Missing Response Times in Survival Analysis

Jennifer M. Bacik

The Pennsylvania State University

The Methodology Center
Technical Report Series
#97-16

College of Health and Human Development
The Pennsylvania State University

The Pennsylvania State University

The Graduate School

Department of Statistics

MISSING RESPONSE TIMES IN SURVIVAL ANALYSIS

A Thesis in Statistics

by

Jennifer Marie Bacik

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 1997

We approve the paper of Jennifer Marie Bacik.

Date of Signature

Susan A. Murphy
Associate Professor of Statistics
Thesis Advisor

Joseph L. Schafer
Assistant Professor of Statistics

Michael G. Akritas
Professor of Statistics
Chair of Graduate Program
in Statistics

Abstract

In prevention studies, it is often of interest to investigate the incidence of initial drug experimentation and its relationships to attributes such as the level of parental monitoring or the rebelliousness of the child. Survival analysis is uniquely suited for understanding such relationships because the response is time until an event occurs, i.e. time until initial drug experimentation.

However, in the course of a longitudinal study, students inevitably miss some assessments and as a result, instead of collecting the age of the initial drug experimentation, the researcher may only know that initial use occurred within a range of ages. As a result, data for these subjects are incomplete. Researchers have responded to this situation with a variety of ad hoc strategies, none entirely acceptable. These methodologies either discard information and/or introduce bias into the estimation of the hazard function. Multiple imputation is a methodology that utilizes all available data and also accounts for the uncertainty due to the missing information.

In a study with intermittent missed assessments, it is necessary to find an appropriate way of using the information from subjects who leave the study but then later return. The goal of this paper is to investigate the use of multiple imputation in survival analysis and evaluate it as a promising methodology of using the information from future assessments. Multiple imputation is compared and contrasted to two traditional methods of handling intermittent missed assessments. Simulations are conducted to evaluate the performances of the three methodologies.

Table of Contents

List of Tables

Acknowledgments

I wish to thank Susan Murphy for her guidance, insight and confidence in me. I also thank Jim Anthony for the use of his data and the Penn State NIDA Center for the Study of Prevention through Innovative Methodology for inviting me to become a part of such a dynamic group.

I extend a very special thank you to Carissa and kb who for the past two years provided a support system filled with advice and encouragement. All of my love and gratitude go out to my parents, Don and Connie, who gave me the gifts of roots and wings, and in so doing, taught me to believe in myself.

This research was supported by Grant #1 P50 DA10075 from
the National Institute on Drug Abuse.

Chapter 1

Introduction

1.1 Survival Analysis

In prevention studies, it is often of interest to investigate the incidence of initial drug experimentation and its relationships to attributes such as the level of parental monitoring or the rebelliousness of the child. Survival analysis is well suited for understanding such relationships as it provides estimates of how the incidence of initial drug experimentation varies with both time and attributes of the child.

In many prevention studies, the time of the event of interest is measured imprecisely. For example, the time of initial drug experimentation may be known only up to a time interval, such as the year of initiation. With this type of data, researchers commonly use discrete-time survival analysis. In discrete-time analyses, the incidence or hazard of the event for a particular time interval is the conditional probability that a subject experiences the event of interest in the time interval, given that the subject did not experience the event prior to this interval. The collection of estimates, each corresponding to a time interval in the study, is called the estimated hazard function.

The hazard for a particular time interval is estimated by dividing the number of people who experience the event of interest in the time interval by the number of people known to be at risk for experiencing the event in that same time interval. This latter group of people is known as the risk set. As time passes, the risk set decreases, so it is possible for the hazard rate to increase even when the number who experience the event decreases.

Survival analysis methods efficiently handle some types of missing data. Two such types of missing data are called *left-truncated* and *right-censored* data. Due to time and economic pressures, longitudinal prevention studies are often set to be a certain length; the study starts and finishes at prespecified calendar times. As a result, subjects may enter the study at different ages. For example, consider a prospective study where some subjects are nine years old at the first assessment and others are ten years old. If it is of interest to estimate the incidence of initial drug experimentation for ages 9, 10, 11, and 12, then only those subjects who have not yet begun drug experimentation at the first assessment of the study will be included in the data set. In this case, the data is said to be left-truncated as the sample includes only the subpopulation of subjects who have not initiated drug experimentation by the first assessment.

In addition, the time to initial drug experimentation is known only for those subjects who begin use before the end of the study. For the remaining subjects, all that is known is that the time until initial drug experimentation is greater than the study period. These subjects' event times are said to be right-censored at the end of the study period. In addition, some subjects may be unwilling or unable to continue participating in the study and providing follow-up information. Again, the time to initial drug experimentation is known only for those subjects who began use prior to their leaving the study. These subjects are termed *lost to follow-up* because it is not known if and when they begin drug use. These subjects' event times are also right-censored.

In contrast to missing data caused by left truncation and right censoring, conventional survival analysis does not deal effectively with intermittent missed assessments. Some students may be absent from school the day the questionnaire

is administered or they may move away from the area one year only to return the following year. Efforts are often made by the school to administer the questionnaire to those students not present at a particular assessment, but some of the efforts may be futile. Therefore, instead of collecting the year of the initial drug experimentation, the researcher may only know that initial use occurred within a range of years. For example, suppose we are interested in assessing time until initial marijuana use in high school. At the end of the school year, we administer a questionnaire in which one of the items asks, "Have you ever smoked marijuana?" Consider a student who reports on the freshman year assessment that he has never smoked marijuana. He is then absent for the sophomore and junior year assessments, but at the senior year assessment, he reports that he has smoked marijuana. The only conclusion that can be made from this data is that sometime in his sophomore, junior, or senior year, he smoked marijuana for the first time. In other words, it is known only that initial use occurred within a range of grades.

An effective but wasteful method used to handle intermittent missed assessments is to censor the event time prior to the first missed assessment. In this method, any information gained from later assessments is discarded. As we will see in Chapter 2, this method is not entirely acceptable because it does not make efficient use of the data.

In a study with intermittent missed assessments, it is necessary to find an appropriate way of using the information from the subjects who leave the study but then later return. The goal of this paper is to investigate the use of multiple imputation in survival analysis and evaluate its effectiveness in using the information from future assessments. The performance of multiple imputation is compared to that of two traditional methods for handling intermittent missed assessments.

1.2 Prototypical Questions

There have been many prevention studies designed and administered with the purpose of investigating and deterring onset of substance use in youth. It is common for these studies to collect information on a number of endpoints, or events of interest, and for each of the endpoints, to use a set of prototypical questions to address the hypotheses of interest. The prototypical questions focus on lifetime use, use in the recent past, and use in the very recent past. For example, the AAPT study [12] used the following three questions to gather information about a subject's cigarette use: "How many cigarettes have you smoked in your whole life?", "How many cigarettes have you smoked in the past month (30 days)?" and "How many cigarettes have you had in the past week (7 days)?" Possible responses ranged from "none" and "only one puff" to "more than one pack" or "more than 5 packs", depending on the time interval that was being addressed. Similarly, the Monitoring the Future study [5] included a general question about overall lifetime cigarette use; possible responses included whether the subject had been a regular user and whether he or she had quit smoking. This general question was followed by a specific question on frequency of use in the past 30 days, where the responses ranged from "not at all" and "less than 1 cigarette per day" to "2 packs or more per day".

AAPT and Monitoring the Future are just two of the many studies in the prevention field that address their hypotheses of interest with these prototypical questions. This line of questioning has particular ramifications for survival analysis, creating difficulty when assessments are missed. For example, consider a study pertaining to time until initial cigarette use by children in 5th, 6th, 7th, and 8th

Table 1.1: Data resulting from prototypical question

	Grade			
Question	5	6	7	8
"Have you ever smoked a cigarette?"	N	?	?	Y

grades. Suppose that the study asks the following question at the end of each school year, "Have you ever smoked a cigarette?" Consider a student who is present for the 5th and 8th grade assessments, but who is absent for the 6th and 7th grade assessments. Suppose the student tries smoking for the first time in 7th grade, although this information is unknown to the researcher. Table 1.1 illustrates the resulting data. As can be seen in this table, this line of questioning in a study with missed assessments leads to the very imprecise conclusion that this student began use sometime during the 6th, 7th, or 8th grades. It is impossible to fill in the data for the missed assessments due to the way the question was formulated. The remainder of this paper investigates alternative methods for analyzing data with this type of missing data.

Chapter 2

Methods of Dealing with Missed Assessments

2.1 Introduction

The following chapter presents three methods of handling the intermittent missed assessments problem. Each of the methods attempts to use a different amount of the information contained in the data.

Each of the three methods to be discussed hinge on the assumption that the data are missing at random (MAR). To understand the meaning of this assumption, consider Panel 2 of the AAPT study [12] in which students were given a questionnaire every year from seventh grade through tenth grade for a total of four assessments. Suppose a goal of the study is to determine whether gender has an effect on time until initiation of smoking and the researcher plans on fitting a regression model including gender as a covariate. Now suppose the data is split into two groups according to the subjects' gender. If within each gender, it can be assumed that the proportion of subjects who miss say, the 8th grade assessment and who initiate during grade 8 or later is approximately equal to the proportion of subjects who miss this assessment and who do not initiate during grade 8, then the MAR assumption holds. If the proportions cannot be assumed to be approximately equal, then missingness and initiation are not independent given gender and the MAR assumption is violated.

2.2 Method 1: Censoring prior to the first missed assessment

One method for dealing with intermittent missed assessments is to artificially censor subjects prior to the first missed assessment. Any information gained at later assessments is discarded. This approach of discarding the additional information does not make efficient use of the data.

Consider Panel 2 of the AAPT study as described above. Suppose a student was present for the 7th and 8th grade assessments, and by his responses to the question regarding lifetime cigarette use, it is determined that at neither of these time points had he previously used cigarettes. Therefore, information about his cigarette-use status is available for the first interval, the time between the 7th and 8th grade assessments. Now suppose that the student missed the 9th grade assessment but was present at the 10th grade assessment and that his response to the question indicates that prior to the 10th grade assessment, he had tried smoking. Because the student reported no prior use at the 7th and 8th grade assessments but did report prior use at the 10th grade assessment, the researcher can conclude that the student began using cigarettes either between the 8th and 9th grade assessments or between the 9th and 10th grade assessments, that is, the student began smoking either during the second or third interval. In the first missing-data method, the researcher disregards this information and censors the student at the end of the first interval. This approach is inefficient because it does not make use of all the information the researcher has about each student – in this case, the information collected at the 10th grade assessment.

An investigation of the hazard estimator determines that this method, although inefficient, introduces no bias in the estimation of the hazard function.

Recall that the hazard function at each time interval can be estimated by the number of students who began cigarette use during that time interval divided by the number of students in the risk set for the same time interval. A student is in the risk set if he/she has not yet begun smoking at the beginning of the interval and is present at the assessment at the end of the interval.

For the particular student in question, it is possible that he began cigarette use during the same interval in which he temporarily left the study, i.e. during the second interval (between the 8th and 9th grade assessments). Therefore, even though the student may have begun cigarette use during this interval, the researcher is not able to observe it because the student missed the 9th grade assessment. For this reason, the student cannot be included in the calculation of the numerator of the hazard function for interval two. Consequently, the student should also not be included in the risk set for that interval. It is also possible that the student began cigarette use during the third interval (between the 9th and 10th grade assessments). Again, he may have begun cigarette use during this interval, but because the researcher does not know if the use actually began in this interval, the student cannot be included in the numerator of the hazard function and therefore neither in the denominator. This issue is discussed further in Malacane [14].

The estimator for the hazard function for the second and third intervals is not biased under the MAR assumption because the student is not included in the risk set if he cannot contribute to the numerator of the estimator. Note this method does not use the knowledge that the student did begin use in one of the two intervals, so the method is necessarily inefficient.

2.3 Method 2: Partial censoring prior to the first missed assessment

The second method utilizes more information than the first method. In the second method, one fills in information for students who miss assessments but then later report no prior cigarette use (because it is known what their answers would have been), but the students who later report cigarette use are censored prior to the first missed assessment.

For example, suppose in the AAPT study, both Student A and Student B were present at the 7th, 8th, and 10th grade assessments but missed the 9th grade assessment. At the three assessments for which he was present, Student A reports no prior cigarette use. Student B, however, reports no prior use at the 7th and 8th grade assessments, but does report prior use at the 10th grade assessment. Because Student A reported no prior use at the 10th grade assessment, the researcher can deduce that if Student A had been present at the 9th grade assessment, he would have reported no prior use. Therefore, it is tempting to include Student A in the risk set for intervals 1, 2, and 3. Student B, on the other hand, is definitely included in the risk set for interval 1 (because she reported no prior use at both the 7th and 8th grade assessments), but she is not included in the risk set for intervals 2 or 3 because it is not known in which interval she began use (between the 8th and 9th grade assessments or between the 9th and 10th grade assessments).

Therefore, in the second method, future statements of students' *not* having used cigarettes are employed to supplement information about students' present cigarette-use habits. However, future statements of students' *having* used cigarettes are not utilized to add information.

