

UNIVERSITY OF CALIFORNIA

Los Angeles

Discrete-Time Survival Mixture Analysis
for Single and Recurrent Events Using Latent Variables

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

by

Katherine Elizabeth Masyn

2003

© Copyright by

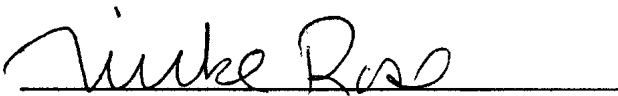
Katherine Elizabeth Masyn

2003

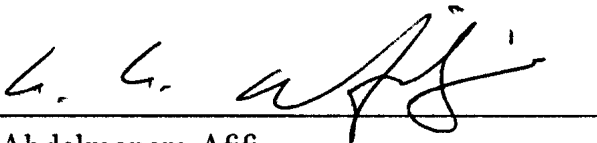
The dissertation of Katherine Elizabeth Masyn is approved.



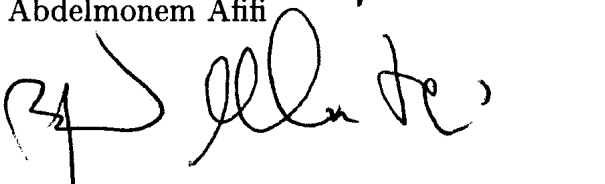
Michael Seltzer



Michael Rose



Abdelmonem Afifi



Bengt Muthén, Committee Chair

University of California, Los Angeles

2003

*To my dad, who taught me to
think independently and follow my heart.*

*To my mom, who taught me to
value the minds and hearts of others.*

Contents

Acknowledgements	xiv
Vita	xvii
Abstract	xxi
1 Introduction	1
1.1 Discrete-time vs. continuous-time survival data	2
1.2 Statement of purpose	4
1.3 Historical background	6
1.3.1 Continuous-time survival analysis	7
1.3.2 Finite mixture models	15
1.3.3 Latent class analysis	17
1.4 Data example	21

2	Single Events	28
2.1	Single event, continuous-time	
	survival analysis	29
2.1.1	Basic notation and event time distributional forms . . .	29
2.1.2	Continuous-time models with covariates	32
2.2	Single event, discrete-time	
	survival analysis	36
2.2.1	Basic notation and event time distributional forms . . .	36
2.2.2	Censoring and truncation	39
2.2.3	Constructing the likelihood	45
2.2.4	Estimation	47
2.2.5	Discrete-time models with covariates	54
2.2.6	Discrete-logit model in a latent variable framework	61
2.2.7	Model assessment	91
3	Unobserved Heterogeneity	93
3.1	Ignoring unobserved heterogeneity	94
3.2	Modeling unobserved heterogeneity	101
3.3	Identifiability	109
3.4	Class enumeration	117

3.4.1	Long-term survivors	137
4	Recurrent Events	156
4.1	Multivariate event histories	157
4.2	Defining risk for recurrent event histories	161
4.3	Basic notation and likelihood	173
4.3.1	Gap time	176
4.3.2	Counting process	180
4.3.3	Total time	184
4.4	Estimation	186
4.4.1	Correcting bias in duration dependence	191
4.5	Unobserved heterogeneity	195
4.6	Example	199
4.6.1	Gap time analysis	203
4.6.2	Counting process analysis	209
4.6.3	Total time analysis	215
4.6.4	Mixture model analysis	216
5	Conclusion	222
5.1	Single events	222
5.2	Unobserved heterogeneity	225

5.3	Recurrent events	229
5.4	Future research	231
Appendix A: Splus Code		235
Appendix B: Mplus input		257
Bibliography		279

List of Figures

2.1	Example hazard and survival probability plots.	40
2.2	Sample hazard and survival probabilities for 12 months post-treatment.	51
2.3	Event history LCA path diagram.	56
2.4	Model 2 estimated hazard and survival probabilities.	68
2.5	Model 2 estimated versus sample survival probabilities by treatment status and wife's education level.	73
2.6	Event history LCR path diagram.	75
2.7	Model 5 estimated hazard and survival probabilities by wife's education level.	84
2.8	Model 5 estimated hazard and survival probabilities by household income.	85
2.9	Model 5 estimated hazard and survival probabilities by % days drinking.	86

3.1	Examples of unobserved heterogeneity.	97
3.2	Examples of unobserved heterogeneity with an observed covariate.	99
3.3	Path diagrams for class enumeration Models 0–3.	125
3.4	Model 6 estimated survival probabilities at baseline.	151
3.5	Model 6 estimated survival probabilities at sample average wife’s education, household income, and % days drinking.	152
3.6	Model 6 estimated survival probabilities for Classes 1 and 2 by wife’s education at sample average household income and % days drinking.	153
3.7	Model 6 estimated survival probabilities for Classes 1 and 2 by household income at sample average wife’s education and % days drinking.	154
3.8	Model 6 estimated survival probabilities for Classes 1 and 2 by % days drinking at sample average wife’s education and house- hold income.	155
4.1	Three subject example of recurrent event observations.	167
4.2	Recurrent event history LCR path diagram.	189
4.3	Estimated contribution of length of relationship (in years) to the logit hazard probabilities of second violent episode.	205

4.4	Model 8a estimated hazard and survival probabilities for second episode by length of relationship.	206
4.5	Model 8b estimated hazard and survival probabilities for third episode by pre-treatment violence and timing of second episode.	210
4.6	Model 10 estimated hazard and survival probabilities for episodes 1-3.	214
4.7	Model 14 estimated hazard and survival probabilities for episodes 1-3.	218
4.8	Model 15 estimated hazard and survival probabilities for first, second, and third episode by latent class.	221

List of Tables

1.1	Descriptive Statistics for Pre- and Post-Treatment % Days Drinking	24
1.2	Frequencies and Proportions for Pre-Treatment Violence Episodes	25
1.3	Descriptive Statistics for Continuous Covariates	25
1.4	Frequencies and Proportions for Post-Treatment Violence Episodes	26
1.5	Frequencies and Proportions for Categorical Covariates	27
2.1	Sample Hazard and Survival Probabilities for First Violence Post-Treatment	50
2.2	Example Data for Discrete-Time Survival Using Event and Risk Indicators	55
2.3	Results for Data Example Model 1	64
2.4	Results for Data Example Model 2	67
2.5	Results for Data Example Model 5	83

3.1	Class Enumeration Models 0–3 Specification	126
3.2	Class Enumeration Populations A–C Definition	127
3.3	Class Enumeration Models 0–3 Results for Population A	128
3.4	Class Enumeration Models 0–3 Results for Population B	129
3.5	Class Enumeration Models 0–3 Results for Population C	130
3.6	Class Enumeration Measures for Population A	131
3.7	Class Enumeration Measures for Population B	132
3.8	Class Enumeration Measures for Population C	133
3.9	Long-term Survivor Populations A–E Definition	140
3.10	Long-term Survivor Models 1–4 Specification	141
3.11	Long-term Survivor Models 1–4 Results for Population A	142
3.12	Long-term Survivor Models 1–4 Results for Population B	143
3.13	Long-term Survivor Models 1–4 Results for Population C	144
3.14	Long-term Survivor Models 1–4 Results for Population D	145
3.15	Long-term Survivor Models 1–4 Results for Population E	146
3.16	1- and 2-Class Model Comparisons	148
3.17	Results for Data Example Model 6	150
4.1	Risk Periods Defined	169
4.2	Example Data for First Event	176
4.3	Example Data for Second and Third Event in GT	179

4.4	Example Data for Second and Third Event in CP	183
4.5	Example Data for Second and Third Event in TT	187
4.6	Sample Frequencies and Proportions for First, Second, and Third Violence Episodes in GT Formulation	200
4.7	Sample Frequencies and Proportions for First, Second, and Third Violence Episodes in CP Formulation	201
4.8	Sample Frequencies and Proportions for First, Second, and Third Violence Episodes in TT Formulation	202
4.9	Results for Data Example Model 8a: Second Episode, GT For- mulation	204
4.10	Results for Data Example Model 8b: Third Episode, GT For- mulation	209
4.11	Results for Data Example Model 10a: Second Episode, CP For- mulation	212
4.12	Results for Data Example Model 10b: Third Episode, CP For- mulation	213
4.13	Results for Data Example Model 14: Combined Model, TT Formulation	217
4.14	Results for Data Example Model 15	220

ACKNOWLEDGEMENTS

My first and primary acknowledgement must go to my advisor and mentor, Bengt Muthén. I can honestly say that what I have accomplished would not have been possible without his guidance and support. It has been the greatest fortune in my education to be able to work with and learn from him. Along with my profound appreciation, he will always have my utmost admiration and respect.

I would also like to thank my other committee members, Mike Seltzer, Mike Rose, and A. Afifi for their attention and thoughtful participation in this process. I would like to include a special thanks to Mike Rose for his good counsel that has served me so well over my years in the doctoral program. I am grateful to the wonderful assistance of Bill Fals-Stewart, who provided the data for this dissertation as well as personal encouragement. I also acknowledge the generous help of Klaus Larsen for his honest and useful critiques of my work.

