of overall health. However, the presence of missing data could lead to problems in analyzing the data and in drawing valid conclusions. In Section 3.1 we discuss some of the situations that lead to missing and/or incomplete data in the AFHS. Section 3.2 gives results of some exploratory analyses related to missingness that we conducted on data from the AFHS.
3.1 Missing Data in the AFHS

3.1.1 Missing Data Due to Non-Compliance

The biggest potential for missing data in the study arises from noncompliance to the study. Veterans eligible for the study can fully comply to the physical examination and to the questionnaire, can partially comply to only the questionnaire, can refuse both the questionnaire and the physical, or can be unlocateable and not have been invited to the study. Subjects who do not participate in the study, for any reason, are missing data on all questionnaire and physical examination variables, both dependent and independent, for study cycles for which they are non-participants. Since nearly all data analyzed in the study comes from the physical examination, only those subjects who are fully compliant to a study cycle are included in the statistical analyses for that cycle. This is the reason why full compliance to the study is so crucial, especially since participation in the AFHS is completely voluntary.

At the beginning of each study cycle, each living Ranch Hand veteran is invited to take part in that cycle. Because the entire Ranch Hand population is invited to the study, there is no way to substitute for Ranch Hands who do not participate. However, when a Comparison who is invited to the study is noncompliant, he is replaced by the next Comparison in the matched set for his respective Ranch Hand veteran who has the same self-perception of health. In this way, the Study Protocol attempts to stem the problem of noncompliance, although not all Comparisons who refuse are replaced, due to such things
as lack of an eligible replacement Comparison in the Ranch Hand’s matched set, or refusal of all eligible replacement Comparisons.

Two other ways missing data arises in the AFHS through compliance / noncompliance is through drop-out and by late arrival into the study. When a subject participates in one or more physical examinations and then, from a certain point onward, fails to participate, he is a “drop-out”. He will be missing data for all study cycles after he drops out of the study, and will have what is known as monotone missing data. A monotone missing data pattern occurs if whenever an observation $Y_j$ is missing, so are $Y_{j'}$ for all $j' > j$. A subject can become a drop-out due to such things as death, becoming unlocateable for the remainder of the study, or refusing to participate in the remaining study cycles.

Late arrival into the study is simply enrolling in the AFHS after the Baseline study cycle in 1982. These subjects will be missing physical examination and questionnaire data for all study cycles that were held up to when they were asked, and they agreed, to participate in the AFHS. Enrolling in the study after Baseline is most often seen in the Comparison group, when replacement Comparisons agree to participate in the study, following the noncompliance of another Comparison in his matched set. Late entry into the study also occurred for five Ranch Hand veterans who were not found in the original search of military records. The last way a veteran can enter the AFHS after the Baseline study cycle is if he had been noncompliant to the Baseline, and possibly one or more follow-up study cycles, and then decides to participate in the study. Regardless of the
reason for a veteran's late arrival into the study, he will be missing data for all study cycles
to which he did not participate.

3.1.2 Missing Data in Fully Compliant Subjects

Even when a subject fully complies to a study cycle, there is still the possibility that
he will not have complete data on all variables, since missing data in the AFHS can arise
during the implementation of the examinations or in the analysis of the results. Subjects
who attend the physical examinations are encouraged to complete the entire physical, but
they are not forced to do so. Sometimes a subject “does not feel like” being weighed or
having blood drawn or answering a specific question. In some instances, when one
variable is missing, it leads to more missing data. For example, the variables weight and
height are used to calculate the body fat variables; thus, if weight and/or height is missing
for a subject at a study cycle, he will also be missing all of his body fat variables for that
cycle. In addition, problems with laboratory procedures or with measurement systems can
lead to missing data. For example, if an insufficient amount of blood is drawn from a
subject or if something happens to the sample between the draw and the laboratory
analyses, there is a chance that some of the dependent and/or independent variables will be
missing. In particular, with regard to measurement systems, the dioxin assay has led to
some subjects having missing data for dioxin results, even if blood has been drawn
specifically for the assay. Although the assay is able to detect minute amounts of dioxin in
serum, some subjects have dioxin levels that are below the level of quantiation. This
means that the assay has determined there is dioxin present in the serum, but the quantity of dioxin is at such a low level, it can not be determined. In these instances, the subject is missing his serum dioxin data, but it is noted that the data is missing due to being below the level of quantitation. Lastly, some subjects are ill or are suffering from health problems, but agree to participate in the physical examination. It is not always in their best interest to complete the entire physical; these subjects may be missing some physical examination variables.

3.1.3 Missing Data Due to Study Design

Missing data is also present because there were no records kept during Operation Ranch Hand as to who was occupationally exposed to herbicides and, of those exposed, the extent of their exposure. Although serum dioxin levels can now be ascertained for subjects, there is no way to accurately quantify all Ranch Hand veterans’ initial exposure to the herbicides. It is assumed in the AFHS that those Ranch Hand veterans with current serum dioxin levels above background were occupationally exposed to herbicides in Vietnam. However, there is no way to accurately determine which, if any, Ranch Hands with current serum dioxin levels below background were ever exposed to herbicides during their Vietnam tour. Different surrogates, such as group membership and military occupation, the Military Records Index, and current and extrapolated initial serum dioxin body burdens, have been used for estimating exposure to the herbicides. Unfortunately,
none of the surrogates can ever fully replace the missing data from the time of each veteran’s Ranch Hand tour.

In addition, all of the AFHS’ serum dioxin analyses that have compared Ranch Hands to Comparisons have excluded Comparisons with serum dioxin levels above background. The models comparing Ranch Hands and Comparisons that use serum dioxin as the exposure surrogate have been used to analyze data from the 1987 and 1992 study cycles. These serum dioxin models categorize Ranch Hands based on initial or current serum dioxin levels, and then define the Comparison category as only those Comparisons with current dioxin levels below background. It is understandable that the AFHS wants to investigate the differences between subjects with background levels of dioxin and subjects with above background dioxin levels. However, by failing to include Comparisons whose dioxin levels are above background, the AFHS appears to be assuming that only veterans who are occupationally exposed to dioxin can have above background levels of dioxin. This is not true, since activities such as smoking can lead to increases in a person’s serum dioxin levels. It is possible that the failure to use all the available data, by excluding some of the Comparisons, is leading to inaccurate results for the dioxin analyses, since the subset of Comparisons included in the analyses may be biased from the total Comparison population and since there could be a great deal of valuable information in the excluded data.

The longitudinal analyses performed in the AFHS are somewhat different from what one might expect, and encounter some missing data of their own. At each follow-up study cycle, the data analyzed longitudinally come from two timepoints: the 1982
Baseline study cycle and the current follow-up examination. Thus, for example, for the 1992 longitudinal analyses, the data was from the 1982 and 1992 examinations. This seems somewhat different from what might be expected. One might expect the longitudinal analyses to investigate the changes between consecutive cycles, or between all study cycles that have been conducted, as well as between the Baseline cycle and the current follow-up study cycle. Subjects who enrolled in the study after 1982, due to any reason, are not included in the longitudinal analyses, even if they have data from every study cycle since the 1982 Baseline examination. Also, those subjects who participated in the early cycle(s) of the AFHS but then dropped out, are excluded from the longitudinal analyses, even if the only data they are missing is from the current cycle. In addition, the discrete longitudinal analyses include only those subjects who were “normal” at the 1982 examination. The reason for this is because the longitudinal analyses are meant to investigate the temporal effects of dioxin between 1982 and the follow-up examination. Because those “abnormal” in 1982 were already abnormal at the start of the study, they are not considered to be at risk.

The longitudinal analyses for each follow-up examination not only exclude subjects who were “abnormal” at Baseline, subjects who are missing Baseline examination data, and subjects who dropped out of the study, they also fail to include any data from timepoints between Baseline and the current examination. It seems plausible that the analyses could be expanded to include more timepoints, more subjects, and more varied analyses, in order for a more comprehensive and thorough longitudinal investigation.
3.2 Preliminary Analyses of Missing Data in the AFHS

Because nearly all the statistical analyses performed for the AFHS are on data that comes from the physical examination, only those subjects who fully comply to a study cycle are included in those analyses. In addition, the longitudinal analyses for a certain follow-up are based on data from the Baseline examination and that follow-up examination, leading to only subjects fully compliant to both the Baseline study cycle and that follow-up cycle being included in the longitudinal analyses. Because I was interested in determining if the subjects in a cycle's longitudinal analysis were possibly biased from the overall population of eligible study participants, most of my exploratory data analyses centered around the compliance status of individuals invited to participate in the AFHS.

3.2.1 Analyses of Compliance Status

The first exploratory data analysis I performed looked at Ranch Hand compliance status, for each pair of consecutive study cycles. I cross-classified each consecutive pair of compliance status categories for members of the Ranch Hand group. Four log-linear models were fit to the data. For each model, Table 3.1 gives the models fit, the pairs of consecutive study cycles, the G² and X² statistics, and the degrees of freedom for the model. Quasi-Independence Model 1 is the model of quasi-independence with the cells for those fully compliant to the two consecutive study cycles fit exactly; this model means that there is independence in compliance status, after accounting for those subjects who
are fully compliant to both of the consecutive study cycles. Quasi-Independence Model 2 is the model of quasi-independence with those fully compliant to both study cycles and those refusing to comply to both study cycles fit exactly; this model means that, after accounting for subjects who fully comply to both study cycles and subjects who refuse to comply to both study cycles, there is independence in compliance status. Quasi-Independence Model 3 is the model of quasi-independence with all subjects with the same compliance status for the two study cycles fit exactly. This model, called the “stayer” model due to the fact that all of the subjects who stay in the same compliance status category are accounted for, means that there is independence in consecutive pairs of compliance status upon accounting for subjects who remain the same in terms of compliance status.

The independence model clearly does not fit any of the three pairs of consecutive study cycles. Also note the large differences between the $G^2$ and $X^2$ statistics for the independence model for each of the pairs of cycles indicating that the small cell counts may be causing a problem. The standardized residuals from each of the independence models (not included) indicated that the lack-of-fit was from those subjects staying in the same category over consecutive study cycles, with the greatest lack-of-fit related to subjects who were fully compliant to consecutive study cycles. Thus, I fit various quasi-independence (QI) models for each of the three pairs of consecutive study cycles. Table 1 results show that QI models fit the compliance data much better than did the independence model. The “stayer” quasi-independence model, QI 3, fits the Ranch Hand cycles 2 and 3 data well (p-value = .4483; d.f.=11), as does the QI model that accounts for full compliers
and noncompliers, QI 2 (p-value = .2194; d.f.=14). Since the fit of the “stayer” QI model did not improve much over the fit of the full complier/noncomplier QI model, for cycles 1 and 2 and for cycles 3 and 4, I examined the standardized residuals from the latter model to determine where the lack-of-fit was originating. The standardized residuals from cycles 3 and 4, after accounting for those fully compliant to both cycles and those noncompliant to both cycles, indicated that more Ranch Hands fully compliant at cycle 3 died between cycles 3 and 4 than was expected, and fewer Ranch Hands noncompliant to cycle 3 died between cycles 3 and 4 than was expected. I do not have an explanation for these results.

For cycles 1 and 2, there were three relatively large standardized residuals from cells relating to Ranch Hands newly compliant at cycle 2. All three of the cells (newly discovered Ranch Hands being either fully or partially compliant at cycle 2, and Ranch Hands who had been noncompliant at cycle 1 becoming partially compliant at cycle 2) had much larger observed counts than expected. This can be explained by the influence of the Ranch Hand Association; the Association strongly encouraged Ranch Hands to take part in the study. The participation of Ranch Hands who were first invited to the study in 1985 or who had been noncompliant to the 1982 study cycle, would have been among those most-strongly supported by the Association. It is interesting to note that, in this same model fit to the data for cycles 2 and 3, these same three cells also had large standardized residuals, indicating this situation is not exclusive to cycles 1 and 2. Upon further accounting for Ranch Hands in these three cells in the full complier/noncomplier QI model, the $G^2$ statistic decreased from 85.87 to 15.67 with 11 d.f.; the p-value for this model is .1539. Overall, this model indicates that, for cycles 1 and 2, upon accounting for
Ranch Hands who are fully compliant to both cycles, who are noncompliers to both cycles, and who noncomply to cycle 1 (either intentionally or not) but comply to cycle 2, there is independence in Ranch Hand compliance status for the two cycles.

I then analyzed the consecutive pairwise compliance status of the Comparison group, in a way identical to the Ranch Hand group. Table 3.2 gives the various models fit, the pairs of study cycles, the $G^2$ and $X^2$ statistics, and the degrees of freedom. I again found the independence model a poor fit to pairwise compliance; the fit of the independence model in the Comparison group, however, was much worse than it had been in the Ranch Hand group. The standardized residuals from each of the independence models suggested, as they had in the Ranch Hand group, that lack-of-fit was in the cells of the contingency table relating to Comparison subjects who had the same compliance status for both of the consecutive cycles. By simply accounting for Comparisons who fully complied at consecutive study cycles, model QI 1, the $G^2$ and $X^2$ statistics decreased dramatically; note that the $G^2$ and $X^2$ values were still relatively high and their difference, especially for cycles 3 and 4, quite large. I also fit symmetry and quasi-symmetry models to each of the pairs of study cycles; however, since these models had an even larger lack-of-fit than the independence and QI models, I did not include the results.

None of the models I fit for the Comparison group, for any of the consecutive pairs of study cycles, fit as well as in the Ranch Hand group. I again looked at the standardized residuals for each pair of consecutive cycles to identify cells in which the fit was poor.

For Comparisons at cycles 1 and 2, there were two cells with large standardized residuals in each of the QI models I fit. The standardized residuals from these two cells
indicated that more Comparisons than expected went from fully compliant at cycle 1 to unlocateable at cycle 2, and more comparisons than expected went from noncompliant at cycle 1 to partially compliant at cycle 2. When a subject is/becomes unlocateable, that simply means that the AFHS has no current address and phone number for that subject, and so cannot contact the subject to participate in the study. It is interesting to note that more cycle 1 fully compliant Comparisons were unlocateable at cycle 2 than was expected, since it was at cycle 1 that not all addresses and phone numbers of Comparisons were known. The fact that more Comparisons than expected changed from noncompliant at cycle 1 to partially compliant at cycle 2 can be explained by the fact that, similar to the Ranch Hand noncompliers at cycle 1, subjects who noncomplied to the Baseline examination cycle in 1982 were strongly encouraged to participate in the AFHS in some way, be it only completing the questionnaire or be it completing both the questionnaire and the physical examination. I re-examined each of the three QI models fit to the cycles 1 and 2 Comparison compliance by also fitting exactly cells for Comparisons who noncomplied at cycle 1 and partially complied at cycle 2, and Comparisons who fully complied at cycle 1 and were unlocateable at cycle 2; the $G^2$ statistic decreased dramatically in all three QI models (an average decrease of 110, with a decrease of 2 degrees of freedom), indicating that it is worthwhile to further account for these two situations. However, the best QI model fit to the data, the “stayer” model that additionally accounted for the two cells, still had a $G^2$ value of 67.24 on 9 degrees of freedom; this is a marked improvement to the previous QI models, but still shows some lack-of-fit.
For Comparison compliance for cycles 2 and 3, the lack-of-fit for each of the three QI models fit was again indicated by unusually large standardized residuals in the two cells for Comparisons who were newly invited to take part in the AFHS. There were many more newly invited Comparisons who either partially or fully complied to cycle 3 than expected. Again, as with the Ranch Hands, this can be explained by the fact that participation in the AFHS was encouraged, especially for those either newly invited to the study or those who had previously noncomplied. In the full complier/noncomplier QI model, additionally fitting exactly those cells for newly invited partial and full compliers led to $G^2 = 66.40$ and $X^2 = 69.27$ with 12 degrees of freedom. Although this, too, is not a good fit, it is a marked improvement over the previous models tried on this data.

Finally, for cycle 3 and 4 Comparison compliance, there was one unusually large standardized residual for each of the three QI models fit, indicating that more Comparisons than expected who were newly invited to the study at cycle 4 were unlocateable. Since the AFHS had been underway for 10 years by the time cycle 4 was conducted, and since these newly invited Comparisons had never been contacted previous to cycle 4, it is not surprising that some of the addresses and phone numbers were no longer current. After accounting for newly invited Comparisons who were unlocateable at cycle 4 in the full complier/noncomplier QI model, model QI 2, $G^2 = 63.20$ and $X^2 = 63.09$ with 9 degrees of freedom, again a marked improvement to previous models.

These exploratory data analyses indicated that, in general, Ranch Hand compliance at two consecutive study cycles is relatively well explained by a quasi-independence model.
that accounts for subjects either fully compliant at both cycles or noncompliant at both cycles. That is, after accounting for Ranch Hand veterans who fully comply to both cycles and Ranch Hands who noncomply to both cycles, there is independence in Ranch Hand compliance status over consecutive pairs of study cycles. In addition, further accounting for Ranch Hands newly compliant at the second of the two consecutive cycles appears to increase the goodness-of-fit even more. The picture does not seem as clear for the Comparison group; however, some patterns do exist. For cycles 1 and 2 and for cycles 2 and 3, under each of the models I fit, there were many more newly compliant Comparisons than was expected, because of previous noncompliers and newly invited Comparisons complying at the second of the two cycles. In addition, for cycles 1 and 2 and for cycles 3 and 4, more Comparisons became unlocateable than was expected.

The results of analyzing Ranch Hand compliance can be understood by three facts. Some Ranch Hand veterans are trying to forget their Vietnam experiences and want nothing to do with anybody or anything related to Operation Ranch Hand. These veterans are among those who noncomply to every cycle of the AFHS; some veterans even go so far as to threaten to sue the Air Force if they are ever contacted again about participating in the AFHS. There are also some Ranch Hand veterans who are very interested in what the AFHS can learn about the effects of dioxin exposure to humans. Among these veterans are those who believe that their health may have been diminished due to their exposure to the herbicides, as well as those veterans who believe that the entire “Agent Orange thing” has been over-exaggerated and that the dioxin exposure experienced by members of Operation Ranch Hand is not causing detrimental health effects. Lastly,
Ranch Hand subjects who were noncompliant or who were late to be classified as a Ranch Hand member have been the main focus of support by the Ranch Hand Association in backing participation in the AFHS. Since the veterans of Operation Ranch Hand are the focus of this study, their full compliance to every physical examination and every questionnaire, at every study cycle, is of the utmost importance.

For Comparison compliance, that the best-fitting models account for fully compliant subjects remaining fully compliant and noncompliant subjects staying noncompliant, indicates that, as in the Ranch Hand group, “compliers comply and noncompliers do not”. Fully compliant subjects, whether Ranch Hand or Comparison, tend to remain fully compliant while noncompliant subjects, again Ranch Hand or Comparison, tend to remain noncompliant. In addition, more newly invited and previously noncompliant Comparisons comply to the study than is expected, a finding also present in the Ranch Hand group. Even though the Comparisons do not have a support group like the Ranch Hand Association, everyone associated with the AFHS realizes the importance of maintaining an adequate “control” population and, thus, encourages Comparisons to take part in the study.

3.2.2 Analyses of Presence of Examination Data

After analyzing compliance status for both Ranch Hands and Comparisons, I then analyzed the presence or absence of physical examination data over the four study cycles. Only fully compliant subjects will have physical examination and, thus, sedimentation rate
data. Subjects eligible to take part in a study cycle, but who either partially comply, noncomply, are unlocateable, or have died since the previous cycle will be missing physical examination and sedimentation rate data. Table 3.3 has the number of eligible subjects for each study cycle, classified by the presence or absence of physical examination data, for both the Ranch Hand and Comparison groups.

I fit a model of independence to each of the two groups to determine whether missing physical examination data was independent of study cycle. For the Ranch Hand group, $G^2 = 17.51$ and $X^2 = 17.34$ with 3 degrees of freedom (p-value=.0006); while for the Comparisons, $G^2 = 6.36$ and $X^2 = 6.42$ with 3 degrees of freedom (p-value=.0954). Independence fits neither the Ranch Hands nor the Comparisons. The lack-of-fit in the independence model for the Ranch Hand group stems from more Ranch Hands than expected having physical examination data for the first two study cycles, and then more Ranch Hands than expected missing physical examination data for the third and fourth study cycles. This agrees with the compliance status analyses, which found that newly invited Ranch Hands were more likely to comply to the study, that Ranch Hands noncompliant to cycle 1 were more likely to comply at cycle 2, and that more Ranch Hands fully compliant to cycle 3 died between cycle 3 and cycle 4 than was expected. Although independence between study cycle and the presence/absence of physical examination data fits much better for the Comparisons, there is some lack-of-fit. More Comparisons than expected are missing physical examination data at cycle 4, and more than expected have physical examination data at cycles 1, 2 and 3. This also reinforces the results from the compliance status models, where more newly invited and previously
noncompliant comparisons than expected were compliant to the study for the early cycles, but more newly invited comparisons than expected were unlocateable at the last study cycle.

### 3.2.3 Analyses of Dioxin versus Compliance

The final exploratory data analysis I performed centered on serum dioxin values and compliance. Body burden dioxin levels were determined for subjects at both the cycle 3 physical examination in 1987 and the cycle 4 physical examination in 1992, giving a clear indication of the subject’s serum dioxin level at that time. Prior to this, there was no way to accurately estimate a subject’s dioxin exposure, regardless of whether or not he participated in the AFHS. However, by fully complying to cycle 3 and/or cycle 4, a subject would have had blood drawn for the dioxin assay and would have learned his body burden dioxin level following the laboratory analyses.

I performed exploratory analyses on cycle 3 serum dioxin values and cycle 4 compliance status; since cycle 5 is being held in 1997, I can not yet analyze cycle 4 dioxin and cycle 5 compliance status. Although the dioxin assay is able to detect minute amounts of dioxin in the blood, there remains the possibility that a sample will have a level of dioxin that is below the level of quantiation. In addition, there can be problems with the blood draw or the sample obtained from the draw, leading to the inability to get a dioxin body burden result. Therefore, there is the possibility that a subject who is fully compliant to the cycle 3 physical examination will be missing cycle 3 dioxin values. For subjects who have dioxin data, those with current lipid-adjusted values below 10 ppt are considered to
have "background" levels of dioxin, while those having current lipid-adjusted dioxin above 10 ppt have "above background" levels of dioxin. Subjects who were fully compliant to the cycle 3 examination and, thus, had blood drawn for a dioxin assay, were defined as having background, above background, or missing dioxin values. These subjects were then cross-classified by their cycle 4 compliance status. Table 3.4 gives the cross-classification for both Ranch Hand and Comparison groups.

The first analysis I performed was to determine whether missing dioxin values are missing at random, for both the Ranch Hand and Comparison groups. I used the EM-algorithm (Dempster, Laird and Rubin; 1977) in both groups to allocate the "missing dioxin" observations into the "background" and "above background" cells, in order to obtain the expected cell probabilities. I assumed a model of random missingness for the missing-data mechanism and a saturated model for the cell probabilities. I determined the expected cell counts for both of the cross-classification tables, and then computed the $X^2$ statistic. For the Comparison group, $X^2 = 3.85$ with 3 degrees of freedom ($p$-value=.2785), indicating that the random missingness model for dioxin values is reasonable in the Comparison group. For the Ranch Hand group, however, $X^2 = 9.39$ with 3 degrees of freedom ($p$-value=.0245), indicating that the cycle 3 missing dioxin values are not missing at random for the Ranch Hands. Two cells in the Ranch Hand contingency table exhibited evidence of lack-of-fit for the missing at random model. More subjects who noncomplied to cycle 4 were missing cycle 3 dioxin values, and more subjects who passed away between cycles 3 and 4 were missing cycle 3 dioxin values, than expected. One possible explanation as to why subjects passing away between cycles 3 and 4 were also
missing cycle 3 dioxin values, is that it is sometimes not in the best interest of fully complaint subjects who are ill to have blood drawn for the dioxin assay. That more Ranch Hand subjects missing cycle 3 dioxin values were noncompliant at cycle 4 than expected could possibly be due to these veterans becoming dissatisfied with the AFHS after their dioxin levels could not be ascertained. It will be interesting to analyze similar data on cycle 4 dioxin and cycle 5 compliance status to see if the same patterns exist there.

The other analysis I performed on cycle 3 dioxin and cycle 4 compliance tested whether cycle 3 dioxin level is independent of cycle 4 compliance. I was concerned that veterans, especially Ranch Hand veterans, would remain in the study until they learned their serum dioxin level, and then would drop out of the study. I was most-interested in determining if Ranch Hand veterans with background levels of dioxin, after learning their dioxin levels are "okay", would drop out of the study. I can only test this for cycle 3 dioxin and cycle 4 compliance status, but will again test this when cycle 5 compliance data becomes available once cycle 5 has been completed. The data in Table 3.4 is again the data needed for this analysis. For Comparisons, the test for independence between cycle 3 dioxin and cycle 4 compliance status gave $X^2 = 5.36$ with 6 degrees of freedom (p-value=.499) and for Ranch Hands, $X^2 = 9.692$ again with 6 degrees of freedom (p-value=.138). Thus, it appears that for both Ranch Hands and Comparisons, cycle 3 dioxin is independent of cycle 4 compliance status. This finding was reassuring, since it indicated that a subject who knows his serum dioxin level is no more or less likely to noncomply to a future study cycle. I will continue to monitor all of the areas on which I ran exploratory data analysis as more study cycles are completed.
Overall, exploratory analyses on the AFHS data indicated that subjects who fully comply to a study cycle are likely to fully comply to other cycles. This is a positive finding in terms of the longitudinal analyses conducted using the AFHS data, since each study cycle’s longitudinal analyses include only those subject who were fully compliant to the Baseline study cycle and to that particular cycle. However, it appears that there are some other patterns to study compliance that could affect the results of the study’s longitudinal analyses. We discovered that there are numerous newly invited or previously noncompliant Ranch Hands and Comparison who have become fully compliant to the study since Baseline (thus having “drop-in” data), in addition to a number of subjects who fully complied to the Baseline cycle but have since dropped out of the study (thus having “drop-out” data). Our preliminary analysis indicated that failing to account for subjects who entered the study after Baseline or who dropped out of the study before it ends could have a substantial effect on the results of any longitudinal analyses conducted. Therefore, we will consider methodologies that will help us to longitudinally analyze the AFHS sedimentation rate data using only complete subjects; only subjects with complete and drop-out data; only complete or drop-in data; and subjects with complete, drop-out or drop-in data.
### Independence Model

<table>
<thead>
<tr>
<th>Cycles</th>
<th>( G^2 )</th>
<th>( X^2 )</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>cycles 1&amp;2</td>
<td>467.02</td>
<td>887.46</td>
<td>16</td>
</tr>
<tr>
<td>cycles 2&amp;3</td>
<td>551.39</td>
<td>793.72</td>
<td>16</td>
</tr>
<tr>
<td>cycles 3&amp;4</td>
<td>484.45</td>
<td>639.76</td>
<td>9</td>
</tr>
</tbody>
</table>

### Quasi-Independence Model 1 (Full compliers fit exactly)

<table>
<thead>
<tr>
<th>Cycles</th>
<th>( G^2 )</th>
<th>( X^2 )</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>cycles 1&amp;2</td>
<td>108.48</td>
<td>121.14</td>
<td>15</td>
</tr>
<tr>
<td>cycles 2&amp;3</td>
<td>51.73</td>
<td>57.15</td>
<td>15</td>
</tr>
<tr>
<td>cycles 3&amp;4</td>
<td>55.79</td>
<td>58.79</td>
<td>8</td>
</tr>
</tbody>
</table>

### Quasi-Independence Model 2 (Full compliers and Noncompliers fit exactly)

<table>
<thead>
<tr>
<th>Cycles</th>
<th>( G^2 )</th>
<th>( X^2 )</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>cycles 1&amp;2</td>
<td>85.87</td>
<td>149.30</td>
<td>14</td>
</tr>
<tr>
<td>cycles 2&amp;3</td>
<td>18.08</td>
<td>17.73</td>
<td>14</td>
</tr>
<tr>
<td>cycles 3&amp;4</td>
<td>26.45</td>
<td>25.56</td>
<td>7</td>
</tr>
</tbody>
</table>

### Quasi-Independence Model 3 (Stayers fit exactly)

<table>
<thead>
<tr>
<th>Cycles</th>
<th>( G^2 )</th>
<th>( X^2 )</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>cycles 1&amp;2</td>
<td>72.25</td>
<td>83.60</td>
<td>11</td>
</tr>
<tr>
<td>cycles 2&amp;3</td>
<td>11.57</td>
<td>10.94</td>
<td>11</td>
</tr>
<tr>
<td>cycles 3&amp;4</td>
<td>23.40</td>
<td>22.84</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 3.1: Models for Ranch Hand Compliance Status for Each Pair of Consecutive Study Cycles

52
<table>
<thead>
<tr>
<th>Independence Model</th>
<th>Cycles</th>
<th>$G^2$</th>
<th>$X^2$</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cycles 1&amp;2</td>
<td>955.63</td>
<td>1079.36</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>cycles 2&amp;3</td>
<td>1184.32</td>
<td>1292.74</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>cycles 3&amp;4</td>
<td>1131.91</td>
<td>1398.79</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quasi-Independence Model 1 (Full compliers fit exactly)</th>
<th>Cycles</th>
<th>$G^2$</th>
<th>$X^2$</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cycles 1&amp;2</td>
<td>219.21</td>
<td>222.40</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>cycles 2&amp;3</td>
<td>149.78</td>
<td>180.36</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>cycles 3&amp;4</td>
<td>183.02</td>
<td>233.30</td>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quasi-Independence Model 2 (Full compliers and Noncompliers fit exactly)</th>
<th>Cycles</th>
<th>$G^2$</th>
<th>$X^2$</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cycles 1&amp;2</td>
<td>204.97</td>
<td>229.23</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>cycles 2&amp;3</td>
<td>130.95</td>
<td>127.59</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>cycles 3&amp;4</td>
<td>106.96</td>
<td>110.40</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quasi-Independence Model 3 (Stayers fit exactly)</th>
<th>Cycles</th>
<th>$G^2$</th>
<th>$X^2$</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cycles 1&amp;2</td>
<td>158.72</td>
<td>161.87</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>cycles 2&amp;3</td>
<td>119.89</td>
<td>117.72</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>cycles 3&amp;4</td>
<td>102.09</td>
<td>111.92</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 3.2: Models for Comparison Compliance Status for Each Pair of Consecutive Study Cycles
<table>
<thead>
<tr>
<th>Cycle</th>
<th>Ranch Hands</th>
<th>Have</th>
<th>Missing</th>
<th>Comparisons</th>
<th>Have</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cycle 1</td>
<td>1045</td>
<td>163</td>
<td>1224</td>
<td>443</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cycle 2</td>
<td>1016</td>
<td>201</td>
<td>1293</td>
<td>447</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cycle 3</td>
<td>995</td>
<td>207</td>
<td>1299</td>
<td>448</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cycle 4</td>
<td>952</td>
<td>235</td>
<td>1281</td>
<td>521</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: Numbers of Ranch Hands and Comparisons with Available and Missing Physical Examination Data

<table>
<thead>
<tr>
<th>Cycle 3 Dioxin</th>
<th>Fully Compliant</th>
<th>Refused</th>
<th>Unlocateable</th>
<th>Deceased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>324</td>
<td>13</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Above Background</td>
<td>414</td>
<td>18</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>Missing</td>
<td>176</td>
<td>14</td>
<td>1</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cycle 3 Dioxin</th>
<th>Fully Compliant</th>
<th>Refused</th>
<th>Unlocateable</th>
<th>Deceased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>990</td>
<td>52</td>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td>Above Background</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Missing</td>
<td>178</td>
<td>12</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 3.4: Cycle 3 Dioxin versus Cycle 4 Compliance Status for Ranch Hands and Comparisons
CHAPTER 4

LITERATURE REVIEW

In longitudinal studies, repeated observations on a response are made on subjects over time. Although the focus of longitudinal studies is frequently on the relationship between the marginal response probabilities and the covariates, the association between responses, and the relationship between the association and the covariates, are often of interest. Regardless of the focus of longitudinal studies, the association between the responses for a single subject must be accounted for in order to make correct inferences.