Although this method uses more information, it produces bias in the esti-

mation of the hazard function. Including only those students who miss assessments but then later report no prior use results in over-representation of non-users. That is, the relative proportion of non-users to eventual users is not maintained, resulting in a negatively biased estimate of the hazard function for those intervals in which the number of students who began use in that interval is unknown. In reference to the above example, information about type A students is included in the estimates of the hazard functions for the second and third intervals, whereas information about type B students is excluded from those estimates. Type A students are therefore overrepresented in the estimate of the hazard function for those intervals.

Comparison of the second-interval hazards for the first and second methods further illustrates this point. In the first method, *all* students (both type A and type B) are censored prior to the first missed assessment so that the relative proportion of non-users to eventual users is preserved. However, by censoring only type B students, the second method induces over-representation of the proportion of type A students relative to the proportion of type B students. Even though additional students have been included in the second method, the number of students who have not yet begun cigarette use at the beginning of the second interval (the numerator of the estimate) is the same for both methods. This is because the additional students included in the second method are precisely those students who have not yet used cigarettes and therefore contribute nothing to the numerator of the estimate. However, the risk set for the second method is larger than the risk set for the first method because the additional students included in the second method are those who have reported no prior use and therefore are at risk for beginning use in that interval. Thus, in the second method, type A students are overrepresented in the risk set, causing the estimator of the hazard function to be smaller than the estimator for the first

method and inducing a negative bias. The estimate of the hazard function for the third interval is affected in a similar way.

Therefore, due to the over-representation of the non-users, the second method of dealing with the intermittent missed assessments introduces a negative bias into the estimate of the hazard function for the second and third intervals. This illustrates the price that is paid for using the additional information.

2.4 Method 3: Multiple Imputation

The third approach for dealing with the missed assessments is a Bayesian procedure called multiple imputation. The idea behind multiple imputation is to solve a missing-data problem by repeatedly solving the complete-data version. In this method, each missing value in the data set is replaced by M plausible values. For example, consider a student from the AAPT study who is present for the 7th grade assessment, misses the 8th and 9th grade assessments, but then returns to the study and is present for the 10th grade assessment. Suppose the student reports no prior use of cigarettes at the 7th grade assessment, but at the 10th grade assessment, he does report prior use. Because of the two missed assessments, it is not known if this student began use between the 7th and 8th grade assessments, between the 8th and 9th grade assessments, or between the 9th and 10th grade assessments; that is, it is unknown whether he began use during interval 1, 2, or 3. In this method, the time of initiation is randomly imputed as one of the three intervals. The imputation process transforms the data set with missing values into a complete-data version. Therefore, repeating the process M times results in M complete data sets. It may be the case that in the first of the M imputations, the time of initiation is imputed

as the time between the 8th and 9th grade assessments (interval 2), whereas in the second imputation, the time of initiation may be imputed as the time between the 7th and 8th grade assessments (interval 1). The chosen or imputed value depends on a model for the distribution of smoking initiation given the observed data. Each of the M data sets is then analyzed using standard survival analysis and the results of the M analyses are subsequently combined to arrive at one overall estimate for each parameter of interest. The resulting estimates incorporate two sources of variability: the within-imputation variation, which is a measure of ordinary sampling variation; and the between-imputation variation, a measure of the uncertainty due to the missing data. In many cases, the number M of imputations needed for valid estimates may be as small as 3 or 5 [19]. This occurs in situations where the fraction of missing information due to nonresponse, denoted by $\hat{\lambda}$, is modest. The fraction of missing information measures the increase in variance of estimation due to the missing values and the ability of the observed values to predict the missing values successfully.

The multiple-imputation procedure, described in Appendix A, uses all available information from all the subjects without introducing substantial bias into the estimation of the hazard function. After the imputation procedure, censoring is due only to the subject not initiating smoking prior to the end of the study. This, of course, is due to the fact that there are no missing data in the imputed data sets. The methodology can also handle left-truncated data because the user of the software can specify which events are logically impossible. In this way, a student's time of initiation will not be imputed as an interval for which the student has not yet entered the study. Further technical explanations of the technique and software, assumptions under which multiple imputation is valid, and the rules for combining the estimates and standard errors are presented in Appendix A.

Chapter 3

Parental Monitoring Data Set

3.1 Sample

The three methods of handling intermittent missed assessments will be illustrated with data collected from an epidemiological study of urban-dwelling children aged 8-12 years in Baltimore, Maryland. One purpose of the study was to test the hypothesis that parental monitoring and supervision may be associated with a reduced risk of drug use in the elementary school years. The sample for this study was drawn from a group of 3,319 children originally recruited for a larger study of the first graders enrolled in 19 urban elementary schools in Baltimore between 1985 and 1986. 1610 children were first interviewed in spring 1989, when they were in grades 3 or 4, and then were reinterviewed each spring from 1990 through 1994. The study population and sample is described in detail by Chilcoat and Anthony [8].

In face-to-face private interviews with each child at each of the six assessments, the interviewer asked questions regarding use of tobacco, alcohol, marijuana, inhalants, and cocaine. Standardized questions were used to assess whether the child had ever used the drug, the age of first use, and frequency of use. Visual aids, such as cartoon sketches of children using a particular drug, were used to help the children understand the questions being asked. For the purposes of comparing the three methods, the focus of this paper will be on first use of cigarettes.

At the first interview in spring 1989, the level of parental monitoring was assessed by use of a 10-item scale. Each child was asked questions about supervisory rules and the level of supervision provided by parents and other caretakers. The

responses to the items were summed to form a continuous measure and then the subjects were categorized according to quartiles of the resulting scale.

An important difference between the parental monitoring study and prevention studies such as AAPT and Monitoring the Future is reflected in the type of questions used to address the hypotheses of interest. Both AAPT and Monitoring the Future indirectly collected information on time of initial use of cigarettes via a set of prototypical questions as described in Chapter 1. It may be of interest to know the age of first use, but this question is not asked directly.

The parental monitoring study, however, uses a different approach to gather information about age of initial cigarette use; it directly asked the following question at each assessment, "How old were you when you first smoked a cigarette?" The implications of asking this retrospective question are dramatically different from those of asking the prototypical questions. For example, consider a student who is nine years old when he enters the parental monitoring study and who begins smoking when he is twelve. Suppose he is present for the assessments at age nine and ten, absent for those at eleven, twelve, and thirteen, and then returns to the study when he is fourteen. Even though he was absent from the study for three of the six assessments, including the assessment for which he would have first reported use (age 12), the exact age of his initiation is (theoretically) known because of the question that was asked. Asking a question regarding age of first use has effectively eliminated the possibility of intermittent missed assessments. It is possible to fill in the data for the missed assessments based on the answer provided at age fourteen. Table 3.1 illustrates the data that can be inferred as a result of asking this retrospective question. The entries in the table represent whether the subject had used cigarettes at or prior to that age. The bold entries represent responses that have been filled

Table 3.1: Data resulting from retrospective question

	Age					
Question	9	10	11	12	13	14
"How old were you when you first smoked a cigarette?"	N	**N**	**N**	**Y**	Y	Y

in. It is recognized that retrospective data collection can suffer from a number of drawbacks, such as recall bias, telescoping, underreporting, or hindsight bias. For the purposes of this paper, these issues will be ignored.

In this paper, it is of interest to assess the risk of initiating cigarette use among those who have never smoked. Therefore, those children who had already tried smoking at the time of the first assessment were eliminated from the data set, resulting in a data set with 1374 subjects. It is also of interest to assess whether parental monitoring has an effect on the risk of initiating cigarette use. The only valid data, then, are those collected after the level of parental monitoring has been assessed. For example, consider a child who is nine years old at the first assessment and who reports that he began smoking when he was nine. In this case, it is unknown whether he initiated smoking before or after the parental monitoring assessment. If he initiated prior to the first assessment, it may be the case that the parents' monitoring style changed as a result of their finding out about the smoking incident. Therefore, to avoid the potential problem of drug use influencing later parental monitoring, only the information collected after the first assessment is included for each child. Therefore, each child has the potential of being in the study for five assessments. This causes 164 subjects to be eliminated from the data set because they were present in the study for only one assessment. The resulting data set consists of 1210 children, 622 females and 588 males. At the first assessment, the

children range in age from eight to twelve, with only one subject being twelve years old.

3.2 Amount of Information in the Complete Data

Recall that due to the retrospective questions, the parental monitoring data set includes no intermittent missed assessments. The only time in which a student's first use of cigarettes is not observed is if he or she did not initiate prior to permanently leaving the study. For this reason, this form of the data will be considered the "complete" version.

To illustrate the amount of information present in the complete data set, Table 3.2 was constructed. Table 3.2 shows how, in the complete data, the 1210 subjects are categorized at each age according to their risk profile, initiation of use, presence in the study and level of parental monitoring measured at the first assessment. Low parental monitoring refers to those students in the lowest 25% of the parental monitoring scale and high parental monitoring refers to all others.

Because the children are at different ages when they enter the study, and because they participate in the study for different lengths of time, they will vary both in the initial age and in the number of intervals for which they have data. Note that for every column, the total number of subjects sums to 1210, indicating that each student is placed in one and only one category for each age. Note also that by summing categories one and two for a particular age, the number of students known to be at risk for initiation at that age is obtained. For example, $103 + 242 + 8 + 8 = 361$ children are at risk for initiation of smoking at age 9, and of these, $8 + 8 = 16$ initiate at age 9. The hazard of initiating of smoking at age 9 is then estimated

Table 3.2: Amount of Information Contained in the Complete Data

Information Category		Parental Monitoring	Age					
			9	10	11	12	13	14
1	Number of subjects in the study and known to be at risk for initiation at end of age i	Low	103	239	267	243	204	90
		High	242	592	717	639	558	263
2	Number of subjects known to initiate at age i	Low	8	17	25	12	16	0
		High	8	36	41	45	27	6
3	Number of subjects known to have initiated prior to age i and are present in the study at age i	Low	0	8	25	48	54	45
		High	0	8	42	79	114	86
4	Number of subjects yet to enter the study at age i	Low	227	71	6	1	0	0
		High	622	230	19	0	0	0
5	Number of subjects who left the study at age i	Low	0	3	15	34	64	203
		High	0	6	53	109	173	517
	Total Number of subjects	Low	338	338	338	338	338	338
		High	872	872	872	872	872	872
		Total	1210	1210	1210	1210	1210	1210

by dividing the number of subjects who initiate smoking at age 9 by the number of subjects at risk for smoking at age 9, or $16/361 = .044$. Similarly, the estimated hazard of initiating at age 11 is $(25 + 41)/(267 + 717 + 25 + 41) = .063$.

Chapter 4

Classical Version of the Parental Monitoring Data Set

4.1 Conversion from Complete to Classical

In the following three chapters, a novel way of using the complete version of the parental monitoring data set described in Chapter 3 will be presented. For the purposes of illustrating the three methods, the complete version of the data set was converted into a classical data set; that is, a data set as it is most often collected in prevention studies; i.e., through the use of the prototypical questions discussed in Section 1.2. This conversion was possible because the complete data set included indicators of whether the student was present or absent at each assessment. The conversion introduces missed assessments in the following way. For example, consider a student who was nine years old at the first assessment, began smoking at twelve years old, and who was present for the first, second, fifth and sixth assessments. Table 4.1 presents the complete and classical data for this student. The question marks in the classical data set denote those assessments for which the student was missing. Here it is impossible to fill in the data for the missed assessments. As the data were converted from the complete to classical version, it was assumed that each subject was asked the prototypical questions on the day before his or her birthday. For example, if the subject was nine years old at assessment one, the assumption was that the subject was interviewed on the day before his or her 10th birthday.

Table 4.2 shows how, in the classical data, the 1210 subjects are categorized at each age. The children will again vary by initial age and number of intervals for which they have data. However, in this version of the data, it is necessary to

Table 4.1: Conversion from Complete to Classical

Age	Assessment	Complete Data	→	Classical Data
9	1	N	→	N
10	2	N	→	N
11	3	N	→	?
12	4	Y	→	?
13	5	Y	→	Y
14	6	Y	→	Y

introduce three additional categories (6, 7, and 8) due to the fact that this data set contains intermittent missed assessments. Here it is not known exactly at what age each subject began smoking. To understand how a subject with missed assessments would be categorized in this table, consider the student's classical data presented in Table 4.1. In Table 4.2, this student is included in category 4 at age 9 since only the information collected after the first assessment is considered for each subject. At age 10, it is known that the student has not yet initiated smoking and therefore is at risk for initiation at the end of age 10 (category 1). The student is missing for the age 11 and age 12 assessments. For age 11, it is not known whether or not the student initiated during that age so he is included in category 6. The student is then placed in category 8 for age 12. This is due to the fact that at age 12, it is not known whether he initiated at age 11, at age 12, or if he has yet to initiate. When the student returns to the study at age 13, he reports prior use of cigarettes. However, because he missed previous assessments, it is not known if he initiated at age 13 or prior to age 13. Therefore, he is included in category 7 for age 13. At age 14, it is known that he initiated prior to this age so he is placed in category 3.