There were others, outside those most closely and academically connected to my dissertation, that also provided important support during this process. Thanks to Tina Christie for all her excellent advice—professional, practical, and personal. Thank you to my other classmates and colleagues with whom I have shared my graduate school experience. Thank you to my

other close friends, Jennifer Anderson, Courtenay Singer, Jennifer Heyob, and Bliss Temple, who have been a source of comfort and cheer throughout my endeavors. My appreciation to my friend and roommate, Kevin Elliott, for study breaks and sanity checks. I am also grateful for the companionship of my cat, Gobblin, and dog, Elvis. Thank you to the members of the Prevention Science Methodology Group and Bengt Muthén's Research Group for their active encouragement and always stimulating exchanges. My gratitude to Anna Tripp for her love and caring, and thanks to Nathaniel and Geneva Tripp for their administrative assistance.

Finally, I would like to acknowledge the particularly influential people that were instrumental in helping me find direction earlier on in my academic career. My thanks to my parents, who instilled in me the love of learning and an understanding of the personal power and freedom that comes with education. Thanks to Mr. Williams, my junior high school math teacher, whose exceptional teaching first inspired my own enthusiasm for mathematics. And thank you to Larry Leemis, my undergraduate advisor, who introduced me to the field of Statistics and the joy of research. His example of skilled and compassionate teaching combined with focused and disciplined research has guided and motivated me during my graduate pursuits.

I end with a general note of thanks to everyone else who has helped me along the way—family, friends, and teachers; I share this accomplishment with many. I am grateful for all the opportunities I have been given that have allowed me the privilege of following my passions in this field.

VITA

- February 10, 1975 Born, Fairfax, Virginia, USA.
- 1991 National Merit Scholarship.
- 1992 Advanced Studies Diploma with
Governor's Seal of Academic Excellence.
- 1992 James Monroe Scholarship,
College of William and Mary,
Williamsburg, Virginia.
- 1994 Class of 1940 Scholarship,
College of William and Mary.
- 1995 B.S. (Summa Cum Laude), Mathematics,
College of William and Mary.
- 1995 Phi Beta Kappa,
College of William and Mary.
- 1995 Benjamin Stoddard Ewell Award,
College of William and Mary.
- 1995 William and Mary Prize in Mathematics,
College of William and Mary.
- 1997–1999 Graduate Student Instructor,
Department of Biostatistics, School of Public Health,
University of California, Berkeley.
- 1998 Graduate Student Researcher,
Department of Epidemiology, School of Public Health,
University of California, Berkeley.
- 1998 Outstanding Graduate Student Instructor Award,
University of California, Berkeley.

- 1999 Teaching Effectiveness Award,
University of California, Berkeley.
- 1999 M.A., Biostatistics,
University of California, Berkeley.
- 1999 Classroom Technologies Grant Recipient,
University of California, Berkeley.
- 1999 UCLA Chancellor's Doctoral Fellowship Award,
University of California, Los Angeles.
- 1999–2002 Graduate Student Researcher,
Graduate School of Education,
University of California, Los Angeles.
- 2001–2002 Graduate Student Instructor,
Graduate School of Education,
University of California, Los Angeles.
- 2001–2002 Research Consultant,
Research Institute on Addictions,
State University of New York, Buffalo.
- 2002–2003 Research Consultant,
The GLOBE Program,
Stanford Research Institute,
Stanford, California.
- 2002–2003 Visiting Assistant Professor,
Department of Education and Child Development,
Whittier College,
Whittier, California.
- 2003 Research Consultant,
Urban Teacher Education Collaborative,
Institute for Democracy, Education and Access,
University of California, Los Angeles.

2003

Leigh Burstein Award for Excellence
in Research Methodology,
University of California, Los Angeles.

PUBLICATIONS AND PRESENTATIONS

Masyn, K. (June, 2003). *Discrete-time survival mixture analysis for recurrent events*. Paper presented at the meeting of the Society for Prevention Research, Washington, DC.

Masyn, K. (June, 2002). *Extensions of discrete-time survival mixture analysis*. Paper presented at the meeting of the Society for Prevention Research, Seattle, WA.

Masyn, K. (June, 2002). *Latent class enumeration revisited: Application of Lo, Mendell, and Rubin (2001) to growth mixture models*. Paper presented at the meeting of the Society for Prevention Research, Seattle, WA.

Masyn, K. (June, 2001). *Discrete-time survival mixture analysis*. Paper presented at the meeting of the Society for Prevention Research, Washington, DC.

Masyn, K. (April, 2001). *Discrete-time survival mixture analysis*. Paper presented at the meeting of the American Educational Research Association, Seattle, WA.

Masyn, K. (June, 2000). *Latent class enumeration for growth mixture models*. Paper presented at the meeting of the Society for Prevention Research, Montreal, Quebec.

Masyn, K., and Brown, E. (April, 2001). *Enumeration of latent classes in general growth mixture modeling*. Paper presented at the meeting of the American Educational Research Association, Seattle, WA.

Muthén, B., Brown, C.H., Masyn, K., Jo. B., Khoo, S., Yang, C., Wang, C., Kellam, S., Carlin, J., and Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics*, 3(4), 459-475.

ABSTRACT OF THE DISSERTATION

Discrete-Time Survival Mixture Analysis
for Single and Recurrent Events Using Latent Variables

by

Katherine Elizabeth Masyn

Doctor of Philosophy in Education

University of California, Los Angeles, 2003

Professor Bengt Muthén, Chair

Survival analysis refers to the general set of statistical methods developed specifically to model the timing of events. This dissertation concerns a subset of those methods that deals with events measured or occurring in discrete-time or grouped-time intervals. A method for modeling single event discrete-time data utilizing a latent class regression (LCR) framework, originally presented by Muthén and Masyn (2001), is further developed and detailed. It is shown that discrete-time data can be represented as a set of binary event indicators and observed risk indicators that allow estimation using a latent class re-

gression specification under a missing-at-random assumption that corresponds to the assumption of noninformative right-censoring. The modeling of the effects of time-dependent and time-independent covariates with constant or time-varying effects is demonstrated along with approaches to model testing. The LCR framework also allows for the modeling of unobserved heterogeneity through finite mixture modeling, i.e., multiple latent classes. The problems of ignoring unobserved heterogeneity and the challenges of discrete-time mixture model identification and specification for single event data are discussed. The LCR model for single event data is extended to recurrent event survival data with a focus on recurrent event processes with a low frequency of recurrences. The gap time, counting process, and total time formulations in the continuous-time setting are all reformulated for discrete-time and model specification and estimation is demonstrated for all three. The proposed model accommodates event-specific baseline hazard probabilities as well as event-specific covariate effects. The model also allows for multiple event occurrences in a single time period for a single subject and accounts for within as well as between subject correlation of event times though the same mixture modeling approach given for single event data. All models are illustrated with data on the event times of domestic violence episodes perpetrated by a sample of married men observed

for 12 months after an alcohol treatment program. Opportunities for future methodology developments for discrete-time models are discussed.

Chapter 1

Introduction

In the fields of social and behavioral science, research questions around specific life course events, such drug use, school drop out, or job loss, are often concerned with both the “if” and “when” of event occurrence. For example, it may be of interest to investigate not only what influences whether a student drops out of school, but when a student drops out. As another example, consider not only whether an individual ever consumes alcohol, but more importantly, the onset age of alcohol use. Traditionally, event data in social research has been treated without regard to event timing, using such modeling techniques as logistic regression, which allows an investigator to explore the relationship between the probability of event occurrence and covariates of interest, including perhaps an intervention or treatment. Survival analysis refers

to the general set of statistical methods developed specifically to model the timing of events. This dissertation concerns a subset of those methods that deals with discrete-time events.

1.1 Discrete-time vs. continuous-time survival data

Time-scales for events can be crudely classified as either *continuous* or *discrete* and the methods applied to one type of time-scale do not necessarily apply to the other, just as regression techniques for continuous outcome variables do not apply directly to categorical outcomes. Applications of continuous-time methods assume that the timing of events is known exactly or that the discrete intervals on which time is measured are sufficiently small that is reasonable to treat the observed times as occurring on a continuous time-scale. Discrete-time events may be of two natures (Allison, 1982): 1) An event may occur at any point in time but only an interval of time during which the event occurred is recorded, e.g., a student may drop-out of school on any particular day of the school year, but data may only be available for the grade-level at which the drop-out occurred. This is sometimes referred to as *grouped-time* survival data, e.g., all the days of the school year are grouped together to form

a nine-month time interval. 2) An event may only occur at discrete points in time, e.g., retention at a certain grade level only occurs at the end of a school year. Although it could be argued that any measure of time is “discrete”, common-sense may distinguish between *grouped-time* and *continuous-time* survival data. Consider, for example, that grouped-time data is likely to have more than one event occur at each measure whereas continuous-time data should have no “ties”. This distinction may also correspond to grouped-time data having notably fewer time periods than the number of individuals under observation.

In the case of (2), continuous-time survival models are clearly inappropriate and it is necessary to apply discrete-time analysis methods. In the case of (1), one approach is to disregard the underlying continuous-time nature of the process, assume that events may only occur at the discrete time points recorded (i.e., events may only occur at the end of each grouped-time interval), and apply the same methods as with (2). Another approach is to assume an underlying continuous-time process and then apply a continuous-time model, adjusting estimates for the discrete nature of the data. For the methodology developments contained in this dissertation, the data are regarded as having been generated by an underlying continuous-time process. However, most models presented for Case 1 are equally applicable to Case 2 data, without any

adjustment. The differences between the two types of discrete-time data tend to relate more to model inference than model specification. There are more assumptions implicit in the Case 1 models with respect to the characteristics of the underlying continuous-time process, e.g., the nature of the duration dependence within each discrete-time interval. Situations where there are model differences due to the nature of the discrete-time data are noted as they arise.