For analysis of continuous response variables, the approach taken by Ware (1985) uses a general class of linear models for the response probabilities, and structural models for the associations between responses. A key assumption is that the responses are approximately Gaussian, which leads to a parameterization only in terms of mean and covariance parameters that vary independently in separable parameter spaces (Ware, 1985). There is no discrete multivariate analogue to this multivariate Gaussian model. In
addition, the missing data, drop-outs, and time-varying covariates present in longitudinal studies make standard multivariate procedures unsuitable (Ware, 1985).

However, the analysis of discrete longitudinal data has been of considerable interest recently, especially in terms of marginal models for binary data. Marginal models for multivariate binary responses separate the model for the marginal expectation of each binary response from the model for the association between the binary responses. Both models are expressed in terms of covariates. The generalized estimating equations (GEE) approach of Liang and Zeger (1986) is one such procedure for analyzing discrete longitudinal data. With GEE, the main interest is on the marginal expectations and the covariates, but they do account for the dependence between responses. GEE is not likelihood-based, and so does not require the complete specification of the joint distribution of the data. Prentice (1988) developed a model for instances where the focus is on the association between responses and between the association and the covariates. He simultaneously modeled the responses and the pairwise products of responses using second-order estimating equations. Although this method leads to estimates of association parameters that are more efficient than those obtained with GEE, it can also become quite computationally intense as the number of response periods gets large.

Fitzmaurice, Laird and Rotnitsky (1993) give a review of recent methods for analyzing repeated binary responses. They describe both non-likelihood-based approaches, such as GEE, and likelihood-based approaches, where the complete joint distribution of the binary response is clearly stated. The authors give advantages and disadvantages of each approach, and compare the two approaches in terms of asymptotic
relative efficiency and bias. The discussants in Liang, Zeger, and Qadish (1992), however, point out that likelihood-based approaches are often preferred. In Section 4.1, we present methodology that can be used when analyzing complete longitudinal binary data; Section 4.2 details methodology for when the data is incomplete.

4.1 Regressions Models for Longitudinal Binary Data

Assume each of N subjects is observed at the same T times. If the responses are binary, we can form the (Tx1) vector of responses for subject i, \( Y_i = (Y_{i1}, ..., Y_{iT})' \), where \( Y_{it} = 1 \) if subject i has response 1 (success) at time \( t = t \), and \( Y_{it} = 0 \) otherwise. In addition, assume that each subject has a (Pxn) covariate vector \( x_i \) at time t; and from this, form the TxP covariate matrix \( X_i = (x_{i1}, ..., x_{iT})' \) for subject i. The marginal distribution of the variable \( Y_{it} \) is Bernoulli, with

\[
f(y_{it} | x_i) = \exp \left\{ y_{it} \theta_{it} - \log \left[ 1 + \exp (\theta_{it}) \right] \right\},
\]

where

\[
\theta_{it} = \text{logit} (\mu_{it}) = \log \left( \frac{\mu_{it}}{1 - \mu_{it}} \right) = x_i \beta,
\]

\[
\mu_{it} = \mu_i (\beta) = E(Y_{it}) = P(Y_{it} = 1 | x_i, \beta) \text{ the probability of success at time } t,
\]

and \( \beta \) is a (Px1) vector of parameters.

Because we have binary responses, the logit link function is natural, but any link function can be used. The (Tx1) vector of marginal success probabilities is \( \mu_i = \mu_i (\beta) = (\mu_{i1}, ..., \mu_{iT})' \).
Note that the only assumptions made so far have been on the marginal success probabilities \( \mu_u(\beta) \).

If the responses for subject \( i \) are assumed to be independent, the joint distribution for subject \( i \) will be

\[
f(y_i | X_i) = \exp\left[ \sum_{r=1}^{T} y_u \theta_u - \sum_{r=1}^{T} \log \{1 + \exp(\theta_u)\} \right],
\]

which, along with results based on the exponential family, gives

\[
\frac{d l_i}{d \beta} = \left( \frac{d \mu_i}{d \beta} \right)' \text{Var}^{-1}(Y_i) (y_i - \mu_i)
\]

\[
= X_i'_i (y_i - \mu_i),
\]

where \( l_i = \ln \{ f(y_i | X_i) \} \). Summing over all \( N \) subjects gives \( \hat{\beta} \) to be the solution of

\[
\sum_{i=1}^{N} X_i'_i (y_i - \mu_i) = 0. \tag{4.0}
\]

We will consider two methods for obtaining estimates for our parameters based on the likelihood equations (4.0).

### 4.1.1 Generalized Estimating Equations

Liang and Zeger (1986), from here on notated L-Z for simplicity, develop moment-based generalized estimating equations (GEE) that give consistent estimates of the regression parameters, \( \beta \), as long as the model on the mean is correct. The GEE for \( \beta \) have the form...
\[ U(\beta) = \sum_{i=1}^{n} \left( \frac{d\mu_i}{d\beta} \right)' V_i^{-1} (y_i - \mu_i), \]

where \( V_i \) is the "working" covariance matrix of \( Y_i \). By re-writing \( V_i = \Delta_i^{-1/2} R_i(\alpha) \Delta_i^{-1/2} \), where \( \Delta_i = \text{diag} \{ \text{var}(Y_{it}) \} \), \( R_i(\alpha) = \text{Corr}(Y_i) \) is a TxT "working" correlation matrix, and \( \alpha \) is a vector of parameters for the model on \( \text{Corr}(Y_i) \), and combining with equation (4.0), the GEE for binary responses is

\[ U(\beta) = \sum_{i=1}^{n} X_i' \Delta_i V_i^{-1} (y_i - \mu_i) = 0. \quad (4.1) \]

By changing the specified models on \( R_i(\alpha) = \text{Corr}(Y_i) \), the association between responses can range from independence to a totally saturated model.

Estimation of \( \alpha \) is through a set of GEE similar to those for \( \beta \). These equations are of the form:

\[ U(\alpha) = \sum_{i=1}^{n} A_i' B_i^{-1} \{ r_i - \rho_i(\alpha) \} = 0, \quad (4.2) \]

where

\[ r_i \]

is the vector of empirical correlations, with \( r_{ut} = \frac{(Y_\mu - \mu_\mu)(Y_u - \mu_u)}{\sqrt{\mu_\mu (1-\mu_\mu) \mu_u(1-\mu_u)}} \),

\[ \rho_i = \rho_i(\alpha) \]

with \( \rho_{ut} = \text{E}(r_{ut}) = \text{Corr}(Y_u, Y_\mu) \),

\[ A_i = \frac{d\rho_i(\alpha)}{d\alpha} \],

and \( B_i \approx \text{Cov}(r_i) \) is a "working" covariance matrix for \( r_i \).

L-Z suggest using a modified Fisher scoring algorithm to obtain \( (\hat{\beta}, \hat{\alpha}) \) from equations (4.1) and (4.2). The estimate \( \hat{\beta} \) is consistent and asymptotically normal as long as the
model for the mean is correct, regardless of whether or not the "working" correlation is correct. If the association has been correctly specified in \( R_i(\alpha) \), the asymptotic variance of \( \hat{\beta} \) can be consistently estimated by

\[
\text{AVar}(\hat{\beta}) = \left( \sum_{i=1}^{N} X_i' \hat{\Lambda}_i^{-1} \hat{V}_i^{-1} \hat{\Lambda}_i X_i \right)^{-1},
\]

where \( \hat{\Lambda}_i \) is \( \Lambda_i \) evaluated at \( \hat{\beta} \), and \( \hat{V}_i \) is \( V_i \) evaluated at \( (\hat{\beta}, \hat{\alpha}) \). However, if the model on the association between responses has been misspecified in the "working" correlation matrix \( R_i(\alpha) \), \( \text{AVar}(\hat{\beta}) \) can be inconsistent. L-Z suggest using the "sandwich" estimate

\[
\{ \text{AVar}(\hat{\beta}) \} \{ \text{AV}(\hat{\beta}) \} \{ \text{AVar}(\hat{\beta}) \},
\]

where \( \text{AV}(\hat{\beta}) = \left( \sum_{i=1}^{N} X_i' \hat{\Lambda}_i \hat{V}_i^{-1} (y_i - \mu_i)(y_i - \mu_i)' \hat{V}_i^{-1} \hat{\Lambda}_i X_i \right) \),

which is robust; that is, consistent even if the "working" covariance \( V_i \) is not the true \( \text{Cov}(Y_i) \) (Liang and Zeger, 1986).

### 4.1.2 Likelihood-Based Regressions Models

The GEE approach to modeling longitudinal binary data is not likelihood-based, and so does not require the joint distribution to be completely specified. Likelihood-based approaches, on the other hand, do require the joint distribution of the longitudinal binary data to be completely specified.
For likelihood-based approaches to longitudinal binary data, a subject with responses at T times is assumed to have a multinomial joint distribution with a $2^T$ probability vector $m_i = \{ m_{ij_1, j_T} \}$ with mean $\mu_i = E(Y_i) = \sum_{j_s \neq t} m_{ij_1, j_T}$, $t=1, ..., T$.

Two parametric models for the joint distribution have been widely used: one in terms of marginal means and marginal correlations (Bahadur, 1961) and the other in terms of conditional logits and log odds-ratios (Cox, 1972; Bishop, Fienberg and Holland, 1975). We will consider the latter model, also known as the log-linear representation, below. We present the methodology in detail because we not only use it to fit models in Chapter 5, but also then expand on it in our work in Chapter 6.

### 4.1.2.1 Log-linear Representation

The log-linear representation, in terms of the conditional logits and log odds-ratios, assumes the joint distribution of $Y_i$ is

$$f(y_i, \Psi_i, \Omega_i) = \exp \{ \Psi_i' y_i + \Omega_i' w_i - A(\Psi_i, \Omega_i) \}, \quad (4.3)$$

where

$W_i = (Y_{i1}, Y_{i2}, ..., Y_{i(T-1)} Y_T, ..., Y_{i1} Y_{i2} Y_T)'$ is a $(Kx1)$ vector of two- and higher-way cross-products of $Y_i$, $\Psi_i = (\psi_{i1}, ..., \psi_{iT})'$ and $\Omega_i = (\omega_{i12}, ..., \omega_{i(T-1)T}, ..., \omega_{iT})'$ are vectors of canonical parameters,
and $A(Ψ_i, Ω_i)$ is the normalizing constant with

$$
\exp\{ A(Ψ_i, Ω_i) \} = \sum \exp \{ Ψ_i \, y_i + Ω_i \, w_i \},
$$

where the summation is over all $2^T$ possible values of $Y_i$. The parameters of $Ψ_i$ and $Ω_i$ can be interpreted in terms of their conditional probabilities and (contrasts of) log conditional odds-ratios, respectively:

$$
Ψ_i = \logit \{ P(Y_{it} = 1|Y_{is} = 0, t \neq s) \} \text{ for } s=1, ..., T
$$

and

$$
\exp(ω_{it}) = \frac{ P\{Y_{is} = 1, Y_{it} = 1|Y_{ir} = 0, v \neq s, t \} P\{Y_{is} = 0, Y_{it} = 0|Y_{ir} = 0, v \neq s, t \} }{ P\{Y_{is} = 1, Y_{it} = 0|Y_{ir} = 0, v \neq s, t \} P\{Y_{is} = 0, Y_{it} = 1|Y_{ir} = 0, v \neq s, t \} }
$$

for $s < t \{ 1, ..., T \}$, and so forth for the higher-order $ω_{it}$ terms. Depending on the model assumed for $Ω_i$, the relationships among the $Y_{it}$ can range from independence to the saturated model.

### 4.1.2.2 Parameterization via Marginal Moments

Zhao and Prentice (1990), hereafter referred to as Z-P, assume that the third- and higher-order association terms of $Ω_i$ are all zero, and are left with a "quadratic exponential" family that has

$$
W_i = (Y_{i1}, Y_{i2}, ..., Y_{i(t-1)}Y_{it})' \text{ and } Ω_i = (ω_{i1}, ..., ω_{i(t-1)})'.
$$
Z-P model the marginal probabilities of success, \( \mu_i = \mu_i(\beta) \), and the marginal covariances \( \sigma_i = (\sigma_{i12}, \ldots, \sigma_{i(t-1)t})' = \sigma_i(\beta, \alpha) \), as a function of covariates, and then make a 1-1 transformation from the canonical parameters \((\Phi_i, \Omega_i)\) to the marginal moments \((\mu_i, \sigma_i)\). Although the canonical parameters \((\Phi_i, \Omega_i)\) are complicated functions of \((\mu_i, \sigma_i)\), the resulting score equations for \(\beta\) and \(\alpha\) are surprisingly simple.

Define \(\gamma_i = \text{E}(W_i)\) and \(V_i = \text{Var}(Y_i)\). The log likelihood function for subject \(i\) is

\[
l_i = \Phi_i' y_i + \Omega_i' w_i - A(\Phi_i, \Omega_i).\]

Then properties of exponential families imply

\[
\frac{d\mu_i}{d \Phi_i} = \text{Cov}(Y_i),
\]

\[
\frac{d\gamma_i}{d \Omega_i} = \text{Cov}(W_i),
\]

\[
\left( \frac{d\mu_i}{d \Omega_i} \right)' = \frac{d\gamma_i}{d \Phi_i} = \text{Cov}(Y_i, W_i),
\]

\[
\frac{dl_i}{d \Phi_i} = y_i - \mu_i
\]

and

\[
\frac{dl_i}{d \Omega_i} = w_i - \gamma_i.
\]

Repeated application of the Chain Rule leads to the likelihood equations:

\[
\begin{pmatrix}
\frac{dl_i}{d \beta} \\
\frac{dl_i}{d \alpha}
\end{pmatrix} =
\begin{pmatrix}
\frac{d\mu_i}{d \beta} & \frac{d\gamma_i}{d \beta} \\
\frac{d\mu_i}{d \alpha} & \frac{d\gamma_i}{d \alpha}
\end{pmatrix}
\begin{pmatrix}
\frac{dl_i}{d \Phi_i} & \frac{dl_i}{d \Omega_i}
\end{pmatrix}
\]

\[
= 0,
\]

so that
\[
\begin{pmatrix}
\frac{d l_i}{d \beta} \\
\frac{d l_i}{d \alpha}
\end{pmatrix} = \begin{bmatrix}
\frac{d \mu_i}{d \beta} & \frac{d \gamma_i}{d \beta} \\
\frac{d \mu_i}{d \alpha} & \frac{d \gamma_i}{d \alpha}
\end{bmatrix} \begin{pmatrix}
\text{Cov}(Y_i) & \text{Cov}(Y_i, W_i) \\
\text{Cov}(Y_i, W_i)' & \text{Cov}(W_i)
\end{pmatrix}^{-1} \begin{pmatrix} y_i - \mu_i \\ w_i - \gamma_i \end{pmatrix} = 0. \tag{4.4}
\]

Z-P let \( S_i = (S_{i12}, \ldots, S_{i(t-1)T})' \), where \( S_{yk} = (y_y - \mu_y)(y_k - \mu_k) \) is the empirical pairwise covariance of \( Y_y \) and \( Y_k \). Noting that \( \sigma_{yk} = E(S_{yk}) \), we get the likelihood equation for subject \( i \) as:

\[
\begin{pmatrix}
\frac{d l_i}{d \beta} \\
\frac{d l_i}{d \alpha}
\end{pmatrix} = \begin{bmatrix}
\frac{d \mu_i}{d \beta} & \frac{d \gamma_i}{d \beta} \\
\frac{d \mu_i}{d \alpha} & \frac{d \gamma_i}{d \alpha}
\end{bmatrix} \begin{pmatrix}
\text{Var}(Y_i) & \text{Cov}(Y_i, S_i) \\
\text{Cov}(Y_i, S_i)' & \text{Var}(S_i)
\end{pmatrix}^{-1} \begin{pmatrix} y_i - \mu_i \\ s_i - \sigma_i \end{pmatrix} = 0.
\]

With \( N \) independent subjects in the longitudinal study, the likelihood equations over the \( N \) subjects is then

\[
\sum_{i=1}^{N} \begin{bmatrix}
\frac{d \mu_i}{d \beta} & \frac{d \gamma_i}{d \beta} \\
\frac{d \mu_i}{d \alpha} & \frac{d \gamma_i}{d \alpha}
\end{bmatrix} \begin{pmatrix} V_i & K_i \\ K_i' & U_i \end{pmatrix}^{-1} \begin{pmatrix} y_i - \mu_i \\ s_i - \sigma_i \end{pmatrix} = 0, \tag{4.5}
\]

where \( V_i = \text{Var}(Y_i) \), \( K_i = \text{Cov}(Y_i, S_i) \) and \( U_i = \text{Cov}(S_i) \).

Solving the equations in (4.5) for \( (\beta, \alpha) \) gives "pseudo-maximum likelihood estimates"; Gourieroux, Monfort and Trognon (1984) showed that when using a quadratic exponential model, the MLE's of the mean and variance parameters will be consistent and asymptotically normal under regularity conditions, regardless of whether the quadratic exponential model really fits. Gourieroux, Monfort and Trognon (1984) differentiated between MLE's obtained using score equations in this way, and those obtained using
sampling likelihood functions, by calling the former "pseudo-maximum likelihood
estimates". These pseudo-MLE's will be MLE's when the three- and higher-order
associations really are zero. \((\hat{\beta}, \hat{\alpha})\), the pseudo-MLE's, are consistent and asymptotically
normal under regularity conditions, even if parameters in \(\begin{pmatrix} V_i & K_i \\ K'_i & U_i \end{pmatrix}\) are estimated by
\(N^{1/2}\)-consistent estimators. However, consistency of the pseudo-MLE's depends on both
the mean and the pairwise correlation models being correctly specified. Therefore, in the
Z-P model, even if the mean is correctly specified, \(\beta\) may be asymptotically biased if the
model on the
pairwise marginal correlations is incorrect. However, there is a "robust" estimate of the
asymptotic covariance matrix, consistent even if the association model is misspecified.
This "robust" estimate is a "sandwich" estimator, like that used by Liang and Zeger
(1986), and has a similar form to the Liang and Zeger estimator.

Since calculating the pseudo-MLE's of \(\beta\) and \(\alpha\) involves computing the matrix
\[C_i = \begin{pmatrix} V_i & K_i \\ K'_i & U_i \end{pmatrix}^{-1},\]
which itself involves repeated summations over all \(2^T\) possible values
of \(y_i\) to compute third and fourth central moments, pseudo-ML estimation is quite
computationally intensive for data sets where \(T\) is even of a moderate size. Therefore, Z-P
suggests an alternative to pseudo-ML estimation by using "working" matrices for \(C_i\) in
which the third and fourth central moments are specified, usually as functions of the first
two moments. The estimating equations

65
with "working" covariance matrices $V_i$, $K_i$, and $U_i$ directly extend the GEE approach of Liang and Zeger (1986) by allowing for joint estimation of the mean and pairwise association parameters. The estimates of $\beta$ and $\alpha$ using the Z-P extension to GEE will be more efficient than those obtained using GEE, but again this is provided that the assumed models for both the mean and the pairwise associations are correctly specified (recall that, with GEE, the estimate of $\beta$ will be consistent as long as the model for the mean is correct). Since this extension to GEE requires correct assumptions about the third and fourth central moments of $y_i$ to obtain consistent estimates of $\beta$, it may be unattractive in studies where the focus is on the marginal response probabilities.

4.1.2.3 Parameterization via Conditional Logits and Log Odds-Ratios

Fitzmaurice and Laird (1993), noted as F-L, offer an alternative approach to that of Z-P. F-L also begin with expressing the joint distribution of $y_i$ in terms of the conditional logits and log odds-ratios in (4.3):

$$f(y_i, \Psi_i, \Omega_i) = \exp\left\{ \Psi_i' y_i + \Omega_i' W_i - A(\Psi_i, \Omega_i) \right\},$$

with $W_i$ containing all two- and higher-order products of the responses for subject $i$, and with $\Omega_i$ containing all two- and higher-order association parameters. Unlike Z-P, who set
all 3- and higher-order association parameters to zero, F-L allow for nonzero higher-order association parameters. Z-P transformed from the canonical parameters to the moment parameters. F-L retain the parameterization as it is in (4.3) and model the association between responses directly, in terms of the log odds-ratios in $Q_i$. This parameterization is more attractive than that of the Z-P model, since the log odds-ratios are not constrained by the marginal probabilities. This parameterization is referred to as a "mixed parameter model", because the model is parameterized in terms of a mix of mean parameters and canonical association parameters (Fitzmaurice, Laird and Rotnitsky, 1993).

F-L assume $\Omega_i$ can be expressed as a function of a $(Q \times 1)$ vector of parameters $\alpha = (\alpha_1, ..., \alpha_Q)'$ and use the linear link function $\Omega_i = Z \alpha$, where $Z_i$ is a $K \times Q$ design matrix. If we let $\Omega_i = Z \alpha$, then the association parameters do not depend on the covariates and the association structure is assumed to be identical for all subjects. Similar computations to those used previously to obtain the likelihood equations in Z-P give the following likelihood equations for the F-L model for subject $i$:

$$
\frac{d\Omega_i}{d\beta} = \left[ \frac{d\mu_i}{d\beta} \frac{d\gamma_i}{d\beta} \right] \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}^{-1} \begin{pmatrix} \gamma_i - \mu_i \\ \mu_i - \gamma_i \end{pmatrix} = 0, 
$$

(4.6)

where $\begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} = \begin{pmatrix} \text{Var}(Y_i) & \text{Cov}(Y_i, W_i) \\ \text{Cov}(Y_i, W_i)' & \text{Var}(W_i) \end{pmatrix}$. Note that these equations are of the exact same form as those used in the Z-P model. However, the matrices $V_{11}$, $V_{12}$, and $V_{22}$ are not "working" matrices in the F-L model (when Z-P generalize the GEE.
equations, they assume the variance/covariance matrices are “working” matrices; they are the covariance matrices for the responses under the assumed model on $\Omega_i$.

Now, recall for $i = 1, \ldots, N$ and $t = 1, \ldots, T$ we assumed

$$
\theta_u = \logit (\mu_u) = \log \left( \frac{\mu_u}{1 - \mu_u} \right) = x'_u \beta,
$$

which gives $\mu_u = \frac{\exp(x'_u \beta)}{1 + \exp(x'_u \beta)}$ and $\text{Var}(Y_u) = \frac{\exp(x'_u \beta)}{[1 + \exp(x'_u \beta)]^2}$. Then, with $\Delta_i = \text{diag}\{\text{Var}(Y_u)\}$ a TxT diagonal matrix with entries being the variances of the $Y_u$'s, it follows that

$$
\frac{d\mu_u}{d\beta} = x'_i \Delta_i,
$$

$$
\frac{d\mu_u}{d\alpha} = 0,
$$

$$
\frac{d\gamma_i}{d\beta} = x'_i \Delta_i \ V_{i11}^{-1} V_{i12},
$$

and

$$
\frac{d\gamma_i}{d\alpha} = Z'_i \left( V_{i22} - V_{i21} V_{i11}^{-1} V_{i12} \right),
$$

where $\gamma_i = E(W_i)$. It follows that the likelihood equations for subject $i$ are:

$$
\begin{pmatrix}
\frac{d l_i}{d\beta} \\
\frac{d l_i}{d\alpha}
\end{pmatrix} =
\begin{pmatrix}
x'_i & x'_i \Delta_i \ V_{i11}^{-1} V_{i12} \\
0 & Z'_i (V_{i22} - V_{i21} V_{i11}^{-1} V_{i12})
\end{pmatrix}
\begin{pmatrix}
V_{i11} & V_{i12} \\
V_{i21} & V_{i22}
\end{pmatrix}^{-1}
\begin{pmatrix}
\gamma_i - \mu_i \\
w_i - \gamma_i
\end{pmatrix}
$$

which after simplification become

$$
\begin{pmatrix}
\frac{d l_i}{d\beta} \\
\frac{d l_i}{d\alpha}
\end{pmatrix} =
\begin{pmatrix}
x'_i \Delta_i \ V_{i11}^{-1} & 0 \\
-Z'_i \gamma_i \ V_{i11}^{-1} & Z'_i
\end{pmatrix}
\begin{pmatrix}
\gamma_i - \mu_i \\
w_i - \gamma_i
\end{pmatrix} = 0.
$$

(4.8)
Assuming independent subjects, we therefore have the following likelihood equations for \((\beta, \alpha)\):

\[
\sum_{i=1}^{\infty} \begin{bmatrix}
X_i' \Delta_i V_{ii1}^{-1} & 0 \\
-Z_i' V_{i21} V_{ii1}^{-1} & Z_i'
\end{bmatrix} \begin{bmatrix} y_i - \mu_i \\ w_i - \gamma_i \end{bmatrix} = 0 ,
\]

which gives

\[
\sum_{i=1}^{\infty} X_i' \Delta_i V_{ii1}^{-1} (y_i - \mu_i) = 0
\]

and

\[
\sum_{i=1}^{\infty} Z_i' \{ w_i - \gamma_i - V_{i21} V_{ii1}^{-1} (y_i - \mu_i) \} = 0 .
\] (4.9)

The MLE's \((\hat{\beta}, \hat{\alpha})\) are the solutions to equations (4.9). The asymptotic variance/covariance matrix of \((\hat{\beta}, \hat{\alpha})\) is approximated by the inverse Fisher information matrix:

\[
\text{Cov}(\hat{\beta}, \hat{\alpha}) \approx \left( \sum_{i=1}^{\infty} E \left( \begin{bmatrix} \frac{d \beta}{d l} \\ \frac{d \beta}{d \alpha} \end{bmatrix} \begin{bmatrix} \frac{d \beta}{d l} \\ \frac{d \alpha}{d \alpha} \end{bmatrix} \right) \right)^{-1}
\]

\[
= \begin{pmatrix}
(\sum_{i=1}^{\infty} X_i' \Delta_i V_{ii1}^{-1} \Delta_i X_i)^{-1} & 0 \\
0 & (\sum_{i=1}^{\infty} Z_i' [V_{i22} - V_{i21} V_{ii1}^{-1} V_{i12}] Z_i)^{-1}
\end{pmatrix} .
\] (4.10)

The \((\beta, \alpha)\) elements of the inverse Fisher information matrix are zero, indicating that \(\beta\) and \(\alpha\) are orthogonal.
F-L suggest using the Fisher scoring algorithm to obtain \((\hat{\beta}, \hat{\alpha})\) by using the formulas:

\[
\hat{\beta}^{(j-1)} = \hat{\beta}^{(j)} + \left( \sum_{i=1}^{N} X_i' \Delta_i^{(j)} V_{i11}^{-1}(j) \Delta_i^{(j)} X_i' \right)^{-1} x\left( \sum_{i=1}^{N} X_i' \Delta_i^{(j)} V_{i11}^{-1}(j) (y_i - \mu_i^{(j)}) \right)
\]

and

\[
\hat{\alpha}^{(j-1)} = \hat{\alpha}^{(j)} + \left( \sum_{i=1}^{N} Z_i' \{ V_{i22}^{(j)} - V_{i21}^{(j)} V_{i11}^{-1}(j) V_{i12}^{(j)} \} Z_i \right)^{-1} x\left( \sum_{i=1}^{N} Z_i' \{ w_i - y_i^{(j)} - V_{i21}^{(j)} V_{i11}^{-1}(j) (y_i - \mu_i^{(j)}) \} \right).
\]  

Because, for any \(\mu_i\) and \(\Omega_i\), the matrices \(y_i, V_{i11}, V_{i21}\), and \(V_{i22}\) depend on two- and higher-order marginal probabilities, and because there is no closed-form expression for these probabilities in terms of \(\mu_i\) and \(\Omega_i\), F-L recommend using an iterative proportional fitting (IPF) algorithm for obtaining estimates of the two- and higher-way marginal probabilities for each step of the Fisher scoring algorithm.

F-L suggest the following IPF algorithm: Given \(\Omega_i\), construct some arbitrary \(2^r\) contingency table, \(S_i(\Omega_i)\), having the assumed conditional log odds-ratios. Then fit the first-order margins \(\mu_i(\beta)\) to \(S_i(\Omega_i)\) using classical IPF, or raking. The cells of \(S_i(\Omega_i)\) will be sequentially adjusted until they satisfy the margins \(\mu_i(\beta)\). The result of the IPF is a \(2^r\) contingency table with cell probabilities \(m_i(\Omega_i, \mu_i(\beta))\) with margins \(\mu_i(\beta)\) and
conditional log odds-ratios $\Omega_i$. Thus, obtaining the MLE's $(\hat{\beta}, \hat{\alpha})$ is a two-step iterative procedure:

step (1): Calculate updated $(\hat{\beta}^{(j-1)}, \hat{\alpha}^{(j-1)})$ by the Fisher scoring algorithm (4.11).

step (2): Use the IPF to fit $\mu_i(\hat{\beta}^{(j-1)})$ to $S_i\{\Omega_i(\hat{\alpha})\}$, and use the fitted cell probabilities $m_i(\Omega_i, \mu_i(\hat{\beta}))$ to obtain $\gamma^{(j-1)}, V_{11}^{(j-1)}, V_{12}^{(j-1)}$, and $V_{22}^{(j-1)}$. Return to step (1) and repeat until convergence.