Table 4.2 illustrates why problems are encountered in a data set collected

Table 4.2: Amount of Information Contained in the Classical Data

	Information Category	Parental Monitoring	Age					
			9	10	11	12	13	14
1	Number of subjects in the study and known to be at risk for initiation at end of age i	Low	103	236	262	238	203	89
		High	240	586	705	625	549	256
2	Number of subjects known to initiate at age i	Low	7	12	21	9	12	0
		High	5	28	31	38	22	3
3	Number of subjects known to have initiated prior to age i and are present in the study at age i	Low	0	7	20	44	50	41
		High	0	5	35	72	107	78
4	Number of subjects yet to enter the study at age i	Low	227	71	6	1	0	0
		High	622	230	19	0	0	0
5	Number of subjects who have left the study before or at age i	Low	0	3	15	34	64	203
		High	0	6	53	109	173	517
6	Number of subjects for which it is unknown if they are still at risk at the end of age i or if they initiated at age i	Low	1	8	6	3	0	0
		High	5	12	16	9	0	1
7	Number of subjects for which it is unknown if they initiated at age i or prior to age i	Low	0	1	5	3	4	4
		High	0	4	10	7	4	10
8	Number of subjects for which it is unknown if they are still at risk at the end of age i, if they initiated at age i, or if they initiated prior to age i	Low	0	0	3	6	5	1
		High	0	1	3	12	17	7
	Total Number of subjects	Low	338	338	338	338	338	338
		High	872	872	872	872	872	872
		Total	1210	1210	1210	1210	1210	1210

via the prototypical questions. Missed assessments introduce uncertainty about which ages the subjects are at risk and if and when they initiate. The researcher must then find an appropriate methodology to deal with this uncertainty.

4.2 Use of the Classical Data in the Three Methods

The three methods of dealing with intermittent missed assessments use to varying degrees the information from the classical data set.

4.2.1 Method 1

Table 4.3 illustrates how the subjects are categorized at each age for Method 1, censoring prior to the first missed assessment. Note in the table there are $23 + 55 = 78$ subjects who are censored at age 9. These are the students who entered the study at age 8 and who were not present at the age 9 assessment. Because only the information collected after the first assessment is used for each student, these 78 students become right-censored at the first assessment for which they are present in the study. This situation occurs at the other ages as well and illustrates one of the penalties of using the prototypical questions to collect information.

To further understand the implementation of this method and how the subjects are categorized at each age, compare the classical data table (Table 4.2) to Table 4.3. In this method, all subjects in categories 6, 7, and 8 at each age from the classic data table become artificially right-censored. In addition, some of the subjects from categories 1, 2, and 3 in the classical data table also become right-censored. For example, consider the student's data presented in Table 4.4. In the classical data table, this student would be included in category 4 at age 9, in

Table 4.3: Data from Method 1

Information Category		Parental Monitoring	Age					
			9	10	11	12	13	14
1	Number of subjects in the study and known to be at risk for initiation at end of age i	Low	81	182	199	174	144	63
		High	190	462	531	442	376	175
2	Number of subjects known to initiate at age i	Low	7	12	20	8	10	0
		High	5	28	29	35	17	3
3	Number of subjects known to have initiated prior to age i and are present in the study at age i	Low	0	7	19	37	40	31
		High	0	5	32	57	85	62
4	Number of subjects yet to enter the study at age i	Low	227	71	6	1	0	0
		High	622	230	19	0	0	0
5	Number of subjects who have left the study before or at age i	Low	0	3	15	34	64	203
		High	0	6	53	109	173	517
6	Number of subjects who are artificially right-censored at age i	Low	23	63	79	84	80	41
		High	55	141	208	229	221	115
	Total Number of subjects	Low	338	338	338	338	338	338
		High	872	872	872	872	872	872
		Total	1210	1210	1210	1210	1210	1210

Table 4.4: Example of Artificial Right-Censoring

	Age					
Question	9	10	11	12	13	14
"Have you ever smoked a cigarette?"	N	N	?	N	Y	Y

category 1 for ages 10, 11, and 12, in category 2 for age 13, and in category 3 for age 14. In the method where all subjects are censored prior to their first missed assessment, however, the student is considered at risk only at the end of age 10 (category 1 in Table 4.3). The student is then right-censored at ages 11, 12, 13, and 14 (category 6 in Table 4.3). Even though it is known that the student initiated at age 13, this information is discarded and the student is censored at age 11.

From the comparison of these two tables, it is seen that Method 1 uses only some of the information from categories 1, 2, and 3 from the classical data table. In addition, Method 1 uses none of the information from categories 6, 7, and 8; all subjects from these categories become artificially right-censored.

Table 4.3 can be used to estimate the hazard of initiating smoking at each of the ages for Method 1. Again, the hazard of initiating smoking at age 9 is estimated by dividing the number of subjects who initiate smoking at age 9 (category 1 in Table 4.3) by the number of subjects at risk at age 9 (sum of categories 1 and 2), or $(7 + 5)/(81 + 190 + 7 + 5) = .042$. Similarly, the hazard at age 11 is estimated by $(20 + 29)/(199 + 531 + 20 + 29) = .063$.

4.2.2 Method 2

In Method 2, only those subjects who temporarily leave the study and then report initiation when they return are artificially censored prior to the first missed

Table 4.5: Data from Method 2

Information Category		Parental Monitoring	Age					
			9	10	11	12	13	14
1	Number of subjects in the study and known to be at risk for initiation at end of age i	Low	103	236	262	238	203	89
		High	240	586	705	625	549	256
2	Number of subjects known to initiate at age i	Low	7	12	21	9	12	0
		High	5	28	31	38	22	3
3	Number of subjects known to have initiated prior to age i and are present in the study at age i	Low	0	7	19	38	42	33
		High	0	5	32	59	89	68
4	Number of subjects yet to enter the study at age i	Low	227	71	6	1	0	0
		High	622	230	19	0	0	0
5	Number of subjects who have left the study before or at age i	Low	0	3	15	34	64	203
		High	0	6	53	109	173	517
6	Number of subjects who are artificially right-censored at age i	Low	1	9	15	18	17	13
		High	5	17	32	41	39	28
	Total Number of subjects	Low	338	338	338	338	338	338
		High	872	872	872	872	872	872
		Total	1210	1210	1210	1210	1210	1210

assessment; information regarding non-initiation is filled in for students who miss assessments but then later report no prior use. This method attempts to use more of the information contained in future assessments than Method 1 does. Table 4.5 illustrates how the subjects are categorized at each age for this method. Notice that in this method, only 6 subjects are artificially right-censored at age 9, compared to the 78 in the previous method. This is due to the fact that even though many of those subjects' first assessments are missing, a majority of them later report no prior use and so the information can be filled in for the missing first assessment.

Comparing Table 4.5 to the classical data table (Table 4.2), note again that all subjects in categories 6, 7, and 8 for each age in the classical data become artificially right-censored in this method. In addition, some of the subjects from category 3 become right-censored. Consider again the student's data in Table 4.4. Even though the student is missing at age 11, the information regarding his non-initiation at age 11 can be filled in due to the fact that he reports no prior use at age 12. In this method then, the information regarding initiation at age 13 is incorporated into the data and subsequent analysis. In Table 4.5, the student would be placed in category 4 for age 9, in category 1 for ages 10 through 12, in category 2 at age 13, and in category 3 at age 14.

Comparing the classical data table to Table 4.5, it is seen that like Method 1, Method 2 uses none of the information from categories 6, 7, and 8 in the classical data table. However, Method 2 does use all the information from categories 1, 2, and 3.

The hazard of initiating smoking at each of the ages for Method 2 can be estimated by using the data in categories 1 and 2 from Table 4.5. The Age 9 hazard is estimated by $(7 + 5)/(103 + 240 + 7 + 5) = .034$ and the Age 11 hazard is

estimated by $(21 + 31)/(262 + 705 + 21 + 31) = .051$.

4.2.3 Method 3

In the above sections, it is seen that Methods 1 and 2 discard some or all of the information gained from those subjects who leave the study but then later return. In contrast, the multiple-imputation procedure utilizes the data from all subjects at all assessments; i.e. it uses all the information from every category in the classical data table.

Chapter 5

Results

5.1 Logistic Regression Model

A hazard model may be fit using PROC LOGISTIC in SAS in order to estimate the effect of parental monitoring measured at the first assessment for each of the ages. Non-proportionality allows for the effect of parental monitoring to fluctuate over time. Recall that the hazard is the conditional probability of a student initiating use of cigarettes at a particular age, given that the student did not begin smoking prior to this age. Denote the hazard probability for child i at age j as h_{ij}. Including the main effect of age, a main effect of parental monitoring, and the interaction between age and parental monitoring leads to the logistic discrete-time hazard model

$$log\frac{h_{ij}}{1-h_{ij}} = \alpha_j + \beta_j Z_i,$$

where α_j represents the baseline profile of risk at age j, Z_i is an indicator of child i's level of parental monitoring at the first assessment and β_j is the regression coefficient for the level of parental monitoring at age j.

Estimates for the parameters of the logistic discrete-time hazard model can be obtained by the method of maximum likelihood. The principle of maximum likelihood is to choose as coefficient estimates those values which maximize the probability of observing what has, in fact, been observed. A derivation of the likelihood function that is maximized via the logistic-regression procedure is given by Singer and Willett [21].

In order to correctly implement the logistic-regression analysis, a separate observational record for each age an individual is known to be at risk must be created. These records form a data set referred to as the "person-period" data set in which each person has multiple records, one per time period of observation. The total number of observations, or person-periods, in the new data set is simply the sum of the number at risk in each of the periods or ages. A response variable indicates whether or not the subject began smoking at that age. Left-truncated and right-censored observations are accounted for in this procedure as well. Appendix D provides SAS code that illustrates the implementation of the analysis.

5.2 Analysis of the Complete Version of the Data

The results of the logistic-regression analysis of the complete version of the data will serve as a benchmark against which to compare the performance of the three methodologies. The results and conclusions from this analysis regarding the effect of parental monitoring on the time of initiation of cigarette use cannot be viewed as the ultimate truth. The truth, after all, would be gathered from an analysis of the entire population, i.e. all urban-dwelling children aged 8-12 years old in Baltimore. However, the estimates from the analysis on the complete version of the data serve as high-quality estimates of the population regression coefficients and can be used for comparison purposes.

Fitting a model with the quartiles of parental monitoring coded as three dummy variables (with the first quartile used as the reference category) showed that each of the three parental monitoring parameter estimates are of the same sign and magnitude. This suggests forming a binary parental monitoring variable where low

Table 5.1: Results from Complete Data

Parameter	Parameter Estimate	Standard Error	p-value
β_9	0.8542	0.5137	0.0963
β_{10}	0.1567	0.3041	0.6063
β_{11}	0.4931	0.2637	0.0615
β_{12}	-0.3549	0.3335	0.2873
β_{13}	0.4830	0.3259	0.1384

parental monitoring consists of those subjects in the first quartile and high parental monitoring consists of those in the second, third, and fourth quartiles. The binary form of the parental monitoring variable is used in the analyses to follow. To allow for the effect of parental monitoring to change over time, the parental monitoring (PM) by age interactions are included. However, because there are no initiators at age 14 in the low parental monitoring group, a parameter for the parental monitoring by age = 14 interaction (β_{14}) cannot be estimated and therefore is not included in the model.

Table 5.1 presents the results of the analysis on the complete data for the interaction terms (the estimated $\beta_j s$ of the hazard model). As can be seen in this table, none of the variables are significant at the .05 level. However, the p-values of $\hat{\beta}_9$ (0.0963) and $\hat{\beta}_{11}$ (0.0615) suggest that there may indeed be some effect of parental monitoring on initiation of smoking for these two ages.

5.3 Analysis of the Classical Data

The classical data set described in Section 4.1 will be used to illustrate the three methods of dealing with intermittent missed assessments.

5.3.1 Method 1

The results of the logistic-regression analysis for this method are shown in Table 5.2. Comparing these results to those in Table 5.1 for the complete data, it is found that the standard error for each estimate has increased. This is expected because these data contain missed assessments. Theoretically, it is known that use of this method introduces no bias into the estimates of the parameters. The difference between the corresponding estimates in the two methods is due to normal variability.

Consider the following 95% confidence interval for each of the j ages, $j = 9, \ldots, 13$,

$$\hat{\beta}_{complete} - \hat{\beta}_{method1} \pm 1.96\sqrt{(s.e.(\hat{\beta}_{complete}))^2 + (s.e.(\hat{\beta}_{method1}))^2}, \qquad (5.1)$$

where $\hat{\beta}_{complete}$ and $s.e.(\hat{\beta}_{complete})$ are the parameter estimate and standard error of one of the βs from the analysis on the complete version of the data, and $\hat{\beta}_{method1}$ and $s.e.(\hat{\beta}_{method1})$ are the corresponding parameter estimate and standard error from Method 1. This interval should have approximately 95% coverage if the estimates $\hat{\beta}_{complete}$ and $\hat{\beta}_{method1}$ are uncorrelated. In practice, it is expected that the estimates are positively correlated, in which case (5.1) would tend to be conservative. An explanation of this confidence interval is given in Appendix F.