1.2 Statement of purpose

Discrete-time survival methods have been in use for as long as continuous-time methods but are somewhat less visible in both the technical and applied literature. The reason for this is unclear. Perhaps, from the technical side, the discrete-time setting does not prove as challenging, statistically speaking, and deals with the less-than-ideal situation of not knowing the exact time-of-event. From the applied side, the reason is most likely a lack of dissemination and accessibility to fully implemented modeling methods. In 1993, Singer and Willett wrote an article of great detail, aimed at the applied social researcher, calling for and demonstrating the use of discrete-time survival analysis using logistic regression. The model they explicate was not new in that this most common approach to modeling discrete-time events was actually suggested by Cox in his seminal 1972 paper. However, their skillful presentation delivered

these models to a previously unexposed applied audience. The use of logistic regression for discrete-time survival has been studied further by Singer and Willett (1993b, 1995, 2003) as well as many others including Prentice and Gloeckler, 1978; Laird and Oliver, 1981; Allison, 1982. There are several competing approaches currently in use including multilevel ordered multinomial regression (Hedeker, Siddiqui, & Hu, in press), mixed Poisson models (Nagin & Land, 1993), log linear models (Vermunt, 1997), and discrete-time Markov chain models (Van de Pol and Langeheine, 1990). These approaches all fall within the larger category of transition models as delimited by Diggle, Liang, and Zeger (1994).

The methodology developments presented in this dissertation approach discrete-time survival analysis from a somewhat different angle by using a latent variable modeling framework. Briefly, the foundation for these developments involves the modeling of discrete-time survival data of varying complexity using latent class regression analysis. This approach is equivalent to the logistic regression survival model in a most basic setting (Muthén & Masyn, 2001). The purpose of this dissertation was as follows:

- To solidify the motivation and explanation of the latent class regression specification of discrete-time models for single event data, including a discussion of link functions, time-independent and time-dependent covariates, model testing, and censored and truncated data (Chapter 2);
- To clarify the issues of model specification and identification for single event models that account for unobserved heterogeneity, such as a person-specific increased or decreased susceptibility to an event occurrence that is not easily measured or directly observed, as well as to evaluate the utility of the long-term survivor model (Chapter 3);
- To present a flexible model for recurrent event data, utilizing the latent class regression framework developed for single events, that allowed for event-specific survival processes and accounted for unobserved heterogeneity (Chapter 4);
- To propose a roadmap for future methodology developments in discrete-time survival analysis (Chapter 5).

1.3 Historical background

The approach of using latent class regression analysis to model discrete-time data can be viewed as lying at the intersection of several well-established areas

of statistical research: 1) survival analysis (continuous-time, in particular); 2) finite mixture modeling; and 3) traditional latent class analysis.¹ What follows is a brief historical overview of these three areas, acknowledging that this is by no means meant to be a comprehensive review of all research in these areas but is intended to provide an adequate backdrop for the methods discussed in the following chapters.

1.3.1 Continuous-time survival analysis

Survival analysis methodology has its roots in life table data—for centuries people have been, both formally and informally, collecting population data on birth rates, morbidity rates, and mortality rates. Typically, with regards to the mortality rates, information would be collected about the age of death, cause of death, etc. Sometimes, auxiliary information would also be included, such as gender, race, occupation, etc. It is not surprising that at some point people wanted to be able to make projections or predictions about life expectancies and also make comparisons across sub-populations. For example, how does the risk of death change across the lifespan? Are men or women

¹Although latent class analysis can be viewed as a subset of finite mixture models, it is treated as distinct in this overview because the historical development of the two areas, statistically speaking, was more separate than intersecting or overlapping.

more likely to live past the age of 80? These questions began to develop into more sophisticated medical research questions about human survival. For example, how does the risk of death from lung cancer progress from the time of diagnosis? Does treatment “A” prolong the time to recurrence of a breast cancer compared to treatment “B”? Does treatment “A” have the same effect on uterine cancer patients? Spurred by such questions, statisticians in the early and mid-20th century set to work on ways to model such data.

Two challenges are immediately evident when working with survival data. One is the presence of missing data. Consider an event such as death. Certainly, if a sample is only observed for a fixed time period, the event of death for each subject will most likely *not* be observed for some portion of the sample. Some will experience the event in the time period and all will eventually experience the event. However, if they do not experience the event during the period of observation, their time of event is unknown—essentially, it is missing. This type of missingness is known as *right-censoring* in survival jargon. There are other types of censoring, e.g., left- and interval-censoring, that are discussed later. Part of the challenge in survival analysis is accounting for this sort of missingness. If one were to simply delete all but those subjects who experienced the event during the period of observation, the survival estimates would look much grimmer than the actual pattern of survival in the

population. It is most desirable to use the information about those subjects that do not experience the event—the exact timing of their events may not be known since they occur after the end of the observation period, but it is known that their event times are larger than the time period of the study. The most recognized point estimate for survival probabilities based on data with right-censoring is the Kaplan-Meier (K-M) estimate (1958), also known as the product-limit estimate. The Greenwood formula for confidence intervals of the K-M point estimates allows the assessment of the precision of those estimates (Greenwood, 1926). An alternative estimator, based on modern counting process techniques, is the Nelson-Aalen (N-A) estimator (Aalen, 1978; Nelson, 1972) . The appeal of these methods of estimation is that they are nonparametric, meaning that the estimation does not require any assumption about the form or shape of the underlying distribution from which the data is drawn. The estimated survival function based on the K-M or N-A estimates is simply a step function with a step occurring at each observed event. This may be contrasted to a model for survival assuming, say, an exponential distribution, where the one defining parameter of the distribution (often denoted by λ) is estimated via maximum likelihood estimation (MLE) in which a value for λ is identified (out of all possible values for λ) that makes the resultant data seem most likely (under the assumption that the exponential distribution is

the “true” underlying distribution). As with any other type of modeling, there is a trade-off between parametric and nonparametric models. Certainly, the nonparametric models have greater flexibility and protect from the dangers of misspecification. However, the parametric models can give gains in parsimony and statistical efficiency if the model is correctly specified (or only “trivially” misspecified).

The second primary challenge in survival analysis is assessing the effects of covariates on survival. Consider first the simple comparison of survival rates across well-defined groups of individuals. With the Kaplan-Meier estimator, the survival function estimates are based on the data from a sample of individuals assuming that all those individuals are drawn from the same population, that is, they all have the same mean survival function. From the K-M approach it is easy to imagine an extension to a two-population scenario where essentially a separate step-function is estimated for each group. One possible comparison would be to look at survival quantiles for the two populations. For example, a comparison could be made between the 50th percentiles of survival for the two groups, i.e., is the time at which 50 percent of population “A” has experienced the event significantly different from the time at which 50 percent of population “B” has experienced the event? Knowing that it is possible to estimate the 50th percentile of each group, complete with confidence intervals,

it is not difficult to imagine the construction of a statistical significance test of the difference. It is a bit more challenging to make a more global statistical comparison between the groups since the event times (and hence the steps of each function) will be different across the groups. What is required in this case is a nonparametric test to correspond to the nonparametric point estimates of survival. Many such tests exist, the most commonly used being the log-rank test. Comparisons of groups when modeling the survival function parametrically have a more solid framework, simply drawing on maximum likelihood theory. One could assume an exponential distribution for both groups and then fit two models, one with different λ 's for the two groups and one with the same λ for both groups, comparing the fit of the two models with respect to the likelihood values.

Now consider a continuous covariate, such as age, that is believed to influence survival. Essentially, most individuals in the sample will not have the same value for the covariate and it is not feasible to estimate a survival function for each individual—it is necessary to assume some sort of commonality between subjects, even if it is conditional upon the covariate. One approach for investigating the influence of this covariate would be to group the subjects with similar covariate values, e.g., age groups, and then carry out one of the aforementioned group comparisons. This is a bit crude in that ignores

the continuous nature of the covariate and essentially discards information. However, it is not without merit. Discretizing the covariate allows for a very flexible model. In the parametric modeling framework, a model can be fit specifying a relationship between certain parameters and the covariate, e.g., $\log(\lambda) = a + bx$. Here, as in the group comparison setting, it is straightforward to compare the likelihoods of models with and without the covariate effect. Keep in mind, however, that with these models there is not only an assumption about the underlying survival distribution, but also an assumption about the functional relationship between the population parameters which define that distribution and the covariates of interest. In the development of conditional survival models what was needed was a method that capitalized on the advantages of a nonparametric survival estimation method without necessitating discretizing the continuous covariates. This need was met by the Cox proportional hazards model. Introduced by Cox in 1972, this model presents a semi-parametric approach to estimating a hazard function with continuous covariates. The basic idea behind the Cox model is breaking the hazard into two parts: 1) the *baseline* hazard, which is a function of time; and 2) the covariate effects, which in its simplest form, does not include a time term. The log of the hazard is assumed to be the sum of these two parts. It is semiparametric in that the baseline hazard function becomes, essentially, a set of nuisance pa-

rameters and is not explicitly modeled, i.e., it does not require any underlying distributional assumptions. The covariate portion is specified parametrically, typically as a linear combination of the covariates. The assumption placed on the covariate effects is then that the log hazard varies as a linear function of the covariates and thus the ratio of hazard function for two specified values of a covariate at any given point in time is constant, i.e., the hazard function values are proportional—hence the nomenclature “Cox proportional hazards model”. Using a full maximum-likelihood approach to obtain estimates for the covariate effects would require some specification of the form of the baseline hazard. The significant breakthrough in estimation presented by Cox was a method called *partial likelihood* that allowed for estimation of covariate portion of the model without any restrictions placed on the baseline hazard portion.