When the covariates $x_{it}, t=1, ..., T$, are all categorical, this procedure will also give a set of estimated expected cell counts under the assumed model (Fitzmaurice and Laird, 1993). The goodness of fit of the model can be computed by comparing observed cell counts to those expected under the model. In addition, $\hat{\beta}$ will be a consistent estimate of $\beta$ as long as the mean structure is correctly specified; this is true regardless of whether the association parameters have been correctly specified in $\Omega_i$. If $\Omega_i$ is incorrectly specified, the inverse Fisher information matrix may be an inconsistent estimate of the asymptotic variance. Similar to both GEE and the Z-P models, a “robust” variance estimate — consistent even when $\Omega_i$ is misspecified— can be formed:

$$\{I_{(\beta)}\} \times \left(\sum_{i=1}^{N} X_i^\prime \Delta_i V_{11}^{-1} (y_i - \mu_i)(y_i - \mu_i) V_{11}^{-1} \Delta_i X_i\right) \times \{I_{(\beta)}\},$$

where

$$I_{(\beta)} = \left(\sum_{i=1}^{N} X_i^\prime \Delta_i V_{11}^{-1} \Delta_i X_i\right)^{-1}.$$

Fitzmaurice, Laird and Rotnitsky (1993) also point out that the orthogonality of the parameter space for $\beta$ and $\alpha$ indicates that information on the time-dependence...
association parameters has no asymptotic effect on $\beta$; the asymptotic variance for $\hat{\beta}$ remains unchanged, whether $\alpha$ is known or estimated (Newey, 1990). $\hat{\beta}$ will also remain asymptotically normal when the number of time-dependence parameters increases with the sample size at an appropriate rate (Fitzmaurice, Laird and Rotnitsky, 1993). A possible disadvantage of the approach taken by F-L is that, unlike marginal association parameters, the conditional log odds-ratios are not easily interpreted. But many longitudinal studies are mainly interested in the regression parameters for the marginal expectations, with the association parameters regarded as a nuisance characteristic of the data. In this case, interpreting the association parameters will not be of interest.

4.2 Regression Models for Incomplete Longitudinal Binary Data

Both Z-P and F-L consider data sets with complete records on all subjects. However, in many longitudinal studies, there is missing data. Fitzmaurice, Laird and Zahner (1996), notated as F-L-Z, discuss a likelihood-based regression model for analyzing incomplete longitudinal binary data based on extending the multivariate logistic regression model given by F-L. F-L-Z assume the missing data have a monotone pattern, so that if a subject is missing the $k^{th}$ response, he is also missing every response $k'>k$.

Using notation from Little and Rubin (1987), let $R_i$, a (Tx1) vector for subject $i$, be the response indicator; that is, for subject $i$ at time $t$, $R_{it}=1$ if $Y_{it}$ is observed, and $R_{it}=0$ if $Y_{it}$ is missing. Then, given $R_i$, we can partition $Y_i$ so that $Y_i=(Y_{it}, Y_{im})$, where
\( Y_{o} \) are the observed responses, and \( Y_{m} \) are the missing responses. \( X_{i}, \mu_{i}, V_{m1} \) and \( \Delta_{i} \) can also be similarly partitioned, so that \( X_{i} = (X_{o}, X_{m}) \), \( \mu_{i} = (\mu_{o}, \mu_{m}) \),
\( V_{m1} = (V_{o11}, V_{m11}) \), and \( \Delta_{i} = (\Delta_{o}, \Delta_{m}) \). The covariate matrix \( X_{i} \) is assumed to have no missing elements, so that, at a given time, while the response for a subject may be missing, the covariates for that subject are observed.

Define \( \lambda = (\beta, \alpha) \) for ease of notation and let \( \nu \) be the vector of parameters of the nonresponse model. The joint distribution of \( R_{i} \) and \( Y_{i} \) can be written as
\[
f(Y_{i}, R_{i} | \lambda, \nu) = f(Y_{i} | \lambda) f(R_{i} | Y_{i}, \nu),
\]
where \( f(Y_{i} | \lambda) = f(Y_{o}, Y_{m} | \lambda) \) is the joint distribution of \( Y_{o} \) and \( Y_{m} \), and \( f(R_{i} | Y_{i}, \nu) \) is the conditional distribution of the missing-data mechanism (Little and Rubin, 1987). The observed data for subject \( i \) is \( (Y_{o}, R_{i}) \). If the distribution of the missing-data mechanism does not depend on \( Y_{i} \), then
\[
f(R_{i} | Y_{i}, \nu) = f(R_{i} | \nu)
\]
and the missing data is missing-completely-at-random (MCAR).

If the distribution of the missing-data mechanism does not depend on the missing data \( Y_{m} \), then
\[
f(R_{i} | Y_{i}, \nu) = f(R_{i} | Y_{o}, \nu)
\]
and the missing data is missing-at-random (MAR). The missing-data mechanism is nonignorable if \( f(R_{i} | Y_{i}, \nu) \) depends on the missing data \( Y_{m} \).

The likelihood of \( \lambda = (\beta, \alpha) \), based on the observed data \( Y_{o} \), ignoring the missing-data mechanism, is any function \( L(\lambda | Y_{o}) \) of \( \lambda \) proportional to
\[
f(Y_{o} | \lambda) = \int f(Y_{o}, Y_{m} | \lambda) dY_{m}.
\]
If the data is either MCAR or MAR, then
\[
f(Y_{o}, R_{i} | \lambda, \nu) = f(Y_{o} | \lambda) f(R_{i} | Y_{o}, \nu),
\]

73
and if \( \lambda \) and \( \nu \) are distinct (the joint parameter space is the product of the two parameter spaces), then likelihood-based inferences for \( \lambda \) from \( L(\lambda, \nu | Y_{at}, R_i) \) will be the same as those for \( \lambda \) from \( L(\lambda | Y_{at}) \). That is, when the data is either MCAR or MAR, the missing-data mechanism is ignorable and likelihood-based inference can be based on \( Y_{at} \), instead of \( (Y_{at}, R_i) \). Valid likelihood-based inference, when the missing-data mechanism is nonignorable, entails specifying \( f ( R_i | Y_{at}, \nu) = f ( R_i | Y_{at}, Y_m, \nu) \).

Both GEE and likelihood-based approaches are popularly used when analyzing missing data. The two methods differ in their approaches for analyzing the incomplete data sets. GEE ignores the missing data and bases analyses only on the observed data. As long as the missing data is MCAR, \( \hat{\beta} \) and the "robust" variance estimate of \( \hat{\beta} \) will continue to be consistent under the GEE approach, as long as the model on the mean has been correctly specified. If the missing data is not MCAR, GEE estimators will be biased. However, if the model on the association parameters has been correctly specified, the MCAR assumption can be unnecessary.

### 4.2.1 A Likelihood-based Regression Model

F-L-Z (1996) use the following definitions and notation: for \( k=1, \ldots, T \) and \( i=1, \ldots, N \)

\[
S_i = \sum_{j=1}^{T} R_{ij} = \text{the number of observed responses for subject } i
\]

\[
\overline{R}_{at} = (R_{i1}, \ldots, R_{it})
\]
and \[ \tilde{Y}_{ik} = (Y_{ik1}, \ldots, Y_{ikk}) . \]

In words, \( \tilde{R}_{ik} \) is the response indicator from time 1 through time \( k \) and \( \tilde{Y}_{ik} \) is the vector of observations from time 1 through time \( k \). These two variables will be used in addition to the variable \( S_i \) to model the drop-out mechanism. Finally, assume

\[ P\{S_i = s | \tilde{Y}_{i1:k}, \nu \} = P\{S_i = s | \tilde{Y}_{i(t+1)}, \nu \} , \]

where

\[ P\{S_i = s | \tilde{Y}_{i(t+1)}, \nu \} = P\{ \tilde{R}_i = 1, \tilde{R}_{i(t+1)} = 0 | \tilde{Y}_{i(t+1)}, \nu \} ; \]

i.e., assume that the probability of a missing response depends on the value of that response and on the observed responses.

Diggle and Kenward (1994) developed a model for monotone missing data for continuous longitudinal data which can be used for nonignorable missing data and that have MCAR and MAR as specific sub-models. Mohlenbergs, Kenward and Lesaffre (1994) extend the Diggle and Kenward model to repeated ordinal data, parameterizing in terms of marginal odds-ratios. This approach can be generalized in terms of conditional logits and conditional odds-ratios.

Using this approach, \( f(Y_i | \lambda) = f(Y_{i1:k} | \beta, \alpha) \) is modeled as in F-L (1993), using a log-linear representation with conditional logits and log odds-ratios. For ML estimation of \( \nu \), F-L-Z consider a logistic regression model:

\[
\text{logit}(P\{ R_{ik} = 1 | \tilde{R}_{i(k-1)} = 1, \tilde{Y}_{ik} \}) = \nu_0 + \sum_{j=1}^{k} \nu_j Y_{i(k+1-j)}. \tag{4.12}
\]

Covariates can be included as additional predictors in the logistic regression model, just as the covariates \( x_{it} \) are included in the logistic regression model for \( \mu_{it} = \mu_{it} (\beta) \). For data that is MCAR, \( \nu_1 = \nu_2 = \ldots = \nu_f = 0 \) in equation (4.12); and for a MAR model, \( \nu_1 = 0 \).
For subject $i$, the incomplete-data log-likelihood is

$$\ln \{ f (y_{oi}, r_i | \lambda, \nu) \} = \ln \left( \sum_{y_{oi}} f (y_{oi} | \lambda) f (r_i | y_{oi}, \nu) \right).$$

Because the incomplete-data likelihood must be maximized with respect to both $\lambda$ and $\nu$ to obtain the MLE's $(\hat{\lambda}, \hat{\nu})$, F-L-Z suggest an EM-algorithm. The E-step entails computing $E \{ \ln [ f (y_i, r_i | \lambda, \nu) ] \}$, based on the current estimates $(\hat{\lambda}, \hat{\nu})$, and the M-step entails maximizing $\sum_{i=1}^{n} E \{ \ln [ f (y_i, r_i | \lambda, \nu) ] \}$ with respect to both $\lambda$ and $\nu$. The maximization in the M-step entails two separate maximizations, one for $\lambda$, and one for $\nu$.

The EM algorithm converges to the MLE's $(\hat{\lambda}, \hat{\nu})$, but does not give estimates of the asymptotic variance-covariance matrix of $(\hat{\lambda}, \hat{\nu})$. In addition, the matrix of second derivatives of the incomplete-data log-likelihood is extremely complicated. One alternative for estimating variances is to use an approximation to the observed information. If the model is correctly specified, the sample empirical covariance matrix of the individual scores is a consistent estimator of the Fisher information, regardless of the completeness of the data, and it entails computing only first derivatives (Fitzmaurice, Laird and Zahner, 1996).

Each of the parameters in any model must be statistically identifiable in order for the parameters to be estimated; estimation of the parameters requires that there be 'sufficient information available in the data' (Fitzmaurice, Laird and Zahner, 1996). Because all ignorable nonresponse models are identifiable, establishing the identifiability of parameters when there is nonresponse in the data is needed only when there is...
nonignorable nonresponse (Fitzmaurice, Laird and Zahner, 1996). In nonignorable nonresponse models, then, it has to be checked that there are no two pairs of parameters \((\lambda_1, \nu_1), (\lambda_2, \nu_2)\) with \((\lambda_1, \nu_1) \neq (\lambda_2, \nu_2)\) that give 
\[ f(y_{oi}, r_i | \lambda_1, \nu_1) = f(y_{oi}, r_i | \lambda_2, \nu_2) \]
for all \(y_{oi}, r_i\). This is not trivial, and thus far, there is no general and useful set of necessary and sufficient conditions for identifiability (Fitzmaurice, Laird and Zahner, 1996).

Because local identifiability is required for global identifiability, F-L-Z suggest examining the Fisher information matrix to determine if the model is locally identifiable. This is only a first step to checking global identifiability. Subject to regularity conditions, if the Fisher information is nonsingular, the model is locally identifiable; if the Fisher information is singular, the model is globally non-identifiable (Rothenberg, 1971). Thus, the Fisher information is used as both an estimate of the variance-covariance matrix of \((\hat{\lambda}, \hat{\nu})\) and to determine identifiability of the model. The Fisher information matrix can be obtained for any values of \((\lambda, \nu)\) relatively easily, since the data are binary. F-L-Z recommend obtaining the Fisher information matrix by generating an artificial sample with one observation for each possible value of \((Y_i, R_i, X_i)\), and then calculating the sample covariance matrix of the individual scores by weighting each individual’s contribution by its joint probability; that is, by taking the expectation of the outer-product of the score equations by summing all of its possible values weighted by their respective probabilities (Fitzmaurice, Laird and Zahner, 1996). The Fisher information obtained by this method is then checked to determine if it is nonsingular.
After determining that the model is locally identifiable, F-L-Z suggest generating another artificial sample with one observation for each possible value of \((Y, R, X)\), using the estimated values of \((\lambda, \nu)\), and then estimating new \((\hat{\lambda}, \hat{\nu})\) in the usual way, by taking the expectation of the likelihood equations by summing all of its possible values, but weighting their values by their respective probabilities. If the new estimate \((\hat{\lambda}, \hat{\nu})\) differs from \((\lambda, \nu)\), the model is not globally identifiable. F-L-Z state that if the new estimate \((\hat{\lambda}, \hat{\nu})\) equals \((\lambda, \nu)\) for an entire grid of reasonable \((\lambda, \nu)\), then the model is most likely identifiable.

An advantage to using the techniques and models developed by F-L-Z is that the sensitivity of the parameter estimates to hypotheses about the missing-data mechanism can be investigated, and various nonresponse assumptions can be examined. Another advantage is that, as with the F-L models, the conditional log odds-ratios used to parameterize the models are not constrained by marginal probabilities. However, one drawback of the F-L-Z model, as with many models dealing with missing data, is the inability to handle general patterns of missing data.

In Chapter 5, we use the methodology we detailed in this chapter to analyze the sedimentation rate data from the AFHS. We fit models to both complete sedimentation rate data and to drop-out sedimentation rate data. Then in Chapter 6, we extend the ideas discussed in this chapter to data that is “drop-in”, when the observations are initially missing but all responses from some point onward are observed.
CHAPTER 5

MODEL FITTING

5.1 Complete Data Models

The first models I fit to the AFHS data were those of Fitzmaurice and Laird (1993), abbreviated F-L. Because the F-L model requires complete data records for each of the subjects, the models were based on only those subjects who have sedimentation rate data from all four physical examinations. That is, the models were fit to only those men who have been fully compliant to all four cycles of the study, since the only way to get physical examination data, and thus sedimentation rate data from a study cycle, is to fully to comply to that cycle. There are a total of 1881 subjects who have sedimentation rate data from all four study cycles; in F-L notation, N=1881 and T=4.

In order to analyze the sedimentation rate data in terms of binary responses, I dichotomized the sedimentation rates as either normal or abnormal, for each study cycle,
as was done in the AFHS for their discrete analyses. For each subject \( i = 1, \ldots, 1881 \), I defined the responses to be

\[ Y_{it} = 0, \quad \text{if subject } i \text{ had a normal sedimentation rate at cycle } t \]

\[ 1, \quad \text{if subject } i \text{ had an abnormal sedimentation rate at cycle } t \]

for \( t = 1, 2, 3, 4 \). Recall from Chapter 2 Section 2.5 that for the cycle 1 Baseline examination, the "abnormal"/"normal" cutpoint was 12 mm/hr; a subject was "normal" if his sedimentation rate was \( \leq 12 \) mm/hr, and "abnormal" if his sedimentation rate was \( > 12 \) mm/hr. For cycles 2 and 3, held in 1985 and 1987, respectively, the cutpoint was 20 mm/hr. Finally, for cycle 4 held in 1992, the cutpoints for abnormal/normal sedimentation rates were age-related: subjects 50 years of age or younger had a cutpoint of 15 mm/hr, and subjects older than 50 years of age had a cutpoint of 20 mm/hr. I chose to use the same "abnormal"/"normal" cutpoints as did the AFHS because I am interested in determining if the study has lost any important information by including in a cycle's longitudinal analyses only those subjects who have data from the Baseline cycle and that specific cycle. Although the AFHS cutpoints may not be "optimal" in some sense, I chose to use them in order to compare results from my sedimentation rate analyses to those of the study.

For each subject \( i = 1, \ldots, 1881 \), \( Y_{i1}, Y_{i2}, Y_{i3}, \) and \( Y_{i4} \) were combined to give a 4x1 vector \( \mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})' \), indicating the "normal"/"abnormal" status for each of the four study cycles, for each subject.
5.1.1 Covariate Selection

There are numerous possible covariates to include in the model at each study cycle. Among the possible candidates for inclusion in the covariate vector at each cycle are group, military occupation, age, race, personality score, percent body fat, current dioxin body burden, time since Vietnam tour, concentration of herbicide during the tour, change in body fat since time of tour, initial dioxin, etc. I chose the following covariates to be included in the model for each subject at each study cycle: group, study year, and age at cycle 1 (1982 Baseline Examination). I included group as a covariate to act as the surrogate for herbicide exposure. It is the easiest, most basic way to categorize subjects into "exposed" and "unexposed" populations. In addition, the other two surrogates for herbicide exposure that have been used by the AFHS, the Military Records Index and dioxin body burden, both have drawbacks. The MRI is known to be inaccurate in detailing individual exposure. The dioxin assay for determining dioxin body burden was not developed until after the second study cycle; thus, there are no dioxin body burden measurements from the times of the 1982 and 1985 study cycles. The group membership variable for a subject was defined as 0 if he was a Comparison, and as 1 if he was a Ranch Hand. Note that this covariate was the same over all four study cycles for a subject.

The second covariate included in the covariate vector was the study year. This is a period effect indicating when, during the course of the AFHS, each cycle's observations were taken. Because the times between study cycles are unequally spaced intervals, simply using cycle number did not seem to be appropriate. Therefore, for each study...
cycle, I defined the study year covariate as the number of years since the AFHS began. Study year was, thus, defined as 0 for cycle 1, as 3 for cycle 2, as 5 for cycle 3, and as 10 for cycle 4. This covariate was the same for all subjects at any given study cycle.

The third covariate I used was the subject's age at the cycle 1 Baseline examination. The AFHS is not consistent with how they include "age" in their models. In cycle 1, the AFHS dichotomizes subjects into those 40 years of age or less at the cycle 1 examination, and those older than 40 years of age at the cycle 1 examination. For the cycle 2 analyses, the subjects were dichotomized into those born in 1942 or later, and those born before 1942. Since the Baseline cycle 1 examination was in 1982, it would appear that these two ways of dichotomizing are the same; but, they are not. Of those subjects who were born in 1942 and were fully compliant to all four study cycles (n=57), 28 were 40 years of age or less at the time of their cycle 1 examination and 29 were older than 40 years of age at their cycle 1 examination. Thus, the AFHS failed to give a clear way of how to include an age effect term in my covariate vector. In the 1982 Baseline, the 1985 first follow-up, and the 1987 second follow-up examinations, the cutpoints for the "abnormal"/"normal" sedimentation rate categories were the same, regardless of the subject's date of birth or his age at the cycle. At the fourth cycle, the 1992 third follow-up examination, the sedimentation rate cutpoints did depend on the subject's age at that study cycle. The cutpoint separated those 50 years of age or younger from those older than 50 years of age. This is an age-related cutpoint like that used in cycle 1, rather than a date-of-birth-related cutpoint like that used in cycle 2. Therefore, to be consistent with how the AFHS dichotomizes sedimentation rates in cycle 4, I chose to use age at cycle 1 as the
“age” variable. In addition, since I wanted to dichotomize the men into “younger” and “older” subjects, I used the number 40 as my cutpoint dichotomizing subjects exactly as the AFHS did in the cycle 1 analyses. Each subject’s age at cycle 1 was defined as 0 if he was 40 years of age or less at cycle 1, and as 1 if he was older than 40 at cycle 1. Again, like the group covariate, each subject’s age-at-cycle-1 covariate was the same across all four study cycles.

For my model, then, I included two time-stationary effects, group membership and age at cycle 1, and one time-varying effect, study year. For subject i, i=1, ..., 1881, the covariate vector $x_{it}$ at occasion $t$, $t=1, ..., 4$, was $x_{it} = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$, with

\[
\begin{align*}
x_{i1} &= 1 \quad \text{for all } i, \text{ for all } t \quad \text{(for overall mean effect)} \\
x_{i2} &= 0 \quad \text{if subject } i \text{ is a Comparison, for all } t \\
&\quad 1 \quad \text{if subject } i \text{ is a Ranch Hand, for all } t \\
x_{i3} &= 0 \quad \text{if } t=1 \text{ (Baseline cycle, 1982), for all } i \\
&\quad 3 \quad \text{if } t=2 \text{ (cycle 2, 1985), for all } i \\
&\quad 5 \quad \text{if } t=3 \text{ (cycle 3, 1987), for all } i \\
&\quad 10 \quad \text{if } t=4 \text{ (cycle 4, 1992), for all } i \\
x_{i4} &= 0 \quad \text{if subject } i \leq 40 \text{ years of age at cycle 1, for all } t \\
&\quad 1 \quad \text{if subject } i > 40 \text{ years of age at cycle 1, for all } t
\end{align*}
\]
The matrix of covariates for individual $i$ was defined to be $X_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})'$.

Because both group membership ($x_{i2}$) and age at cycle 1 ($x_{i4}$) were dichotomized, I ended up with four different possible combinations of covariates at each of the four study cycles. These four combinations at each cycle were:

<table>
<thead>
<tr>
<th>$x_{i2}$</th>
<th>$x_{i4}$</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>Comparison, ≤ 40 years of age at cycle 1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>Comparison, &gt; 40 years of age at cycle 1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>Ranch Hand, ≤ 40 years of age at cycle 1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Ranch Hand, &gt; 40 years of age at cycle 1</td>
</tr>
</tbody>
</table>

The cross-classification of sedimentation rate for subjects, over the four study cycles, for each of the combinations of covariates, is given in Table 5.1. Thus, for example, there were 332 subjects who were Comparisons less than or equal to 40 years old at cycle 1 who had normal sedimentation rates at all four study cycles.

### 5.1.2 Initial Model

Again, following the F-L paper, I assumed the marginal distribution of the sedimentation rate, $Y_u$, to be binary

$$f(y_u | x_u) = \exp(y_u \theta_u - \log(1 + \exp(\theta_u)))$$,
with
\[ \mu_u = P\{Y_{it} = 1|\mathbf{x}_i, \beta\}, \]

\[ \theta_{it} = \log \left( \frac{\mu_u}{1 - \mu_u} \right) \]

\[ = x_{it}' \beta = \beta_0 + \beta_1 x_{i12} + \beta_2 x_{i13} + \beta_3 x_{i14} \]

and

\[ \beta_0 = \text{overall mean effect} \]

\[ \beta_1 = \text{effect due to group membership} \]

\[ \beta_2 = \text{effect due to time period} \]

\[ \beta_3 = \text{effect due to age (at cycle 1)} \]

The joint distribution of \( Y_i \) for my longitudinal analysis was

\[ f(\mathbf{y}_i, \Psi_i, \Omega_i) = \exp(\Psi_i'y_i + \Omega_i'w_i - \Lambda(\Psi_i', \Omega_i)), \]

with

\[ \Psi_i = (\psi_{i1}, \psi_{i2}, \psi_{i3}, \psi_{i4})', \]

and

\[ \Omega_i = (\omega_{i12}, \omega_{i13}, \omega_{i14}, \omega_{i23}, \omega_{i24}, \omega_{i34}, \omega_{i123}, \omega_{i124}, \omega_{i134}, \omega_{i234}, \omega_{i1234})'. \]

To explore the dependence structure among the \( Y_{it} \), I fit numerous models to the data, including independence, first-order Markov chain, second-order Markov chain, all pairs, among others. In all, I fit 22 different association structures to the complete data for this specific set of covariates. As in F-L, in each of the models I assumed \( \Omega_i \) was of the form
12. \( \Omega = Z \alpha \), indicating that the association structure is assumed to be independent of the covariates and is the same for all subjects.

For starting values \( \hat{\beta}^{(0)} \) and \( \hat{\alpha}^{(0)} \), I used the following formulas:

\[
\hat{\beta}^{(0)} = \left( \sum_{i=1}^{16} X_i' \Delta_i^{(0)} V_{i11}^{-1(0)} \Delta_i^{(0)} X_i \right)^{-1} \times \left( \sum_{i=1}^{16} X_i' \Delta_i^{(0)} V_{i11}^{-1(0)} (y_i - \mu_i^{(0)}) \right)
\]

and

\[
\hat{\alpha}^{(0)} = \left( Z' \sum_{i=1}^{16} (V_{i22}^{(0)} - V_{i21}^{(0)} V_{i11}^{-1(0)} V_{i21}^{(0)}) Z \right)^{-1} \times \left( Z' \sum_{i=1}^{16} [w_i - y_i^{(0)} - V_{i21}^{(0)} V_{i11}^{-1(0)} (y_i - \mu_i^{(0)})] \right),
\]

where \( \Delta_i^{(0)} \), \( V_{i11}^{(0)} \), \( V_{i22}^{(0)} \), and \( V_{i11}^{(0)} \) were the variance/covariance matrices calculated from values observed in each of the four different tables. \( \mu_i^{(0)} \) and \( y_i^{(0)} \) were calculated based on observed values in the collapsed (over the four covariates) table, because the matrices \( X_i \), \( \Delta_i \), \( V_{i11} \), \( V_{i22} \), and \( V_{i11} \) depend on the subject’s covariates but not on his specific \( y_i \) and \( w_i \) values. If I had used only each covariate’s observed values to determine starting values for \( \mu_i \) and \( y_i \), the sums over \( (y_i - \mu_i^{(0)}) \) and over \( (w_i - y_i^{(0)}) \) would have been zero for each of the four covariates and, therefore, \( \hat{\beta}^{(0)} \) and \( \hat{\alpha}^{(0)} \) would also have been zero.

The iterative procedure described in Subsection 4.1.2.3 was used to fit the models to the data. In each of the models fit, I used the following convergence criteria: the IPF loop was considered to have converged when the sum of the absolute relative differences of expected cell counts of all 64 cells of the tables \( 2^4 = 16 \) cells for each of the 4
covariates) from two consecutive loops of the IPF was less then 0.01; the parameter estimates of $\beta$ and $\alpha$ were considered to have converged at step $J$ when the sum of the absolute relative differences between the four terms of $\beta^{(j-1)}$ and $\beta^{(j)}$ was less than 0.01 and the (sum of the) absolute relative difference between the term(s) of $\alpha^{(j-1)}$ and $\alpha^{(j)}$ was less than 0.01. I explored using different starting values and different convergence criteria for fitting the various models to the sedimentation rate data. However, changing the starting values and/or the convergence criteria had no effect on the results.

The computer programs for fitting the models were written in FORTRAN and were run on a UNIX-HP workstation. The programs converged to the MLE's quite rapidly; on average, it took 10 to 15 seconds for a program to run. However, the computer programs themselves were extremely programming intensive. In addition, changing from one association model to another or from one marginal model to another required intensive revision of the computer code, especially if the new model had a different number of parameters than the previous model.

Table 5.2 gives the parameter estimates, chi-squared statistics and degrees of freedom for some of the 22 models fit to the sedimentation rate data; all the $\beta$'s and $\alpha$'s in the table are estimates and the “hats” are not included for ease of notation. The standard errors given in the table are the square roots of the estimated variances which were computed using the formula in (4.10). The models included in Table 5.2 were selected because of their intuitive appeal or their fit. For example, one might expect a first- or second-order Markov chain to be a reasonable model for the association between sedimentation rate responses. I included in the table two models involving the (234) term
since the physical examinations from cycles 2, 3, and 4 were all conducted at the same location. The model (23,134) is in the table because cycle 1, 3, and 4 were equally spaced (five years apart), while cycle 3 was closest in time to cycle 2. In terms of the model fits, the All-Pair model (12,13,14,23,24,34) appeared to have the best fit out of all the association models tried, followed by model (23,34,124). The relatively large values for the $X^2$ and $G^2$ statistics appear to indicate that there is still some lack-of-fit present in the association models. However, we suspect that asymptotics may not hold in our situation due to the sparse cells in our contingency table. We are planning simulations in our future work to get an exact test for our data and to determine how accurate the asymptotic statistics are. Of the association models not listed in the table, none fit the sedimentation rate data any better than those in the table.

From Table 5.2, one can see that each of the models fit to the sedimentation rate data indicated an increase in sedimentation rate abnormalities over the course of the study as well as an increase in sedimentation rate abnormalities in those subjects who are older. These findings agree with what is known about sedimentation rates: a person’s sedimentation rate usually increases with age, and sedimentation rates are generally higher in older cohorts. In addition, each model indicates a very slight non-statistically significant increase in abnormal sedimentation rates in the Ranch Hand group. This would be consistent with the beliefs of those who say that occupational exposure to herbicides causes some sort of change in health. However, relative to the age and period effects, the effect due to being a member of the Ranch Hand, versus the Comparison, group is quite small.
5.1.3 Other Models

None of the 22 association structures appear to fit very well for this choice of $\beta$. Therefore, I explored different models for $\text{logit}(\mu_u)$, including adding the interaction of group and study year into the marginal model, and adding the interaction of group and age at cycle 1 into the marginal model. Table 5.3 gives a comparison of some of the different models for $\text{logit}(\mu_u)$ for the same association structures listed in Table 5.2. Again, the "hats" are left off the estimates for ease of notation. Because the $X^2$ and $G^2$ statistics are what we used in assessing the fit of the original marginal model, we will also use them here in order to make a comparison of the various models. For each association structure, none of these marginal models showed a significant improvement in fit over the original marginal model that included only the main effects of group, study year and age at cycle 1.