Five versions of this confidence interval can be calculated, one for each of the age groups. The confidence interval is constructed to assess if the estimator $\hat{\beta}_{complete}$, from the analysis of the complete version of the data, and the estimator $\hat{\beta}_{method1}$ from Method 1 are actually estimating the same parameter. If they are indeed estimating the same parameter, then over repeated sampling from the population, 95% or more of the intervals should contain zero. By calculating this confidence interval for each of the five components of β, it is found that all five of

Table 5.2: Results from Method 1

Parameter	Parameter Estimate	Standard Error	p-value
β_9	1.1890	0.6004	0.0477
β_{10}	0.0843	0.3560	0.8129
β_{11}	0.6099	0.3023	0.0436
β_{12}	-0.5437	0.4020	0.1762
β_{13}	0.4291	0.4104	0.2957

them contain zero. Thus, there is no evidence to suggest that $\hat{\beta}_{complete}$ and $\hat{\beta}_{method1}$ are estimating different parameters.

From Table 5.2, it is also seen that use of Method 1 has caused a shift in the significance of $\hat{\beta}_9$ and $\hat{\beta}_{11}$. Both estimates have changed from being marginally insignificant to marginally significant. Two reasons why this has occurred are either initiators in the high-monitored group are missing assessments or non-initiators in the low-monitored group are missing assessments. Examination of Table 5.3 illustrates the reason significance at age 9 is introduced. The entries in this table represent percent reduction in the number of subjects who missed the age 9 assessment, moving from the complete data version to the method in which subjects are censored prior to the first missed assessment. For example, the number of high-monitored initiators consists of 8 subjects in the complete data. As a result of converting to the classical version, this group drops to 5 subjects, or a 37% reduction. From Table 5.3, it is seen that the same percentage (21%) of non-initiators at age 9 in both the high- and low-monitored groups have missed assessments. In contrast, a much greater percentage of high-monitored initiators missed the age 9 assessment (37%) compared to the low-monitored initiators (12.5%). The reduction

Table 5.3: Percent Missing at Age 9

	High PM	Low PM
Non-initiators	$\frac{242-190}{242} = 21\%$	$\frac{103-81}{103} = 21\%$
Initiators	$\frac{8-5}{8} = 37\%$	$\frac{8-7}{8} = 12.5\%$

Table 5.4: Percent Missing at Age 11

	High PM	Low PM
Non-initiators	$\frac{717-531}{717} = 26\%$	$\frac{267-199}{267} = 25.5\%$
Initiators	$\frac{41-29}{41} = 29\%$	$\frac{25-20}{25} = 20\%$

in the number of high-monitored initiators explains the introduction of significance at age 9.

The introduction of significance at age 11 may also be due to the reduction in the number of initiators in the high-monitored group. Table 5.4 illustrates this idea.

These tables, especially Table 5.3, also indicate that the missing at random (MAR) assumption may be violated. Missing at random in this context refers to the situation in which within a parental monitoring group, the proportion of subjects who miss say, the age 9 assessment and who initiate at age 9 is approximately equal to the proportion of subjects who miss the assessment and who do not initiate at

age 9. By comparing the number of high-monitored initiators at age 9 in Table 5.3 (21%) to the number of high-monitored non-initiators at age 9 (37%), it is seen that there is a large difference between these two proportions, suggesting that the MAR assumption may be violated. This situation can also be seen in the low-monitored group at age 9 and to a lesser degree for both parental monitoring groups at age 11.

5.3.2 Method 2

The results of the analysis for this method are shown in Table 5.5. Again, it is noted that because the data are not complete, the standard errors of the estimates have increased compared to those in the analysis of the complete version of the data. However, they are not as large as those presented in Table 5.2 for the method in which *all* subjects are censored prior to the first missed assessment. No significant changes between the two censoring methods occurred in the ratio of the estimates to their standard errors; the p-values did not change dramatically. However, theoretically, it is known that bias has been introduced by implementing this method.

A conservative 95% confidence interval similar to (4.1) can be constructed for this method for each of the j ages, $j = 9, \ldots, 13$,

$$\hat{\beta}_{complete} - \hat{\beta}_{method2} \pm 1.96\sqrt{(s.e.(\hat{\beta}_{complete}))^2 + (s.e.(\hat{\beta}_{method2}))^2}. \qquad (5.2)$$

The confidence intervals for each of the five components of β contain zero. Thus, there is no evidence to suggest that $\hat{\beta}_{complete}$ and $\hat{\beta}_{method2}$ are estimating different parameters.

Table 5.5: Results from Method 2

Parameter	Parameter Estimate	Standard Error	p-value
β_9	1.1824	0.5973	0.0477
β_{10}	0.0622	0.3535	0.8604
β_{11}	0.6004	0.2917	0.0396
β_{12}	-0.4749	0.3785	0.2096
β_{13}	0.3888	0.3682	0.2910

5.3.3 Method 3

In multiple imputation, the missing-data problem is addressed by creating M plausible versions of the complete data set. This method was applied to the parental monitoring data set to create 5 imputations. Details regarding implementation of the method are given in Appendix B. Two of the five imputations are presented in Tables 5.6 and 5.7. Imputations 3, 4, and 5 are located in Appendix E. In each of the tables, the 1210 subjects are categorized according to their risk profile, initiation of use, presence in the study and the level of parental monitoring measured at the first assessment. Because there are no missed assessments in the imputed data sets, there is no need for categories 6, 7, and 8 as in the classical data table (Table 4.2). Essentially, the subjects in categories 6, 7, and 8 from the classical data table have been reallocated to categories 1, 2, and 3 in the imputed data tables. For each of the imputed data sets, it is possible to estimate the hazard of initiating smoking at each of the ages. For example, the hazard of initiating smoking at age 9 in imputation 1 is estimated by $(7 + 6)/(104 + 244 + 7 + 6) = .036$. For imputation 2, the estimated hazard of initiating smoking for age 11 is $(31 + 40)/(264 + 717 + 31 + 40) = .067$.

Across the five imputations, the number of subjects in each category may vary. For example, in imputation 1, the number of subjects who initiate at age 12 in the low-monitored group is 12 and the number in the high-monitored group is 50. This is in contrast to imputation 2, where the number of subjects initiating at age 12 in the low- and high-monitored groups are 11 and 47, respectively. The variation in the number of people in each category is due to the fact that we are sampling from a probability distribution, that is, the distribution of smoking initiation given the observed data.

The results of the logistic-regression analyses from this method are presented in Table 5.8. The first five columns of this table display the parameter estimates and standard errors of β_j, $j = 9, \ldots, 13$, from the logistic-regression analyses on each of the 5 imputed data sets. The final column of the table gives the combined data estimates and standard errors, where the estimates were combined according to the rules presented in Appendix A. From Table 5.8, it is seen that for any one parameter, the estimates and standard errors vary across the five imputations. The combined parameter estimate is simply the average of the estimates from the five imputations. The combined standard error incorporates two sources of variability, the within-imputation variability and the between-imputation variability, so that the combined standard error for the parameter is greater than any of the individual standard errors.

It is possible to construct a confidence interval similar to (4.1) or (4.2) for each of the ages j, $j = 9, \ldots, 13$, to assess whether $\hat{\beta}_{complete}$ and $\hat{\beta}_{method3}$ are estimators of the same parameter of the population. Again, no evidence was found to suggest that they are estimating different parameters, because each of the five confidence intervals contains zero.

Table 5.6: Amount of Information Contained in Imputation 1

Information Category		Parental Monitoring	Age					
			9	10	11	12	13	14
1	Number of subjects known to be at risk for initiation at end of age i	Low	104	242	266	242	204	90
		High	244	595	717	634	555	262
2	Number of subjects known to initiate at age i	Low	7	15	29	12	15	0
		High	6	35	44	50	25	4
3	Number of subjects known to have initiated prior to age i and are present in the study at age i	Low	0	7	22	49	55	45
		High	0	6	39	79	119	89
4	Number of subjects yet to enter the study at age i	Low	227	71	6	1	0	0
		High	622	230	19	0	0	0
5	Number of subjects who have left the study before or at age i	Low	0	3	15	34	64	203
		High	0	6	53	109	173	517
	Total Number of subjects	Low	338	338	338	338	338	338
		High	872	872	872	872	872	872
		Total	1210	1210	1210	1210	1210	1210

Table 5.7: Amount of Information Contained in Imputation 2

Information Category		Parental Monitoring	Age					
			9	10	11	12	13	14
1	Number of subjects known to be at risk for initiation at end of age i	Low	104	242	264	241	203	89
		High	245	591	717	637	555	262
2	Number of subjects known to initiate at age i	Low	7	15	31	11	15	0
		High	5	40	40	47	28	4
3	Number of subjects known to have initiated prior to age i and are present in the study at age i	Low	0	7	22	51	56	46
		High	0	5	43	79	116	89
4	Number of subjects yet to enter the study before or at age i	Low	227	71	6	1	0	0
		High	622	230	19	0	0	0
5	Number of subjects who have left the study at age i	Low	0	3	15	34	64	203
		High	0	6	53	109	173	517
	Total Number of subjects	Low	338	338	338	338	338	338
		High	872	872	872	872	872	872
		Total	1210	1210	1210	1210	1210	1210

Table 5.8: Results from Method 3: Imputations and Combined

Parameter	Imp 1	Imp 2	Imp 3	Imp 4	Imp 5	Combined
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
β_9	1.0069 (0.5685)	1.1933 (0.5971)	0.8487 (0.5472)	1.1933 (0.5971)	1.1933 (0.5971)	1.0871 (0.6063)
β_{10}	0.0523 (0.3179)	-0.0880 (0.3122)	0.1542 (0.3041)	-0.0780 (0.3229)	0.0638 (0.3088)	0.0209 (0.3329)
β_{11}	0.5747 (0.2497)	0.7442 (0.2499)	0.5752 (0.2580)	0.6681 (0.2463)	0.4810 (0.2585)	0.6086 (0.2755)
β_{12}	-0.4640 (0.3302)	-0.4803 (0.3434)	-0.2334 (0.3138)	-0.3226 (0.3343)	-0.4056 (0.3321)	-0.3812 (0.3497)
β_{13}	0.4900 (0.3367)	0.3816 (0.3303)	0.3144 (0.3374)	0.3131 (0.3374)	0.4745 (0.3259)	0.3947 (0.3463)

Table 5.9: Results from Method 3: Combined

Parameter	Parameter Estimate	Standard Error	p-value
β_9	1.0871	0.6063	0.0734
β_{10}	0.0209	0.3329	0.9501
β_{11}	0.6086	0.2755	0.0286
β_{12}	-0.3812	0.3497	0.2764
β_{13}	0.3947	0.3463	0.2546

Comparing the results of the multiple-imputation method presented in Table 5.9 to the results of the analysis on the complete version of the data presented in Table 5.1 reveals an interesting result. The p-value of $\hat{\beta}_{11}$ decreased by half from 0.0615 in the complete data analysis to 0.0286 in the multiple-imputation analysis. Although the change in p-value may not be statistically important, this situation is addressed because there is a tendency for data analysts to become excited when the p-value is less than 0.05. An investigation of the imputations for age 11 reveals the reason for this occurrence.

Comparing the complete data table (Table 3.2) to the tables of the five imputations (Tables 5.6, 5.7, E.1, E.2, E.3) shows that across the five imputations, the number of subjects in most of the information categories (ages $\neq$ 11) is approximately the same as the number of subjects in the corresponding category for the complete data. For example, consider the low-monitored initiators for age 12 (information category 2 in the tables). In the complete data, it is known that there are 12 initiators in this group. Twelve subjects are low-monitored initiators in the first imputed data set, 11 in the second, 14 in the third, 12 in the fourth, and 12 in the fifth.

Now consider the data for age 11. In the complete data, it is known that there are 41 initiators in the high-monitored group and 25 initiators in the low-monitored group. The number of initiators in the high-monitored group for the five imputations range from 40 to 44 (40, 44, 41, 43, 43) and the number of initiators in the low-monitored group range from 26 to 31 (29, 31, 27, 31, 26). The imputation procedure then is consistently imputing more initiators in the low-monitored group than are known to be initiators from the complete version of the data. This points to the reason that the p-value of $\hat{\beta}_{11}$ dropped. It is not clear why the imputation procedure is performing in this manner. A possible explanation may lie in the fraction of missing information for this particular variable or in the evidence supporting a violation of the missing at random assumption discussed in Section 5.3.1. Further investigation of this issue is needed.