The Cox model has also been extended to accommodate other challenges related to modeling covariate effects in the survival setting. These extensions include time-varying covariates, where the values on a covariate set for an individual may change over time, and time-varying covariate effects, where the *influence* of a given covariate may change over the time period of observation. Consider, for example, parent supervision as a predictor of school drop-out. The level of parent supervision may change as a child progresses through school (time-varying covariate) and the impact of parent supervision

on a child's behavior may change as the child ages (time-varying effects). Additional extensions of the Cox model have been made in the last three decades including, but certainly not limited to, group-varying baseline hazards, model tests, residual diagnostics, combination with repeated measures models, etc. (See, for example, Klein and Moeschberger, 1997, for an overview of modern continuous-time survival analysis and Hougaard, 2000, for an in-depth coverage of the latest advances in multivariate survival data.) Included in the more recent extensions are advances in modeling population heterogeneity with respect to survival. In the original Cox model, there was no error term as is found in traditional regression models. The implicit assumption is that there is no variability in survival probabilities beyond that which is explained by the covariates included in the model. This is not necessarily the most realistic assumption, but one that was applied until statisticians, with the help of estimation algorithms and computing advancements, were able to create a more general model by including a random effect term in the hazard function. These models are typically referred to as *frailty models* that allow for random variation in the population around a “mean” survival curve (Vuapel, Manton, and Stallard, 1979).

1.3.2 Finite mixture models

There is another way to conceive and model population heterogeneity and that is with the use of *finite mixture models*. The idea behind these models is that there are two or more sub-populations within the population from which the sample is drawn with characteristically different distributions in outcome (in this case, survival). However, membership in these sub-populations is not directly observed—it must be inferred. So rather than simple random variation about a single mean curve, such as with frailty models, there is a “mixing” of two or more groups, each with distinct survival functions. Finite mixture modeling was actually born out of a need to model skewness and non-normality in data distributions—virtually any distribution can be approximated by a mixing of k normal distributions if you allow k to be indiscriminately large. One of the first demonstrations of mixture modeling was done by Karl Pearson in 1894 when he fit a two-component univariate normal mixture model to crab measurement data. He used the method-of-moments to estimate his model. There were few others that immediately followed suit because the moments-based fitting was far too computationally intensive. Some struggled to find more viable, as well as superior, alternative estimation procedures. Tan and Chang (1972) were among the researchers of the time that proved the maximum likelihood solution to be better for mixture models than the

method-of-moments. Following on the heels of this insight was the release of the landmark article by Dempster, Laird, and Rubin (1977) that explicated in general terms an iterative estimation scheme for maximum-likelihood estimation from incomplete data. The recognition that finite mixture models could be easily reconceived as missing data problems (and thus estimated via the EM algorithm) and the rapid and widespread computer advancements in speed and processing marked a true advent in finite mixture modeling. Indeed, the last 20 years have seen a remarkable increase in the development, extension, application, and understanding of mixture modeling. (For a in-depth treatment of the current state of mixture modeling, see McLachlan and Peel, 2000.) Not surprisingly, the applications of mixture modeling reaching into many more complex modeling settings beyond the original univariate normal case have recently begun to include survival analysis, although survival mixtures were first suggested by Heckman and Singer in 1984. As with mixture models in general, survival mixture models can be applied to data where a simpler single parametric model does not suffice, e.g., mixture of Weibull distributions (Gupta & Gupta, 1996). In addition, the previously mentioned idea of frailty can be incorporated and estimated nonparametrically using mixture models rather than specifying a parametric distribution on the frailty factor in the survival model, e.g., the sample is modeled to derive from two subpopulations,

one with a higher susceptibility to the event than the over (Heckman & Singer, 1984a, 1984b). A special case of this use of mixtures to model frailty is what is known as “cure-rate” models in survival analysis. In some data applications, individuals may be cured, that is to say, there may be a fraction of the sample that, in reality, is not at risk for the event in question. These individuals are commonly referred to as *long-term survivors* (LTS) and a two-component survival mixture model accommodates the presence of such a subpopulation. (See Maller and Zhou, 1996, for a full range of LTS applications.)

1.3.3 Latent class analysis

Latent class models can be considered a special class of mixture models formulated as a mixture of generalized linear models. However, latent class analysis (LCA) has a rich history somewhat independent of the development of finite mixture models that is worth reviewing. LCA was born of the field of latent variable modeling. The idea is that there are two sorts of variables: observed or *manifest* variables and unobserved or *latent* variables. A specified set of observed variables (also called indicator variables) are assumed to be imperfect measures of one or more underlying latent variables, that is, the relationship (covariance) of the observed variables is attributable to each manifest variable’s relationship to the latent variable(s). An example of such a relationship

would be a set of diagnostic questions on a psychological survey (manifest variables) designed to measure depression (latent variable). The work of Jöreskog and Sörbom (1979) established the solid methodological framework for factor analysis—latent variable models with continuous observed variables regressed on one or more continuous latent variables. LCA can be characterized as the categorical data analogue to traditional factor analysis. The most basic latent class model can be traced back to Lazarsfeld in his discussion of a broader class of what he coined as *latent structure analysis* (Lazarsfeld & Henry, 1968). The fundamental assumption in LCA is that the relationship among the observed categorical variables is “explained” by an underlying categorical latent variable (latent class variable), i.e., the observed variables are conditionally (locally) independent given latent class. Essentially, the latent class variable defines a so-called typology or profile based on the clustering of individual response patterns across the observed items. The basic formulation of a latent class model with binary indicators follows.

Let \mathbf{u} be a vector of J binary indicator variables, scored 0 or 1, and let c be a categorical latent variable with K classes, where $c \in \{1, \dots, K\}$ and $c = k$ indicates membership in class k . For \mathbf{u} , the assumption of conditional independence is given by

$$P(\mathbf{u} | c) = \prod_{j=1}^{j=J} P(u_j | c), \quad (1.1)$$

which can alternately be expressed as

$$P(\mathbf{u} | c) = \prod_{j=1}^{j=J} \left[P(u_j = 1 | c)^{u_j} P(u_j = 0 | c)^{1-u_j} \right]. \quad (1.2)$$

The distribution of c is defined by a multinomial logistic regression,

$$P(c = k) = \frac{\exp(\alpha_{0k})}{\sum_{m=1}^K \exp(\alpha_{0m})}, \quad (1.3)$$

where $\alpha_{0K} = 0$ for the reference class, K , and the distribution of each u_j can be defined by a simple logistic regression,

$$P(u_j = 1 | c = k) = \frac{1}{1 + \exp(-(\nu_{jk}))}. \quad (1.4)$$

In equation (1.3), the α_k 's capture the probabilities of the K classes and in (1.4) the ν_{jk} 's are the *logit* of the u_j 's for each class k . The logistic regressions for c and the u 's also allow the inclusion of covariates, specifying *logit* functional relationships for both c and \mathbf{u} with respect to the covariates. This is termed a latent class regression model (LCR) (Formann, 1992; Huang & Bandeen-Roche, in press). The equations for the regression of the u 's and c on a set of covariates, z , are given by

$$P(u_j = 1 \mid c = k, z) = \frac{1}{1 + \exp(-(\nu_{jk} + \beta'z))}, \quad (1.5)$$

and

$$P(c = k \mid z) = \frac{\exp(\alpha_{0k} + \alpha'_k z)}{\sum_{m=1}^K \exp(\alpha_{0m} + \alpha'_m z)}, \quad (1.6)$$

where $\alpha_{0K} = 0$ and $\alpha_K = \mathbf{0}$. The observed-data likelihood for a single individual i is then given by

$$P(\mathbf{u}_i \mid z_i) = \sum_{k=1}^K P(\mathbf{u}_i \mid c_i = k, z_i)P(c_i = k \mid z_i). \quad (1.7)$$

Goodman (1974) resolved the problem of parameter estimation for LCA models and his algorithm was implemented by Clogg in 1977. Subsequently, many advancements in LCA have been made by Goodman and Clogg as well as others. (For an overview of the most recent developments in LCA, see Hagenaars and McCutcheon, 2002.) Muthén and Shedden (1999) offered the broadest framework yet for latent class models by fully integrating LCA with more general latent variable models. Bandeen-Roche, Miglioretti, Zeger, and Rathouz (1997) provided methods for assumption-checking and model diagnostics for LCR. Larsen (in press) has combined latent class modeling with the Cox proportional hazards model so that classes defined by categorical indicators in an LCA delineate different survival typologies.