The association structures I fit to each of the various marginal models ranged from the independence model to models that allowed for a high degree of association among the variables. Because none of the models that included group, age at cycle 1, and study year as covariates in the marginal model had a particularly good fit, I considered including more and/or different covariates in the marginal model. One variable that has been shown in the AFHS to be significantly related to serum dioxin is military occupation. Among the Ranch Hands, ground crew have the highest levels of dioxin, followed by flight engineers, with officers having the lowest levels of dioxin. If exposure to TCDD-containing
herbicides were to have a negative effect on health, one would expect that a sensitive nonspecific indicator of overall general health, such as sedimentation rate, would indicate groundcrew members as having the most abnormalities and officers the least, among the Ranch Hands. I defined each subject's military occupation as officer (officer-pilot, officer-navigator, officer-nonflyer), flight engineer (flying airmen), or groundcrew (non-flying airmen). I then cross-classified sedimentation rate for subjects over the four study cycles, for each combination of group (Ranch Hand, Comparison), age (≤40 at cycle 1, > 40 at cycle 1), and military occupation. However, the cross-classification tables were quite sparse; each of the four tables in Table 5.1 was further broken down into the three military occupations. Therefore, I collapsed over the age-at-cycle 1 variable, in order to get tables that were less-sparse; the cross-classification of sedimentation rate for subjects over the four cycles was then over group and military occupation.

I fit various association models for the marginal model that included group, study year, military occupation and the interaction of group and military occupation as the covariates. However, none of the models were a significant improvement over the previous models. Therefore, I decided to use the initial marginal model, with group, age at cycle 1, and study year as covariates, as the one to which future models would be fit.

5.2 Drop-out Models

In the first models I fit, each subject had complete data, with sedimentation rate information from all four cycles. Because my exploratory data analyses indicated that
subjects fully compliant to one cycle are likely to fully comply at other cycles, and subjects who noncomply to one cycle are likely to noncomply at other cycles, it is possible that subjects with complete sedimentation rate data are a biased sample from the overall Ranch Hand and Comparison populations. This would be true if, for instance, noncompliant subjects were prevented from attending physical examinations because of health problems, while compliant subjects had no such health problems to interfere with their participation. The approach given by Fitzmaurice, Laird and Zahner (1996), abbreviated F-L-Z, allows the F-L model based only on subjects with complete sedimentation rate data, to be extended to subjects who have dropped out of the AFHS before its completion in 2002 and, hence, do not have a complete vector of sedimentation rates. The F-L-Z extension also allows for the investigation of assumptions about the missing-data mechanism for drop-outs.

Thus, I used the method of F-L-Z to extend the models I had fit on subjects with complete sedimentation rate data to include subjects who were fully compliant at the Baseline study cycle in 1982, but have since dropped out of the AFHS. There are a total of 103 men who were fully compliant to only the first cycle, the 1982 Baseline examination; 52 men who fully complied to only cycles 1 and 2, in 1982 and 1985, respectively; 139 men who were fully compliant to cycles 1, 2, and 3; and 1881 men who were fully compliant to all four study cycles. This gives a total of $N^\prime=2175$ subjects having either complete response vectors or response vectors with a monotone missing-data pattern.
For each subject \( i \), \( i = 1, \ldots, 2175 \), and for each \( t \), \( t = 1,2,3,4 \), I defined the covariate vector \( x_i^t \), for the model on \( \text{logit}(\mu_i) \), as I had previously. That is, for each subject \( i \) and at each time \( t \), \( x_i^t \) was a 4x1 vector with a 1 for the first element (for the overall mean), and whose second element indicated the subject’s group membership, third element indicated the study year of cycle \( t \), and fourth element indicated the subject’s age at cycle 1. Regardless of whether or not a subject was missing sedimentation rate data for a specific study cycle, his covariate vector for that cycle was always completely known.

For the \( N' = 2175 \) subjects, I then defined the missing-data indicator \( R_i = (R_{i1}, R_{i2}, R_{i3}, R_{i4}) \) where \( R_{it} = 1 \) if \( Y_{it} \) was observed, and \( R_{it} = 0 \) if \( Y_{it} \) was missing. Note that if \( R_{it} = 0 \) then \( R_{it'} = 0 \) for all \( t' > t \). I also partitioned \( Y_i = (Y_{i1}, Y_{i2}) \) where \( Y_{i1} \) were the observed elements of \( Y_i \), and \( Y_{i2} \) were the missing elements of \( Y_i \), for \( i = 1, \ldots, 2175 \). Again, as before, elements of \( Y_{i1} \) corresponding to “abnormal” sedimentation rates were represented by a 1, and elements corresponding to “normal” sedimentation rates were represented by a 0. Table 5.4 gives the cross-classification of \( Y_{i1} \), over the four study cycles, for each of the combinations of the group and age covariates.

Finally, as in F-L-Z, I defined \( S_i = \sum_{t=1}^{4} R_{it} \), the number of observed responses for subject \( i \), and let \( \overline{R}_i = (R_{i1}, \ldots, R_{it}) \) and \( \overline{Y}_i = (Y_{i1}, \ldots, Y_{it}) \), for \( i = 1, \ldots, 2175 \) and for \( t = 1,2,3,4 \). I continue to use the models for \( f(y_i | \beta, \alpha) \) proposed by Fitzmaurice and
Laird (1993). Thus, I needed only to consider models for \( f(r_i | y_i, \nu) \), where \( r_i \) is the observed value of \( R_i \) and \( \nu \) is the vector of parameters for the model on \( R \).

I followed F-L-Z, considering models for \( f(r_i | y_i, \nu) \) that were logistic regression models. However, unlike F-L-Z, I considered models of the form

\[
\text{logit}(P(R_i = 1 | \overline{R}_{(t-1)} = 1, \nu, d_u)) = \nu_u' d_u,
\]

where \( d_u \) was a vector of completely observed covariates. There were no observations \( Y_j \), for \( j \leq t \) included in \( d_u \) nor any unobserved covariates in \( d_u \). F-L-Z point out that ensuring the identifiability of parameters in nonignorable nonresponse models is not an easy task. However, since all of the models I fit to the nonresponse mechanism depended only on completely observed covariates, the models were missing-at-random. Since MAR models are ignorable, I did not have to deal with ensuring the identifiability of parameters in my nonresponse models. In addition, since all of the drop-out models were ignorable models, the maximizations for \( \lambda = (\beta, \alpha) \) and for \( \nu \) are completely separate. That is,

\[
f(y_i, r_i | \lambda, \nu) = f(y_i | \lambda) f(r_i | y_i, \nu) = f(y_i | \lambda) f(r_i | d_i, \nu)
\]

The maximization for \( \lambda = (\beta, \alpha) \) was done via the E-M algorithm, and the maximization for \( \nu \) was a simple logistic regression, completely separate from the maximization of \( \lambda \).

For maximizing \( \lambda = (\beta, \alpha) \) using the E-M algorithm, the E-step computed

\[
E_{y_i \nu_i \lambda_i} \left\{ \frac{d \ln f(y_i | \lambda)}{d \lambda} \right\} \text{ given the current value of } \hat{\lambda} \text{ and the M-step maximized}
\]
\[
\sum_{i=1}^{n} E_{x_i \mid y_i} \left\{ \frac{d \ln f(y_i \mid \lambda)}{d \lambda} \right\}
\]
with respect to \( \lambda \). In words, the E-step gave us "complete" data \( e(y_i) = E_{x_i \mid y_i} (Y_i) \) and \( e(w_i) = E_{x_i \mid y_i} (W_i) \) based on the current \( \hat{\lambda} \). The M-step then maximized the likelihood with respect to \( \hat{\lambda} \) using this "complete" data.

I first fit various logistic regression models to \( \logit(P(Y_{it} = 1 \mid \lambda_{it} = 1, \nu, d_{it})) \), in order to determine what variables might be useful in modeling the missing-data mechanism. Among the variables I included in the covariate vector \( d_{it} \) were group, age at cycle 1, study year, and military occupation. As with the marginal models fit to the \( Y_{it} \), the first element of each covariate vector was a 1 (for the overall mean), and the variables group, age at cycle 1, and study year were defined as in the \( x_{it} \) (e.g. the group variable was dichotomized as 1 for Ranch Hands and as 0 for Comparisons, etc.). Table 5.5 shows the results of some of the various logistic regressions models fit; the "hats" are left off the estimates for ease of notation.

In all of the models fit to the drop-out mechanism, the group variable was nonsignificant, indicating that group membership (Ranch Hand, Comparison) has no impact on a subject dropping out of, or remaining in, the study. This means that group membership does not appear to be a significant factor in determining whether or not subjects remain fully compliant to the AFHS. Since sedimentation rate can be obtained only for those subjects who are fully compliant to the study, group membership is not an important variable in whether a subject has sedimentation rate data over the course of the study.
The variables study year and military occupation, however, appeared to be important terms in all the logistic regression models, and the coefficients for both variables had negative estimates. Because the drop-out model is of the form

\[
P \{ R_u = 1 | R_{(u-1)} = 1, ν, d_u \} = \frac{\exp \{ d_u' ν \}}{1 + \exp \{ d_u' ν \}},
\]

a negative estimate for the study year variable indicates that the longer the study has been underway, the more likely a subject is to drop out. This is not a surprising finding, since men are dying, losing interest in the study, becoming older and possibly more infirm, and becoming unlocateable, among other things. By collapsing the data in Table 3.3 from Chapter 3 over group and then determining the proportions of eligible subjects who have physical examination data for each study cycle, one can see that the percentage of subjects with physical examination, and thus sedimentation rate, data decreases over the course of the AFHS. Hence, the results of the logistic regression analyses for the drop-out process agree with previous findings.

The variable military occupation was included in various logistic regression models for the drop-out mechanism, and always appeared to be important in the model. The ordinal military occupation variable was defined as 1 for subjects who were officers, as 2 for subjects who were flight engineers, and as 3 for groundcrew; this same sort of definition is used in the AFHS to classify the different military occupations. A negative estimate for the variable indicates that groundcrew members are the most likely, and officers the least likely, to drop out between two consecutive study cycles. There are two things of note here. The first is that military occupation appears to be an important factor
in a subject dropping out, while group membership does not. The second is that, within the Ranch Hand group, the groundcrew members had the most opportunity for exposure to the herbicides, and officers the least such opportunity. If it is true that herbicides have a negative health consequence and if this negative consequence is the reason for a subject dropping out of the study, it would not be surprising to see the groundcrew members dropping out at the highest rate, followed by the flight engineers, and finally the officers. However, an additional explanation for this finding of higher drop-out among the groundcrew is education level. Groundcrew members in general were less-educated than officers and less-educated people typically have higher nonresponse rates. I do need to note that adding group and the interaction of group and military occupation to the model with study year and military occupation as explanatory variables did not improve the fit of the model. If herbicide exposure did lead to negative health consequences that in turn led subjects to drop out of the study, we would expect to find Ranch Hands dropping out at a much higher rate than the Comparisons and, among the Ranch Hands, the groundcrew dropping out at the highest rate and the officers at the lowest rate.

By the above definition of military occupation, I imposed a spacing between the different occupation levels that may not be accurate. Therefore, I also fit models to the drop-out process that had military occupation defined via indicator variables. That is, I defined the first military occupation indicator variable as 1 for officers and as 0 otherwise, the second military occupation indicator variable as 1 for flight engineers and as 0 otherwise, and the third military occupation indicator variable as 1 for groundcrew and as 0 otherwise. I then added the two of the three (i.e.: the officer and the flight engineer)
indicator variables to the drop-out model. The results of the indicator variable models paralleled those of the previous models that included military occupation defined on the ordinal scale. The indicator variables always appeared to be important terms in the drop-out model and the coefficients for the indicator variables indicated that officers have the lowest drop-out rate, followed by the flight engineers, with the groundcrew having the highest drop-out rate. In addition, adding or deleting the group variable from models barely changed the fit.

The models that defined military occupation via indicator variables had fits that were comparable to the models that defined military occupation through an ordinal scale, with nearly exact $X^2$ values and only one degree of freedom difference in the two models. Because the ordinal definition imposes an ordering and a spacing between the different occupations that may not be accurate and because the results of the model fitting for the two definitions were basically the same, we preferred the indicator variable models when considering fitting the full model to the complete and drop-out data.

Whenever interaction terms were added into a drop-out model (e.g. when the military occupation-year interaction was added to the model with military occupation and year as explanatory variables), previously important variables became unimportant in the new model and the $X^2$ statistic did not improve.

One area of interest to us was that of the model of random missingness; that is, $\logit (P\{R_{it} = 1| \bar{R}_{it-1} = 1, \nu, d_{it}\}) = \nu_0$. The random missingness model is comparable to ignoring the missing-data process. We computed various $G^2$ statistics and partitioned the statistics from nested models in order to determine if our MAR models were better than
the random missingness model. When we added the study year term to the random missingness model, the difference in $G^2$ was 19.32 with a loss of only one degree of freedom. This indicated that adding study year to the drop-out model is worthwhile. In addition, partitioning various other models indicated that adding the military occupation variable (either ordinal or indicator) and the age variable to the drop-out model were also worthwhile. The results of partitioning the $G^2$ statistic indicated to us that random missingness is not a reasonable assumption for the missing-data mechanism and, therefore, that we should not ignore the missing-data process.

In general, no one drop-out model had a markedly better fit than any of the other models. Although I found variables that appear to be important in the drop-out process, there is still some unexplained variation as to why subjects drop out of the study. This may partly be explained by the large sample size: there were 2175 $R_{12}$’s, 2072 $R_{13}$’s, and 2020 $R_{14}$’s (along with their respective covariate vectors) contributing information into the drop-out model. This gives a total of 6267 $R_i$’s in the drop-out model fitting.

After fitting models to determine what factors appear important in the drop-out process, I used some of these variables to fit the full model for both complete and drop-out data. We are interested in using variables that appear to be related to drop-out in our full model for sedimentation rate because factors related to drop-out may also be related to sedimentation rate (in particular, abnormal sedimentation rate). Table 5.6 gives results of the full model for the complete and drop-out sedimentation rate data for drop-out Models 4 and 8 from Table 5.5. The “hats” again are left off the parameters for ease of notation. Drop-out Model 4 has group, study year and age at cycle 1 as variables, while
Model 8 uses study year, age at cycle 1 and the indicator variables. The marginal model for $\beta$ is that which has group, study year and age at cycle 1 as covariates and the association models are the same nine association models shown in Tables 5.2 and 5.3.

When determining expected cell counts in the full model, drop-out models containing variables that are covariates in the marginal model for $\beta$ (i.e., group or study year) are much more attractive computationally than drop-out models that include variables not in the marginal model for $\beta$ (i.e., military occupation). Because the results for the complete data based on our maximization of $(\beta, \alpha)$ are already classified by the covariates in the $\beta$ model, drop-out models that only include (some or all of) these same variables are easily combined with the $(\beta, \alpha)$ models to give us the expected cell counts in the full model for both complete and drop-out data. When the drop-out model includes variables not in the model for $\beta$, the computations for the expected cell counts for the full model become much more complicated.

Again, as with the complete data models, we suspect that asymptotics of the $X^2$ and $G^2$ statistics are not wholly appropriate here since we have sparse tables; our future work will also include simulations in order to determine an exact test. However, we will use the asymptotic statistics here in order to compare the full models to the complete data models.

In contrasting results of the full models in Table 5.6 with results of the complete data models in Table 5.2, we see that the parameter estimates and standard errors are quite similar. Study year and age at cycle 1 again appear to be important in the full model,
just as they were in the complete data models. In addition, group still does not seem to have an impact on sedimentation rate. The positive estimates for study year and age at cycle 1 in the full models indicate, as they did in the complete data models, that the probability of an abnormal sedimentation rate has increased over the course of the study and that older subjects have a higher probability of an abnormal sedimentation rate than do younger subjects.

In terms of the association models, the results show the full models to again be comparable to the models fit to the complete data in terms of the parameter estimates and their standard errors. In addition, the All-Pair Model (12,13,14,23,24,34) appears to fit the full data best, just as when we worked with only the complete data. In terms of the fit of the full models, there was very little difference among the numerous drop-out mechanisms tried. This is probably due to the fact that no one drop-out model stood out from any of the others when we were modeling the drop-out process. Unfortunately, it appears that additionally accounting for subjects who drop out of the study does not help us to more accurately determine what factors impact the chances of an abnormal sedimentation rate, at least for the models I worked with.
$$ (x_{a2}, x_{a4}) = (0, 0) : $$

<table>
<thead>
<tr>
<th>Cycle 1</th>
<th>Cycle 2</th>
<th>Cycle 3</th>
<th>Cycle 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>332 54</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1 5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1 3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0 0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4 5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>2 2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0 0</td>
</tr>
</tbody>
</table>

$$ (x_{a2}, x_{a4}) = (0, 1) : $$

<table>
<thead>
<tr>
<th>Cycle 1</th>
<th>Cycle 2</th>
<th>Cycle 3</th>
<th>Cycle 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>476 72</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>9</td>
<td>7 7</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>5</td>
<td>6 6</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>6 6</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>12 8</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1 1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4 4</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>8 8</td>
</tr>
</tbody>
</table>

$$ (x_{a2}, x_{a4}) = (1, 0) : $$

<table>
<thead>
<tr>
<th>Cycle 1</th>
<th>Cycle 2</th>
<th>Cycle 3</th>
<th>Cycle 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>276 45</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>4 4</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2 2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>3 3</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0 0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1 1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0 0</td>
</tr>
</tbody>
</table>

* $(x_{a2}, x_{a4}) = \text{(group of subject } i, \text{ age at cycle 1 of subject } i)$

(continued)

Table 5.1: Erythrocyte Sedimentation Rate by Group and Age at cycle 1 for Subjects with Complete Data
Table 5.1 (continued)

\[(x_{u2}, x_{t4}) = (1,1)\] *

<table>
<thead>
<tr>
<th>Cycle 1</th>
<th>Cycle 2</th>
<th>Cycle 3</th>
<th>Cycle 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>403</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>60</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>

* \[(x_{u2}, x_{u4}) = (\text{group of subject i, age at cycle 1 of subject i})\]
<table>
<thead>
<tr>
<th>Model</th>
<th>Marginal Parameters</th>
<th>Dependence Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, Independence</td>
<td>$\beta_0 = -3.9035$ (.1222)</td>
<td>$G^2 = 571.19$</td>
</tr>
<tr>
<td></td>
<td>$\beta_1 = 0.0609$ (.0915)</td>
<td>$X^2 = 13,327.00$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.1989$ (.0130)</td>
<td>d.f. = 56</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.4089$ (.0998)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\text{Indepen­}dence y_g, = 0.0609$ (.0915)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$J T' = 13,327.01$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\gamma^2 = 0.1989$ (.0130)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\gamma^3 = 0.4089$ (.0998)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d.f. = 56$</td>
<td></td>
</tr>
<tr>
<td>2, 1st-order Markov Chain</td>
<td>$\beta_0 = -3.8626$ (.1409)</td>
<td>$G^2 = 197.91$</td>
</tr>
<tr>
<td></td>
<td>$\beta_1 = 0.0138$ (.1085)</td>
<td>$X^2 = 316.390$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.2151$ (.0121)</td>
<td>d.f. = 53</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.3066$ (.1132)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{12} = 2.6990$ (.3252)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{23} = 3.5286$ (.2478)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{34} = 2.8723$ (.2108)</td>
<td></td>
</tr>
<tr>
<td>3, 2nd order Markov Chain</td>
<td>$\beta_0 = -3.8392$ (.1437)</td>
<td>$G^2 = 121.97$</td>
</tr>
<tr>
<td></td>
<td>$\beta_1 = 0.0203$ (.1133)</td>
<td>$X^2 = 189.55$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.2131$ (.0116)</td>
<td>d.f. = 49</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.2649$ (.1177)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{12} = 2.1533$ (.6087)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{13} = 2.2447$ (.4276)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{23} = 2.0488$ (.6365)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{34} = 1.9288$ (.3439)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{23} = 2.2823$ (.2409)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{34} = 0.3842$ (.7042)</td>
<td></td>
</tr>
<tr>
<td>4, (12,13,23, 24,34)</td>
<td>$\beta_0 = -3.8470$ (.1438)</td>
<td>$G^2 = 124.28$</td>
</tr>
<tr>
<td></td>
<td>$\beta_1 = 0.0223$ (.1132)</td>
<td>$X^1 = 198.28$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.2136$ (.0117)</td>
<td>d.f. = 51</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.2667$ (.1177)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{12} = 1.6104$ (.4137)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{13} = 1.9362$ (.3908)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{23} = 2.1584$ (.2859)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{34} = 1.9929$ (.2900)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{24} = 2.3302$ (.2235)</td>
<td></td>
</tr>
<tr>
<td>5, All-Pair</td>
<td>$\beta_0 = -3.8434$ (.1443)</td>
<td>$G^2 = 92.88$</td>
</tr>
<tr>
<td></td>
<td>$\beta_1 = 0.0522$ (.1152)</td>
<td>$X^1 = 103.50$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.2095$ (.0113)</td>
<td>d.f. = 50</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.2512$ (.1197)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{12} = 1.1115$ (.4135)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{13} = 1.2591$ (.4057)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{14} = 1.7338$ (.3838)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{23} = 2.3509$ (.2844)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{24} = 1.8290$ (.2917)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{34} = 2.1060$ (.2259)</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2: Parameter Estimates, $G^2$ and $X^2$ Statistics and Degrees of Freedom for Sedimentation Rate Models Fit to Complete Data (Standard errors in parentheses)
| Model         | Marginal Parameters | Dependence Parameters |  
|--------------|---------------------|-----------------------|---
| 6            | $\beta_0 = -3.8076$ (.1397) | $\alpha_{13} = 2.8197$ (.6862) | $G^2 = 144.87$
| (23,134)     | $\beta_1 = 0.0559$ (.1118) | $\alpha_{14} = 1.9703$ (.4019) | $X^2 = 275.54$
|              | $\beta_2 = 0.2037$ (.0112) | $\alpha_{23} = 3.5408$ (.2456) | d.f. = 51
|              | $\beta_3 = 0.2860$ (.1166) | $\alpha_{34} = 2.5679$ (.2157) | 
|              | $\gamma_{10} = 0.2037$ (.0112) | $\alpha_{13} = -0.5659$ (.7822) | 
| 7            | $\beta_0 = -3.8521$ (.1425) | $\alpha_{12} = 1.7182$ (.3613) | $G^2 = 105.63$
| (12,14,234)  | $\beta_1 = 0.0424$ (.1148) | $\alpha_{14} = 1.9760$ (.3568) | $X^2 = 132.59$
|              | $\beta_2 = 0.2114$ (.0111) | $\alpha_{23} = 2.3168$ (.5808) | d.f. = 50
|              | $\beta_3 = 0.2595$ (.1194) | $\alpha_{34} = 1.6249$ (.3513) | 
|              | $\gamma_{10} = 0.2037$ (.0112) | $\alpha_{23} = 0.2914$ (.6670) | 
| 8            | $\beta_0 = -3.8164$ (.1414) | $\alpha_{13} = 1.7239$ (.3463) | $G^2 = 101.78$
| (13,14,234)  | $\beta_1 = 0.0478$ (.1150) | $\alpha_{14} = 1.8756$ (.3626) | $X^2 = 141.99$
|              | $\beta_2 = 0.2066$ (.0109) | $\alpha_{23} = 2.2938$ (.5704) | d.f. = 50
|              | $\beta_3 = 0.2566$ (.1196) | $\alpha_{34} = 1.9265$ (.2445) | 
|              | $\gamma_{10} = 0.2037$ (.0112) | $\alpha_{23} = 0.3509$ (.6565) | 
| 9            | $\beta_0 = -3.8510$ (.1425) | $\alpha_{12} = 2.0247$ (.9246) | $G^2 = 105.60$
| (23,34,124)  | $\beta_1 = 0.0416$ (.1148) | $\alpha_{14} = 2.0197$ (.3797) | $X^2 = 125.99$
|              | $\beta_2 = 0.2113$ (.0111) | $\alpha_{23} = 2.5602$ (.2800) | d.f. = 50
|              | $\beta_3 = 0.2596$ (.1194) | $\alpha_{34} = 1.7148$ (.3055) | 
|              | $\gamma_{10} = 0.2037$ (.0112) | $\alpha_{34} = 2.3133$ (.2250) | 
|              |                           | $\alpha_{12} = -0.3565$ (.10016) | 

104
<table>
<thead>
<tr>
<th>Model</th>
<th>( G^2 )</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>571.19</td>
<td>571.46</td>
<td>571.24</td>
<td></td>
</tr>
<tr>
<td>Independence</td>
<td>( X^2 = 13327.00 )</td>
<td>13077.67</td>
<td>13319.41</td>
<td></td>
</tr>
<tr>
<td>d.f. = 55</td>
<td>55</td>
<td>55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>197.91</td>
<td>197.87</td>
<td>197.83</td>
<td></td>
</tr>
<tr>
<td>1st-order Markov Chain</td>
<td>( X^2 = 316.39 )</td>
<td>315.91</td>
<td>315.88</td>
<td></td>
</tr>
<tr>
<td>d.f. = 52</td>
<td>52</td>
<td>52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>121.97</td>
<td>121.93</td>
<td>121.95</td>
<td></td>
</tr>
<tr>
<td>2nd-order Markov Chain</td>
<td>( X^2 = 189.55 )</td>
<td>189.18</td>
<td>189.02</td>
<td></td>
</tr>
<tr>
<td>d.f. = 48</td>
<td>48</td>
<td>48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>124.28</td>
<td>124.24</td>
<td>124.26</td>
<td></td>
</tr>
<tr>
<td>(12,13,23,24,34)</td>
<td>( X^2 = 198.28 )</td>
<td>198.09</td>
<td>197.80</td>
<td></td>
</tr>
<tr>
<td>d.f. = 50</td>
<td>50</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>92.88</td>
<td>92.86</td>
<td>92.76</td>
<td></td>
</tr>
<tr>
<td>All Pair</td>
<td>( X^2 = 103.50 )</td>
<td>103.50</td>
<td>102.67</td>
<td></td>
</tr>
<tr>
<td>d.f. = 49</td>
<td>49</td>
<td>49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>144.87</td>
<td>144.87</td>
<td>144.49</td>
<td></td>
</tr>
<tr>
<td>(23,134)</td>
<td>( X^2 = 275.54 )</td>
<td>275.34</td>
<td>264.83</td>
<td></td>
</tr>
<tr>
<td>d.f. = 50</td>
<td>50</td>
<td>50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fits of different models on \( \beta \) where \( \mu_{it} = P\{Y_{it} = 1|x_{it}, \beta\} \)

a) \( \logit (\mu_{it}) = \beta_0 + \beta_1 (gp_i) + \beta_2 (studyr_i) + \beta_3 (age@cyl_i) \)

b) \( \logit (\mu_{it}) = \beta_0 + \beta_1 (gp_i) + \beta_2 (studyr_i) + \beta_3 (age@cyl_i) + \beta_4 (gp_i * age_i) \)

c) \( \logit (\mu_{it}) = \beta_0 + \beta_1 (gp_i) + \beta_2 (studyr_i) + \beta_3 (age@cyl_i) + \beta_4 (gp_i * studyr_i) \)

(continued)

Table 5.3: Parameter Estimates, \( G^2 \) and \( X^2 \) Statistics and Degrees of Freedom for Different Covariate Structures Fit to Complete Data
Table 5.3 (continued)

<table>
<thead>
<tr>
<th></th>
<th>( G^2 )</th>
<th>( X^2 )</th>
<th>( d.f. )</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>105.63</td>
<td>105.61</td>
<td>105.52</td>
</tr>
<tr>
<td></td>
<td>132.59</td>
<td>132.33</td>
<td>132.41</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>8</td>
<td>101.78</td>
<td>101.76</td>
<td>101.62</td>
</tr>
<tr>
<td></td>
<td>141.99</td>
<td>140.67</td>
<td>140.46</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>9</td>
<td>105.60</td>
<td>105.58</td>
<td>105.48</td>
</tr>
<tr>
<td></td>
<td>125.99</td>
<td>125.82</td>
<td>125.50</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>49</td>
<td>49</td>
</tr>
</tbody>
</table>

Fits of different models on \( \beta \) where \( \mu_u = P(Y_u|X_u, \beta) \)

a) \( \text{logit} (\mu_u) = \beta_0 + \beta_1 (gp_i) + \beta_2 (studyyr_r) + \beta_3 (age@cy1_r) \)
b) \( \text{logit} (\mu_u) = \beta_0 + \beta_1 (gp_i) + \beta_2 (studyyr_r) + \beta_3 (age@cy1_r) + \beta_4 (gp_i * age_r) \)
c) \( \text{logit} (\mu_u) = \beta_0 + \beta_1 (gp_i) + \beta_2 (studyyr_r) + \beta_3 (age@cy1_r) + \beta_4 (gp_i * studyyr_r) \)
\((x_{a2}, x_{a4}) = (0,0)\) *

Subjects with:

<table>
<thead>
<tr>
<th>(Y_1)</th>
<th>(Y_2)</th>
<th>(Y_3)</th>
<th>(Y_4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>332 54 18</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1  11  22</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3  0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0  1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4  5  1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2  0  0  2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0  0</td>
</tr>
</tbody>
</table>

\((x_{a2}, x_{a4}) = (0,1)\) *

Subjects with:

<table>
<thead>
<tr>
<th>(Y_1)</th>
<th>(Y_2)</th>
<th>(Y_3)</th>
<th>(Y_4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>476 72 44</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>9</td>
<td>7  3  12  30</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>5</td>
<td>6  1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>6  2  2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>12  8  0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4  2  1  4</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4  0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>8  0  0</td>
</tr>
</tbody>
</table>

* \((x_{a2}, x_{a4}) = \) (group of subject i, age at cycle 1 of subject i)

Table 5.4: Erythrocyte Sedimentation Rate by Group and Age at cycle 1 for Subjects with Drop-out Data
Table 5.4 (continued)

\[(x_{i2}, x_{i4}) = (1,0) \ast\]

Subjects with : $Y_1, Y_2, Y_3, Y_4$ $Y_1, Y_2, Y_3$ $Y_1, Y_2$ $Y_1$

<table>
<thead>
<tr>
<th></th>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$Y_3$</th>
<th>$Y_4$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>276</td>
<td>45</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\[(x_{i2}, x_{i4}) = (1,1) \ast\]

Subjects with : $Y_1, Y_2, Y_3, Y_4$ $Y_1, Y_2, Y_3$ $Y_1, Y_2$ $Y_1$

<table>
<thead>
<tr>
<th></th>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$Y_3$</th>
<th>$Y_4$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>403</td>
<td>60</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>9</td>
<td>4</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