5.4 Simulation Results

Simulations were conducted to evaluate the performance of the three methods of handling intermittent missed assessments. It was first determined the percentage of the risk set that was missing at each age for each parental monitoring group in the classical version of the data set. Then, 1000 different data sets were generated from the complete version of the data, so that in each of the data sets, approximately the same percentage of the risk set at each age for each parental monitoring group was missing as was found in the corresponding risk set in the original classical data set. Table 5.10 shows the percent of the risk set missing for the original classical data set and the average percent of the risk set missing for the simulated data sets. As can be seen from this table, the amount of simulated miss-

Table 5.10: Percent of Risk Set Missing at Each Age

		Age					
	PM	9	10	11	12	13	14
Original	Low	20.7	16	8.9	7.5	3.2	1.1
	High	22	14	12.3	12.1	7.5	2.6
Simulated	Low	21.0	15.9	9.0	7.5	3.2	1.1
	High	22.0	14.0	12.3	12.1	7.5	2.7

ingness averaged over the 1000 data sets is approximately the same as that found in the original classical data set.

Each of the 1000 data sets was then analyzed by the three methods. Within each method, the parameter estimates and standard errors from the 1000 analyses were averaged. The results of the simulations for the three methods are shown in Table 5.11 along with the results from the analysis on the complete version of the data.

As described in Chapter 2, it is expected that the implementation of Method 2 will introduce a negative bias into the estimation of the hazard. Recall that α_j represents the baseline profile of risk at age j, $j = 9, \ldots, 14$. The negative bias can be seen in Table 5.11 by comparing each $\hat{\alpha}_j$ in Method 2 to the corresponding $\hat{\alpha}_j$ from the complete-data analysis; $\hat{\alpha}_j$ from Method 2 is consistently lower than $\hat{\alpha}_j$ from the complete-data analysis. This is not the case for the other methods.

Each of the $\hat{\beta}_j s$ represents the ratio of the hazard at age j for low-monitored individuals to the hazard at age j for high-monitored individuals. If the percentage of the risk set that is missing differs greatly by parental monitoring, then some bias

Table 5.11: Simulation Results

Parameter	Complete	Method 1	Method 2	Method 3
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
α_9	-3.4095 (0.3593)	-3.4314 (0.4137)	-3.6621 (0.4124)	-3.5930 (0.4121)
α_{10}	-2.8000 (0.1717)	-2.7792 (0.1917)	-2.9995 (0.1906)	-2.8085 (0.1812)
α_{11}	-2.8615 (0.1606)	-2.7895 (0.1805)	-3.0511 (0.1775)	-2.8236 (0.1697)
α_{12}	-2.6532 (0.1542)	-2.6203 (0.1878)	-2.8917 (0.1744)	-2.6486 (0.1659)
α_{13}	-3.0285 (0.1971)	-2.9657 (0.2465)	-3.2004 (0.2187)	-3.0189 (0.2131)
α_{14}	-4.0747 (0.4117)	-4.0797 (0.5103)	-4.1633 (0.4362)	-4.0429 (0.4253)
β_9	0.8542 (0.5137)	0.8628 (0.5899)	0.8697 (0.5868)	0.8656 (0.5862)
β_{10}	0.1567 (0.3041)	0.1945 (0.3389)	0.1683 (0.3365)	0.1766 (0.3217)
β_{11}	0.4931 (0.2637)	0.5145 (0.295)	0.5093 (0.2890)	0.4859 (0.2754)
β_{12}	-0.3549 (0.3335)	-0.3347 (0.3967)	-0.2756 (0.3664)	-0.3631 (0.3577)
β_{13}	0.4830 (0.3259)	0.5591 (0.3893)	0.5628 (0.3505)	0.4803 (0.3417)

in the $\hat{\beta}_j s$ is expected in Method 2. This bias would appear because one parental monitoring group is overrepresented in the estimation of the hazard compared to the other. For example, if the high-monitored individuals were much more likely to be missing at each age than the low-monitored individuals, then the $\hat{\beta}_j s$ from Method 2 would be higher than the corresponding $\hat{\beta}_j s$ from the analysis of the complete version of the data. It would be possible then to conclude that a positive bias was introduced in the $\hat{\beta}_j s$ by using Method 2. As is seen in Table 5.10, for all ages (except age 10), the percentage of the risk set that is missing is higher for high-monitored individuals than for low-monitored individuals. However, this difference is quite small and explains why, in the simulation results for Method 2, a bias is not seen in the $\hat{\beta}_j s$.

Chapter 6

Summary and Future Problems

Intermittent missed assessments are unavoidable in longitudinal drug use prevention research. This problem necessitates finding an appropriate way of using the information gained from the subjects who leave the study but then later return. This paper lays a foundation for assessing whether multiple imputation is an appropriate methodology for this situation.

Each of the three methods of handling intermittent missed assessments uses a different amount of that "future" information. The method in which all subjects are censored prior to the first missed assessment uses none of the information. The method in which only those who eventually initiate are censored uses some future information, but pays a price for that use; bias is introduced. In contrast, the multiple-imputation procedure uses all of the information gained from future assessments.

It was seen, however, that the future assessments in the parental monitoring data set contained very little information, because there was little distinction in the results across the three methods. Of interest, then is to perform simulations in which the levels of information in future assessments vary. The results of these simulations will help to assess at what point the amount of information in future assessments makes multiple imputation worthwhile.

Another question that arises from this work is whether it is possible to see a difference among the three methods when it is known that there is little or no significance in the parameters of the data model. It would be useful to evaluate the

three methods when in fact one or more of the parameters shows strong statistical significance. Future simulations include altering the complete version of the parental monitoring data set to introduce a greater effect of parental monitoring on time until initiation of smoking. Interesting conclusions may then be drawn about the performance of the three methodologies under a variety of conditions.

In many situations, the usual tools survival analysis provides for intermittent missed assessments, i.e. censoring prior to the first missed assessment, may be adequate. However, there are some scenarios where this will not be the case. For example, consider the situation in which at the initial assessment, multiple covariates, such as level of parental monitoring and gender, are measured on each subject along with the subject's response to the question regarding prior use of cigarettes. Censoring prior to the first missed assessment cannot be used when the values of some of the covariates are missing. Multiple imputation provides a way of handling this situation for it can "fill in" the missing covariates as well as the missing responses.

Another situation for which the usual techniques of survival analysis may not be adequate is the situation in which the missing at random assumption is violated. In general, one way in which researchers attempt to ensure that the MAR assumption holds is by including multiple covariates in their models. By cross-classifying the subjects according to the different levels of the covariates, the subjects will be segmented into smaller, more homogeneous groups and the proportion of missing initiators and the proportion of missing non-initiators in each of the groups will tend to be approximately equal, thereby allowing the MAR assumption to hold.

Consider, however, the following typical setting in which survival analysis is used. A longitudinal study in which subjects have been randomized to two intervention programs is administered to determine which program is more effective

in delaying initiation of smoking. As is usually the case with longitudinal studies, suppose there are missing responses for some of the subjects at some of the assessments. It is tempting to include additional variables, such as level of commitment after program-end, in the model because they may help explain the missingness patterns and may help ensure that the MAR assumption holds. For example, it may be that those who were less committed to the program are more likely to miss assessments than those who were more committed. However, because the goal of this study is to determine the causal effect of the intervention program, level of commitment may be a venue via which the intervention program causally effects smoking cessation. Therefore, the researcher would not want to include level of commitment in the regression model.

Use of multiple imputation allows both of the above goals (ensuring that missing at random holds and determining the effect of intervention program) to be met since it differentiates between the *imputation model* and the *analysis model.* Including the additional variables, such as level of commitment, in the imputation model helps to ensure that the MAR assumption holds. Then, when the imputed data are analyzed to determine whether the intervention program has an effect on delaying time until initiation of smoking, level of commitment does not need to be included as a covariate in the analysis model. In this way, using multiple imputation in place of the usual techniques survival analysis provides for the randomized trial setting helps to account for problems that may develop if it is thought that the missing at random assumption has been violated.

The above two situations point to reasons why further evaluation of multiple imputation in the context of survival analysis must be done.

As mentioned in Section 2.3, the fraction of missing information, denoted

by $\hat{\lambda}$, is a measurement of the increase in variation of estimation due to the missing values and the ability of the observed values to predict the missing values successfully. The formula for estimating this quantity is given in Appendix A. The calculation of $\hat{\lambda}$ has been suggested as a useful diagnostic tool because it supplies the user with information about how the missing data contributes to the inferential uncertainty about the parameters of the model [19]. However, it is necessary to evaluate the appropriateness of $\hat{\lambda}$ in the survival analysis setting. In this setting, the amount of missingness at a particular assessment must be determined as a percentage of the size of the risk set at that assessment. Consider the situation in which a student misses an assessment after he has initiated smoking. In the usual sense of missingness, the student can indeed be considered missing at that assessment. However, because he is no longer part of the risk set for that assessment, his absence has no effect on the estimation of time until initiation of smoking. This scenario points to the need to assess the meaningfulness of $\hat{\lambda}$ in this setting.

Another interesting issue for future work involves the confidence interval (4.1) which assesses whether $\hat{\beta}_{method1}$ and $\hat{\beta}_{complete}$ are estimating the same parameter. As discussed in Appendix F, this confidence interval is too conservative because it does not contain a covariance term. One possible way to estimate the covariance between $\hat{\beta}_{method1}$ and $\hat{\beta}_{complete}$ is by bootstrapping. In this procedure, 1000 data sets are sampled without replacement from the original complete data set discussed in Chapter 3 ($n = 1210$). The result of this procedure is the *ith* version of the original complete data set (also with $n = 1210$), which is then converted to a classical data set as described in Section 4.1. The logistic-regression analysis is run on the *ith* version of the complete data to produce $\hat{\beta}_{i,complete}$ and then on the corresponding classical data set to produce $\hat{\beta}_{i,method1}$. Repeating this process 1000 times results

in 1000 $\hat{\beta}_{method1}s$ and 1000 $\hat{\beta}_{complete}s$. The sample covariance $\hat{\beta}_{method1}$ and $\hat{\beta}_{complete}$ is then calculated by using these two sets of numbers. Future work includes implementing this idea for each of the k methods, $k = 1, 2, 3$.

Appendix A

Theory behind Multiple Imputation

A.1 Overview

The multiple imputation methods used in this paper and discussed in the following appendices are those methods described by Schafer [19]. Consider an n x K data matrix Y, where the n rows of the matrix represent subjects and the K columns represent variables measured on each of the subjects. The rows of Y are modeled as independent, identically distributed (iid) draws from a multivariate distribution with a parameter vector denoted by θ. It is not always possible to observe all of the entries of Y. Let the missing parts of Y be denoted by Y_{mis} and the observed parts of Y by Y_{obs}.

Because multiple imputation is a Bayesian method, a distribution for the complete data Y must be specified. This distribution depends on an unknown parameter θ. Next, a probability distribution for θ, called the prior distribution, must be specified. The prior distribution is a subjective distribution and represents the researcher's knowledge of or belief about the distribution of θ before the data is observed. In many cases, no strong prior information is available about θ, and a noninformative prior is chosen. A sample is then taken from the population and the prior distribution is updated with the information from the sample. The updated prior is called the posterior distribution of θ given the data. The specification of the distributions of Y and θ can be used to deduce the distribution of Y_{mis} given Y_{obs}. Then, each imputation is a random draw from the distribution of Y_{mis} given Y_{obs}.

In Section A.1, some important assumptions underlying the multiple-imputation

procedure are discussed. In Section A.2, the multivariate distribution and prior used in this paper are given. In many cases, the distribution of Y_{mis} given Y_{obs} is difficult to compute, so a method called data augmentation is used to generate draws from this distribution. This is discussed in Section A.3.

The imputation procedure essentially fills in the missing parts of Y to create an alternate version of the complete data. Repeating the procedure several times yields several versions of the complete data, each of which is analyzed using standard complete-data techniques. The results of these analyses are then combined. The rules for combining the estimates and standard errors from the complete-data analyses are given in Section A.4.

In typical settings where fractions of missing information are modest, the number of imputations needed for valid inferences is as small as 3 or 5. It is possible to create more than this, but the resulting gain in efficiency is typically unimportant. This concept is discussed by Schafer [19] and Rubin [17].

A.2 Assumptions

In this multiple-imputation method, the rows of the data matrix Y are assumed to be independent, identically distributed draws from a known multivariate distribution with parameter θ, where θ has a known prior distribution. If the variables are continuous, a common model used is the multivariate normal distribution. For categorical data, the multinomial model may be used. There also exist multiple imputation routines for a class of models for mixed normal and categorical data.

In missing-data problems, knowledge about the mechanism that determines which values are missing is an important factor in choosing an appropriate analysis

and interpreting the results. This multiple-imputation procedure works under the assumption that the data are missing at random, as defined by Rubin [16]. Missing at random refers to a situation in which the probability that an observation is missing may depend on Y_{obs} but not on Y_{mis}. If the data are assumed to be missing at random in a longitudinal study, then whether a response is missing, say at assessment 3, may depend on a subject's response at assessments 1 and 2, but not at assessment 3. In the context of time to initial drug experimentation for students in grades 7 through 10, missingness at the grade 9 assessment may depend on a student's responses at the grade 7 and grade 8 assessments, but not on the response at the grade 9 assessment. In other words, missingness at grade 9 may depend on whether the student had initiated between grades 7 and 8 and on the observed values of the covariates collected at grade 7, but not on whether the student had initiated between grades 8 and 9.