1.4 Data example

The data set used throughout this dissertation for methodology motivations and illustrations is from a study conducted by researchers at the Research Institute on Addictions at SUNY, Buffalo.² The purpose of the original study was to evaluate the effectiveness of behavior couples therapy (BCT) on the drinking behavior of alcohol dependent males. A secondary objective of this study was to examine the relationship between drinking and domestic violence in men who have undergone treatment for alcohol dependence. One-hundred and seventy males, either married or cohabitating,³ undergoing intake for alcohol abuse, were randomized to three counseling treatment regimes: 1) marital treatment (BCT), 2) individual-based treatment with no spouse involvement, or 3) attention control treatment, with spouse involvement limited to informational sessions and lectures on substance use and not any active couples treatment as in (1). Subjects were administered a three month follow-back survey at entrance to treatment regarding the time prior to treatment and completed the same survey at three, six, nine, and twelve months post-treatment.

²Special thanks to William Fals-Stewart at RIA for making this data available.

³Although each male subject may be either married or cohabiting, for the purposes of this discussion, the female partner will be referred to as “spouse” or “wife”, regardless of the actual marital status of the couple.

Questions regarded drinking occurrences as well as occurrences of domestic violence. For the examples in this dissertation, the data is discretized into six two-month post-treatment periods. In addition to being alcohol abusers, all subjects in the original sample reported committing at least one domestic violence offense in the three months prior to treatment. Additional information was collected on each subject, including age, wife's age, education level, wife's education level, marital status, length of relationship, annual family income, race, and wife's race. All but 19 (11%) of the subjects satisfied the DSM-III-R criteria for alcohol dependence. Twenty subjects had been referred to treatment because of a DWI offense. All of the subjects were surveyed at the end of each follow-up period. Thus, there was no loss to follow-up and all subjects had complete data for the pre-treatment period as well as the full 12 month post-treatment period of observation.

Table 1.1 displays the descriptive summary statistics for the measures of drinking during each two-month period, defined as percent-days-drinking, with drinking on a given day defined as having at least one drink on that day. Tables 1.2 and 1.4 give the frequencies and relative frequencies for the number of violent episodes in the three month pre-treatment period and the six post-treatment follow-up periods. Table 1.3 gives the descriptive summary statistics for the two continuous scale covariates, husband's age (in years) and

length of relationship (in years). The husband's age and wife's age variables were strongly correlated ($r=0.96$) and so only husband's age was included in the analyses. Table 1.5 gives the frequencies and relative frequencies for the categorical covariates. Several of these displayed categories are the result of combining categories defined in the original data. The husband's race variable originally included the categories "White", "Black", "Hispanic", and "other". Due to small counts in the non-White race categories, these groups were combined into a single category. Similar category recombinations were performed for the annual family income, husband's education, and wife's education variables. As with husband's and wife's age, only the husband's race variable was included in analyses because of the near-perfect correspondence between husband's race and wife's race—only 5% of couples were interracial.

Table 1.1: Descriptive Statistics for Pre- and Post-Treatment % Days Drinking

Period	M (SD)	Range
3 months pre-tx	0.72 (0.23)	0.01 - 1.00
Month 1-2 post-tx	0.22 (0.27)	0.00 - 1.00
Month 3-4 post-tx	0.23 (0.30)	0.00 - 1.00
Month 5-6 post-tx	0.31 (0.34)	0.00 - 1.00
Month 7-8 post-tx	0.30 (0.32)	0.00 - 1.00
Month 9-10 post-tx	0.30 (0.34)	0.00 - 1.00
Month 11-12 post-tx	0.36 (0.35)	0.00 - 1.00

All example data manipulation, summaries, plots, and data simulations for this dissertation were done in Splus Professional, Version 6.1, which is produced and distributed by the Insightful Corporation. Select code is provided in Appendix A. All models presented in this dissertation were estimated using Mplus, Version 2.14, which is produced and distributed by Muthén & Muthén. The input files for select models fit to the real data example are provided in Appendix B.

Table 1.2: Frequencies and Proportions for Pre-Treatment Violence Episodes

Period	# of episodes	Frequency	Proportion
3 months pre-tx	1	55	0.32
	2	35	0.21
	3	24	0.14
	4	14	0.08
	5-10	20	0.12
	11+	22	0.13

Table 1.3: Descriptive Statistics for Continuous Covariates

Variable	M (SD)	Range
Age (in years)	43.24 (12.62)	22 - 70
Length of relationship (in years)	11.09 (7.89)	1 - 34

Table 1.4: Frequencies and Proportions for Post-Treatment Violence Episodes

Period	# of episodes	Frequency	Proportion
Month 1-2 post-tx	0	136	0.80
	1	26	0.15
	2	7	0.04
	3	1	0.01
Month 3-4 post-tx	0	152	0.89
	1	15	0.09
	2	2	0.01
	3	1	0.01
Month 5-6 post-tx	0	152	0.89
	1	12	0.07
	2	6	0.04
	3	0	0.00
Month 7-8 post-tx	0	153	0.90
	1	14	0.08
	2	2	0.01
	3	1	0.01
Month 9-10 post-tx	0	155	0.91
	1	11	0.06
	2	1	0.01
	3	1	0.01
	4	2	0.01
Month 11-12 post-tx	0	152	0.89
	1	7	0.04
	2	8	0.05
	3	2	0.01
	4	1	0.01

Table 1.5: Frequencies and Proportions for Categorical Covariates

Variable	Categories	Frequency	Proportion
Treatment group	BCT	56	0.33
	Individual-based	56	0.33
	Attention control	58	0.34
Husband's education	H.S. diploma or less	64	0.38
	Some college	54	0.32
	College graduate	31	0.18
	Some graduate school	21	0.12
	or graduate degree		
Wife's education	H.S. diploma or less	71	0.42
	Some college	54	0.32
	College graduate	32	0.19
	Some graduate school	13	0.08
	or graduate degree		
Husband's race	White	140	0.82
	Non-White	30	0.18
Marital status	Married	161	0.95
	Cohabiting	9	0.05
Annual family income	\$0-20,000	13	0.08
	\$20,001-25,000	21	0.12
	\$25,001-30,000	36	0.21
	\$30,001-35,000	40	0.24
	\$35,001-40,000	29	0.17
	\$40,001+	31	0.18
DWI referral	Yes	20	0.12
	No	150	0.88
Alcohol dependence criteria	Met	151	0.89
	Not met	19	0.11

Chapter 2

Single Events

This chapter covers the specification and estimation of single event survival models, with no unmeasured covariates, i.e., no unobserved heterogeneity. It begins with the foundations of single event survival analysis (also referred to as univariate survival analysis) for continuous-time data and then gives the reformulations for discrete-time data. It also covers the modeling of covariate predictors of survival, assumption checking, and assessments of model fit.

2.1 Single event, continuous-time survival analysis

2.1.1 Basic notation and event time distributional forms

Let the given sample consist of n independent individuals i , with $i = 1, \dots, n$. Let T_i be the event time for individual i relative to a known start time, say $t = 0$, common for all individuals, and assume T is a non-negative continuous random variable from an unknown distribution, again common for all individuals, such that $T_i \sim T$.

The survival function, describing the probability of an individual surviving beyond time t , i.e., experiencing the event after time t , is defined as

$$S(t) = P(T > t). \quad (2.1)$$

The survival function has the following three properties: 1) $S(0) = 1$; 2) $\lim_{t \rightarrow \infty} S(t) = 0$; and 3) $S(t)$ is a monotonic, nonincreasing, and nonnegative function. Notice the relationship between $S(t)$ and the probability density function, $f(\cdot)$:

$$S(t) = 1 - F(t) = \int_t^{\infty} f(v)dv. \quad (2.2)$$

The most common representation of the event time distribution is the hazard function (also known as the hazard rate or intensity), defined as

$$h(t) = \lim_{\Delta t \rightarrow \infty} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \quad (2.3)$$

From this definition, one can see that the hazard rate can be thought of as the *instantaneous* failure rate, that is to say, $h(t)\Delta t$ is the “approximate” probability¹ of an individual who has not experienced the event by time t experiencing the event in the next instant following t .² The hazard rate may also be interpreted as the average number of events in a one unit interval of time.

¹Technically, the hazard rate is not a probability since it may take on values greater than one.

² $f(t)\Delta t$ can be thought of as the “approximate” probability of failure at time t when looking at risk from time zero rather than conditional on survival to the moment right before t . In understanding the difference between $f(t)$ and $h(t)$, consider the different answer one may get when asking a doctor to estimate the probability of death for a patient on the fifth day post-surgery; the doctor may give a very different estimate on the first day post-surgery than on the fourth.

With T as a continuous random variable, the following relationships between $S(t)$ and $h(t)$ hold:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d \ln S(t)}{dt}, \quad (2.4)$$

and

$$S(t) = \exp \left[- \int_0^t h(v) dv \right]. \quad (2.5)$$

The survival and hazard functions are the most commonly used distributional representations in survival analysis. However, under certain conditions, other representations can prove useful, in terms of their statistical properties. For example, some methods use the hazard function formulation for estimation purposes and then use the estimated cumulative hazard for assumption checking.

As discussed in Chapter 1, there are parametric, nonparametric, and semiparametric approaches to estimating various distributional quantities associated with the different distributional forms for survival time. The most common parametric models for continuous-time survival data are the exponential, Weibull, log logistic, log normal, and generalized gamma. The most common nonparametric estimators are the Kaplan-Meier and the Nelson-Aalen. The following section goes into more detail about conditional survival models, including the semiparametric regression model of Cox (1972).