* $(x_{i2}, x_{i4}) = (\text{group of subject } i, \text{ age at cycle } 1 \text{ of subject } i)$
<table>
<thead>
<tr>
<th>Drop-out Model</th>
<th>Parameter Estimates</th>
<th>$X^2$</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 $\nu_0 + \nu_1$ (group)</td>
<td>$\nu_0 = 3.0278 (.0809)$</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>$\nu_1 = -0.0359 (.1187)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 $\nu_0 + \nu_1$ (study year)</td>
<td>$\nu_0 = 3.5574 (.1434)$</td>
<td>$X^2 = 24.44$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\nu_1 = -0.0871 (.0197)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 $\nu_0 + \nu_1$ (m.o.)</td>
<td>$\nu_0 = 3.5459 (.1659)$</td>
<td>$X^2 = 45.26$</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>$\nu_1 = -0.2457 (.0683)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 $\nu_0 + \nu_1$ (group) + $\nu_2$ (study year) + $\nu_3$ (age at cycle 1)</td>
<td>$\nu_0 = 3.6961 (.1744)$</td>
<td>$X^2 = 29.40$</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>$\nu_1 = -0.0353 (.1200)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\nu_2 = -0.0872 (.0197)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\nu_3 = -0.1934 (.1255)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 $\nu_0 + \nu_1$ (m.o.) + $\nu_2$ (study year)</td>
<td>$\nu_0 = 4.1006 (.2122)$</td>
<td>$X^2 = 25.98$</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>$\nu_1 = -0.2489 (.0685)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\nu_2 = -0.0877 (.0197)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Dropout model $P\{R_{it}=1|d_{it}, \nu, \overline{R}_{it-1}=1\} = d_{it}' \nu$

$\nu_0 =$ coefficient of overall mean
m.o. = military occupation (defined as 1 for officers, 2 for flight engineers, 3 for groundcrew)

(continued)

Table 5.5: Parameter Estimates, $X^2$ Statistics and Degrees of Freedom for Logistic Regression Models Fit to the Drop-out Mechanism (Standard Errors in Parentheses)
Table 5.5 (continued)

<table>
<thead>
<tr>
<th>Drop-out Model</th>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \nu_0 + \nu_1 \text{ (m.o.)} + \nu_2 \text{ (study year)} + \nu_3 \text{ (age at cycle 1)} )</td>
<td>( \nu_0 = 4.6105 \ (0.2637) ) ( \nu_1 = -0.3495 \ (0.0784) ) ( \nu_2 = -0.0881 \ (0.0197) ) ( \nu_3 = -0.4544 \ (0.1363) ) ( X^2 = 38.23 ) ( \text{d.f.} = 14 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Drop-out Model</th>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \nu_0 + \nu_1 I_{\text{officer}} + \nu_2 I_{\text{flt.engineer}} + \nu_3 \text{ (study year)} )</td>
<td>( \nu_0 = 3.3459 \ (0.1530) ) ( \nu_1 = 0.4945 \ (0.1380) ) ( \nu_2 = 0.2604 \ (0.1685) ) ( \nu_3 = -0.0877 \ (0.0197) ) ( X^2 = 25.75 ) ( \text{d.f.} = 5 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Drop-out Model</th>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \nu_0 + \nu_1 I_{\text{officer}} + \nu_2 I_{\text{flt.engineer}} + \nu_3 \text{ (study year)} + \nu_4 \text{ (age at cycle 1)} )</td>
<td>( \nu_0 = 3.5556 \ (0.1699) ) ( \nu_1 = 0.6927 \ (0.1487) ) ( \nu_2 = 0.4472 \ (0.1766) ) ( \nu_3 = -0.0882 \ (0.0197) ) ( \nu_4 = -0.4673 \ (0.1379) ) ( X^2 = 38.49 ) ( \text{d.f.} = 13 )</td>
</tr>
</tbody>
</table>

Note: Dropout model \( P \{ R_i = 1 | d_i, \nu, \overline{R_i}, i-1 = 1 \} = d_i' \nu \)

\( \nu_0 \) = coefficient of overall mean

m.o. = military occupation (defined as 1 for officers, 2 for flight engineers, 3 for groundcrew)
Table 5.6: Parameter Estimates, $G^2$ and $X^2$ Statistics and Degrees of Freedom for Full Model Fit to Complete and Drop-out Sedimentation Rate Data (Standard Errors in Parentheses)
<table>
<thead>
<tr>
<th>Model</th>
<th>Marginal Association</th>
<th>Drop-out</th>
<th>Model 4</th>
<th>Model 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$\beta_0 = -3.7165 \ (1.291)$</td>
<td>$\alpha_{12} = 1.0226 \ (3.407)$</td>
<td>$G^2 = 224.85$</td>
<td>$G^2 = 224.99$</td>
</tr>
<tr>
<td>All-Pair</td>
<td>$\beta_1 = 0.1468 \ (10.46)$</td>
<td>$\alpha_{13} = 1.3349 \ (3.372)$</td>
<td>$X^2 = 243.63$</td>
<td>$X^2 = 245.24$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.1903 \ (0.096)$</td>
<td>$\alpha_{14} = 1.7414 \ (3.262)$</td>
<td>d.f. = 102</td>
<td>d.f. = 101</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.3679 \ (11.50)$</td>
<td>$\alpha_{23} = 2.3647 \ (2.470)$</td>
<td>$\alpha_{24} = 1.8083 \ (2.544)$</td>
<td>$\alpha_{34} = 2.0809 \ (2.020)$</td>
</tr>
<tr>
<td>6</td>
<td>$\beta_1 = -3.6916 \ (1.254)$</td>
<td>$\alpha_{13} = 2.2468 \ (5.544)$</td>
<td>$G^2 = 277.19$</td>
<td>$G^2 = 277.34$</td>
</tr>
<tr>
<td>(23,134)</td>
<td>$\beta_1 = 0.1476 \ (1.015)$</td>
<td>$\alpha_{14} = 1.9542 \ (3.493)$</td>
<td>$X^2 = 412.53$</td>
<td>$X^2 = 415.81$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.1850 \ (0.097)$</td>
<td>$\alpha_{23} = 3.5313 \ (2.142)$</td>
<td>d.f. = 103</td>
<td>d.f. = 102</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.3936 \ (1.076)$</td>
<td>$\alpha_{24} = 2.5218 \ (1.926)$</td>
<td>$\alpha_{34} = -0.5377 \ (0.6397)$</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>$\beta_1 = -3.7323 \ (1.276)$</td>
<td>$\alpha_{12} = 1.6677 \ (2.981)$</td>
<td>$G^2 = 240.96$</td>
<td>$G^2 = 241.11$</td>
</tr>
<tr>
<td>(12,14,234)</td>
<td>$\beta_1 = 0.1373 \ (1.043)$</td>
<td>$\alpha_{14} = 2.0390 \ (3.034)$</td>
<td>$X^2 = 283.01$</td>
<td>$X^2 = 284.16$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.1930 \ (0.094)$</td>
<td>$\alpha_{23} = 2.3993 \ (4.839)$</td>
<td>d.f. = 102</td>
<td>d.f. = 101</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.3755 \ (1.103)$</td>
<td>$\alpha_{24} = 1.6049 \ (3.094)$</td>
<td>$\alpha_{34} = 2.3120 \ (2.179)$</td>
<td>$\alpha_{234} = 0.1764 \ (0.5617)$</td>
</tr>
<tr>
<td>8</td>
<td>$\beta_1 = -3.7034 \ (1.270)$</td>
<td>$\alpha_{13} = 1.7736 \ (2.909)$</td>
<td>$G^2 = 233.70$</td>
<td>$G^2 = 233.85$</td>
</tr>
<tr>
<td>(13,14,234)</td>
<td>$\beta_1 = 0.1423 \ (1.045)$</td>
<td>$\alpha_{14} = 1.8915 \ (3.104)$</td>
<td>$X^2 = 284.00$</td>
<td>$X^2 = 286.21$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.1888 \ (0.093)$</td>
<td>$\alpha_{23} = 2.3585 \ (4.772)$</td>
<td>d.f. = 102</td>
<td>d.f. = 101</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.3734 \ (1.105)$</td>
<td>$\alpha_{24} = 1.9289 \ (3.031)$</td>
<td>$\alpha_{34} = 1.9097 \ (2.203)$</td>
<td>$\alpha_{234} = 0.2676 \ (0.5552)$</td>
</tr>
<tr>
<td>9</td>
<td>$\beta_1 = -3.7311 \ (1.276)$</td>
<td>$\alpha_{12} = 2.0333 \ (7.527)$</td>
<td>$G^2 = 240.78$</td>
<td>$G^2 = 240.93$</td>
</tr>
<tr>
<td>(23,34,124)</td>
<td>$\beta_1 = 0.1360 \ (1.042)$</td>
<td>$\alpha_{14} = 2.0852 \ (3.250)$</td>
<td>$X^2 = 278.18$</td>
<td>$X^2 = 279.35$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.1931 \ (0.094)$</td>
<td>$\alpha_{23} = 2.5436 \ (2.432)$</td>
<td>d.f. = 102</td>
<td>d.f. = 101</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.3749 \ (1.102)$</td>
<td>$\alpha_{24} = 1.6812 \ (2.682)$</td>
<td>$\alpha_{34} = 2.3446 \ (2.009)$</td>
<td>$\alpha_{124} = -0.4200 \ (0.8168)$</td>
</tr>
</tbody>
</table>
CHAPTER 6

DROP-IN DATA

Because sedimentation rate is determined from blood drawn during the physical examination portion of the AFHS, only those subjects who are fully compliant to a study cycle have observed sedimentation rate data for that cycle. As more cycles of the AFHS are completed, the patterns of observed and missing sedimentation rate data for subjects have become increasingly complicated. Fortunately, there are three patterns that are most prevalent in the sedimentation rate data. The first pattern is that of complete data. This type of data is seen for those subjects who fully comply to each physical exam and, thus, have a sedimentation rate variable for all study cycles. The second pattern is of "drop-out". This pattern is seen in subjects who were fully compliant to the 1982 Baseline examination, and possibly one or more follow-up examination, but who fail to comply from some point onward. The third pattern that is seen quite often in the AFHS sedimentation rate data is of "drop-in". This type of data is observed in subjects who become fully compliant to the AFHS at some point after the 1982 Baseline examination
and, once fully compliant, remain fully compliant to every cycle from then on. The "drop-in" pattern is opposite of the "drop-out" pattern, although both are monotone. For drops-ins, if an individual's \( t^{th} \) response is missing, then every response \( t' < t \) is also missing.

Drop-in data is seen most often in Replacement Comparisons who are asked to participate in the AFHS to replace a refusing Comparison and who fully comply to all cycles once they enroll in the study. Since many of these Replacement Comparisons were first asked to join the AFHS at one of the follow-up examinations, they are missing at least the 1982 Baseline sedimentation rate data. Other subjects who have "drop-in" data are those who had previously been unlocateable (the AFHS had no way to contact these subjects to ask them to participate in the study), but are now both locatable and fully compliant to all cycles of the study. Subjects who had previously refused to participate in the AFHS but who agree to fully participate in the study from some point on also have drop-in data.

I have already used the approach proposed by Fitzmaurice and Laird (1993), abbreviated F-L, to analyze those subjects who have complete data. The models proposed by Fitzmaurice, Laird, and Zahner (1996), abbreviated F-L-Z, were then used to analyze subjects who have drop-out data. The approach, an extension of that given by F-L, assumes a logistic regression model for the missing-data mechanism that allows the missingness to depend both on covariates and on the observed and missing responses. However, I assumed only missing-at-random (MAR) models for the drop-out mechanism and, thus, avoided the arduous task of establishing the identifiability of the parameters in the models.
I have further extended both the F-L and the F-L-Z approaches to accommodate those subjects who "drop-in" the AFHS. Like F-L-Z, I assume a likelihood-based regression model that uses a logistic regression model for the missing-data mechanism for the drop-ins. However, I assume that the model for the missingness depends only on observed covariates, while F-L-Z allow their model to depend on covariates and on observed and missing responses. By assuming that the nonresponse model depends only on observed covariates, I obtain a MAR model for the missingness. Since MAR models are ignorable, I do not have the major issue of determining the identifiability of the parameters in my models. In Section 6.1, we develop the methodology for analyzing "drop-in" subjects. We develop the approach for a general number of time periods $T$, and then in Section 6.2 apply the methods to the AFHS sedimentation rate data using $T=4$. In Section 6.3, we extend the approach of Section 6.1 to accommodate subjects with complete, drop-in or drop-out data. Again, we use a general number of time periods $T$ in developing the methodology, and then in Section 6.4 apply it to the AFHS sedimentation rate data with $T=4$.

6.1 The Likelihood Function for Drop-In Data

Assume there are $T$ observation timepoints and $N'$ total subjects. Associate with each of the $N'$ subjects a $T \times 1$ vector $Y_i = (Y_{i1}, \ldots, Y_{iT})'$ of responses and a $T \times 1$ vector $R_i = (R_{i1}, \ldots, R_{iT})'$ of response indicators. For subject $i$, $R_i$ is defined such that $R_{it} = 1$ if $Y_{it}$ is observed, and $R_{it} = 0$ if $Y_{it}$ is missing. Given $R_i$, we can partition $Y_i$ such that $Y_i =$
(Y_m, Y_o), where Y_m are the missing values of Y, and Y_o are the values of Y that are observed. Elements of Y_o that correspond to "success" are denoted by a 1, while those that correspond to "failure" are denoted by a 0.

For our approach, we will consider models for data that follows the monotone missing-data pattern where if response t is missing, so are all the responses t'<t. Thus, the subjects either have complete data or they have "drop-in" data. Once a subject is observed, he is observed at all future time periods. The joint distribution of (Y, R) can be written
\[ f(y, r) = f(r | y) f(y) \]
where y and r are the observed values of Y and R, respectively. We will now develop models for both \( f(r | y) \) and \( f(y) \), beginning with \( f(y) \).

6.1.1 Complete-Data Likelihood

Assume that at each time t, subject i has a Jx1 vector of covariates \( x_t \). For every subject i and every time t, \( x_t \) is always completely observed, regardless of whether \( Y_t \) is observed. Let \( X_i = (x_{i1}, \ldots, x_{iT})' \) be the TxP matrix of covariates for subject i.

As in Section 4.1, we assume the marginal model for our binary data \( Y_t \) is a logistic regression model:
\[ f(y_t | x_t) = \exp[y_t \theta_t - \log(1 + \exp(\theta_t))], \]
with \[ \theta_u = \log \left( \frac{\mu_u}{1 - \mu_u} \right) = x_u' \beta \]

where \( \mu_u = \mu_u(\beta) = P(Y_u = 1|x_u, \beta) \) and \( \beta \) is a Px1 vector of parameters.

We then parameterize the joint distribution of \( Y \) in terms of conditional odds-ratios, as in Section 4.1.2.1:

\[
f(y, \Psi, \Omega) = \exp \{ \Psi y + \Omega w - A(\Psi, \Omega) \},
\]

where \( W = (Y_1, Y_{12}, \ldots, Y_{T-1}, Y_T, \ldots, Y_{1T}, Y_{12} \ldots Y_{TT}) \) is the Kx1 vector of two-way and higher-order cross-products of \( Y \),

\[
\Psi = (\psi_1, \ldots, \psi_{TT})' \quad \text{and} \quad \Omega = (\omega_{11}, \ldots, \omega_{T(T-1)} \ldots, \omega_{12 \ldots TT})' \quad \text{are vectors of canonical parameters and}
\]

\( A(\Psi, \Omega) \) the normalizing constant,

all three of which are defined following equation (4.3) in Chapter 4. Likewise, we again assume \( \Omega \) can be expressed as a function of a Qx1 vector of parameters

\( \alpha = (\alpha_1, \ldots, \alpha_Q)' \) with a linear link \( \Omega_i = Z_i \alpha \), where \( Z_i \) is a KxQ design matrix.

For ease of notation, let \( \lambda = (\beta, \alpha) \) and define \( \gamma_i = E(W_i), V_{i11} = \text{Cov}(Y_i), V_{i12} = \text{Cov}(Y_i, W_i), V_{i22} = \text{Cov}(W_i) \) and \( \Delta_i = \text{diag}(\text{Cov}(Y_i)) \). From equations (4.7) and (4.8) in Chapter 4, we see that if we always had completely observed data, the MLE \( \hat{\lambda} = (\hat{\beta}, \hat{\alpha}) \) would be the solution to the equations:

\[
\sum_{i=1}^N X_i' \Delta_i V_{i11}^{-1} (y_i - \mu_i) = 0 \quad \text{and} \quad \sum_{i=1}^N Z_i' \{ w_i - \gamma_i - V_{i12} V_{i11}^{-1} (y_i - \mu_i) \} = 0.
\]

117
6.1.2 Drop-In Likelihood

Let $\phi$ be the vector of parameters of the nonresponse model for the drop-ins, and let $f(r_i|Y_i, \phi)$ be the joint distribution of $R_i$ given $Y_i$ and $\phi$. Also let $a_i$ be the vector of covariates related to the drop-in mechanism for subject $i$ at time $t$, and form $a_i = (a_{it}, \ldots, a_{it})$. Similar to F-L-Z, we define a single random variable to characterize $R_i$; let $U_i = \sum_{t=1}^{T} R_t$, the number of observed responses for subject $i$. Further, let $\overline{R}_i = (R_{1i}, \ldots, R_{Ti})$ and $\overline{Y}_i = (Y_{1i}, \ldots, Y_{Ti})$. That is, for subject $i$, $\overline{R}_i$ is the vector of response indicators from time $t$ through the end of the study, while $\overline{Y}_i$ is the vector of responses from time $t$ through the end of the study. Finally, collectively denote the elements of $W_i$ that are observed by $W_{oi}$, and the missing elements by $W_{mo}$; for ease of notation, let $W_i = (W_{oa}, W_{om})'$.

We will only consider MAR models for the missing-data mechanism. Hence, the probability of response status at time $t$ for subject $i$ given that subject $i$ is a drop-in will have the form:

$$P(R_{it} = r_{it} | Y_{it}, a_{it}, \phi) = P(R_{it} = r_{it} | a_{it}, \phi).$$

Since we are looking at drop-in monotone data, we are interested in

$$P(R_{it} = r_{it} | \overline{R}_{i(t-1)} = 1, a_{it}, \phi);$$

that is, we want to find out what variables are important in determining whether/when why a subject is observed from some time point on. Finally, we want to consider logistic
regression models for the drop-in mechanism. Putting all of this together, we have models of the form:

\[
\text{logit}(P(R_i = 1|R_{i-1} = 1, a, \phi)) = a \cdot \phi.
\]

Or in other words, we have

\[
P(R_i = 1|R_{i-1} = 1, a, \phi) = \frac{\exp(a \cdot \phi)}{1 + \exp(a \cdot \phi)}
\]

and

\[
P(R_i = 0|R_{i-1} = 1, a, \phi) = \frac{1}{1 + \exp(a \cdot \phi)}.
\]

It follows, then, that

\[
P(U = u_i | a, \phi) = P(R_1 = 0, \ldots, R_{i(T-1)} = 0, R_{i(T-U)} = 1, \ldots, R_T = 1 | a, \phi)
\]

\[
= P(R_{i(U)} = 0, R_{i(U-1)} = 1 | a, \phi)
\]

\[
= P(R_{i(T-U)} = 0, R_{i(U-1)} = 1 | a, \phi)
\]

\[
= P(R_{i(T-U)} = 0, R_{i(U-1)} = 1, a, \phi) P(R_{i(U-1)} = 1 | a, \phi)
\]

\[
= \left( \frac{1}{1 + \exp(a \cdot \phi)} \right)^{\sum_{U_i < T}} \left( \prod_{j=U_i}^{T-1} P(R_j = 1 | R_{j-1} = 1, a, \phi) \right)
\]

\[
= \left( \frac{1}{1 + \exp(a \cdot \phi)} \right)^{\sum_{U_i < T}} \left( \prod_{j=U_i}^{T-1} \left( \frac{\exp(a \cdot \phi)}{1 + \exp(a \cdot \phi)} \right) \right)
\]

Since \( U_i = \sum_{t=1}^{T} R_i \), it follows that

119


6.1.3 Maximizing the Incomplete-Data Likelihood

The contribution to the incomplete-data likelihood, which is observed, for subject \(i\)

\[
P(R_i = r_i | \alpha, \phi) = \left( \frac{1}{1 + \exp\{\alpha_{(T_n-r_i)} \phi\} \prod_{j=T_n-r_i+1}^{T-1} \frac{\exp(\alpha_j \phi)}{1 + \exp(\alpha_j \phi)} \right).
\]

(6.3)

so that the contribution to the log-likelihood is then

\[
\ln f(y_{\alpha}, r_i | \lambda, \phi) = \ln \{ f(r_i | \alpha, \phi) \} + \ln \{ \sum_{y_m} f(y_m, y_{\alpha} | \lambda) \}. 
\]

(6.4)

We want to maximize this incomplete-data (log-) likelihood with respect to \(\lambda\) and \(\phi\).

Taking derivatives with respect to \(\lambda\) and \(\phi\) of the \(i^{th}\) individual's contribution to the log-likelihood, we get

\[
\frac{d \ln f(y_{\alpha}, r_i | \lambda, \phi)}{d \phi} = \frac{d \ln f(r_i | \alpha, \phi)}{d \phi}
\]

(6.5)

\[
\frac{d \ln \{ \sum_{y_m} f(y_m, y_{\alpha} | \lambda) \}}{d \lambda} = \frac{d \ln f(y_{\alpha}, r_i | \lambda, \phi)}{d \lambda}.
\]
Combining (6.3) and (6.4), we get

\[
\frac{d \ln f(y_m, r_i | \lambda, \phi)}{d \phi} = \frac{d}{d \phi} \ln \left( \frac{1}{1 + \exp \{a_i (r_i - \eta_i) \phi \}} \right)^{\eta_i} \left( \prod_{j=T-\eta_i+1}^{T-1} \left( \frac{\exp \{a_j \phi \}}{1 + \exp \{a_j \phi \}} \right) \right) .
\]  

(6.6)

To help us determine the expression in (6.5), note that

\[
E_{y_m | y_{m_i}} \left\{ \frac{d \ln f(y_m | \lambda)}{d \lambda} \right\} = E_{y_m | y_{m_i}} \left\{ \frac{d}{d \lambda} \ln f(y_m | y_{m_i}, \lambda) f(y_{m_i} | \lambda) \right\}
\]

\[
= E_{y_m | y_{m_i}} \left\{ \frac{d}{d \lambda} \ln f(y_m | y_{m_i}, \lambda) + \frac{d}{d \lambda} \ln f(y_{m_i} | \lambda) \right\}
\]

\[
= E_{y_m | y_{m_i}} \left\{ \frac{d}{d \lambda} \ln f(y_m | y_{m_i}, \lambda) \right\} + \frac{d}{d \lambda} \ln \left( \sum_{y_{m_i}} f(y_{m_i}, y_{m_i} | \lambda) \right)
\]

\[
= \sum_{y_{m_i}} \left\{ \frac{d}{d \lambda} \ln f(y_m | y_{m_i}, \lambda) \right\} * f(y_m | y_{m_i}, \lambda) + \frac{d}{d \lambda} \ln \left( \sum_{y_{m_i}} f(y_{m_i}, y_{m_i} | \lambda) \right)
\]

\[
= \sum_{y_{m_i}} \left\{ \frac{1}{f(y_m | y_{m_i}, \lambda)} \left[ \frac{d}{d \lambda} f(y_m | y_{m_i}, \lambda) \right] * f(y_m | y_{m_i}, \lambda) \right\}
\]

\[
+ \frac{d}{d \lambda} \ln \left( \sum_{y_{m_i}} f(y_{m_i}, y_{m_i} | \lambda) \right)
\]

\[
= \sum_{y_{m_i}} \frac{d}{d \lambda} \left\{ f(y_m | y_{m_i}, \lambda) \right\} + \frac{d}{d \lambda} \ln \left( \sum_{y_{m_i}} f(y_{m_i}, y_{m_i} | \lambda) \right)
\]

\[
= 0 + \frac{d}{d \lambda} \ln \left( \sum_{y_{m_i}} f(y_{m_i}, y_{m_i} | \lambda) \right)
\]
\[
\frac{d}{d\lambda} \ln \left\{ \sum_{y_m} f(y_m, y_n | \lambda) \right\}.
\]

(6.7)

The third line comes from the fact that \( f(y_m | \lambda) \) is free of \( y_m \), and the second from last line comes from the fact that \( f(y_m | y_n, \lambda) \) is also a member of an exponential family.

Combining (6.5) and (6.7), we have

\[
\frac{d \ln f(y_m, r_n | \lambda, \phi)}{d\lambda} = E_{y_m | r_n} \left\{ \frac{d \ln f(y_n | \lambda)}{d\lambda} \right\}.
\]

(6.8)

As stated earlier, since we know \( f(y_n | \lambda) \) is of the form (6.1), if we had only complete data, the equation in (6.8) would become that in (6.2). However, we know from the equations (4.7) in Chapter 4 that

\[
\frac{d}{d\beta} \ln f(y_i | \beta, \alpha) = X_i \Delta, V_n^{-1} \{ y_i - \mu_i \}
\]

and

\[
\frac{d}{d\alpha} \ln f(y_i | \beta, \alpha) = Z_i \{ y_i - \mu_i \}. \]

It follows then that

\[
E_{y_m | r_n} \left\{ \frac{d \ln f(y_n | \lambda)}{d\lambda} \right\} = E_{y_m | r_n} \left\{ \frac{d \ln f(y_n | \beta, \alpha)}{d(\beta, \alpha)} \right\}
\]

\[
= E_{y_m | r_n} \left\{ X_i \Delta, V_n^{-1} \{ y_i - \mu_i \} \right\}
\]

\[
= E_{y_m | r_n} \left\{ Z_i \{ e(y_i) - e(\mu_i) \} \right\}, \tag{6.9}
\]

where \( e(y_i) = (e(y_m), y_n) = E_{y_m | r_n} \{ y_i \} \) and \( e(w_i) = (e(w_m), w_n) = E_{y_m | r_n} \{ W_i \} \).
Our logistic regression model for the drop-in response mechanism is

\[
\text{logit}(P(R_i = 1 | \overline{R}_{(t-1)}) = 1, a_{it}, \phi) = a_{it} \phi
\]

\[= a_{i0} \phi_0 + a_{it} \phi_1 + \ldots + a_{ikt} \phi_k.\]

For ease of notation, assume \(a_{i0} = 1\) for all \(i\) and for all \(t\). Also define

\[
\overline{p}_{it} = P(R_i = 1 | \overline{R}_{(t-1)}) = 1, a_{it}, \phi = \frac{\exp(a_{it} \phi)}{1 + \exp(a_{it} \phi)},
\]

the probability that subject \(i\) is observed at time \(t\) given that subject \(i\) is observed at every time after \(t\). \(\overline{p}_{it}\) may seem awkward at first, but recall that we are working with a data set in which we know the last response(s) is(are) observed. Then, from (6.3), it follows that for person \(i\),

\[
\frac{dI_i(\phi | r_i, a_i)}{d\phi} = \frac{d \ln f(r_i | a_i, \phi)}{d\phi}
\]

\[= \frac{d}{d\phi} \left[ \sum_{t=T \to u-1} [a_{it} \phi - \ln(1 + \exp(a_{it} \phi))] - I_{(u < T)} \ln(1 + \exp(a_{it} \phi)) \right]
\]

\[= \sum_{t=T \to u-1} \frac{a_{it} \phi - \ln(1 + \exp(a_{it} \phi))}{1 + \exp(a_{it} \phi)} - I_{(u < T)} \frac{\exp(a_{i} \phi)}{1 + \exp(a_{it} \phi)} \frac{a_{it}}{1 + \exp(a_{it} \phi)}
\]

\[= \sum_{t=T \to u-1} a_{it} \left( r_i - \overline{p}_{it} \right) + I_{(u < T)} a_{it} \overline{r}_{(T-1)} - I_{(u < T)} a_{it} \overline{y}_{(T-1)}
\]

\[= \sum_{t=T \to u-1} a_{it} \left( r_i - \overline{p}_{it} \right), \quad (6.10)
\]
where $\phi_i$ is a term in the vector of (drop-in) nonresponse parameters. The next to the last equation stems from if $U_i = u_i < T$ then $r_{it} = 0$ for $t=1,\ldots,T-u_i$ and $r_{it} = 1$ for $t=T-u_i+1,\ldots,T$; if $U_i = T$ then $I_{(u_i,T)} = 0$. Either way, the added term is 0.

Summing over the $N'$ subjects, then, the MLE $\hat{\lambda}$ is the solution to

$$\sum_{i=1}^{N'} \left( X_i' \Delta_i V_{i1}^{-1} \{ e(y_i) - \mu_i \} \right) = 0 \quad (6.11)$$

and the MLE $\hat{\phi}$ is the solution to

$$\sum_{i=1}^{N'} \frac{d l_i(\phi | r_i, a_i)}{d\phi} = 0, \quad (6.12)$$

where $\frac{d l_i(\phi | r_i, a_i)}{d\phi}$ is given in (6.10).

There are two things to note here. The first is that for subject $i$, the matrix $X_i$ is fully specified, regardless of whether the subject has complete or drop-in data. The second is that the estimators for $\hat{\beta}$ will be consistent only when both the marginal model and the association model have been correctly specified (Fitzmaurice et al., 1993). Let

$S_i(\beta) = X_i' \Delta_i V_{i1}^{-1} \{ y_i - \mu_i \}$ in (4.9) from Chapter 4, and

$S_i^*(\beta) = X_i' \Delta_i V_{i1}^{-1} \{ e(y_i) - \mu_i \}$ from (6.9). When we have complete data and a correctly specified marginal model, we have that $E(S_i(\beta)) = 0$ since $E(y_i - \mu_i) = 0$. This leads to consistency of $\hat{\beta}$, regardless of whether the association model is correct (Godambe and Kale, 1991; Fitzmaurice et al., 1993). However, with incomplete data and
misspecification of the association structure, we have that \( E \{ S_t^* (\beta) \} \neq 0 \) (Fitzmaurice et al., 1993). Thus, when we are dealing with only complete data, as long as the marginal model is correct, so that \( E(Y_i) = \mu_i \), then \( \hat{\beta} \) will be consistent regardless of whether the association structure is correct. However, with missing data, \( \hat{\beta} \) will be consistent only if the full distribution is correct. Hence we not only need \( E \{ e(Y_i) \} = \mu_i \), but also that \( E \{ e(W_i) \} = \gamma_i \).