Another assumption of this multiple-imputation procedure is that the parameters θ of the data model and the parameters ξ of the missingness mechanism are distinct. Intuitively, distinctness of θ and ξ refers to the situation in which knowledge about the parameters of the data model and knowledge about the mechanism that determines which subjects are present at the assessments supply little information about each other. Under the assumptions of missing at random and distinctness, the process that creates the missing data can essentially be ignored and inferences about θ can be made without regard to the missing data mechanism. For a mathematical discussion of ignorable missing data mechanisms, refer to Rubin [17].

Inference by multiple imputation can be viewed as a two-stage procedure, where each stage operates under a specified model. The first stage consists of the

creation of the M imputed data sets and the second stage consists of the analyses of each of those data sets and the subsequent combining of the results. It is important, however, that the imputation model and the analysis model are consistent with one another in order for inferences to be valid. In particular, the analysis model should be contained in the imputation model. For example, if the analysis model includes a particular three-way interaction, the imputation model should also contain that interaction. In addition, the unbiasedness of this multiple-imputation method relies on the assumption that the models are consistent with one another.

A.3 Multinomial Model and Dirichlet prior

The following discussion derived in Schafer [19] assumes the n x K complete-data matrix Y can be modeled by a multinomial distribution on the cells of a K-dimensional contingency table. Let $Y_1, Y_2, \ldots, Y_K$ be a set of categorical variables, where Y_j takes possible values $1, 2, \ldots, d_j, j = 1, 2, \ldots, K$. In the context of this paper, Y_1 is the parental monitoring variable and Y_2,...,Y_K are the variables measuring whether the subjects had initiated smoking prior to each of the K - 1 assessments. Since each variable has two levels, $d_j = 2$, $j = 1, 2, \ldots, K$. In this setting, Y can be reduced to a contingency table with D cells where $D = \prod_{j=1}^{K} d_j$ = the number of distinct combinations of the levels of Y_1, Y_2,...,Y_K and where the cells are indexed by $d = 1, 2, \ldots, D$. In other words, each cell refers to a particular response pattern within a level of parental monitoring. For example, one of the cells of the contingency table is initiation at age 10 in the high-monitored group. Another cell might be initiation at age 14 in the low-monitored group.

To clarify the idea of reducing the complete data matrix Y to a contingency

Table A.1: Data Matrix to be Reduced to a Contingency Table

ID	PM	Age9	Age10
1	2	2	2
2	1	2	1
3	2	1	1
4	2	2	2
5	1	2	2
6	1	2	1
7	2	2	2
8	1	2	2

table, consider a small data set ($n = 8$) in which each subject is eight years old at the first assessment and has not initiated smoking prior to that assessment. Each subject is then measured on level of parental monitoring at the first assessment and whether he/she initiated smoking prior to age 9 and prior to age 10. Table 4.4 gives the data for this example. For the parental monitoring variable (PM), 1 indicates low parental monitoring and 2 indicates high parental monitoring. For Age9 and Age10, 1 = prior cigarette use and 2 = no prior use. Recall that only the information collected after the first assessment is included for each subject; therefore, for simplicity, a column for Age8 is not included in the table.

The data matrix in Table A.1 can be reduced to the contingency table presented in Table A.2. Since there are three variables that are being measured, each with 2 levels, $K = 3$ and the number of cells $D = 2^3 = 8$. Note that each cell corresponds to a different response pattern, where the first value in the pattern refers to the level of parental monitoring. For example, response pattern "1 2 1" refers to those subjects who are in the low parental monitoring group and who initiate at age 10. Response pattern "2 1 1" refers to those subjects who are in the high parental

Table A.2: Contingency Table

Response Pattern	Number of Subjects
1 2 1	2
1 1 1	0
1 2 2	2
2 1 1	1
2 2 1	0
2 2 2	3
1 1 2	0
2 1 2	0

monitoring group and who initiate at age 9. Note also that the contingency table contains cells for response patterns that are logically impossible (patterns "1 1 2" and "2 1 2"). For example, pattern "1 1 2" refers to low-monitored subjects who report prior use at age 9 and no prior use at age 10. These cells are called structural zeros and are described in more detail in Appendix B.

Assuming that the subjects are independent and identically distributed and the sample size n is fixed, the model for this data is multinomial with parameter $\theta = (\theta_1, \theta_2, ..., \theta_D)$, where θ_d is the probability that a subject falls into cell or response pattern d. In the example from Table A.2, $D = 2^3$ and $\theta_1 =$ the probability that the response pattern "1 2 1" occurs, i.e. that a subject is in the low parental monitoring group and initiates at age 10. Let y_d be the number of subjects who fall into cell d so that $Y = (y_1, y_2, ..., y_D)$ is the set of cell frequencies ($y_1 = 2$). The probability distribution for Y is then

$$P(Y|\theta) = \frac{n!}{y_1!y_2!\cdots y_D!}\theta_1^{y_1}\theta_2^{y_2}\cdots\theta_D^{y_D}$$

for $\sum_{d=1}^{D} y_d = n$ and 0 otherwise.

The model is assumed to be saturated, that is, there are no restrictions on the probabilities except that they are in the interval [0,1] and they sum to 1. All two-way and higher order associations are permitted.

One of the features of Bayesian inference is the use of prior distributions in order to introduce additional information about the parameter θ. Adding even a small amount of information about θ can be useful in obtaining unique and stable estimates, especially in situations where the data are considered "abnormal". Examples of data abnormalities include multiple modes, ridges, saddlepoints, and boundary solutions [19].

The Dirichlet distribution is often used as a prior distribution for the multinomial model. The Dirichlet prior density is

$$\pi(\theta) = k(\alpha_1, \alpha_2, \ldots, \alpha_K)\theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1} \cdots \theta_K^{\alpha_K - 1},$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ are the user-specified hyperparameters and

$$k(\alpha_1, \alpha_2, \ldots, \alpha_K) = \frac{\Gamma(\alpha_1 + \alpha_2 + \ldots + \alpha_K)}{\Gamma(\alpha_1)\Gamma(\alpha_2) \cdots \Gamma(\alpha_K)},$$

where $\Gamma(\cdot)$ denotes the gamma function.

For a thorough discussion of the multinomial model, Dirichlet prior, and related issues, refer to Schafer, Chapters 7 and 8 [19].

A.4 Data Augmentation

The multiple-imputation methodology can be implemented via a data augmentation algorithm. Data augmentation is a member of the Markov chain Monte Carlo family of techniques, which are iterative methods for creating pseudorandom draws from probability distributions. Data augmentation is often used in situations

in which the conditional distribution of the missing data given the observed data is intractable, but the conditional distribution of the parameter θ given only the complete data and the conditional distribution of the missing data given both the observed data and the parameter θ are easily simulated. Data augmentation can be used to create imputations of the missing data because it generates draws from the distribution of the missing data given the observed data.

In the algorithm, a starting value for the parameter θ is designated. The starting value may be the maximum likelihood estimate or posterior mode for the model, such as the one that is produced by the EM algorithm [10]. Each iteration of data augmentation then consists of a random imputation of the missing data given the observed data and the current parameter value (Imputation- or I-step), followed by a draw from the posterior distribution of the parameter given the observed data and the imputed data (Posterior- or P-step). By subsampling the chain M times at say, every 100th I-step, M approximately-independent imputations of the missing data can be created. Each of these M imputations are essentially a complete-data version of Y; that is, an n x K matrix with no missing values. Each imputed data set can then be analyzed using standard complete-data methods. After the M analyses are performed, the results can be combined into one overall inferential statement about the parameters of interest according to the rules explained below [19].

A.5 Rules for forming the overall inferential statement

The following rules for combining the complete-data estimates and variances are presented in Schafer [19] and in Rubin [17].

Let $\hat{Q}$ be the complete-data point estimate for Q, where Q is any function

of the parameters of the model. For example, Q may be the regression parameter β_9 in the logistic discrete-time hazard model explained in Section 5.1. Let U be the variance estimate associated with Q. With M imputations, M versions of $\hat{Q}$ and U can be calculated. Letting $\hat{Q}^{(j)}$ be the point estimate of the j^{th} set of imputed data and $U^{(j)}$ be the variance estimate of the j^{th} set of imputed data, $j = 1, \ldots, M$, the following estimates can be calculated.

The point estimate for Q is the average of the complete-data estimates,

$$\bar{Q} = \frac{1}{M}\sum_{j=1}^{M}\hat{Q}^{(j)}.$$

The within-imputation variance $\bar{U}$ is the average of the complete-data variance estimates,

$$\bar{U} = \frac{1}{M}\sum_{j=1}^{M}U^{(j)}.$$

The between-imputation variance B is the variance of the complete-data point estimates,

$$B = \frac{1}{M-1}\sum_{j=1}^{M}(\hat{Q}^{(j)} - \bar{Q})^2.$$

The total variance T is then computed as

$$T = \bar{U} + (1 + M^{-1})B.$$

Inferences for Q are based on

$$T^{-1/2}(Q - \bar{Q}) \sim t_\nu,$$

where the degrees of freedom are given by

$$\nu = (M-1)\left[1 + \frac{\bar{U}}{(1+M^{-1})B}\right]^2.$$

The fraction of missing information is estimated by

$$\hat{\lambda} = \frac{r + 2/\nu + 3}{r + 1},$$

where

$$r = \frac{(1 + M^{-1})B}{\bar{U}}.$$

Appendix B

Implementation of Multiple Imputation

Splus software developed by J.L. Schafer [19] for the implementation of the multiple-imputation procedure described in Appendix A is available via the Internet and the World Wide Web. It can be found at http://www.stat.psu.edu/~jls/misoftwa.html. There are four different Splus software packages available: NORM, CAT, MIX, and PAN. Each package contains a README file, which provides instructions on installation and use of the program, so this file should be read first.

The following discussion presents some of the issues encountered when implementing the multiple-imputation procedure in Splus for the classical parental monitoring data set presented in Chapter 4. The analysis of the data set used CAT due to the categorical nature of the variables.

B.1 Data Matrix

Recall that for the classical version of this data set, it is of interest to fill in the intermittent missed assessments. Included in the data matrix then, are each subject's responses to the question regarding prior use of cigarettes at each assessment for which the student was present. Also included is the value of the parental monitoring variable indicating whether the subject was categorized in the low or high parental monitoring group at the first assessment. Table B.1 shows an example of how the data may appear for eight subjects. ID is an identification code for each subject, and PM is the level of parental monitoring where 1 = low parental monitoring and 2 = high parental monitoring. The variables Age8 - Age14 refer to

Table B.1: Example of Data Matrix

ID	PM	Age8	Age9	Age10	Age11	Age12	Age13	Age14	Startage	Endage
1	2	NA	NA	2	2	2	2	NA	10	13
2	1	NA	2	2	2	2	2	NA	9	13
3	2	NA	2	2	2	2	2	NA	9	13
4	2	NA	NA	2	NA	2	2	2	10	14
5	2	NA	2	2	2	2	2	NA	9	13
6	2	2	NA	2	2	2	2	NA	8	13
7	2	NA	2	1	1	1	1	1	9	14
8	2	NA	NA	NA	2	2	2	NA	11	13

the subject's cigarette-use status at each of the assessments, where 1 = prior use, 2 = no prior use, and NA = missing. There are three reasons why in this data set, a student may have a missing value (NA) at any particular assessment: it may be that the student has not yet entered the study, or that the student has already left the study, or that the student actually missed that assessment. The first two situations can be identified by using the variables Startage and Endage. For example, consider student 8. This student has four missing values (ages 8, 9, 10, and 14); however, these missing values are due to the fact that he did not enter the study until the age 11 assessment and left the study after the age 13 assessment. In contrast, student 4 has missing values for 3 assessments; two for which he had not yet entered the study (ages 8 and 9) and one for which he actually missed the assessment (age 11).

B.2 Empty Cells

The K-dimensional contingency table referred to in Section A.3 may contain empty cells for two reasons. If the event corresponding to the empty cell is

logically impossible, then the cell is said to contain a *structural zero.* (This is the case for the cell corresponding to response pattern "1 1 2" in the example from Section A.3). If the cell is empty simply by chance, that is, it is empty not because the event could not occur, but because by a matter of happenstance, it did not occur in this particular data set, then this cell is considered a *random zero.* (This is the case for the cell corresponding to response pattern "2 2 1" in the example from Section A.3). Both situations require special handling in the implementation of the EM and/or data augmentation algorithms.

B.2.1 Specification of Structural Zeros

Consider θ to be an array, where the dimensions of the array are determined by the number of variables and the number of levels of each variable in the data set. Because there are 8 variables of interest in the data set (the parental monitoring variable and 7 response variables), each with 2 levels, θ is a 2^8 array.

In order to specify the structural zeros for the implementation of the EM and data augmentation algorithms, another array of the same dimension as θ needs to be declared. The elements or cells of this array are the hyperparameters for the Dirichlet prior distribution. If structural zeros are present, it is necessary to determine which cells of the array correspond to the structural zeros and then to assign the value of "NA" to these cells.