2.1.2 Continuous-time models with covariates

The previous sections have mainly addressed the distributional forms of continuous event time under the assumption that all individuals in the sample were independent and identically distributed. However, it is rare that a study does not collect, in addition to measures of outcome, measures of covariates, also called explanatory variables, that may describe heterogeneity in the survival process across the sample. And, it is often the case that the interest of the researcher is not exclusively or even primarily in predicting absolute risk across time but in comparing the relative risk across time between groups of individuals, defined by a given set of covariates, and in making inferences about differences in subpopulation risks. For example, consider a clinical trial of a new cancer treatment designed to prolong time-to-death. It would be crucial to determine the magnitude and significance of the reduction in the risk of death in the treatment group compared to the control group, regardless of the overall risk in the general population.

As in growth modeling, covariates of survival time can be categorical or continuous and time-dependent or time-independent. Consider first the case of time-independent covariates. Let z be a $p \times 1$ vector of covariates. There are many different ways to specify the dependence of event time on a set of covariates. Here, only the two most common categories of event time regression

models are considered: accelerated life models and multiplicative hazard rate models.

The survival function for the accelerated life model is given by

$$S(t \mid z) = S_0(t \cdot \psi(z)), \quad (2.6)$$

where $S_0(T)$ is the *baseline* survival function, i.e., the survival function when $z = 0$, and $\psi(z)$ is a *link function*. Another way to represent this model in terms of the random variable itself is

$$T = \frac{T_0}{\psi(z)}, \quad (2.7)$$

where T_0 has the survival distribution, $S_0(t)$. The accelerated life model can be expressed equivalently in terms of the hazard by

$$h(t \mid z) = \psi(z) h_0(t \cdot \psi(z)). \quad (2.8)$$

By this model, the effect of the covariates is to change the time scale by a factor of $\psi(z)$. Essentially, the covariates *accelerate* or *decelerate* the movement of a subject through time. If $\psi(z_i) = 5$, subject i moves through time five times more quickly than a subject with $z = 0$; alternatively, subject i has an expected lifetime one fifth of that of a subject under baseline conditions. This model does have some intuitive appeal, particularly when considering wear-out or burn-out scenarios for the survival process. However, the model

may be intractable without parametric assumptions imposed on $S_0(t)$ and does not as readily accommodate censoring and time-dependent covariates as the multiplicative hazard rate model.

The hazard function for the multiplicative hazard rate model is given by

$$h(t | z) = h_0(t) \cdot \psi(z) \tag{2.9}$$

where, as before, $h_0(t)$ is the *baseline* hazard, that is, $h_0(t) = h(t | z = 0)$. This model is commonly referred to as the *proportional hazards* model because of a key feature under the absence of time-dependent covariates and the assumption of time-independent covariate effects: the hazard rates of two individuals with covariate values z_i and z_j are proportional. This means that the ratio of hazard rates for two subjects, i and j , is a constant, independent of time:

$$\frac{h(t | z_i)}{h(t | z_j)} = \frac{h_0(t) \cdot \psi(z_i)}{h_0(t) \cdot \psi(z_j)} = \frac{\psi(z_i)}{\psi(z_j)}. \tag{2.10}$$

The most commonly used form for the link function is $\psi(z) = \exp(\beta'z)$. This formulation is known as the Cox proportional hazards model or the Cox regression model³. Each regression coefficient has the interpretation as the difference in the log hazard rates for a one unit change in the corresponding covariate, that is, the log hazard rate ratio for a one unit increase in the

³The Cox “regression model” is a more appropriate label since Cox’s formulation allows for relaxation of the proportionality assumption.

corresponding covariate. Cox's seminal paper, published in 1972, can be credited with the dominance of this model in survival analysis applications, not because of its novel specification or its indisputable and ubiquitous applicability for survival data, but because of the most significant contribution of the paper regarding the estimation of such a model. Cox suggested an estimation method he termed *partial likelihood*, which essentially constructs a likelihood based on a conditioning principle where there is a likelihood contribution *only* at each event time. Right censoring times do not enter the likelihood which results in no bias in the coefficient estimates and standard errors as long as censoring time is truly independent of event time. The purpose of this model is to estimate the regression coefficients. In this formulation, the baseline hazard function is a set of nuisance parameters that drop out of the likelihood. Leaving the baseline hazard unspecified also places this model in the *semiparametric* category. This feature prevents bias in the covariate effect estimation due to misspecification of the baseline hazard. However, it is possible to obtain nonparametric estimates of the baseline hazard function using a post-hoc computation involving the cumulative hazard function and the estimated link function values for each subject. (For more standard text

on univariate continuous-time models, see Kalbfleisch and Prentice, 1980; Cox and Oakes, 1984; Fleming and Harrington, 1991; Yamaguchi, 1991; Collett, 1994; and Klein and Moeschberger, 1997.)

2.2 Single event, discrete-time survival analysis

2.2.1 Basic notation and event time distributional forms

Consider data that were referred to as “Case 1” in the introduction: an underlying continuous-time process with event times grouped into discrete intervals. Let the given sample consist of n independent individuals i , with $i = 1, \dots, n$ with corresponding survival times $T_i \sim T$, as previously defined.

In this setting, event time is only observed in J grouped intervals $[t_j, t_{j+1})$ where $j = 0, \dots, J - 1$, $t_0 = 0$, and $t_J = \infty$. Let Γ_i represent the time interval in which T_i falls, so that Γ is a discrete random variable with the event set $\{1, 2, \dots, J\}$. Then $\Gamma_i = \gamma$ if $t_{\gamma-1} \leq T_i < t_\gamma$. Given these definitions, the following distributional forms, parallel to those in the continuous-time setting, can be specified along with their relations to their continuous-time counterparts.

The survival probability, describing the probability of an individual surviving beyond the interval γ , i.e., experiencing the event after time t_γ , is defined as

$$P_S(\gamma) = P(\Gamma > \gamma) = P(T \geq t_\gamma) = S(t_\gamma^-). \quad (2.11)$$

The survival probability has the following relationship to the probability mass function, $P_f(\cdot)$:

$$P_S(\gamma) = 1 - P_F(\gamma) = \sum_{j=\gamma+1}^J P_f(j), \quad (2.12)$$

where $P_S(J) = 0$ and $\sum_{j=1}^J P_S(j) = 1$.

The discrete-time equivalent to the hazard function is the hazard probability. The hazard probability for period γ is the probability that an individual experiences an event in period γ provided that she has not experienced the event in an earlier period. This can be expressed by

$$\begin{aligned} P_h(\gamma) &= P(\Gamma = \gamma \mid \Gamma \geq \gamma) \\ &= P(t_{\gamma-1} \leq T_i < t_\gamma \mid T_i \geq t_{\gamma-1}) \\ &= \frac{S(t_{\gamma-1}^-) - S(t_\gamma^-)}{S(t_{\gamma-1}^-)} \\ &= \frac{P_f(\gamma)}{P_S(\gamma - 1)}, \end{aligned} \quad (2.13)$$

where $P_f(\gamma)$ is the probability mass function, $P_f(\gamma) = P(\Gamma = \gamma)$. The following relationships between $P_S(\gamma)$ and $P_h(\gamma)$ then hold:

$$P_h(\gamma) = \frac{P_S(\gamma - 1) - P_S(\gamma)}{P_S(\gamma - 1)} \quad (2.14)$$

and

$$\begin{aligned} P_S(\gamma) &= P(\Gamma > \gamma) \\ &= P(\Gamma \neq 1 \mid \Gamma \geq 1) \cdot P(\Gamma \neq 2 \mid \Gamma \geq 2) \cdots P(\Gamma \neq \gamma \mid \Gamma \geq \gamma) \\ &= \prod_{j=1}^{\gamma} (1 - P_h(j)). \end{aligned} \quad (2.15)$$

Examining plots of the hazard probabilities is quite useful in understanding how risk for an event changes over time and how those changes influence the corresponding survival probabilities. For example, most electronic devices have what it called a *bathtub-shaped* or *U-shaped* hazard function; risk of failure is high in the beginning *burn-in* period, due to manufacturing defects, etc., and then much later in the life of the surviving devices due to *wear-out*. The lifetimes of many living creatures follow a bathtub hazard as well: we are most vulnerable at the beginning and end of our lifespan. Relating the shape of the hazard to the survival function, when the hazard is zero, the survival function is constant; when the hazard is high, the survival function is decreasing quickly; when the hazard is low, the survival function is decreasing slowly. The shape of the hazard function and survival function are useful to examine together as the survival function not only reflects the

cumulative risk impact on the population in each time period, it quantifies the proportion of the population susceptible to the risk defined by the hazard for each time period. Figure 2.1 displays six hypothetical example sets of hazard probabilities and the corresponding survival probabilities, plotted over time.

2.2.2 Censoring and truncation

Missing data is endemic to longitudinal study settings; survival analysis is no exception. The various mechanisms for missing data in the survival context are usually grouped under the encompassing term, *censoring*. Most generally, censoring occurs when the exact survival time is only known for a portion of the sample, with event times for the remaining subjects only known to occur in certain intervals. There are three categories of censoring: right, left, and interval censoring.

Left censoring occurs when a subject in the sample has experienced the event of interest prior to the onset of observation. In this case, all that is known about the event timing is that it occurred sometime between $t = 0$ and the beginning of the study.⁴

⁴This assumes that retrospective data is not available to obtain the exact timing of the event.