6.1.4 Obtaining the Maximum Likelihood Estimates

The EM-algorithm can be used to obtain the MLE \( \hat{\lambda} = (\hat{\beta}, \hat{\alpha}) \). In the E-step,

\[
E_{y_{iw}^{iaw}} \left\{ \frac{d \ln f(y_i | \lambda)}{d \lambda} \right\}
\]

is computed, based on current value of \( \hat{\lambda} \). For the M-step,

\[
\sum_{i=1}^{N} E_{y_{iw}^{iaw}} \left\{ \frac{d \ln f(y_i | \lambda)}{d \lambda} \right\}
\]

is maximized with respect to \( \lambda \), with the maximization done as in F-L (1993). The maximization is a two-step procedure that uses the Fisher scoring algorithm along with an IPF algorithm to determine the second- and higher-order equations; this procedure is given in Section 4.1.2.3 in Chapter 4. In words, the E-step gives us "complete" data \((e(y_{im}), y_{om})\) and \((e(w_{im}), w_{om})\) for subject i, based on the current estimate of \( \lambda \). Then the M-step maximizes the likelihood function, based on our "complete" data from the E-step. The maximization is carried out exactly as when we had only complete data, since the E-step has given us a "complete" data set.
The EM-algorithm leads to the MLE's $\hat{\lambda} = (\hat{\beta}, \hat{\alpha})$. However, it does not give an estimate of the asymptotic variance-covariance matrix of $\hat{\lambda} = (\hat{\beta}, \hat{\alpha})$. The usual inverse Fisher information matrix is a consistent estimator of the asymptotic variance-covariance matrix only when the non-response is MCAR. Even though the inverse of the observed information matrix can be used to estimate the asymptotic variance-covariance matrix of $\hat{\lambda}$ when the non-response is MAR, the matrix of second derivatives of the log-likelihood based on the incomplete data is very complicated. Fortunately, the sample empirical covariance matrix of the scores, in any correctly specified model, is a consistent estimator of the Fisher information regardless of whether the data is complete, and is readily available from computations performed during each iteration of the EM-algorithm. We have, then, a consistent estimator of the asymptotic variance-covariance matrix of $\hat{\lambda} = (\hat{\beta}, \hat{\alpha})$ given by the sample empirical covariance matrix:

$$
\text{cov}(\hat{\beta}, \hat{\alpha}) = \sum_{i=1}^{N} \left[ \begin{array}{c} \frac{dl_{\alpha}}{d\beta} \\ \frac{dl_{\beta}}{d\beta} \\ \frac{dl_{\alpha}}{d\alpha} \\ \frac{dl_{\beta}}{d\alpha} \end{array} \right] \left[ \begin{array}{c} \frac{dl_{\alpha}}{d\beta} \\ \frac{dl_{\beta}}{d\beta} \\ \frac{dl_{\alpha}}{d\alpha} \\ \frac{dl_{\beta}}{d\alpha} \end{array} \right]^T \right]^{-1}
$$

$$
= \sum_{i=1}^{N} \left[ E \left( \frac{dl_{\beta}}{d\beta} \right) \frac{dl_{\alpha}}{d\alpha} \right] \left( E \left( \frac{dl_{\beta}}{d\beta} \right) \right)^{-1}
$$

(6.13)

where
and \( E\left( \frac{dl}{d\beta} \right) = X_i', \Delta_i V_{111}^{-1} \{ e(y_i) - \mu \} \)

and \( E\left( \frac{dl}{d\alpha} \right) = Z_i \left[ e(w_i) - y_i - V_{112} V_{111}^{-1} \{ e(y_i) - \mu \} \right] \). Note that if both the marginal model and the association model are correctly specified, \( E\{ \text{cov}(\hat{\beta}, \hat{\alpha}) \} \) is (4.10) from Chapter 4.

To obtain the MLE \( \hat{\phi} \) from (6.10) and (6.12), recall that our model for \( P \{ R_t = 1 \mid R_{t-1} = 1 \} \) is simply a logistic regression. It may be easiest to understand each subject's contribution to (6.12), which is given in (6.6), using Table 6.1. This table shows each possible drop-in pattern with its respective likelihood contribution. In words, a subject who has complete data contributes to the drop-in likelihood function for all time periods. A subject who does not have complete data contributes to the drop-in likelihood function for each time period he is observed and for the time period just prior to when he drops into the study. Once each subject's contribution to (6.12) is correctly determined, the MLE \( \hat{\phi} \) is obtained through ordinary logistic regression. Any standard statistical package (Splus, SAS, etc.) can be used to determine \( \hat{\phi} \).

### 6.2 Model Fitting to the Drop-In Data

I used the above drop-in data methodology to fit models to the AFHS sedimentation rate data. My drop-in data analyses were conducted on subjects who had
sedimentation rate data from at least the 1992 study cycle and who were either observed at every other cycle (thus having complete data) or had data such that if observation t was missing then so was every previous observation. There are a total of 1,881 subjects who “dropped in” at cycle 1 and, thus, have data from all four cycles; 139 subjects who “dropped in” at cycle 2 and have data from cycle 2, 3 and 4; 59 subjects who “dropped in” at cycle 3, thus having observations from cycles 3 and 4; and 61 subjects who “dropped in” at cycle 4, to have only cycle 4 sedimentation rate data. This gives a total of N′ = 2,140 subjects with drop-in data and a total number of timepoints T = 4.

I defined the covariate vector $x_{i, t}$, for the model on the marginal response probability, $\mu_{i, t}$, similar to the way I had in the first models fit to the complete data. The first three elements of $x_{i, t}$ were defined exactly as before so that, for subject i and time t, the first element of $x_{i, t}$ was a 1, the second element indicated the group membership of subject i and the third element denoted the study year at time t. However, for the fourth element, that which indicated the “age” of subject i, I could not use the subject’s age at his cycle 1 examination because subjects who dropped in after the Baseline examination were missing this variable. Since each drop-in subject’s age at the time of his cycle 4 examination was known, I could use this as my “age” variable. Recall that the age at cycle 1 variable had been defined as a 0 for subjects ≤ 40 years old at their cycle 1 examination, and as a 1 for those > 40 years old at their cycle 1 examination, and that cycle 1 was conducted in 1982 while cycle 4 was in 1992. To define “age” in a way parallel to the previous definition, I defined age at cycle 4 to be 0 for subjects ≤ 50 at their cycle 4 examination, and as 1 for those > 50 at the time of their cycle 4 examination.
There were 14 subjects out of the 1881 subjects who had complete data whose age at
cycle 1 variable was defined differently from their age at cycle 4 variable. However, this
difference in definition had no effect on the results of the models fit. Note that by defining
“age” by using subjects’ age at cycle 4, each subject’s covariate vector was always
completely observed for every study cycle, regardless if his sedimentation rate data was
observed.

I then defined each subject’s missing-data indicator \( R_i = (R_{i1}, R_{i2}, R_{i3}, R_{i4}) \)’ such
that \( R_{it} = 1 \) if \( Y_{it} \) was observed, and \( R_{it} = 0 \) if \( Y_{it} \) was missing. I then partitioned
\( Y_i = (Y_m, Y_o) \) where \( Y_m \) were the missing elements of \( Y_i \), and \( Y_o \) the observed elements.

Elements of \( Y_o \) that represented “abnormal” sedimentation rates were denoted by a 1, and
those that represented “normal” sedimentation rates were denoted by a 0. The
“abnormal”/”normal” cutpoints for the sedimentation rate data are the same as used
previously and are given in Section 2.5 and Section 5.1. Table 6.2 gives the cross-
classification of \( Y_o \) over the four study cycles for each combination of the group and age-
at-cycle-4 covariates. Note that this table is “backward” of the cross-classifications given
previously in Chapter 5, in that the subjects in Table 6.2 are classified first by their
\( Y_4 \) value, and then their \( Y_3 \) value, and so forth.

I then defined \( U_i = \sum_{t=1}^{4} R_{it} \), the number of observed responses for subject \( i \), and let
\( \overline{R}_i = (R_{i1}, \ldots, R_{i4}) \) and \( \overline{Y}_i = (Y_{i1}, \ldots, Y_{i4}) \), for \( i=1, \ldots, 2140 \) and \( t=1, \ldots, 4 \). Again, as when I
analyzed the drop-out data, I used the models for \( f(y_i|\beta, \alpha) \) that were proposed by
Fitzmaurice and Laird (1993). The main task, then, was determining what variables might be useful in modeling the drop-in mechanism. Recall that I assumed models for the drop-in process that were logistic regressions of the form \( \text{logit}(P(R_u = 1|\bar{R}_{u(t-1)} = 1, \alpha_u, \phi}) \), where \( \alpha_u \) is a vector of completely observed covariates and \( \phi \) is the vector of parameters for the drop-in model.

Since we assumed a MAR drop-in model, the maximizations for \( \lambda = (\beta, \alpha) \) and for \( \phi \) were completely separate. I used the EM-algorithm to find the MLE's of \( \lambda = (\beta, \alpha) \), just as I had done in the drop-out case. With the drop-in data, though, the beginning observation(s) were those computed in the E-step (rather than the later observation(s) as in the drop-out data case); that is, the E-step gave us "complete" data \( e(y_i) = E_{y_{im} \mid \alpha} (Y_i) = (e(y_{im}), y_{i\alpha}) \) and \( e(w_i) = E_{w_{im} \mid \alpha} (W_i) = (e(w_{im}), w_{i\alpha}) \) based on the current \( \hat{\lambda} \). The M-step used this "complete" data to maximize the likelihood with respect to \( \hat{\lambda} \).

For models fit to the drop-in mechanism, I considered numerous different variables to include in the covariate vector \( \alpha_u \), including group, military occupation, age at cycle 4, and study year. I used logistic regression models with the response indicator as the dependent variable to help determine what covariates are related to responding. The first element of each covariate vector was a 1 (for the over all mean), and, when included, the variables group, age at cycle 4, and study year were defined as in the \( x_u \). Recall that the assumed model gives us probabilities of the form:
As we will see below, when interpreting the results one must remember that it is assumed that these subjects drop into the AFHS at some time point. Table 6.3 presents the results of some of the various models fit to the drop-in mechanism.

The group variable (Ranch Hand, Comparison) appeared to be an important factor in every drop-in model fit. In addition, the estimate for the group variable was always positive. This indicates that a subject who is observed at a specific time point is more likely to be observed at an earlier time point if he is a Ranch Hand. In other words, Ranch Hand subjects are more likely to drop into the study at an earlier cycle, while Comparisons are more likely to drop into the study at a later cycle. These results agree with what was found in the Chapter 3 analyses. Table 3.3 of physical examination data shows that a higher percentage of eligible Ranch Hands than eligible Comparisons “dropped into” the study at cycle 1. In addition, the exploratory analyses indicated that Ranch Hands who were noncompliant to cycle 1 or who were newly invited to cycle 2 were fully compliant to cycle 2 at a very high rate. This finding once again appears to demonstrate the influence of the Ranch Hand Association, which strongly encouraged Ranch Hand participation in the AFHS. The positive group coefficient also makes sense in terms of what is known about the Comparisons. In order to ensure an adequate sample size in the control (Comparison) group, any refusing Comparison is replaced by another Comparison from the same matched set for that Ranch Hand. As more Comparisons drop out over the
course of the study, due to whatever reason, we would expect to see an increase in Replacement Comparisons who drop into the study.

The study year, age at cycle 4 and military occupation variables also always appeared to be important terms in the drop-in model. The coefficient for the age-at-cycle-4 variable was positive in every model fit. This would indicate that a subject observed at a specific study cycle is more likely to also be observed at a previous study cycle if he is older. This means that older subjects drop into the study at an earlier time, while younger subjects drop-in later in the course of the study. This is logical, since older subjects may become less likely to drop-into a study as the study continues on, due to such things as poor health or death.

The military occupation variable also appears to be important in the drop-in process. As with the drop-out models, I fit models that included military occupation as an ordinal variable and as indicator variables. For the ordinal military occupation variable models, military occupation was always an important term in the model for the drop-in mechanism and the coefficients always indicated that a subject who is observed at a study cycle is more likely to be observed at a previous cycle if he is a member of the groundcrew. In other words, given that a subject drops into the AFHS, he will tend to drop-in at an earlier cycle if he is a member of the groundcrew, at a somewhat later cycle if he is a flight engineer and at an even later cycle if he is an officer. For the models that defined military occupation via indicator variables, the parameter estimate and standard error for the flight engineer variable indicated that groundcrew and flight engineers are basically the same with respect to dropping into the study. We thus only need to include
one military occupation variable, that for officers, in our drop-in models. In addition, the
models indicated that groundcrew and flight engineers drop-in at earlier cycles than do the
officers. In addition, the models that used indicator variables for military occupation had a
slightly better fit than those that used the ordinal military occupation variable.

Finally, another variable that always appeared to be important in explaining
dropping into the AFHS was study year. The positive estimate for study year indicated
that if a subject is observed at time t, he is more likely to be observed at time t-1 for larger
values of t, t=2,3,4. That is, subjects who drop into the study are likely to drop-in at an
earlier, rather than later, study cycle. One might expect this, since potential study subjects
may be more willing to join a study that has just begun or has been in progress for only a
short time, rather than a study that has been running for 10 or 15 years. It also agrees
with the Chapter 3 analyses that found fully compliant subjects to remain fully compliant
across pairs of study cycles, so that dropping in early (say, cycle 1 or 2) and then staying
in the study appears highly likely.

As before when we fit models to the drop-out mechanism, we were interested here
in whether the model of random missingness was an appropriate model for the drop-in
mechanism. That is, is logit( \( P(R_u = 1| \bar{R}_{(t-1)} = 1, a_u, \phi) \))=\( \phi \) is a reasonable model for
the drop-in process. We computed various \( G^2 \) statistics and partitioned the statistics
from nested models in order to determine if our MAR models were better than the random
missingness model and, if so, what variables are worthwhile adding to the model. When
we added the group variable to the random missingness model, the difference in \( G^2 \) was
68.7 with a loss of only one degree of freedom. This indicated that adding group to the

133
random missingness model is worthwhile. In addition, partitioning the $G^2$ statistics from various other models indicated that adding variables such as age, study year, the group-by-year interaction and military occupation is also worthwhile. In general, the results of partitioning the $G^2$ statistic indicated to us that random missingness is not a reasonable model for the drop-in mechanism and that we should not ignore the missing-data mechanism.

Overall, the models fit to the drop-in mechanism were good. Note that the $X^2$ statistics for the drop-in models from Table 6.3 were closer to their respective degrees of freedom than the drop-out models we had fit in Chapter 5. We have quite a large sample size for fitting our drop-in model, which may partially account for some of the unexplained variability in the models. Due to this large sample, we feel that the drop-in models do an adequate job in determining what variables are important in a subject dropping into the study. Again, as pointed out above, the models that included the indicator military occupation variable had a slightly better fit than the same models that used the ordinal military occupation variable (refer to Models 5 and 7 for an example). Since there is an implied spacing in the ordinal military occupation variable that may be inaccurate and since there is a slightly better fit with the indicator variables, we prefer to use the drop-in models that define military occupation through indicator variables.

We used the above results in the next step in our model fitting, which was to propose some models for the sedimentation rate data from both completely observed and drop-in subjects. As with the drop-out models, the drop-in models containing variables that are covariates in the marginal model for $\beta$ (i.e., group or study year) are much more
attractive computationally than drop-in models that include variables not in the marginal model for $\beta$ (i.e., military occupation). Because the results for the complete data based on our maximization of $(\beta, \alpha)$ are already classified by the covariates in the $\beta$ model, drop-in models that only include (some or all of) these same variables are easily combined with the $(\beta, \alpha)$ models to give us the expected cell counts for the full model for both complete and drop-in data. When the drop-in model includes variables not in the model for $\beta$, the computations for the expected cell counts for the full model become much more complicated.

Table 6.4 gives the results of fitting the full model for both complete and drop-in sedimentation rate data, for drop-in Models 3 and 7 from Table 6.3. Drop-in Model 3 has group, study year, age at cycle 4 and the interaction of group and study year, while Model 7 uses group, study year, age at cycle 4 and the military occupation indicator variable for officers. The marginal model for $\beta$ is our model that includes group, study year and age at cycle 4 as covariates, and the association models are the same nine association models we have illustrated thus far in our results.

The parameter estimates in Table 6.4 agree with the earlier analyses on the complete, and the complete and drop-out data. Study year ($\beta_2$) and age at cycle 4 ($\beta_3$) continue to be important factors in modeling the probability of a subject having an abnormal sedimentation rate. The positive estimates for both variables indicate, as before, that older subjects have a higher chance of an abnormal sedimentation rate, and that the probability of an abnormal sedimentation rate increases over the course of the study. These results also agree with what is known about sedimentation rate: older subjects have
more sedimentation rate abnormalities, and subjects’ sedimentation rates increase over
time. In addition, the group variable still does not appear to be an important factor in
models for the probability of a subject having an abnormal sedimentation rate. In fact, the
estimates of the group effect are very close to those from the complete data models, while
the full model for the complete and drop-out data showed the group effect to be much
larger. Regardless, throughout all of the analyses conducted so far, a subject’s group
membership (Ranch Hand, Comparison) does not appear to be a factor in determining his
chances of an abnormal sedimentation rate. In general, all of the parameter estimates for
both β and α are quite similar to those from the complete data models; they are
relatively close to the parameter estimates of the models based on the complete and drop-out
data.

Again, we suspect that the asymptotics for the $X^2$ and $G^2$ statistics may not hold
here due to sparseness in our contingency tables. Our future work plan also includes
simulations to determine an exact test for our full models fit to the complete and drop-in
data. We will for now use the $X^2$ and $G^2$ statistics in order to assess fit of the full model
for the various association structures and for comparisons across the complete data, and
complete and drop-out data scenarios.

The first thing to note in Table 6.4 in regards to the model fit is that, unlike in the
models fit to the drop-out mechanism, it appears that the specific model used for the drop-in
mechanism makes a difference. In our Table 6.4 drop-in Model 7 that had group, study
year, age at cycle 4 and the officer military occupation indicator variable gives us the best
fit, for this choice of β and for all of our association models. The difference in $X^2$ values
on average is about 30 and the difference in $G^2$ values about 13 between drop-in Model 3 to Model 7 in Table 6.4.

In addition, we see that the All-Pair Model again appears to fit the data best, in terms of our $X^2$ and $G^2$ values. In fact, the difference in $X^2$ and $G^2$ is less than one for the All-Pair Model under drop-in Model 7. The next best model in terms of $X^2$ and $G^2$ is the (12,14,234) model, which also has a difference in $X^2$ and $G^2$ of less than one.

We see that, in general, the differences between $X^2$ and $G^2$ are smaller in the full model for the complete and drop-in data than they are in the complete data models and in the full model for complete and drop-out data. However, the $X^2$ and $G^2$ values are smaller in the full model for the complete and drop-out data (with little or no difference in degrees of freedom). The results of our simulations for the exact tests will help us to better compare the fits of the models. Although we found models on the drop-in process that appear to have an impact on the fit of the full model to the complete and drop-in data, we were unable to improve over the models that were fit using only the complete data, at least in terms of the $X^2$ and $G^2$ statistics.

6.3 The Likelihood Function for Drop-In or Drop-Out Data

So far, we have investigated methods that can handle complete data; drop-out data; and drop-in data. However, the ability to analyze data from all three patterns at once is probably more of interest to us. The models of Fitzmaurice and Laird (1993) and
Fitzmaurice, Laird and Zahner (1996) can be further extended to accommodate subjects with complete data, drop-in data or drop-out data. As with the drop-in models previously discussed, we will only consider models for the missingness that are MAR. Both the drop-in mechanism and the drop-out mechanism will have models that are MAR and, thus, ignorable. The main advantage of assuming a MAR model for the missingness is that we avoid the daunting task of determining the identifiability of the parameters in the models.

For the following, we assume that the data set contains two distinct groups: drop-ins and potential drop-outs. The drop-in group consists of those subjects who were not present at the time of the first observation. However, these subjects “dropped-into” the study at some time point and have been observed at every time point since then. Drop-in subjects have a monotone missing-data pattern such that if they are missing at time \( t \), then they are also missing at every time point \( t' < t \). The potential drop-out group consists of those subjects who were present at the time of the first observation. These subjects are either observed at all the remaining time periods (and, thus, have complete data), or are observed until they drop out of the study are not observed again after the time of drop out (and, thus, have drop-out data). The potential drop-out subjects have a monotone missing-data pattern such that if they are missing a response at time \( t \), then they are also missing responses at every time \( t' > t \). It is natural to include subjects with complete data into the “potential drop-out” group, and not the “drop-in” group, since the two groups are based on whether a subject was observed at the first time point. For ease, we will call the two groups drop-in and drop-out.
For analyzing the drop-in and drop-out men together, we need to make the following definitions and notations:

\[ \bar{R}_i = (R_{i1}, \ldots, R_{it}) \]
\[ \bar{R}_u = (R_{ut}, \ldots, R_{ur}) \]

\[ S_i = \sum_{t=1}^{T} R_{it} \text{ for each subject } i \text{ in the drop-out group} \quad (6.14a) \]

\[ U_i = \sum_{t=1}^{T} R_{it} \text{ for each subject } i \text{ in the drop-in group} \quad (6.14b) \]

In words, \( \bar{R}_i \) is the vector of response indicators for subject \( i \) from the beginning of the study through time \( t \), while \( \bar{R}_u \) is the vector of response indicators for subject \( i \) from time \( t \) through the end of the study. In addition, both \( S_i \) and \( U_i \) indicate the total number of observed responses for a subject. Since the models for the drop-out process most likely differ from those for the drop-in process, we use different names for these variables in order to distinguish between the drop-in and drop-out groups. We use the variables \( S_i \) and \( U_i \) instead of the variable \( R_i \) because they are easier to work with. Since there is a direct correspondence between a subject's \( R_i \) variable and his \( S_i \) or \( U_i \) variable, the missing-data model can be directly applied to \( S_i \) or \( U_i \). Also, let

\[ d_u \text{ be a vector of covariates at time } t \text{ for subject } i \text{ in the drop-out group} \]
\[ a_i \text{ be a vector of covariates at time } t \text{ for subject } i \text{ in the drop-in group} \]
\[ \lambda = (\beta, \alpha) \text{, where } \beta, \alpha \text{ are the parameters for the marginal and association models for } Y_i = (Y_{i1}, \ldots, Y_{it})' \text{ from Subsection 4.1.2.1.} \]
\( \nu \) be a vector of parameters of the drop-out process
and \( \phi \) be a vector of parameters of the drop-in process.

Note that the covariates included in the drop-out vector \( d_u \) are not necessarily the same
covariates included in the drop-in vector \( a_u \). That is, the factors that are related to a
subject dropping into the study may not be the same factors that are related to a subject
dropping out of the study. Finally, given \( R_i \), we partition \( Y_i \) and \( W_i \) for each subject \( i \)
such that

\[
Y_i = (Y_{(m1)_i}, Y_{oa}) \quad \text{if subject } i \text{ is a drop-in}
\]
\[
= Y_{oa} \quad \text{if subject } i \text{ has complete data}
\]
\[
= (Y_{oa}, Y_{(m2)_i}) \quad \text{if subject } i \text{ has drop-out data}
\]

and \( W_i = (W_{oa}, W_{mi}) \) if subject \( i \) has drop-in or drop-out data

\[
= W_{oa} \quad \text{if subject } i \text{ has complete data}
\]

where \( m1 \) indicates that the beginning observation(s) is(are) missing, \( m2 \) indicates that the
last observation(s) is(are) missing, the \( Y_{oa} \) correspond to the observed elements of \( Y_i \), and
the \( W_{oa} \) correspond to the elements of \( W_i \) that are observed and \( W_{mi} \) to the elements of \( W_i \)
that are missing. Likewise, let

\[
e(Y_i) := E_{Y_{(m1)_i}Y_{oa}} (Y_i) = (e(Y_{(m1)_i}), Y_{oa}) \quad \text{if subject } i \text{ is a drop-in}
\]
\[
= Y_{oa} \quad \text{if subject } i \text{ has complete data}
\]
\[
= (Y_{oa}, e(Y_{(m2)_i})) \quad \text{if subject } i \text{ has drop-out data}
\]

and

\[
(6.15)
\]
Let \( N_{di} \) be the total number of men with drop-in data and \( N_{do} \) the total number of men with drop-out data (drop-outs and complete subjects), with \( N = N_{di} + N_{do} \).

Since we are distinguishing between the drop-in and drop-out groups based on whether the first observation was missing, we need to make a slight change in the logistic regression model for the drop-in group to account for this conditioning. The model for the drop-out group is already based on the fact that the first observation is present, so no such change is needed. For the drop-out group, we assume the same logistic regression model as we had previously, so that \( \logit(P\{R_{it} = 1 | R_{t-1} = 1, \nu, d_{it}\}) = d_{it}' \nu \), for \( t=2,...,T \).

This implies that

\[
\tilde{p}_{it} := P\{R_{it} = 1 | R_{t-1} = 1, \nu, d_{it}\} = \frac{\exp\{d_{it}' \nu\}}{1 + \exp\{d_{it}' \nu\}}.
\]

(6.16)

For the drop-in group, we will assume a logistic regression model as we had previously, but conditional on the fact that the first observation is missing, in addition to the fact that the observation from the last time point is observed. Whereas our previous model, which handled both complete and drop-in cases, was based on \( P\{R_{it} = 1 | R_{t-1} = 1, \phi, a_{it}\} \), we will now assume

\[
\logit(P\{R_{it} = 1 | R_{t-1} = 1, R_{it} = 0, \phi, a_{it}\}) = a_{it}' \phi.
\]
for $t=1,...,T-1$, which leads to

$$
\overline{p}_{u} := \overline{P} \{ R_{u} = 1 | R_{(u-1)} = 1, R_{11} = 0, \phi, a_{u} \} = \frac{\exp \{ a_{u} \phi \}}{1 + \exp \{ a_{u} \phi \}} \quad \text{(6.17)}
$$

Finally, let $p_{i} = P \{ R_{i1} = 1 \}$ . Table 6.5 gives the observed patterns in the data, along with their respective contributions to the response probability, $f ( r_{i} | \nu, \phi )$ . Note that $m_{u}$ in Table 6.5 denotes observation $t$ as missing, while $y_{it}$ denotes it is observed.

For subject $i$, then, we have that the likelihood of the data is

$$
f ( y_{\alpha}, r_{i} | \lambda, \nu, \phi ) = f ( y_{\omega} | \lambda ) f ( r_{i} | \nu, \phi )
$$

which leads to

$$
\prod_{i=1}^{N} f ( y_{\omega}, r_{i} | \lambda, \nu, \phi ) = \left[ \prod_{i=1}^{N} f ( y_{\omega} | \lambda ) \right] \left[ \prod_{i=1}^{N} f ( r_{i} | \phi ) \right] \left[ \prod_{i=1}^{N} f ( r_{i} | \nu ) \right]
$$

$$
= \left[ \prod_{i=1}^{N} f ( y_{\omega} | \lambda ) \right] \left[ \prod_{i=1}^{N} f ( u_{i} | \phi ) \right] \left[ \prod_{i=1}^{N} f ( s_{i} | \nu ) \right]
$$

where the product for $f ( u_{i} | \phi )$ is over all the subjects with drop-in data, that for $f ( s_{i} | \nu )$ is over all those with drop-out data, and where $U_{i}$ and $S_{i}$ are given in (6.14a) and (6.14b).