For the parental monitoring data set, let "priorem" and "priorda" each represent a 2^8 array. The default Dirichlet prior distribution for the EM algorithm is one in which all hyperparameters are equal to 1. Assuming there are no random zeros, "priorem" is first declared to be a 2^8 with all cells in the array equal to 1 by the Splus command

```
priorem_array(1,c(2,2,2,2,2,2,2,2)).          (B.1)
```

The subscripts of the array follow the order of the variables in the data set. For example, the first subscript of the array refers to the parental monitoring variable, the second subscript to the Age8 variable, the third to the Age9 variable, and so on.

For the data augmentation algorithm, the default Dirichlet prior distribution is one in which all hyperparameters are equal to 0.5 (Jeffreys prior) for reasons given in Schafer [19]. The Splus command used to declared "priorda" as a 2^8 array with all cells equal to 0.5 is

```
priorda_array(0.5,c(2,2,2,2,2,2,2,2)).
```

Because it is impossible for a student to report no prior use following a response of prior use, it is necessary to specify these situations as structural zeros. For example, it is impossible for a student to have the following data.

PM	Age8	Age9	Age10	Age11	Age12	Age13	Age14	Startage	Endage
2	NA	2	1	2	2	2	2	9	14

This is a situation in which a student would have reported no prior use at the age 9 assessment, prior use at the age 10 assessment, and no prior use at the age 11 assessment; an event that is not logically possible within the context of this study. Therefore, this event is assigned a structural zero for the EM algorithm in Splus by

```
priorem[ , ,1,2, , , ,]_NA.
```

This command tells Splus that the situation in which Age10 = 1 and Age11= 2 is impossible regardless of the values of the other cells in the array. Other structural

zeros are specified in a similar way for both the EM and the data augmentation algorithms. For further details regarding arrays in S, refer to Becker et al [6].

For those assessments in which the student has a missing value because he/she has not yet entered the study, the imputation procedure will fill in a "2" for no prior use. This is another example of a structural zero. Because all students considered in this study have not initiated smoking at the first assessment (i.e. they have a "2" for the first interval), the imputation procedure will not fill in prior use (a "1") for those assessments because this situation has been specified to be logically impossible. For those assessments for which the student has already left the study, it is possible that the imputation procedure may fill in prior use. In the logistic-regression analysis to be discussed in Appendix C, the last age for which a subject is included in the analysis is the last age at which he is at risk; thus, imputations of later use are discarded.

B.2.2 Potential Problem regarding Random Zeros

The classical parental monitoring data set contains a random zero since there are no initiators in the low parental monitoring group for the age 14 assessment (see Table 4.2). This event is possible, but by happenstance, it did not occur in this data set. As a result, using the EM algorithm (with the Dirichlet hyperparameters = 1 and the structural zeros specified as above) to obtain a reasonable starting value for θ produces an estimate of θ located on the boundary of the parameter space; i.e. the estimated probability of a low-monitored student initiating smoking at age 14 is zero. If this value of θ from the EM algorithm was used as a starting value for the data augmentation algorithm, a warning message would be issued by Splus: "Warning message: Starting value on the boundary".

There are two ways to remedy this situation. One is to change the starting value of θ slightly so that the element of θ corresponding to the random zero actually has a value slightly greater than zero. This can be done by reallocating probability from say, the high-monitored group at age 14 to the low-monitored group at age 14. In this way, the estimated probabilities of the cells still sum to one, as required by the multinomial model, yet the starting value of θ will not be on the boundary.

Another way in which the situation can be remedied is to apply, in the implementation of EM, a Dirichlet prior distribution in which all hyperparameters are greater than one. For example, one could apply a Dirichlet prior with all hyperparameters equal to 1.1 as was done for this analysis. This is done by replacing (B.1) in the Splus code by the command

```
priorem_array(1.1,c(2,2,2,2,2,2,2,2)).
```

Employing this prior distribution is intuitively equivalent to adding 0.1 prior observations to each cell. In this way, the event corresponding to initiation in the low parental monitoring at age 14 has a positive (but very small) probability of occurring.

B.3 Splus code

The following code illustrates how the multiple imputation procedure was implemented for the parental monitoring data.

```
s_prelim.cat(y)     #performs preliminary manipulations on
                    #the matrix y which contains missing values


s$nmis              #to view the number of missing values
                     #for each variable in y


priorem_array(1.1,c(2,2,2,2,2,2,2,2))     #prior is a 2^8 array
priorem[,1,1,1,1,1,1,1]_NA
priorem[,1,2,,,,,]_NA                     #structural zeros set to NA
priorem[,1, ,2,,,,]_NA
priorem[,1, , ,2,,,]_NA                   #Dirichlet prior with
priorem[,1, , , ,2,,]_NA                  #hyperparameters = 1.1
priorem[,1, , , , ,2,]_NA
priorem[,1, , , , , ,2]_NA

priorem[,,1,2,,,,]_NA
priorem[,,1, ,2,,,]_NA
priorem[,,1, , ,2,,]_NA
priorem[,,1, , , ,2,]_NA
priorem[,,1, , , , ,2]_NA

priorem[,,,1,2,,,]_NA
```

```
priorem[,,,1, ,2,,]_NA
priorem[,,,1, , ,2,]_NA
priorem[,,,1, , , ,2]_NA

priorem[,,,,1,2,,]_NA
priorem[,,,,1, ,2,]_NA
priorem[,,,,1, , ,2]_NA

priorem[,,,,,1,2,]_NA
priorem[,,,,,1, ,2]_NA

priorem[,,,,,,1,2]_NA

thetahat_em.cat(s,prior=priorem)        #EM algorithm

rngseed(12345)                          #random number seed is initialized
                                        #for the data augmentation algorithm


priorda_array(.5,c(2,2,2,2,2,2,2,2))      #structural zeros set to NA
priorda[,1,1,1,1,1,1,1,1]_NA
priorda[,1,2,,,,,,]_NA                    #Jeffreys prior used
priorda[,1, ,2,,,,]_NA
priorda[,1, , ,2,,,]_NA
priorda[,1, , , ,2,,]_NA
priorda[,1, , , , , ,2,]_NA
priorda[,1, , , , , , ,2]_NA

priorda[,,1,2,,,,]_NA
priorda[,,1, ,2,,,]_NA
```

```
priorda[,,1, , ,2,,]_NA
priorda[,,1, , , ,2,]_NA
priorda[,,1, , , , ,2]_NA

priorda[,,,1,2,,,]_NA
priorda[,,,1, ,2,,]_NA
priorda[,,,1, , ,2,]_NA
priorda[,,,1, , , ,2]_NA

priorda[,,,,1,2,,]_NA
priorda[,,,,1, ,2,]_NA
priorda[,,,,1, , ,2]_NA

priorda[,,,,,1,2,]_NA
priorda[,,,,,1, ,2]_NA

priorda[,,,,,,1,2]_NA



theta_da.cat(s,start=thetahat,prior=priorda,    #data augmentation
steps=1000,showits=T)
                                                #burn-in period of
                                                #1000 steps

                                                #start data augmentation
                                                #at thetahat


imp1_imp.cat(s, theta)                          #imputation 1
```

```
theta2_da.cat(s,start=theta,prior=prior,
steps=100,showits=T)
imp2_imp.cat(s,theta2)                          #imputation 2


theta3_da.cat(s, start=theta2,prior=prior,
steps=100,showits=T)
imp3_imp.cat(s,theta3)                          #imputation 3

theta4_da.cat(s,start=theta3,prior=prior,
steps=100,showits=T)
imp4_imp.cat(s,theta4)                          #imputation 4

theta5_da.cat(s,start=theta4,prior=prior,
steps=100,showits=T)
imp5_imp.cat(s,theta5)                          #imputation 5
```

After the analysis is performed on each of the 5 imputed data sets using the SAS code in Appendix D, the following Splus command is used to combine the estimates:

```
combo_mi.inference(est,se)
```

Here "est" is a list of five vectors of estimated regression coefficients from the logistic-regression analyses of the five imputed data sets and "se" is a list of five vectors containing standard errors from the regression analyses corresponding to the estimates in "est".

Appendix C

Computing the Parameter Estimates of the Discrete Time Hazard Model in SAS

PROC LOGISTIC in SAS produces estimates of the parameters of the discrete-time hazard model for each of the imputed data sets. Note those subjects who have already initiated cigarette use at the beginning of the study are not included in the estimation process. Again, the risk set for a time interval is considered to be those students who are in the study at the beginning of the interval, have not yet begun cigarette use at the beginning of the interval, and who are present at the assessment at the end of the interval. Once a student initiates smoking, the student is no longer part of the risk set.

Each of the imputed data sets takes the form of one line of data for each subject. Each line contains information about the subject's cigarette use over time, along with the subject's level of parental monitoring at the first assessment. Table C.1 shows an example of how the imputed data may appear for 3 subjects. ID is an identification code for each student, PM (parental monitoring) is a time-invariant predictor with two levels, 1 and 2, and Age8 - Age14 are the subject's cigarette use status at each assessment, where a 1 indicates prior cigarette use at that assessment and a 2 indicates no prior use. Note in this data set, both left truncation and right censoring occurs. Subjects 2 and 3 have delayed entry times (evident from the value of Startage) and therefore the data is left-truncated. In addition, subjects 1 and 3 are right-censored since neither initiate smoking during the time they were present in the study. Even though the imputation routine filled in a "1" for age 14 for

Table C.1: Imputed Data Set

ID	PM	Age8	Age9	Age10	Age11	Age12	Age13	Age14	Startage	Endage
1	2	2	2	2	2	2	2	2	8	13
2	2	2	2	2	1	1	1	1	9	14
3	1	2	2	2	2	2	2	1	10	13

subject 3, note that the subject was only present in the study until age 13 (evident from the value of Endage). Therefore, the imputed initiation of smoking for subject 3 at age 14 is not included in the analysis.

In order to perform the logistic regression, the data set with one line of data for each subject needs to be converted into a person-period data set in which each subject has multiple lines of data, one for each time interval the subject is considered at risk. Table C.2 displays the data converted into the person-period format. Here, ID is again used to identify the student, T9-T14 are dummy variables

Table C.2: Person-Period Data Set

ID	PM	T9	T10	T11	T12	T13	T14	Y
1	2	1	0	0	0	0	0	2
1	2	0	1	0	0	0	0	2
1	2	0	0	1	0	0	0	2
1	2	0	0	0	1	0	0	2
1	2	0	0	0	0	1	0	2
2	2	0	1	0	0	0	0	2
2	2	0	0	1	0	0	0	1
3	1	0	0	1	0	0	0	2
3	1	0	0	0	1	0	0	2
3	1	0	0	0	0	1	0	2

indicating the age, PM is a time-invariant predictor with two levels, 1 and 2, and Y is an indicator variable which equals 1 if the student began smoking during that age and 2 if the subject did not. Recall that the only data considered is data collected

after the first assessment. For example, even though student 1 was present in the study at age 8 (Startage = 8), the first age he is included in the person-period data set is age 9.

After conversion of each imputed data set into the person-period format, PROC LOGISTIC can be applied to arrive at parameter estimates of the discrete-time hazard model. The code in Appendix D shows how the conversion of the data into a person-period dataset and the analysis using PROC LOGISTIC can be implemented.

Word of Caution

Recall the logistic regression model presented in Section 5.1,

$$log\frac{h_{ij}}{1-h_{ij}} = \alpha_j + \beta_j Z_i.$$

As mentioned in Section B.2.2, the parental monitoring data set contains no initiators at age 14 in the low-monitored group. This situation forces the researcher to reconsider the model to be fit in PROC LOGISTIC. If β_{14} were included in the logistic regression model, the resulting estimate $\hat{\beta}_{14}$ would be infinite and the maximum likelihood estimate would not exist. SAS attempts to warn the user by issuing a message regarding infinite parameters or a complete or quasi-complete separation in the data. Therefore, β_{14} should not be included in the fitting of the model. For more details regarding infinite parameters and complete and quasi-complete separations, refer to *A Tutorial on Logistic Regression* [22].

In the parental monitoring study, data was collected on some individuals for ages 15 and 16. However, a frequency table of initiation by age showed that there were no initiators at ages 15 and 16 in either parental monitoring group. Therefore,

neither the α or β parameters representing the responses at age 15 and age 16 (α_{15}, α_{16}, β_{15}, β_{15}) can be included in the model for similar reasons as explained above.

The absence of initiation at ages 15 and 16 also raises another issue concerning the creation of the person-period data set. Including lines of data for ages 15 and 16 in the person-period data set can have severe repercussions when fitting a model in which parental monitoring is treated as time invariant. The resulting estimates are incorrect and cannot be used to make sound statistical conclusions. As a rule of thumb, the person-period data set should only include lines of data for ages at which the subject is considered at risk and for which the corresponding parameters have been included in the model statement.

Each of the situations described above reflects a lack of information in the data set about one or more of the parameters in the model. Because these situations are common in real data sets, it is crucial that the researcher is familiar with his or her data and knows how to account for the idiosyncracies of the data in the analysis.