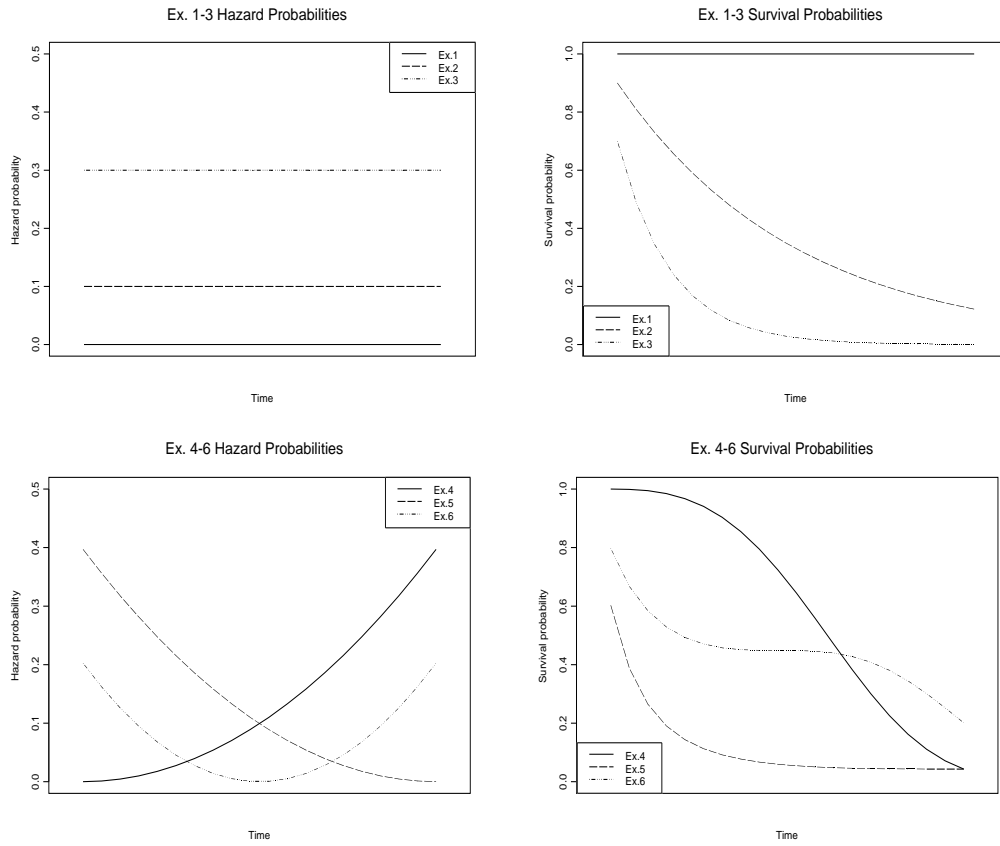


Figure 2.1: Example hazard and survival probability plots.

Right censoring occurs when a subject in the sample has not experienced the event of interest at the cessation of observation. In this case, all that is known about the event timing is that it occurs sometime after the end of the study. In this scenario, it is assumed that, eventually, each subject will experience the event. That is to say, if each subject were followed indefinitely, all would at some point experience the event of interest. There are three primary right censoring schemes. The first is referred to as *Type I censoring*. This censoring occurs when observation of each subject is halted at a pre-specified time. These times may be the same or different across subjects. A common scenario that results in Type I censoring is for a study to conclude on a specific calendar date, prior to all subjects experiencing the event of interest. *Type II censoring* results when a sample is observed until a pre-specified number of events has occurred. A common study design that results in Type II censoring involves animal experiments where the study is stopped after k deaths occur, where k was determined to be the minimum number of event times needed for sufficient statistical power. *Random censoring* is similar to Type I censoring except that the time of censoring is itself a random variable, that is, it is not a fixed or pre-specified value. Study attrition can result in random censoring, i.e., unforeseen circumstances, independent of each subject's event time, may

cause subjects to no longer be under the observation of the researcher while still being at-risk for the event.

Interval censoring occurs when a subject is only known to have experienced the event of interest within a given time interval but the exact time is unknown. Interval censoring is common in clinical trials and longitudinal studies with regularly timed follow-up assessments. Imagine a clinical outcome of interest that can only be determined by physician examination: at one assessment, the subject is considered disease-free and at the next assessment, the subject is diagnosed as having the disease; in these cases, it can be difficult, if not impossible, to determine when during the time between the two assessments the subject actually developed the disease. Discretely measured survival data can be considered a special case where *all* subjects that experience the event during the observation period are interval censored and that the possible intervals of censoring are common to all subjects, e.g., all subjects are assessed at the same follow-up times. Both Type I and random right censoring as well as left censoring can be reformulated as special cases of interval censoring, with left censored individuals experiencing the event in the interval from zero to the time of first observation and right censored individuals experiencing the event in the interval from the time of last observation to infinity.

As previously noted, censoring is simply the general term used for missing data mechanisms in survival data. The range of assumptions about the censoring parallel the more general missing data assumptions applied to other data settings.⁵ The assumption of *noninformative censoring* corresponds to the assumption of *ignorable missingness*, which includes both missing-completely-at-random (MCAR) and missing-at-random (MAR). If the distribution of censoring times is independent of event times, censored observations may be treated as MCAR. If the distribution of censoring times is independent of event times, conditional on the set of observed covariates, then censored observations may be treated as MAR. The case of *informative censoring* corresponds to *nonignorable missingness*. In these situations, censoring times depend upon event times. Interval censoring, as it defines discretely measured survival times, is implicitly addressed. For Type II censoring, order statistic techniques may be applied.

Truncation is another feature of survival data, often discussed or presented in conjunction with the topic of censoring. As censoring can be thought of as a missing data feature, truncation can be thought of as a selective sampling feature. *Left truncation* occurs when individuals must experience a certain event (not the event of interest) and/or not have experienced the event of

⁵For a complete discussion of missing data techniques, see Little and Rubin (2002).

interest to be observed by the researcher. For example, when studying teacher retention, taking a cross-sectional sample of current teachers would result in left-truncated data since all teachers who had left the profession prior to the beginning of the study would not be observed. Another example of left truncation is often referred to as *delayed entry* or *late entry*, meaning that observation on a given subject does not begin at the origin of event time, e.g., subjects of different ages 18–35 enter the study at the same time but only subjects who have not experienced the event prior to study commencement are known to the researcher. Note the difference between left truncation and left censoring: in the case of left censoring, individuals who experience the event prior to the first observation still have the possibility of being included in the sample; in the case of left truncation, selection to the sample itself is conditional on the event having not occurred prior to the first observation. *Right truncation* occurs when individuals must experience the event to be observed by the researcher. For example, in a recidivism study sampling exclusively from current inmates, only those who had already been rearrested would be known to the researcher. In the case of right truncation, there can be no right censoring. For both types of truncation, the conditional nature of each observation, be it an event or censoring time, must be accounted for in construction of the likelihood. For the purposes of this dissertation, the absence of truncation is

assumed unless otherwise specified. Also, only noninformative, Type I and random right-censoring is assumed unless otherwise indicated.

2.2.3 Constructing the likelihood

Construction of the likelihood can be approached in a systematic manner by considering the information each subject contributes to the understanding of the overall survival process. For an individual who experiences the event in interval γ_i , it is known that $\Gamma_i = \gamma_i$; the likelihood for that observation is then $P(\Gamma_i = \gamma_i) = P_f(\gamma_i)$. For an individual right-censored during the interval C_{ri} , it is known only that $\Gamma_i > C_{ri} - 1$;⁶ the likelihood for that observation is then $P(\Gamma_i > C_{ri} - 1) = P_S(C_{ri} - 1)$. For an individual left-censored at the interval C_{li} , it is known that $\Gamma_i < C_{li}$; the likelihood for that observation is then $P(\Gamma_i < C_{li}) = 1 - P_S(C_{li} - 1)$. For an individual interval-censored during the time spanned by intervals I_{li} to I_{ri} , it is known that $I_{li} \leq \Gamma_i \leq I_{ri}$; the likelihood for that observation is then $P(I_{li} \leq \Gamma_i \leq I_{ri}) = P_S(I_{li} - 1) - P_S(I_{ri})$. Assume that the censoring is noninformative. Then the full likelihood equation for $1, \dots, n$ observations can be written as

⁶Prentice and Gloeckler (1978) showed that even if the right-censoring occurs during the time period C_r , that the maximum likelihood estimates are consistent if treating the censoring time as the end of the interval $C_r - 1$.

$$L = \left(\prod_{i \in U_E} [P_f(\gamma_i)] \right) \left(\prod_{i \in U_R} [P_S(C_{ri} - 1)] \right) \left(\prod_{i \in U_L} [1 - P_S(C_{li} - 1)] \right) \left(\prod_{i \in U_I} [P_S(I_{li} - 1) - P_S(I_{ri})] \right), \quad (2.16)$$

where U_E is the set of indices for exact event times in the sample, U_R is the set of indices for right-censored times, U_L is the set of indices for left-censored times, and U_I is the set of indices for interval-censored times.