Note that, for a drop-out, $S_{i} = \sum_{r=1}^{T} R_{ri} \leq T$. Then, for drop-out,

$$
P \{ R_{i} = 1 \} = P \{ R_{1}, R_{2} = 1, \ldots, R_{n} = 1, R_{(n-1)} = 0, \ldots, R_{T} = 0 \}
$$

142
\begin{align*}
&= P(R_{r_2} = 1, \ldots, R_{r_\eta} = 1, R_{\eta(r-\eta+1)} = 0, \ldots, R_{r_T} = 0 | R_{r_1} = 1) P(R_{r_1} = 1) \\
&= P(R_{r_2} = 1, \ldots, R_{r_\eta} = 1, R_{\eta(r-\eta+1)} = 0 | R_{r_1} = 1) p_1 \\
&= p_1 P(R_{r_2} = 1 | R_{r_1} = 1) P(R_{r_3} = 1 | R_{r_2} = 1) \ldots P(R_{\eta(r-\eta+1)} = 0 | R_{r_\eta} = 1) \\
&= p_1 \left( \prod_{j=2}^{\eta} P(R_{r_j} = 1 | R_{r_{j-1}} = 1) \right) \left( P(R_{\eta(r-\eta+1)} = 0 | R_{r_\eta} = 1) \right)^{T(r-\eta)} \\
&= p_1 \left( \prod_{j=2}^{\eta} \frac{\exp(d'_{y_j} \nu)}{1 + \exp(d'_{y_j} \nu)} \right) \left( \frac{1}{1 + \exp(d'_{T(r-\eta+1)} \nu)} \right)^{T(r-\eta)}.
\end{align*}

Since \( S_i = \sum_{r=1}^{T} R_{r} \), it follows that

\begin{align*}
P(S_i = s_i) &= p_1 \left( \prod_{j=2}^{\eta} \frac{\exp(d'_{y_j} \nu)}{1 + \exp(d'_{y_j} \nu)} \right) \left( \frac{1}{1 + \exp(d'_{T(r-\eta+1)} \nu)} \right)^{T(r-\eta)} \\
&= p_1 P(S_i = s_i). \quad (6.18)
\end{align*}

And for a drop-in, \( U_i = \sum_{r=1}^{T} R_{r} \) < \( T \) and

\begin{align*}
P(R_i = r_i) &= P(R_{r_1} = 0, \ldots, R_{\eta(r-\eta)} = 0, R_{\eta(r-\eta+1)} = 1, \ldots, R_{r_T} = 1) \\
&= P(R_{r_1} = 0) \\
&\times P(R_{r_2} = 0, \ldots, R_{\eta(r-\eta)} = 0, R_{\eta(r-\eta+1)} = 1, \ldots, R_{r_T} = 1 | R_{r_1} = 0) \\
&= (1 - p_1) P(R_{\eta(r-\eta)} = 0, R_{\eta(r-\eta+1)} = 1, \ldots, R_{r_T} = 1 | R_{r_1} = 0) \\
&\times P(R_{r_T} = 1 | R_{r_1} = 0) \\
&= (1 - p_1) P(R_{\eta(r-\eta)} = 0 | R_{r_1} = 0) \\
&\times P(R_{r_T} = 1 | R_{r_1} = 0) \\
&= (1 - p_1) P(R_{\eta(r-\eta)} = 0 | R_{r_1} = 0) \\
&\times P(R_{r_T} = 1 | R_{r_1} = 0).
\end{align*}

143
\[
\begin{align*}
\mathbf{P}(R_{(T-\eta+1)} &= \| R_{(T-\eta+2)} = 1, R_{11} = 0) \ldots P(R_{(T-1)} = \| R_{T+1} = 1, R_{11} = 0) \\
= \left(1 - p_1\right) P(R_{(T-\eta)} = 0 | R_{(T-\eta+1)} = 1, R_{11} = 0) \\
&= \left(1 - p_1\right) \left(\prod_{j=T-\eta+1}^{T-1} P(R_j = \| R_{(j+1)} = 1, R_{11} = 0) \right) \\
&= \left(1 - p_1\right) \left(\prod_{j=T-\eta+1}^{T-1} \frac{\exp(a_j \phi)}{1 + \exp(a_j \phi)} \right) \left(\frac{1}{1 + \exp(a_{(T-\eta)} \phi)} \right)^{I(T < T-1)}
\end{align*}
\]

which gives

\[
P(U_i = u_i) = \left(1 - p_1\right) \left(\prod_{j=T-\eta+1}^{T-1} \frac{\exp(a_j \phi)}{1 + \exp(a_j \phi)} \right) \left(\frac{1}{1 + \exp(a_j(T-\eta) \phi)} \right)^{I(U_i = T-1 )}
\]

Putting this all together, we have the likelihood function for the drop-in and drop-out data as

\[
\prod_{i=1}^{N} f(y_{\omega}, r_{i} | \lambda, \phi, \nu) = \left\{ \prod_{i=1}^{N} f(y_{\omega} | \lambda) \right\} \left\{ \prod_{i=1}^{N} f(u_{i} | \phi) \right\} \left\{ \prod_{i=1}^{N} f(s_{i} | \nu) \right\}
\]

\[
= \left\{ \prod_{i=1}^{N} f(y_{\omega} | \lambda) \right\} (1 - p_1)^{N_{\omega}} p_{i}^{N_{\omega}} \left[ \prod_{j=1}^{N_{\omega}} P(U_j = u_j) \right] \left[ \prod_{k=1}^{N_{\omega}} P^\prime (S_k = s_k) \right],
\]

so that

\[
I = \ln \left\{ \prod_{i=1}^{N} f(y_{\omega}, r_{i} | \lambda, \phi, \nu, p_{i}) \right\}
\]
\[ \sum_{i=1}^{N} \ln f(y_{ai} | \lambda) + N_{DI} \ln(1 - p_{1}) + N_{DO} \ln(p_{1}) + \sum_{j=1}^{N_{DO}} \ln P^\prime \{ U_{j} = u_{j} \} \]

\[ + \sum_{k=1}^{N_{DO}} \ln P^\prime \{ S_{k} = s_{k} \} , \text{where} \]

\[ P^\prime \{ U_{j} = u_{j} \} \text{ and } P^\prime \{ S_{k} = s_{k} \} \text{ are given in (6.18) and (6.19).} \]

The maximizations of \( \lambda, p_{1}, \phi, \) and \( \nu \) can all be done separately. The maximization for \( \lambda = (\beta, \alpha) \) is done the same way as before. The details of this maximization, which uses the EM-algorithm, are given in Section 6.1. Here, though, \( e(y_{i}) \) and \( e(w_{i}) \) are defined in (6.15). As before, \( X_{i} \) is completely specified for each subject \( i \), regardless if he has complete data, drop-in data, or drop-out data. The estimators \( \hat{\beta} \) will be consistent only if both the mean and the association models are correct. Thus, we need \( E(e(y_{i})) = \mu_{i} \) and \( E(e(w_{i})) = \gamma_{i} \). Again, as before, we can use the sample empirical covariance matrix as the estimator of the asymptotic variance-covariance matrix of \( (\hat{\beta}, \hat{\alpha}) \). The formula for this sample empirical covariance matrix is given in (6.13).

The maximization of \( p_{1} = P\{ R_{1} = 1 \} \) is quite easy. The likelihood equation for \( p_{1} \) is given by \( \frac{d l}{dp_{1}} = \frac{-N_{DI}}{1 - p_{1}} + \frac{N_{DO}}{p_{1}} = 0 \), which leads to \( \hat{p}_{1} = \frac{N_{DO}}{N_{DO} + N_{DI}} = \frac{N_{DO}}{N} \) as the MLE.

The maximizations for both \( \phi \) and \( \nu \) are carried out as they were previously, and are carried out separately. The details for maximizing \( \phi \) are given in Section 6.1. However, when we are considering data that is either drop-in or drop-out (i.e.: "drop-in" or "potential drop-out"), we need to make one adjustment to the maximization given in 145.
Section 6.1. In terms of subjects included in the drop-in mechanism analyses, we will consider only those subjects who drop-in sometime between time 2 and time T. Subjects with complete data are now considered to be part of the drop-out group. A subject with drop-in data contributes to the drop-in likelihood function for each time period he is observed and for the time period just prior to when he drops into the study. That is, a subject who drops-in at time $t$ contributes $(1 - \bar{p}_{(t-1)}) \bar{p}_u \bar{p}_{(t-1)} \cdots \bar{p}_{t}$ to the drop-in likelihood function. Again, as in Section 6.1, once each subject's contribution to the drop-in mechanism is determined, the MLE $\hat{\phi}$ is obtained through ordinary logistic regression, as in Section 6.1.

For maximizing $\nu$, we proceed as we did previously. When we had worked with drop-out data in Chapter 4, we included subjects with complete data in the drop-out likelihood function. Thus, we do not need to change how we proceed in determining the MLE $\hat{\nu}$. The one thing to remember here is that, unlike F-L-Z, we are only considering MAR models for the drop-in mechanism; this makes determining $\hat{\nu}$ much easier. Subjects with complete data contribute to the drop-out likelihood function for each time period, while those who drop-out contribute to the drop-out likelihood function for every time period for which they were observed and for the time period just following when they drop-out. Again, the MLE $\hat{\nu}$ is obtained through ordinary logistic regression. Any standard statistical package can be used to determine both $\hat{\phi}$ and $\hat{\nu}$.
6.4 Model Fitting to Drop-In or Drop-Out Data

I used the methodology that I developed in Section 6.3 to analyze sedimentation rate data from subjects who had either drop-in or drop-out data. From Section 6.2, we know there are 139 subjects who dropped into the study at cycle 2, 59 subjects who dropped in at cycle 3, and 61 subjects who dropped in at cycle 4, giving us a total of 259 subjects who are in the “drop-in” group. In addition, we know from Chapter 5 that there are 1881 subjects with sedimentation rate data from all 4 study cycles; 103 subjects who were observed at only cycle 1; 52 men observed at cycles 1 and 2; and 139 men observed at cycles 1, 2 and 3. This gives a total of 2175 subjects in the “potential drop-out” (or “drop-out”, for ease) group. In our notation from Section 6.3, we have $N_{D_{I}}=259$. $N_{D_{O}}=2175$ and $N = N_{D_{O}} + N_{D_{I}} = 2434$. We also have the total number of timepoints $T=4$.

As before, for subject $i$ at time $t$, the first element of the covariate vector $x_{it}$ used in the marginal response model was a 1 (for overall mean), the second element indicated the group of subject $i$, and the third element indicated the study year at time $t$. Group was defined as 1 for Ranch Hand subjects and as 0 for Comparison subjects, while study year was defined as 0 for the Baseline cycle in 1982, as 3 for cycle 2 in 1985, as 5 for cycle 3 in 1987 and as 10 for cycle 4 in 1992. The fourth element, the “age” variable, was based on age at cycle 1 if the subject was a drop-out and on age at cycle 4 if the subject was a drop-in. This ensured that $x_{it}$ was completely observed for all $i$ and all $t$. For drop-outs, “age”
was defined as 0 for subjects \( \leq 40 \) years of age at the time of their cycle 1 examination and as 1 for subjects \( >40 \) years of age at their cycle 1 examination. Since cycle 4 was 10 years after cycle 1, the analogous age cutpoint based on age at cycle 4 is 50. This led to the age variable being defined as 1 for a drop-in subject \( >50 \) at the time of his cycle 4 examination and as 0 for a drop-in subject \( \leq 50 \) years of age at his cycle 4 examination.

Each subject’s missing data indicator \( R_i = (R_{i1}, R_{i2}, R_{i3}, R_{i4})' \) was defined as in earlier analyses, with \( R_{it} = 1 \) when \( Y_{it} \) was observed and \( R_{it} = 0 \) when \( Y_{it} \) was missing. I then partitioned \( Y_i \) for subject \( i \) so that

\[
Y_i = (Y_{i1}, Y_{i2}) \quad \text{for drop-out subjects}
\]

\[
(Y_{i3}, Y_{i4}) \quad \text{for drop-in subjects.}
\]

Note that \( Y_i = Y_{i1} \) for drop-out subjects with complete data. Elements of \( Y_{i2} \) that represent “abnormal” sedimentation rates were denoted by a 1 and those representing “normal” sedimentation rates were denoted 0. The “abnormal” / “normal” cutpoints for the sedimentation rate data are given in Sections 2.5 and 5.1. The cross-classification of drop-out subjects is shown in Table 5.4. The cross-classification of drop-ins is the right three columns of each table in Table 6.2.

The number of observed responses for subject \( i \), if he was a drop-out, was defined by the variable \( S_i = \sum_{t=1}^{4} R_{it} \) and, if he was a drop-in, was defined by the variable

\[
U_i = \sum_{t=1}^{4} R_{it} . \quad \text{For each drop-out subject } i , \text{ I defined } \bar{R}_{it} = (R_{i1} , \ldots , R_{i4})' \text{ and}
\]
\[ \bar{Y}_t = (Y_{1t}, \ldots, Y_{nt})' \] and for each drop-in subject, I defined \( \bar{R}_t = (R_{1t}, \ldots, R_{nt})' \) and
\[ \bar{Y}_t = (Y_{1t}, \ldots, Y_{nt})'; \quad t=1, \ldots, 4. \]

Since we determined the MLE's for \( \lambda = (\beta, \alpha) \), \( \nu \) and \( \phi \) from separate maximizations of the likelihood function, we were able to somewhat follow what we had done in fitting models to the complete data; complete and drop-out data; and complete and drop-in data. I again used the models proposed by Fitzmaurice and Laird (1993) for \( f(y_i|\beta, \alpha) \). The EM-algorithm gave us "complete" data at each E-step of the algorithm and maximized the likelihood function with respect to \( \lambda = (\beta, \alpha) \) at each M-step. The "complete" data in the drop-in or drop-out case was more complicated than before since \( y_i \) could have missing observations in the initial time period(s), in the ending time period(s), or in no time periods. After the E-step gave us "complete" data
\[ e(y_i) = E_{\lambda, \nu}(Y_i) \quad \text{and} \quad e(w_i) = E_{\lambda, \nu}(W_i) \] based on the current value of \( \lambda = (\beta, \alpha) \), we used the two-step iterative procedure outlined in Section 4.1.2.3 to do the maximization in the M-step. As in all of the maximizations performed thus far to determine \( \hat{\lambda} \), I used the following convergence criteria: the IPF loop in the second step of the iterative procedure was considered to have converged when the sum of the absolute relative differences of expected cell counts of the 64 cells of the tables from two consecutive loops of the IPF was less than 0.01; the parameter estimates of \( \beta \) and \( \alpha \) were considered to have converged at step \( J \) when the sum of the absolute relative differences between the four terms of \( \beta^{(J-1)} \) and \( \beta^{(J)} \) was less than 0.01 and the (sum of the) absolute relative difference(s) between the term(s) of \( \alpha^{(J-1)} \) and \( \alpha^{(J)} \) was less than 0.01.
In determining \( \hat{\nu} \), recall that the model assumed for the drop-out process gives us probabilities of the form:

\[
\bar{p}_t := P(R_u = 1| \overline{R}_{t-1} = 1, d_u, \nu) = \frac{\exp(d_u^t \nu)}{1 + \exp(d_u^t \nu)},
\]

\( t=2,3,4 \), where \( d_u \) is a vector of completely observed covariates and \( \nu \) is a vector of parameters associated with the drop-out mechanism. Numerous covariates such as group, study year, age, and military occupation were included in the drop-out model to determine what variables are important in explaining a subject dropping out of the AFHS.

Since the models fit to the drop-out mechanism in Section 5.2 used both the complete and drop-out data to determine \( \hat{\nu} \), the results of fitting the logistic regression models in Table 5.5 are the same results we obtain here.

As we see from Table 5.5, group does not appear to be an important factor in a subject dropping out of (remaining fully compliant to) the AFHS. Again, since sedimentation rate data can only be determined for subjects who are fully compliant to the study, group membership is not important in whether a subject has sedimentation rate data over the entire course of the study. The study year and military occupation variables did appear to be important terms in the logistic regression models. The coefficient for the study year variable always had a negative estimate, indicating that the longer the study has been underway, the more likely a subject is to drop out. This is not surprising and also agrees with previous findings, as discussed in Section 5.2. The military occupation variable was included in the drop-out model as both an ordinal variable and as indicator variables. Regardless of the definition for military occupation, results always indicated
that groundcrew members are the most likely, and officers the least likely, to drop out of the study. Because the indicator variable models were comparable in terms of fit to the ordinal variable models and because the ordinal variable imposes a spacing between the occupations that may not be true, we prefer the indicator variable models.

In general, no one drop-out model had a markedly better fit than any of the other models. Although I found variables that appear to be important in the drop-out process, there is still some unexplained variability as to why subjects drop out of the study. However, we have quite a large sample size for fitting our drop-out model (6267 total \( R_{it} \)'s), which may partially account for some of this unexplained variability. Due to our large sample, we feel that the drop-out models do an adequate job in determining what variables are important in subjects dropping out of the AFHS.

For models fit to the drop-in mechanism, recall that we assumed the drop-in probabilities were of the form:

\[
\bar{p}_u := P(R_{it} = 1 | \bar{R}_{it-1} = 1, R_{it} = 0, a, \phi) = \frac{\exp\{a_\phi\}}{1 + \exp\{a_\phi\}},
\]

\( t = 2, 3 \), where \( a_u \) was a vector of completely observed covariates and \( \phi \) a vector of parameters associated with the drop-in process. To determine what factors are important in a subject dropping into the study, I included numerous different variables in the covariate vector \( a_u \), such as group, study year, age and military occupation. When we were fitting the drop-in models in Section 6.2, we considered complete and drop-in data in our drop-in mechanism. Here, however, our drop-in data set consists of only those subjects who have dropped into the study since the Baseline examination in 1982. Thus,
our data set for fitting the drop-in process was much smaller here. In addition, we only need to fit \( \overline{\beta}_n \) for \( t=2 \) and \( 3 \) (not for \( t=1,2, \) and \( 3 \) as in Section 6.2) because the \( \overline{\beta}_{t1} \) term from Section 6.2 is replaced by the \( p_1 = P(R_s = 1) \) term here.

Table 6.6 gives the results of some of the various logistic regression models fit to the drop-in mechanism using only those subjects who were not in the study at time one but dropped in later. The "hats" are left off the parameters for ease of notation. As we see from Table 6.6, group appears to be an important factor in subjects dropping into the AFHS. The positive coefficient for the parameter estimate indicates that Ranch Hand subjects drop into the study at earlier cycles and Comparisons at later cycles. This result is not surprising and agrees with earlier findings, as discussed in Section 6.2. It appears that the study year, age and military occupation variables are not important factors in subjects dropping into the study. This finding is opposite that in Section 6.2. In our present analyses, though, we are using only drop-in subjects (not complete and drop-in, as in Section 6.2) and we are only considering data based on cycles 2 and 3 (not cycles 1,2 and 3, as in Section 6.2). The much larger sample size when the complete data are used apparently hides this relationship.

In general, the model fits are good. As with the drop-in models fit in Section 6.2, we see that the models that include military occupation as indicator variables are comparable to those that include military occupation as an ordinal variable, in terms of the \( X^2 \) statistic. Since there is not much difference in the model fit and since the ordinal variable imposes a spacing that may not be true, we again prefer the drop-in models that define military occupation through indicator variables.
Again, as previously, we were interested in determining whether the model of random missingness was appropriate for the drop-in mechanism. That is, we wanted to know if \( \logit(P(R_2 = 1|\bar{R}_{i(t+1)} = 1, R_{i1} = 0, a, \phi)) = \phi_0 \) is reasonable. We computed the \( G^2 \) statistics for various nested models and then partitioned the statistics to determine what variables, if any, were worthwhile adding to the random missingness model. The group variable was the only variable that led to any substantial decrease in the value of the \( G^2 \) statistic. Although fewer variables than previously were added to the random missingness model, it does still appear that random missingness is not an appropriate assumption for the drop-in mechanism. Therefore, we should not ignore the missing-data (drop-in) process.

The full model for the drop-in or drop-out data combines the three sets of parameter estimates, those for \( \lambda = (\beta, \alpha) \), for \( \nu \) and for \( \phi \), in order to give expected cell counts under the specific model fit. As before, the drop-out models and the drop-in models that contain variables that are covariates in the marginal model for \( \beta \) (i.e., group or study year) are much more attractive computationally than those that include variables not in the marginal model for \( \beta \) (i.e., military occupation). Because the results for the complete data based on our maximization of \( (\beta, \alpha) \) are already classified by the covariates in the \( \beta \) model, drop-in and drop-out models that only include (some or all of) these same variables are easily combined with the \( (\beta, \alpha) \) models to give us the estimated expected cell counts for the full model for both the drop-in and drop-out data. When the drop-in and/or drop-out model includes variables not in the model for \( \beta \), the
computations for the expected cell counts for the full model become much more complicated.

Table 6.7 gives the results of the full model for drop-in and drop-out sedimentation rate data, for drop-in Model 1 from Table 6.6 and drop-out Models 4 and 8 from Table 5.5. Drop-in Model 1 contains the group variable. Drop-out Model 4 has group, study year and age, while drop-out Model 8 includes study year, age and the military occupation indicator variables. The marginal model for $\beta$ is the model that includes group, study year and age as covariates, and the association models are the same nine association models we have illustrated thus far in our results.

The parameter estimates in Table 6.7 agree with the earlier analyses on the complete, and the complete and drop-out data. Study year ($\beta_1$) and age ($\beta_2$) still appear to be important factors in modeling the probability of a subject having an abnormal sedimentation rate. The positive estimates for both variables indicate that older subjects have a higher probability of an abnormal sedimentation rate than do younger subjects, and that the probability of an abnormal sedimentation rate has increased over the course of the study. Again, these results agree with what is known about sedimentation rate: older subjects have more sedimentation rate abnormalities, and subjects' sedimentation rates tend to increase over time. The group variable still does not appear to be an important factor in models for the probability of a subject having an abnormal sedimentation rate. It is interesting to note that the coefficients for the group effect are very close to those from models fit to the complete and drop-out data (Table 5.6). These estimates are much larger than those from the models fit to the complete data (Table 5.2) and to the complete and
drop-in data (Table 6.4). Throughout all of the analyses so far, a subject’s group membership does not appear to be a factor in whether he has an abnormal sedimentation rate. We can see from Tables 5.2, 5.6, 6.4, 6.7 that, in general, the parameter estimates for $\beta$ and $\alpha$ are similar; although there are slight fluctuations in going from one set of data to another (e.g. complete data set to complete and drop-in data set), there are no marked differences in the parameter estimates across the nine association structures.

We suspect that the asymptotics for the $X^2$ and $G^2$ statistics may not hold here due to sparseness in our contingency tables. Our future work plan also includes simulations to determine an exact test for our full models fit to the drop-in and drop-out data. To assess fit of the full model for the various association structures and to compare to the previous models fit, we will use the $X^2$ and $G^2$ statistics for now.

First, note that the results in Table 6.7 indicate that the specific model used for the drop-out mechanism makes a difference in the fit of the model. This is in direct opposition to the results from fitting the full model to the complete and drop-out data in Section 5.2; there, the specific model used for the drop-out mechanism did not appear to have an impact on the goodness-of-fit. In Table 6.7, drop-out Model 4 that had group, year and age gives us the best fit for this choice of $\beta$ and all association models. The difference in $X^2$ values on average is about 40 and the difference in $G^2$ values about 22 between drop-out Model 8 and drop-out Model 4 in Table 6.7. In addition, when our drop-in model included variables other than the group variable (i.e. the drop-in model with group and military occupation as variables), the fit of the full model barely changed. This is again in opposition to previous results. When we fit the full model to the complete and drop-in
data in Section 6.2, the specific model used for the drop-in mechanism had an impact on the goodness-of-fit. Overall, for fitting the full model to the drop-in and drop-out sedimentation rate data from the AFHS, it appears that the drop-in model that includes the group variable and the drop-out model that had group, year and age give us the best fit for this choice of \( \beta \) and all association models.

As in all previous results, the All-Pair association model appears to fit the drop-in or drop-out data best, in terms of our \( X^2 \) and \( G^2 \) statistics. The difference between \( X^2 \) and \( G^2 \) is less than one for the All-Pair Model under drop-out Model 4 (and drop-in Model 1). The second-best fitting association model is not so clear-cut, with either the (12,14,234) model or the (13,14,234) model as possibilities. It is interesting to note that the (12,14,234) model has the second-best fit in the full model for the complete and drop-in data, while the (13,14,234) model has the second-best fit in the full model for the complete and drop-out data.

We also see that the differences between \( X^2 \) and \( G^2 \) are smaller for the first four association models in the full model for the drop-in or drop-out data than in the full model for the complete and drop-in data, but not the last four association models. The difference between \( X^2 \) and \( G^2 \) is basically the same for the All-Pair association model for the two different full models. The differences between \( X^2 \) and \( G^2 \) are smaller in all nine association models in the full model for the drop-in or drop-out data than they are in the full model for complete and drop-out data. However, in going from the full model for either complete and drop-in or complete and drop-out data to the full model for drop-in or drop-out data, for each specific association model, the increase in degrees of freedom is
more than offset by the increase in the $X^2$ and $G^2$ values (e.g. an increase in degrees of freedom of approximately 55 along with an increase in $X^2$ of approximately 160 and an increase in $G^2$ of approximately 180 in going from complete and drop-out to drop-in or drop-out. Again, the results of our simulations for the exact tests will help us to better compare the fits of the models. It appears, for now, that although we found models for the missing-data mechanisms that impact the fit of the full model to the drop-in or drop-out data, we did not improve over the models that were fit to sedimentation rate using only complete data.
### Time Periods Likelihood Contribution to the Drop-In Model

<table>
<thead>
<tr>
<th>i</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>T-2</th>
<th>T-1</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_{i1}$</td>
<td>$y_{i2}$</td>
<td>$y_{i3}$</td>
<td>...</td>
<td>$y_{i(T-2)}$</td>
<td>$y_{i(T-1)}$</td>
<td>$y_{iT}$</td>
<td>$\overline{P}<em>{i1} \cdot \overline{P}</em>{i2} \cdot \overline{P}<em>{i3} \cdot \ldots \cdot \overline{P}</em>{i(T-1)}$</td>
</tr>
<tr>
<td>$m_{i1}$</td>
<td>$y_{i2}$</td>
<td>$y_{i3}$</td>
<td>...</td>
<td>$y_{i(T-2)}$</td>
<td>$y_{i(T-1)}$</td>
<td>$y_{iT}$</td>
<td>$(1 - \overline{P}<em>{i1}) \cdot \overline{P}</em>{i2} \cdot \overline{P}<em>{i3} \cdot \ldots \cdot \overline{P}</em>{i(T-1)}$</td>
</tr>
<tr>
<td>$m_{i1}$</td>
<td>$m_{i2}$</td>
<td>$y_{i3}$</td>
<td>...</td>
<td>$y_{i(T-2)}$</td>
<td>$y_{i(T-1)}$</td>
<td>$y_{iT}$</td>
<td>$(1 - \overline{P}<em>{i1}) \cdot \overline{P}</em>{i2} \cdot \overline{P}<em>{i3} \cdot \ldots \cdot \overline{P}</em>{i(T-1)}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$m_{i1}$</td>
<td>$m_{i2}$</td>
<td>$m_{i3}$</td>
<td>...</td>
<td>$m_{i(T-2)}$</td>
<td>$y_{i(T-1)}$</td>
<td>$y_{iT}$</td>
<td>$(1 - \overline{P}<em>{i(T-2)}) \cdot \overline{P}</em>{i(T-1)}$</td>
</tr>
<tr>
<td>$m_{i1}$</td>
<td>$m_{i2}$</td>
<td>$m_{i3}$</td>
<td>...</td>
<td>$m_{i(T-2)}$</td>
<td>$m_{i(T-1)}$</td>
<td>$y_{iT}$</td>
<td>$(1 - \overline{P}_{i(T-1)})$</td>
</tr>
</tbody>
</table>

Note: $\overline{P}_{it} = P(R_{it} = 1) = 1$ by design of the drop-in data set.

- $m_{it}$ indicates subject i is missing observation at time t

Table 6.1: Drop-in Patterns and Likelihood Contributions
\((x_{n2}, x_{n4}) = (0,0)\) *

Subjects with: \(Y_1, Y_2, Y_3, Y_4\) \(Y_2, Y_3, Y_4\) \(Y_3, Y_4\) \(Y_4\)

<table>
<thead>
<tr>
<th>(Y_4)</th>
<th>(Y_3)</th>
<th>(Y_2)</th>
<th>(Y_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>322</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

\((x_{n2}, x_{n4}) = (0,1)\) *

Subjects with: \(Y_1, Y_2, Y_3, Y_4\) \(Y_2, Y_3, Y_4\) \(Y_3, Y_4\) \(Y_4\)

<table>
<thead>
<tr>
<th>(Y_4)</th>
<th>(Y_3)</th>
<th>(Y_2)</th>
<th>(Y_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>486</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>72</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

\* \((x_{n2}, x_{n4}) = \) (group of subject i, age at cycle 1 of subject i)

(continued)

Table 6.2: Erythrocyte Sedimentation Rate by Group and Age at cycle 1 for Subjects with Drop-in Data

159
Table 6.2 (continued)

\[(x_{i2}, x_{i4}) = (1,0) \]

Subjects with : \[Y_1, Y_2, Y_3, Y_4 \quad Y_2, Y_3, Y_4 \quad Y_3, Y_4 \quad Y_4\]

<table>
<thead>
<tr>
<th>(Y_1)</th>
<th>(Y_4)</th>
<th>(Y_3)</th>
<th>(Y_2)</th>
<th>(Y_4)</th>
<th>(Y_3)</th>
<th>(Y_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>273</td>
<td>0</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[(x_{i2}, x_{i4}) = (1,1) \]

Subjects with : \[Y_1, Y_2, Y_3, Y_4 \quad Y_2, Y_3, Y_4 \quad Y_3, Y_4 \quad Y_4\]

<table>
<thead>
<tr>
<th>(Y_1)</th>
<th>(Y_4)</th>
<th>(Y_3)</th>
<th>(Y_2)</th>
<th>(Y_4)</th>
<th>(Y_3)</th>
<th>(Y_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>406</td>
<td>3</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[(x_{i2}, x_{i4}) = (group\ of\ subject\ i,\ age\ at\ cycle\ 1\ of\ subject\ i)\]

160
<table>
<thead>
<tr>
<th>Drop-in Model</th>
<th>Parameter Estimates</th>
<th>$\chi^2$</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\phi_0 = 2.3282 (.1010)$</td>
<td>7.46</td>
<td>2</td>
</tr>
<tr>
<td>$\phi_0 + \phi_1 \text{ (group)}$</td>
<td>$\phi_1 = 0.9406 (.2028)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+$\phi_2 \text{ (study year)}$</td>
<td>$\phi_2 = 0.1796 (.0346)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+$\phi_3 \text{ (group*study year)}$</td>
<td>$\phi_3 = 0.2100 (.0884)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$\phi_0 = 1.9911 (.1140)$</td>
<td>21.57</td>
<td>8</td>
</tr>
<tr>
<td>$\phi_0 + \phi_1 \text{ (group)}$</td>
<td>$\phi_1 = 1.2672 (.1607)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+$\phi_2 \text{ (study year)}$</td>
<td>$\phi_2 = 0.2170 (.0315)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+$\phi_3 \text{ (age at cycle 4)}$</td>
<td>$\phi_3 = 0.4817 (.1280)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$\phi_0 = 2.0638 (.1200)$</td>
<td>18.04</td>
<td>7</td>
</tr>
<tr>
<td>$\phi_0 + \phi_1 \text{ (group)}$</td>
<td>$\phi_1 = 0.9382 (.2031)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+$\phi_2 \text{ (study year)}$</td>
<td>$\phi_2 = 0.1810 (.0347)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+$\phi_3 \text{ (age at cycle 4)}$</td>
<td>$\phi_3 = 0.4805 (.1283)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+$\phi_4 \text{ (group*study year)}$</td>
<td>$\phi_4 = 0.2093 (.0885)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>$\phi_0 = 1.8038 (.2005)$</td>
<td>98.4</td>
<td>32</td>
</tr>
<tr>
<td>$\phi_0 + \phi_1 \text{ (group)}$</td>
<td>$\phi_1 = 1.2469 (.1606)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+$\phi_2 \text{ (age at cycle 4)}$</td>
<td>$\phi_2 = 0.6896 (.1398)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+$\phi_3 \text{ (m.o.)}$</td>
<td>$\phi_3 = 0.2897 (.0759)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Drop-in model $P\{R_{ij} = 1| a \mu, \phi, \overline{R}_{ij-1} = 1\} = a \mu^{\prime} \phi$

$\phi_0 = \text{coefficient of overall mean}$

m.o. = military occupation (defined as 1 for officers, 2 for flight engineers, 3 for groundcrew)