Appendix D

SAS Code for Implementation of PROC LOGISTIC

```
/*************************************************************
This program is an adaptation of Singer and Willett's code [21]

This program creates the person period data set for the first
     imputed dataset and performs analysis using PROC LOGISTIC

*************************************************************/

OPTIONS PS=60 LS=80 NODATE NONUMBER;  *  Set page size to 60 lines
                                         and line size to 80 characters
                                         and do not show the date or
                                         page number in output;



data start;
        infile "imp1.dat";
        input idcode pm age9 age10 age11 age12 age13 age14 startage endage;

                                    /* imp1.dat is the first imputed */
                                    /* data set from the multiple    */
                                    /* imputation procedure          */

        if pm = 2 then pmbin = 0;     /* recode parental monitoring  */
```

```
            else pmbin = 1;          /*     variable for easier     */
                                     /*       interpretation        */

/*  The next seven data steps determine the imputed age of initiation
    where itobage = age of initation if subject initiated during study
                  = 99 if subject did not initiate                    */
data one;
    set start;
    array age[7] age8-age14;
    itobage=.;
    i = min(startage-7,endage-7,14-7);
       if age[i] = 1 then itobage = i + 7;
         else itobage = .;

data two;
    set one;
    array age[7] age8-age14;
    i = min(startage + 1 -7, endage-7,14-7);
    if itobage = . then
              if age[i] = 1 then itobage = i + 7;
                   else itobage = .;

 data three;
    set two;
    array age[7] age8-age14;
    i = min(startage + 2 -7, endage-7,14-7);
    if itobage = . then
              if age[i] = 1 then itobage = i + 7;
                   else itobage = .;
```

```
 data four;
    set three;
    array age[7] age8-age14;
    i = min(startage + 3 -7, endage-7,14-7);
    if itobage = . then
              if age[i] = 1 then itobage = i + 7;
                   else itobage = .;

data five;
    set four;
    array age[7] age8-age14;
    i = min(startage + 4 -7, endage-7,14-7);
    if itobage = . then
              if age[i] = 1 then itobage = i + 7;
                   else itobage = .;

data six;
    set five;
    array age[7] age8-age14;
    i = min(startage + 5 -7, endage-7,14-7);
    if itobage = . then
              if age[i] = 1 then itobage = i + 7;
                   else itobage = .;

data itobage;
    set six;
         if itobage= . then itobage=99;
```

```
/*  This data step determines if the subject is censored */

data censor;
    set itobage;
    if itobage = 99 then censor=1;
         else censor=0;

/*  This data step determines the last age for which the subject is
                                                  considered at risk   */
data lastint;
    set censor;
         if censor = 0 then lastint = itobage;
              else lastint = endage;



/*  This data step creates the person period data set */

data persper;
    set lastint;
                                                   /* y=1 begin use  */
                                                   /* y=2 not begin use */

array age[7] age8-age14;
array pmage[7] pmage8-pmage14;                    /* pmage is the interaction
do period = (startage + 1) to min(lastint,14);       between age and pm*/
         if period = lastint and censor = 0 then y = 1;
              else y = 2;
    do index = 1 to 7;
         if (index+7) = period and (index+7) > startage and (index+7)<= lastint
```

```
            then age[index] =1;
                 else age[index]=0;
        pmage[index]=pmbin*age[index];
    end;
    output;
end;



data analysis;
    set persper;

/* This proc fits baseline model with time indicators only  */

    proc logistic out=estimate;
        model y =  age9 age10 age11 age12 age13 age14 / noint;
        title "Baseline Model";

 /*  This proc fits the model with main effect of parental monitoring */

     proc logistic data=persper  out=estimate;
     model y = age9 age10 age11 age12 age13 age14 pmbin / noint;
     title "Model with Main Effect of Parental Monitoring with binary coding";


 /*  This proc fits the model with interaction between monitoring and times */

 proc logistic data=persper  out=estimate;
     model y = age9 age10 age11 age12 age13 age14
```

```
pmage9 pmage10 pmage11 pmage12 pmage13  / noint;
title "Model with Interaction between Parental Monitoring and Time";


run;
```

Appendix E

Data from Imputations 3, 4, and 5

Table E.1: Amount of Information Contained in Imputation 3

Information Category		Parental Monitoring	Age					
			9	10	11	12	13	14
1	Number of subjects known to be at risk for initiation at end of age i	Low	104	240	266	240	203	89
		High	243	593	718	638	556	264
2	Number of subjects known to initiate at age i	Low	7	17	27	14	14	0
		High	7	36	41	47	28	3
3	Number of subjects known to have initiated prior to age i and are present in the study at age i	Low	0	7	24	49	57	46
		High	0	7	41	78	115	88
4	Number of subjects yet to enter the study at age i	Low	227	71	6	1	0	0
		High	622	230	19	0	0	0
5	Number of subjects who have left the study before or at age i	Low	0	3	15	34	64	203
		High	0	6	53	109	173	517
	Total Number of subjects	Low	338	338	338	338	338	338
		High	872	872	872	872	872	872
		Total	1210	1210	1210	1210	1210	1210

Table E.2: Amount of Information Contained in Imputation 4

Information Category		Parental Monitoring	Age					
			9	10	11	12	13	14
1	Number of subjects known to be at risk for initiation at end of age i	Low	104	243	265	241	204	90
		High	245	594	717	640	558	264
2	Number of subjects known to initiate at age i	Low	7	14	31	12	14	0
		High	5	37	43	44	28	5
3	Number of subjects known to have initiated prior to age i and are present in the study at age i	Low	0	7	21	50	56	45
		High	0	5	40	79	113	86
4	Number of subjects yet to enter the study at age i	Low	227	71	6	1	0	0
		High	622	230	19	0	0	0
5	Number of subjects who have left the study before or at age i	Low	0	3	15	34	64	203
		High	0	6	53	109	173	517
	Total Number of subjects	Low	338	338	338	338	338	338
		High	872	872	872	872	872	872
		Total	1210	1210	1210	1210	1210	1210

Table E.3: Amount of Information Contained in Imputation 5

Information Category		Parental Monitoring	Age					
			9	10	11	12	13	14
1	Number of subjects known to be at risk for initiation at end of age i	Low	104	241	268	244	205	90
		High	245	594	717	637	556	263
2	Number of subjects known to initiate at age i	Low	7	16	26	12	16	1
		High	5	37	43	47	27	4
3	Number of subjects known to have initiated prior to age i and are present in the study at age i	Low	0	7	23	47	53	44
		High	0	5	40	79	116	88
4	Number of subjects yet to enter the study at age i	Low	227	71	6	1	0	0
		High	622	230	19	0	0	0
5	Number of subjects who have left the study before or at age i	Low	0	3	15	34	64	203
		High	0	6	53	109	173	517
	Total Number of subjects	Low	338	338	338	338	338	338
		High	872	872	872	872	872	872
		Total	1210	1210	1210	1210	1210	1210

Appendix F

Explanation of Confidence Interval

For each of the ages j, $j = 9, \ldots, 13$, consider the following conservative 95% confidence interval,

$$\hat{\beta}_{complete} - \hat{\beta} \pm 1.96\sqrt{(s.e.(\hat{\beta}_{complete}))^2 + (s.e.(\hat{\beta}))^2}, \tag{F.1}$$

where $\hat{\beta}_{complete}$ and $s.e.(\hat{\beta}_{complete})$ are the parameter estimate and standard error of β from the analysis on the complete version of the data and $\hat{\beta}$ and $s.e.(\hat{\beta})$ are the corresponding parameter estimate and standard error for β from one of the three methods. The impetus for forming this confidence interval is to address the question, "Do $\hat{\beta}_{complete}$ and $\hat{\beta}$ estimate the same parameter?"

It is difficult to estimate the correlation between the complete and simulated data sets and as a result, a covariance term is not included in the confidence interval. If the covariance term were included, it would appear under the square root as a negative quantity. It is clear, however, that across repeated samples from the population, $\hat{\beta}_{complete}$ and $\hat{\beta}$ should be positively correlated and so the confidence interval (F.1) is a conservative 95% confidence interval for 0 (because $\hat{\beta}_{complete}$ and $\hat{\beta}$ are estimates of the same parameter).

In the simulations of size 1000 described in Section 5.4, put for the *ith* confidence interval,

$$I_i = \begin{cases} 1 & \text{if (F.1) contains zero} \\ 0 & \text{otherwise} \end{cases}$$

The I_i, $i = 1, \ldots, 1000$, are identically distributed, but they are not independent. This is due to the fact that the 1000 simulated data sets are not independent since they are generated from the same complete data set.

It is of interest to estimate the probability the confidence interval contains 0 ($P[I_i = 1]$). If $\hat{\beta}_{complete}$ and $\hat{\beta}$ are estimating the same parameter, then this probability should be .95. Let the estimator A $= \frac{1}{1000} \sum_{i=1}^{1000} I_i$. Then A is an unbiased estimate of $P[I = 1]$ because

$$\begin{aligned} E\left(\frac{1}{1000}\sum_{i=1}^{1000} I_i\right) &= \frac{1}{1000}\sum_{i=1}^{1000} E(I_i) \\ &= \frac{1}{1000}\sum_{i=1}^{1000} P[I_i = 1](1) + P[I_i = 0](0) \qquad \text{(F.2)} \\ &= \left(\frac{1}{1000}\right)(1000)P[I = 1] \qquad \text{(F.3)} \\ &= P[I = 1], \end{aligned}$$

where equality between (F.2) and (F.3) is due to the fact that all $I_i s$ are identically distributed. Thus, by calculating A and comparing this to 0.95 for each of the three methods, it is possible to assess if $\hat{\beta}_{complete}$ and $\hat{\beta}$ are estimating the same parameter.

References

[1] A. Agresti. *Categorical Data Analysis.* J. Wiley & Sons, New York, 1990.

[2] P.D. Allison. *Event History Analysis: Regression for Longitudinal Event Data.* Number 46 in Sage University Paper Series. Sage Publications, 1984.

[3] P.K. Anderson, O. Borgan, R.D. Gill, and N. Keilding. *Statistical Models Based on Counting Processes.* Springer-Verlag New York, Inc., 1993.

[4] S.F. Arnold. *Mathematical Statistics.* Prentice-Hall, Inc., Englewood Cliffs, NJ, 1990.

[5] J.G. Bachman, K.N. Wadsworth, P.M. O'Malley, L.D. Johnston, and J.E. Schulenberg. *Smoking, Drinking, and Drug Use in Young Adulthood: The Impacts of New Freedoms and New Responsibilities.* Research Monographs in Adolescence. Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey, 1997.

[6] R.A. Becker, J.M. Chambers, and A.R. Wilks. *The New S Language: A Programming Environment for Data Analysis and Graphics.* Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, CA, 1988.

[7] G. Casella and R.L. Berger. *Statistical Inference.* Wadsworth, Inc., Belmont, CA, 1990.

[8] H. D. Chilcoat and J. C. Anthony. Impact of parent monitoring on initiation of drug use through late childhood. *Journal of the American Academy of Child and Adolescent Psychiatry*, 35(1):91–100, January 1996.

[9] H. D. Chilcoat, T. J. Dishion, and J. C. Anthony. Parent monitoring and the incidence of drug sampling in urban elementary school children. *American Journal of Epidemiology*, 141(1):25–31, 1995.

[10] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–22, 1977.

[11] B. Efron. Logistic regression, survival analysis, and the kaplan-meier curve. *Journal of the American Statistical Association*, 83(402):414–425, June 1988.

[12] W. B. Hansen and J. W. Graham. Preventing alcohol, marijuana, and cigarette use among adolescents: Peer pressure resistance training versus establishing conservative norms. *Preventive Medicine*, 20:414–430, 1991.

[13] J.P. Klein and M.L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data.* Statistics for Biology and Health. Springer, New York, 1997.

[14] L. A. Malacane. Determining the risk set in event history analysis. Master of science, The Pennsylvania State University, May 1996.

[15] M. K. B. Parmar and D. Machin. *Survival Analysis: A Practical Approach.* John Wiley & Sons, Chichester, England, 1995.

[16] D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[17] D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons, 1987.

[18] SAS Institute Inc., Cary, NC. *SAS/STAT User's Guide, Version 6*, fourth edition, 1990.

[19] J.L. Schafer. *Analysis of Incomplete Multivariate Data.* Chapman and Hall, London, 1997.

[20] J. D. Singer and J. B. Willett. Designing and analyzing studies of onset, cessation, and relapse: Using survival analysis in drug abuse prevention research. In L.M. Collins and L.A. Seitz, editors, *Advances in Data Analysis for Prevention Research*, volume 142 of *NIDA Research Monograph Series*, pages 196–263. National Institutes of Health, 1994.

[21] J. D. Singer and John B. Willett. It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, 18:155–195, 1993.

[22] Ying So. *A Tutorial on Logistic Regression.* SAS Institute Inc., Cary, NC. Paper from SUGI 18 Proceedings.

[23] M.E. Stokes, C.S. Davis, and G.G. Koch. *Categorical Data Analysis Using the SAS System.* SAS Institute Inc., Cary, NC, 1995.

[24] M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540, 1987.