Table 6.3: Parameter Estimates, $\chi^2$ Statistics and Degrees of Freedom for Logistic Regression Models Fit to the Drop-in Mechanism (Standard Errors in Parentheses)
<table>
<thead>
<tr>
<th>Drop-in Model</th>
<th>Parameter Estimates</th>
<th>( X^2 )</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 ( \phi_0 + \phi_1 ) (group)</td>
<td>( \phi_0 = 1.2661 (.2127) )</td>
<td>54.81</td>
<td>31</td>
</tr>
<tr>
<td>+( \phi_2 ) (study year)</td>
<td>( \phi_1 = 1.2694 (.1607) )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+( \phi_3 ) (age at cycle 4)</td>
<td>( \phi_2 = 0.2188 (.0315) )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+( \phi_4 ) (m.o.)</td>
<td>( \phi_3 = 0.7075 (.1406) )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \phi_0 ) (group)</td>
<td>( \phi_4 = 0.2981 (.0763) )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| 6 \( \phi_0 + \phi_1 \) (group) | \( \phi_0 = 1.3400 (.2163) \) | 48.13 | 30 |
| +\( \phi_2 \) (study year) | \( \phi_1 = 0.9410 (.2034) \) |  |  |
| +\( \phi_3 \) (age at cycle 4) | \( \phi_2 = 0.1828 (.0348) \) |  |  |
| +\( \phi_4 \) (m.o.) | \( \phi_3 = 0.7061 (.1406) \) |  |  |
| +\( \phi_5 \) (group*study year) | \( \phi_4 = 0.7061 (.1406) \) |  |  |

| 7 \( \phi_0 + \phi_1 \) (group) | \( \phi_0 = 2.1309 (.1204) \) | 36.27 | 19 |
| +\( \phi_2 \) (study year) | \( \phi_1 = 1.2676 (.1607) \) |  |  |
| +\( \phi_3 \) (age at cycle 4) | \( \phi_2 = 0.2191 (.0315) \) |  |  |
| +\( \phi_4 \) \( I_{(officer)} \) | \( \phi_3 = 0.6828 (.1374) \) |  |  |
| +\( \phi_5 \) | \( \phi_4 = 0.5928 (.1377) \) |  |  |

Note: Drop-in model \( P \{ R_{i} = 1 | a_{\mu}, \phi, \overline{R}_{i-1} = 1 \} = a_{\mu} \cdot \phi \)

\( \phi_0 \) = coefficient of overall mean
m.o. = military occupation (defined as 1 for officers, 2 for flight engineers, 3 for groundcrew)
Parameter Estimates

<table>
<thead>
<tr>
<th>Model</th>
<th>Marginal Estimate</th>
<th>Association Estimate</th>
<th>Drop-in Model 3</th>
<th>Drop-in Model 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\beta_0 = -3.8995 (1.130)$</td>
<td></td>
<td>$G^2 = 780.45$</td>
<td>$G^2 = 766.87$</td>
</tr>
<tr>
<td></td>
<td>$\beta_1 = 0.0865 (0.0864)$</td>
<td>$\alpha_1 = 0.0865 (0.0864)$</td>
<td>$X^2 = 14322.95$</td>
<td>$X^2 = 14309.10$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.2046 (0.0122)$</td>
<td>$\alpha_2 = 0.2046 (0.0122)$</td>
<td>d.f. = 107</td>
<td>d.f. = 107</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.3196 (0.0912)$</td>
<td>$\alpha_3 = 0.3196 (0.0912)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$\beta_0 = -3.8407 (1.307)$</td>
<td>$\alpha_1 = 2.7192 (3.093)$</td>
<td>$G^2 = 367.96$</td>
<td>$G^2 = 354.38$</td>
</tr>
<tr>
<td>1st-order</td>
<td>$\beta_1 = 0.0291 (1.031)$</td>
<td>$\alpha_2 = 3.6061 (2.360)$</td>
<td>$X^2 = 487.58$</td>
<td>$X^2 = 449.03$</td>
</tr>
<tr>
<td>Markov</td>
<td>$\beta_2 = 0.2171 (0.0115)$</td>
<td>$\alpha_3 = 2.9171 (0.0208)$</td>
<td>d.f. = 104</td>
<td>d.f. = 104</td>
</tr>
<tr>
<td>Chain</td>
<td>$\beta_3 = 0.2191 (1.056)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$\beta_0 = -3.8154 (1.333)$</td>
<td>$\alpha_1 = 2.1856 (5.847)$</td>
<td>$G^2 = 297.11$</td>
<td>$G^2 = 283.53$</td>
</tr>
<tr>
<td>2nd order</td>
<td>$\beta_1 = 0.0343 (1.072)$</td>
<td>$\alpha_2 = 2.2638 (2.079)$</td>
<td>$X^2 = 366.71$</td>
<td>$X^2 = 332.46$</td>
</tr>
<tr>
<td>Markov</td>
<td>$\beta_2 = 0.2149 (0.0111)$</td>
<td>$\alpha_3 = 2.7744 (0.4995)$</td>
<td>d.f. = 100</td>
<td>d.f. = 100</td>
</tr>
<tr>
<td>Chain</td>
<td>$\beta_3 = 0.1816 (1.095)$</td>
<td>$\alpha_4 = 1.9038 (3.315)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_5 = 2.4465 (0.2327)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_6 = -1.0401 (0.7420)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_7 = -0.3349 (0.5773)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>$\beta_0 = -3.8225 (1.333)$</td>
<td>$\alpha_1 = 1.6148 (0.3964)$</td>
<td>$G^2 = 299.64$</td>
<td>$G^2 = 286.04$</td>
</tr>
<tr>
<td>(12,13,23,</td>
<td>$\beta_1 = 0.0364 (1.072)$</td>
<td>$\alpha_2 = 1.9409 (0.7350)$</td>
<td>$X^2 = 386.77$</td>
<td>$X^2 = 354.41$</td>
</tr>
<tr>
<td>24,34)</td>
<td>$\beta_2 = 0.2152 (0.0111)$</td>
<td>$\alpha_3 = 2.3009 (0.2727)$</td>
<td>d.f. = 102</td>
<td>d.f. = 102</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.1829 (1.095)$</td>
<td>$\alpha_4 = 1.8276 (0.2737)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_5 = 2.3813 (0.2118)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Models for Drop-in Mechanism:

3. $P\{R_{it} = 1|\text{group}, \text{study year, age@cycle 4, } \phi, \bar{R}_{i,t+1} = 1\} = \phi_0 + \phi_1 (\text{group}) + \phi_2 (\text{study year}) + \phi_3 (\text{age@cycle 4}) + \phi_4 (\text{group*study year})$

7. $P\{R_{it} = 1|\text{group}, \text{study year, age@cycle 4, } I_{\text{officer}}, \phi, \bar{R}_{i,t+1} = 1\} = \phi_0 + \phi_1 (\text{group}) + \phi_2 (\text{study year}) + \phi_3 (\text{age@cycle 4}) + \phi_4 I_{\text{officer}}$

(continued)

Table 6.4: Parameter Estimates, $G^2$ and $X^2$ Statistics and Degrees of Freedom for Full Model Fit to Complete and Drop-in Sedimentation Rate Data (Standard errors in parentheses)
<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter Estimates</th>
<th>Drop-in</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Model 3</td>
<td>Model 7</td>
</tr>
<tr>
<td>5</td>
<td>$\beta_0 = -3.8162 \quad (.1335)$</td>
<td>$\alpha_{12} = 1.1206 \quad (.3966)$</td>
<td>$G^2 = 267.87$</td>
<td>$G^2 = 254.29$</td>
</tr>
<tr>
<td>All-Pair</td>
<td>$\beta_1 = 0.0608 \quad (.1090)$</td>
<td>$\alpha_{13} = 1.2418 \quad (.3907)$</td>
<td>$X^2 = 287.30$</td>
<td>$X^2 = 253.41$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.2116 \quad (.0108)$</td>
<td>$\alpha_{14} = 1.7430 \quad (.3634)$</td>
<td>d.f. = 101</td>
<td>d.f. = 101</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.1654 \quad (.1113)$</td>
<td>$\alpha_{23} = 2.4938 \quad (.2710)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_{24} = 1.6638 \quad (.2765)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_{34} = 2.1680 \quad (.2139)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>$\beta_0 = -3.7822 \quad (.1293)$</td>
<td>$\alpha_{13} = 2.2140 \quad (.6623)$</td>
<td>$G^2 = 314.59$</td>
<td>$G^2 = 301.01$</td>
</tr>
<tr>
<td>(23,134)</td>
<td>$\beta_1 = 0.0642 \quad (.1061)$</td>
<td>$\alpha_{14} = 1.9808 \quad (.3825)$</td>
<td>$X^2 = 446.08$</td>
<td>$X^2 = 406.01$</td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.2065 \quad (.0107)$</td>
<td>$\alpha_{23} = 3.6194 \quad (.2339)$</td>
<td>d.f. = 102</td>
<td>d.f. = 102</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.1953 \quad (.1087)$</td>
<td>$\alpha_{24} = 2.6308 \quad (.2049)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_{34} = 0.5950 \quad (.7521)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>$\beta_0 = -3.8254 \quad (.1318)$</td>
<td>$\alpha_{12} = 1.7456 \quad (.3421)$</td>
<td>$G^2 = 279.76$</td>
<td>$G^2 = 266.18$</td>
</tr>
<tr>
<td>(12,14,</td>
<td>$\beta_1 = 0.0532 \quad (.1086)$</td>
<td>$\alpha_{13} = 1.9930 \quad (.3355)$</td>
<td>$X^2 = 302.56$</td>
<td>$X^2 = 266.64$</td>
</tr>
<tr>
<td>234)</td>
<td>$\beta_2 = 0.2133 \quad (.0106)$</td>
<td>$\alpha_{14} = 3.0091 \quad (.4627)$</td>
<td>d.f. = 101</td>
<td>d.f. = 101</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.1734 \quad (.1110)$</td>
<td>$\alpha_{23} = 1.5993 \quad (.3391)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_{24} = 2.4383 \quad (.2339)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_{34} = 0.4087 \quad (.5596)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>$\beta_0 = -3.7850 \quad (.1306)$</td>
<td>$\alpha_{13} = 1.7246 \quad (.3288)$</td>
<td>$G^2 = 276.68$</td>
<td>$G^2 = 263.09$</td>
</tr>
<tr>
<td>(13,14,</td>
<td>$\beta_1 = 0.0565 \quad (.1088)$</td>
<td>$\alpha_{14} = 1.8876 \quad (.3447)$</td>
<td>$X^2 = 316.26$</td>
<td>$X^2 = 279.85$</td>
</tr>
<tr>
<td>234)</td>
<td>$\beta_2 = 0.2082 \quad (.0103)$</td>
<td>$\alpha_{23} = 2.9862 \quad (.4548)$</td>
<td>d.f. = 101</td>
<td>d.f. = 101</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.1710 \quad (.1112)$</td>
<td>$\alpha_{24} = 1.9034 \quad (.3321)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_{34} = 2.1026 \quad (.2344)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_{34} = 0.3520 \quad (.5510)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>$\beta_0 = -3.8257 \quad (.1318)$</td>
<td>$\alpha_{12} = 2.0082 \quad (.8667)$</td>
<td>$G^2 = 279.97$</td>
<td>$G^2 = 266.39$</td>
</tr>
<tr>
<td>(23,34,</td>
<td>$\beta_1 = 0.0525 \quad (.1087)$</td>
<td>$\alpha_{13} = 2.0284 \quad (.3597)$</td>
<td>$X^2 = 309.81$</td>
<td>$X^2 = 275.84$</td>
</tr>
<tr>
<td>124)</td>
<td>$\beta_2 = 0.2134 \quad (.0106)$</td>
<td>$\alpha_{14} = 2.6904 \quad (.2677)$</td>
<td>d.f. = 101</td>
<td>d.f. = 101</td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.1729 \quad (.1111)$</td>
<td>$\alpha_{23} = 1.5448 \quad (.2888)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_{24} = 2.3682 \quad (.2133)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\alpha_{34} = 0.3117 \quad (.9411)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Table 6.5: Drop-in and Drop-out Data Patterns and Likelihood Contributions

<table>
<thead>
<tr>
<th>Time Periods</th>
<th>Likelihood Contribution to the Missing-Data Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>T-2</td>
<td></td>
</tr>
<tr>
<td>T-1</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>

- $m_{i1}$, $m_{i2}$, $m_{i3}$, ..., $m_{i(T-2)}$, $m_{i(T-1)}$, $y_{iT}$
- $(1 - p_1)(1 - \overline{p}_{i(T-1)})$
- $m_{i1}$, $m_{i2}$, $m_{i3}$, ..., $m_{i(T-2)}$, $y_{i(T-1)}$, $y_{iT}$
- $(1 - p_1)\overline{p}_{i(T-1)}(1 - \overline{p}_{i(T-2)})$
- ...
- $m_{i1}$, $y_{i2}$, $y_{i3}$, ..., $y_{i(T-2)}$, $y_{i(T-1)}$, $y_{iT}$
- $(1 - p_1)\overline{p}_{i(T-1)}\overline{p}_{i(T-2)} ... \overline{p}_{i2}$
- $y_{i1}$, $y_{i2}$, $y_{i3}$, ..., $y_{i(T-2)}$, $y_{i(T-1)}$, $y_{iT}$
- $p_1\overline{p}_{i2}\overline{p}_{i3} ... \overline{p}_{i(T-1)}\overline{p}_{iT}$
- $y_{i1}$, $y_{i2}$, $y_{i3}$, ..., $y_{i(T-2)}$, $y_{i(T-1)}$, $m_{iT}$
- $p_1\overline{p}_{i2}\overline{p}_{i3} ... \overline{p}_{i(T-1)}(1 - \overline{p}_{iT})$
- ...
- $y_{i1}$, $m_{i2}$, $m_{i3}$, ..., $m_{i(T-2)}$, $m_{i(T-1)}$, $m_{iT}$
- $p_i(1 - \overline{p}_{i2})$

**Note:** $m_{it}$ indicates subject $i$ is missing observation at time $t$.
### Drop-in Model Parameter Estimates

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter Estimates</th>
<th>$X^2$</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\phi_0 = 0.9201 (0.1159)$</td>
<td>$X^2 = 6.29$</td>
<td></td>
</tr>
<tr>
<td>$\phi_0 + \phi_1 \text{ (group)}$</td>
<td>$\phi_1 = 0.6380 (0.2984)$</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>$\phi_0 = 0.2070 (0.4541)$</td>
<td>$X^2 = 3.71$</td>
<td></td>
</tr>
<tr>
<td>$\phi_0 + \phi_1 \text{ (group)}$</td>
<td>$\phi_1 = 0.6610 (0.2996)$</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>$+ \phi_2 \text{ (study year)}$</td>
<td>$\phi_2 = -0.1731 (0.1075)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$\phi_0 = 0.1771 (0.4682)$</td>
<td>$X^2 = 12.57$</td>
<td></td>
</tr>
<tr>
<td>$\phi_0 + \phi_1 \text{ (group)}$</td>
<td>$\phi_1 = 0.6595 (0.2997)$</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>$+ \phi_2 \text{ (study year)}$</td>
<td>$\phi_2 = 0.1737 (0.1075)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$+ \phi_3 \text{ (age at cycle 4)}$</td>
<td>$\phi_3 = 0.0563 (0.2144)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>$\phi_0 = 0.8656 (0.2487)$</td>
<td>$X^2 = 12.89$</td>
<td></td>
</tr>
<tr>
<td>$\phi_0 + \phi_1 \text{ (group)}$</td>
<td>$\phi_1 = 0.6392 (0.2985)$</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>$+ \phi_2 \text{ (m.o.)}$</td>
<td>$\phi_2 = 0.0282 (0.1141)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$\phi_0 = 0.9157 (0.1734)$</td>
<td>$X^2 = 12.21$</td>
<td></td>
</tr>
<tr>
<td>$\phi_0 + \phi_1 \text{ (group)}$</td>
<td>$\phi_1 = 0.6579 (0.2996)$</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>$+ \phi_2 I_{\text{officer}}$</td>
<td>$\phi_2 = -0.0515 (0.2267)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$+ \phi_3 I_{\text{flight engineer}}$</td>
<td>$\phi_3 = 0.2351 (0.3706)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** Drop-in model $P\{ R_{it} = 1 | \alpha_i, \phi, \vec{R}_{i,t-1} = 1, R_{i1} = 0 \} = \alpha_i \cdot \phi$

- $\phi_0$ = coefficient of overall mean
- m.o. = military occupation (defined as 1 for officers, 2 for flight engineers, 3 for groundcrew)

Table 6.6: Results of Logistic Regression Models Fit to the Drop-in Mechanism in Fitting the Full Model to the Drop-in or Drop-out Data (Standard Errors in Parentheses)
### Parameter Estimates

<table>
<thead>
<tr>
<th>Model</th>
<th>Marginal</th>
<th>Association</th>
<th>Model 4</th>
<th>Model 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\beta_0 = -3.7783 (.1020)$</td>
<td>$G^2 = 967.25$</td>
<td>$G^2 = 989.38$</td>
<td></td>
</tr>
<tr>
<td>Independence</td>
<td>$\beta_1 = 0.1685 (.0784)$</td>
<td>$X^2 = 10677.45$</td>
<td>$X^2 = 10766.15$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.1816 (.0109)$</td>
<td>d.f. = 162</td>
<td>d.f. = 161</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.4232 (.0851)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$\beta_0 = -3.7109 (.1174)$</td>
<td>$G^2 = 511.13$</td>
<td>$G^2 = 533.26$</td>
<td></td>
</tr>
<tr>
<td>1st-order</td>
<td>$\alpha_{12} = 2.6903 (.2566)$</td>
<td>$X^2 = 668.40$</td>
<td>$X^2 = 726.57$</td>
<td></td>
</tr>
<tr>
<td>Markov</td>
<td>$\beta_1 = 0.1132 (.0940)$</td>
<td>d.f. = 159</td>
<td>d.f. = 158</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{23} = 3.5740 (.2068)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chain</td>
<td>$\beta_2 = 0.1958 (.0101)$</td>
<td>$G^2 = 431.89$</td>
<td>$G^2 = 454.02$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{34} = 2.8499 (.1750)$</td>
<td>d.f. = 155</td>
<td>d.f. = 154</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.3373 (.0977)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$\beta_0 = -3.6883 (.1197)$</td>
<td>$G^2 = 431.89$</td>
<td>$G^2 = 454.02$</td>
<td></td>
</tr>
<tr>
<td>2nd order</td>
<td>$\alpha_{12} = 2.2583 (.4790)$</td>
<td>$X^2 = 466.49$</td>
<td>$X^2 = 499.86$</td>
<td></td>
</tr>
<tr>
<td>Markov</td>
<td>$\beta_1 = 0.1230 (.0980)$</td>
<td>d.f. = 157</td>
<td>d.f. = 156</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{13} = 2.4101 (.3362)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chain</td>
<td>$\beta_2 = 0.1947 (.0096)$</td>
<td>$G^2 = 436.29$</td>
<td>$G^2 = 458.42$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{24} = 1.8752 (.2947)$</td>
<td>d.f. = 156</td>
<td>d.f. = 155</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{34} = 2.4195 (.2104)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{123} = -1.2912 (.2393)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{234} = -0.3546 (.4933)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>$\beta_0 = -3.6956 (.1198)$</td>
<td>$\alpha_{12} = 1.5086 (.3281)$</td>
<td>$G^2 = 436.29$</td>
<td>$G^2 = 458.42$</td>
</tr>
<tr>
<td>(12,13,23,24,34)</td>
<td>$\alpha_{13} = 2.0243 (.3118)$</td>
<td>$X^2 = 493.37$</td>
<td>$X^2 = 528.05$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_2 = 0.1949 (.0096)$</td>
<td>d.f. = 157</td>
<td>d.f. = 156</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{23} = 2.3163 (.2329)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\beta_3 = 0.3087 (.1017)$</td>
<td>$\alpha_{24} = 1.7920 (.2388)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\alpha_{34} = 2.3451 (.1894)$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Drop-in model: $P\{ R_{it}=1|\phi, R_{i,t-1}=1, R_{i1}=0 \} = \phi_0 + \phi_1 (\text{group})$

Drop-out models:

4. $P\{ R_{it}=1|\text{group, study year, age, }\nu, R_{i,t-1}=1\} = \nu_0 + \nu_1 (\text{group}) + \nu_2 (\text{study year}) + \nu_3 (\text{age})$

8. $P\{ R_{it}=1|\text{study year, age, military occupation, }\nu, R_{i,t-1}=1\} = \nu_0 + \nu_1 I_{(\text{officer})} + \nu_2 I_{(\text{engineer})} + \nu_3 (\text{study year}) + \nu_4 (\text{age})$

(continued)

Table 6.7: Parameter Estimates, $G^2$ and $X^2$ Statistics and Degrees of Freedom for Full Models Fit to Sedimentation Rate Data (Standard errors in parentheses)
<table>
<thead>
<tr>
<th>Model</th>
<th>Marginal</th>
<th>Association</th>
<th>Model 4</th>
<th>Model 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>( \beta_0 = -3.6975 (.1200) )</td>
<td>( \alpha_{12} = 1.0300 (.3301) )</td>
<td>( G^2 = 404.22 )</td>
<td>( G^2 = 426.35 )</td>
</tr>
<tr>
<td>All-Pair</td>
<td>( \beta_1 = 0.1502 (.0997) )</td>
<td>( \alpha_{13} = 1.3181 (.3280) )</td>
<td>( X^2 = 404.98 )</td>
<td>( X^2 = 440.61 )</td>
</tr>
<tr>
<td></td>
<td>( \beta_2 = 0.1926 (.0092) )</td>
<td>( \alpha_{14} = 1.7491 (.3118) )</td>
<td>d.f. = 156</td>
<td>d.f. = 155</td>
</tr>
<tr>
<td></td>
<td>( \beta_3 = 0.2965 (.1034) )</td>
<td>( \alpha_{23} = 2.4911 (.2377) )</td>
<td>( \alpha_{24} = 1.6457 (.2432) )</td>
<td>( \alpha_{34} = 2.1423 (.1929) )</td>
</tr>
<tr>
<td>6</td>
<td>( \beta_0 = -3.6732 (.1167) )</td>
<td>( \alpha_{13} = 2.2650 (.5403) )</td>
<td>( G^2 = 451.21 )</td>
<td>( G^2 = 473.34 )</td>
</tr>
<tr>
<td>(23,134)</td>
<td>( \beta_1 = 0.1504 (.0970) )</td>
<td>( \alpha_{14} = 1.9633 (.3352) )</td>
<td>( X^2 = 585.58 )</td>
<td>( X^2 = 615.13 )</td>
</tr>
<tr>
<td></td>
<td>( \beta_2 = 0.1880 (.0093) )</td>
<td>( \alpha_{23} = 3.5977 (.2059) )</td>
<td>d.f. = 157</td>
<td>d.f. = 156</td>
</tr>
<tr>
<td></td>
<td>( \beta_3 = 0.3192 (.1009) )</td>
<td>( \alpha_{34} = 2.5854 (.1846) )</td>
<td>( \alpha_{134} = -0.5670 (.6211) )</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>( \beta_0 = -3.7122 (.1187) )</td>
<td>( \alpha_{13} = 1.6969 (.2853) )</td>
<td>( G^2 = 419.17 )</td>
<td>( G^2 = 441.30 )</td>
</tr>
<tr>
<td>(12,14,234)</td>
<td>( \beta_1 = 0.1431 (.0993) )</td>
<td>( \alpha_{14} = 2.0497 (.2876) )</td>
<td>( X^2 = 440.35 )</td>
<td>( X^2 = 482.98 )</td>
</tr>
<tr>
<td></td>
<td>( \beta_2 = 0.1949 (.0091) )</td>
<td>( \alpha_{23} = 3.0607 (.3960) )</td>
<td>d.f. = 156</td>
<td>d.f. = 155</td>
</tr>
<tr>
<td></td>
<td>( \beta_3 = 0.3036 (.1031) )</td>
<td>( \alpha_{24} = 1.5789 (.3007) )</td>
<td>( \alpha_{34} = 2.4711 (.2124) )</td>
<td>( \alpha_{234} = -0.5029 (.4829) )</td>
</tr>
<tr>
<td>8</td>
<td>( \beta_0 = -3.6791 (.1179) )</td>
<td>( \alpha_{13} = 1.7774 (.2787) )</td>
<td>( G^2 = 412.83 )</td>
<td>( G^2 = 434.96 )</td>
</tr>
<tr>
<td>(13,14,234)</td>
<td>( \beta_1 = 0.1462 (.0995) )</td>
<td>( \alpha_{14} = 1.9001 (.2972) )</td>
<td>( X^2 = 440.73 )</td>
<td>( X^2 = 473.42 )</td>
</tr>
<tr>
<td></td>
<td>( \beta_2 = 0.1903 (.0089) )</td>
<td>( \alpha_{23} = 3.0197 (.3895) )</td>
<td>d.f. = 156</td>
<td>d.f. = 155</td>
</tr>
<tr>
<td></td>
<td>( \beta_3 = 0.3020 (.1033) )</td>
<td>( \alpha_{24} = 1.9072 (.2945) )</td>
<td>( \alpha_{34} = 2.0722 (.2131) )</td>
<td>( \alpha_{234} = -0.4150 (.4764) )</td>
</tr>
<tr>
<td>9</td>
<td>( \beta_0 = -3.7130 (.1187) )</td>
<td>( \alpha_{12} = 2.0258 (.7083) )</td>
<td>( G^2 = 419.49 )</td>
<td>( G^2 = 441.62 )</td>
</tr>
<tr>
<td>(23,34,124)</td>
<td>( \beta_1 = 0.1417 (.0994) )</td>
<td>( \alpha_{14} = 2.0898 (.3106) )</td>
<td>( X^2 = 444.86 )</td>
<td>( X^2 = 487.62 )</td>
</tr>
<tr>
<td></td>
<td>( \beta_2 = 0.1952 (.0091) )</td>
<td>( \alpha_{23} = 2.6645 (.2348) )</td>
<td>d.f. = 156</td>
<td>d.f. = 155</td>
</tr>
<tr>
<td></td>
<td>( \beta_3 = 0.3032 (.1032) )</td>
<td>( \alpha_{24} = 1.5146 (.2556) )</td>
<td>( \alpha_{34} = 2.3941 (.1921) )</td>
<td>( \alpha_{124} = -0.3898 (.7715) )</td>
</tr>
</tbody>
</table>
CHAPTER 7

SUMMARY AND FUTURE WORK

In this dissertation, a methodology for analyzing incomplete longitudinal binary data was proposed. Under our assumptions, the data followed a “drop-in” monotone missing-data pattern where if observation \( t \) was missing for a subject, then so was every observation \( t' < t \). We considered a likelihood-based regression model where the marginal expectation of the binary response was related to a set of observed covariates by a known link function. We assumed models for the missing-data mechanism depend on observed covariates, giving us MAR models. By using models that are MAR, and thus known to be identifiable, we did not have to undertake the arduous task of establishing the identifiability of parameters in a non-ignorable response model. In addition, the MAR models led to likelihood equations that could be maximized separately to determine the MLE’s of the parameters of the drop-in model and the parameters of the marginal and association models. The EM-algorithm was used to determine the MLE’s for the parameters of the marginal and association models, while the MLE’s for the parameters of the drop-in model were determined through simple logistic regression.
We further extended our drop-in methodology to data that is either drop-in or drop-out. The two groups of subjects were differentiated by whether they were observed at the first timepoint. Subjects in the drop-in group were missing the first observation(s) but have been observed at every time point since they entered into the study, while subjects in the drop-out group (the "potential drop-out group") were observed at the first timepoint and have either been observed at every timepoint since (have complete data) or have been observed until a timepoint after which they are no longer observed (have drop-out data). We again assumed a likelihood-based regression model for the marginal and association structure and assumed MAR models for both the drop-in mechanism and the drop-out mechanism. Both the drop-in model and the drop-out model were logistic regression models that depended only on observed covariates. We did not have to establish the identifiability of the parameters in our missing-data models since we used MAR models. In addition, our likelihood function again led to separate maximizations for the marginal and association parameters, the drop-in parameters, and the drop-out parameters. The EM-algorithm was again used to obtain the MLE’s for the marginal and association parameters, while (separate) logistic regressions were used to determine the MLE’s of the drop-in parameters and the drop-out parameters.

Sedimentation rate data from the Air Force Health Study (AFHS) was the motivation for developing the drop-in methodology and the drop-in or drop-out methodology. After developing the two methodologies, we used them to analyze sedimentation rate for subjects with drop-in data and then drop-in or drop-out data. We
also demonstrated previous methods for complete and for drop-out data using sedimentation rate data from the AFHS.

There are numerous topics that are of interest to us for future work. The main topic relates to assessing the fit of the models to the AFHS sedimentation rate data. We used the $X^2$ and $G^2$ statistics to assess the fit of various models to sedimentation rate data for subjects having complete data; drop-out data; drop-in data; and drop-in or drop-out data. However, we suspect the asymptotics of the statistics may not be appropriate due to sparseness in the tables. We plan to simulate the exact distributions of the four statistics in order to get exact tests for the models fit to these four data sets.

There are many possible extensions to the methodologies proposed here. One possibility would be an extension to those subjects who drop-in and then drop-out. These subjects are not only missing the initial observation(s) but also the final observation(s). Whereas the drop-in or drop-out method conditioned on whether the first observation was present, a method that additionally deals with drop-in and then drop-out data would further condition on whether the last observation was present. The models presented here for subjects who only drop-in or only drop-out easily determine where in the study the subjects dropped in or dropped out. However, when a subject can drop-in and then drop-out, determining when the drop-in and drop-out occurred becomes a much harder task. In addition, as the number of observation timepoints increase, the missing-data models become increasingly complicated.

Another extension, to both the drop-in methodology and the drop-in or drop-out methodology, is to allow for missing-data models other than those based only on observed
covariates. We might consider missing-data models where one or more of the covariates is missing, or where the missingness depends on the value of the unobserved response. However, when we start considering nonignorable missing-data models, we run into the problem of establishing the identifiability of the parameters in the model. In addition, when the missing-data model depends on unobserved responses, we no longer have separate maximizations of the likelihood function to determine the MLE's for the parameters of the marginal and association models and the parameters of the missing-data models. The maximizations then become much more complicated computationally.

The last two topics of future work are related specifically to models fit to the AFHS sedimentation rate data. In all of the association models we fit so far, we assumed that the association structure was the same for every subject regardless of covariates. Another possibility is letting the association structure depend on covariates of subject i. It is plausible to assume that certain characteristics of a subject (i.e. age, group, military occupation) have an effect on the associations among his sedimentation rate responses. However, when we allow for different association structures among the responses for subjects, the models become quite computationally intense.

The final area of future interest is to fit exactly the cells in which subjects have normal sedimentation rate data for all study cycles. In Table 5.1, one can see that the cell in each of the four covariate tables indicating subjects who remained normal for all four study cycles, \( Y_{11} = Y_{12} = Y_{13} = Y_{14} = 0 \), has a very large count compared to the other cells in the tables. These four cells could be drastically influencing the models fit to the data. One option is to consider models where cells indicating subjects with
$Y_{t1} = Y_{t2} = Y_{t3} = Y_{t4} = 0$ are fit exactly. These models would be fit to the sedimentation rate data after accounting for subjects who had normal sedimentation rates at every cycle. However, it is not immediately clear how this should be done and appears to be quite intense computationally.
LIST OF REFERENCES