Consider the most common scenario with only complete and Type I or random right-censored observations. Let the observed data be represented by $\{A, \delta\}$ where $A_i = \min(\Gamma_i, C_{ri})$ and $\delta_i = \mathbb{I}(\Gamma_i \leq C_{ri})$. Essentially, a_i is the last time period during which the subject is observed⁷ and δ_i is the indicator of whether an event or censoring occurred during that final period. The likelihood can then be expressed as

$$L = \prod_{i=1}^n [P_f(a_i)]^{\delta_i} [P_S(a_i - 1)]^{1-\delta_i}. \quad (2.17)$$

Using the relationship between the probability mass and survival functions and the hazard functions given in Equations 2.13 and 2.15, the likelihood can

⁷This does not mean the subject is observed to be at-risk for the entirety of the time period.

alternatively by expressed in terms of the hazard probabilities by

$$L = \prod_{i=1}^n \left\{ [P_h(a_i)]^{\delta_i} \prod_{j=1}^{a_i-1} [1 - P_h(j)] \right\}. \quad (2.18)$$

In the case of right-truncated data, only events are observed. Suppose the data are right-truncated at the interval (Q_{RT}) where $Q_{RT} < J$; the likelihood for each right-truncated observation it then

$$P(\Gamma_i = \gamma_i \mid \Gamma_i \leq Q_{RT}) = \frac{P_f(\gamma_i)}{1 - P_S(Q_{RT})} \quad (2.19)$$

Similarly, for data left-truncated at the interval Q_{LT} , the likelihood for each uncensored individual left-truncated at the interval Q_{LT_i} is then

$$P(\Gamma_i = \gamma_i \mid \Gamma_i \geq Q_{LT_i}) = \frac{P_f(\gamma_i)}{P_S(Q_{LT_i} - 1)} \quad (2.20)$$

and for each right-censored left-truncated individual is

$$P(\Gamma_i > C_{ri} - 1 \mid \Gamma_i \geq Q_{LT_i}) = \frac{P_S(C_{ri} - 1)}{P_S(Q_{LT_i} - 1)}. \quad (2.21)$$

2.2.4 Estimation

In the continuous time setting, no two subjects share the same event time (theoretically) but in the discrete time setting, multiple subjects may share the same time interval of event occurrence. Because of this difference, there is not the same distinction to be made between nonparametric and parametric approaches to estimation. Certainly, there are more and less restricted models

for the baseline hazard probabilities as well as for the covariate effects, but all of the conditional models presented in this dissertation utilize full likelihood techniques (as opposed to the partial likelihood employed for the Cox model).

The most straight-forward method for estimation of the unconditional survival probabilities for complete or right-censored data is analogous to the Product-Limit estimator. Essentially, the hazard probability is estimated for each time interval by taking the ratio of the number of subjects experiencing the event in a given interval over the number of subjects *observed* to be “at-risk” for a given time interval. A subject is observed to be at-risk in period j if she has not experienced the event prior to period j and is not censored in period j or before. In other words, a subject i , $i = 1, \dots, n$ is considered at-risk for interval j if $(a_i \geq j, \delta_i = 1)$ or $(a_i > j, \delta_i = 0)$. Let n_j be the number at-risk for interval j , d_j be the number of events during interval j , and c_j be the number of subjects censored during interval j . Then the estimate for the hazard probability for a specific time period is given by

$$\hat{P}_h(\gamma) = \frac{d_\gamma}{n_\gamma}, \quad (2.22)$$

where

$$n_\gamma = n_{\gamma-1} - d_{\gamma-1} - c_\gamma = n - \left[\sum_{j=1}^{\gamma-1} (d_j + c_j) + c_\gamma \right]. \quad (2.23)$$

So, the number at-risk in period j is equal to the total sample size minus the total number of events occurring up to period j and the total number of

subjects censored through period j . Using the relationship between the hazard probability and the survival probability given in Equation 2.15, the estimated survival probability is given by

$$\hat{P}_S(\gamma) = \prod_{j=1}^{\gamma} (1 - \hat{P}_h(j)) = \prod_{j=1}^{\gamma} \left(\frac{n_j - d_j}{n_j} \right). \quad (2.24)$$

Efron(1988) showed that the above equation is the discrete-time limit to the continuous-time Kaplan-Meier estimate. The estimate for the hazard probability given in Equation 2.23 is also the maximum likelihood estimate for complete and noninformative right-censored discrete-time data.

For the data example presented in Chapter 1, the estimated hazard probabilities for occurrence of first violence in each of the six post-treatment time periods and the corresponding survival probabilities are given in Table 2.1. For example, in the second time period, 136 subjects are at-risk for their first episode of violence during the post-treatment period and 13 commit an act of violence during that period, yielding an estimated hazard probability for that time period of $\frac{13}{136} = 0.10$. Survival beyond the second period is then the product of the complements of the hazard probabilities for the first two time periods: $\hat{P}_S(2) = (1 - \frac{34}{170})(1 - \frac{13}{136}) = 0.72$. Figure 2.2 displays the plots for the hazard and survival probabilities. The hazard probabilities are low overall and show some decrease over the twelve months. Computing the survival probabilities translate the hazard probabilities into the more intuitive

Table 2.1: Sample Hazard and Survival Probabilities for First Violence Post-Treatment

Months	1-2	3-4	5-6	7-8	9-10	11-12
# at-risk	170	136	123	114	108	104
# first episodes	34	13	9	6	4	3
Hazard	0.20	0.10	0.07	0.05	0.04	0.03
Survival	0.80	0.72	0.67	0.64	0.61	0.59

survival rates: 41% of the subjects have committed a domestic violence offense within the first year after treatment.

This estimation approach suggests an alternative representation of the hazard probabilities in terms of event history and risk indicators. That is,

$$P_h(\gamma) = P(\Gamma = \gamma \mid \Gamma \geq \gamma) = P(E_\gamma = 1 \mid R_\gamma = 1), \quad (2.25)$$

where

$$E_\gamma = \mathbf{I}(\Gamma = \gamma) \quad (2.26)$$

and

$$R_\gamma = \mathbf{I}(\Gamma \geq \gamma). \quad (2.27)$$

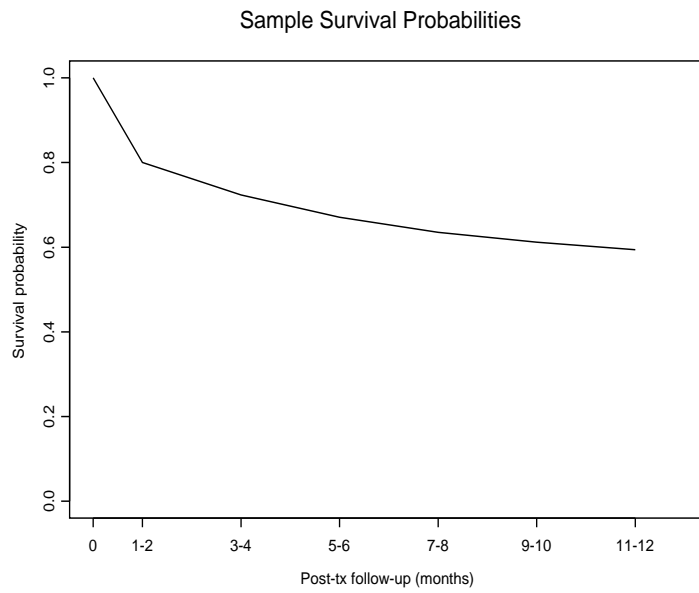
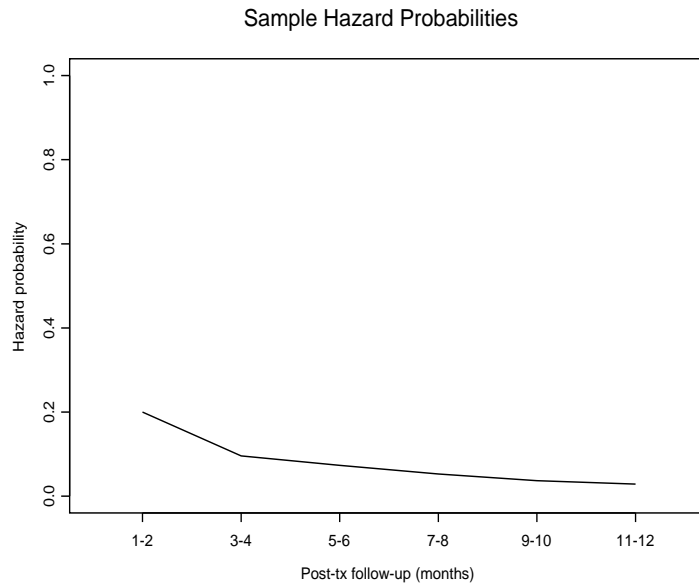


Figure 2.2: Sample hazard and survival probabilities for 12 months post-treatment.

Thus, E_{ji} indicates whether subject i experienced an event in period j and R_{ji} indicates whether subject i is at-risk for an event in period j , where $j = 1, \dots, J$. Substituting into the likelihood from Equation 2.18 gives

$$\begin{aligned} L_i &= [P_h(a_i)]^{\delta_i} \prod_{j=1}^{a_i-1} [1 - P_h(j)] \\ &= [P(E_{a_i} = 1 \mid R_{a_i} = 1)]^{\delta_i} \prod_{j=1}^{a_i-1} [1 - P(E_j = 1 \mid R_j = 1)]. \end{aligned} \quad (2.28)$$

The above likelihood is put in terms of E , R , δ and A . Now consider an indicator, similar to R but indicating *observed* risk. Remember that a subject is only observed to be at risk in period j if she has not experienced the event prior to period j and she is not censored in period j or before. Let R_j^o be an indicator of observed risk in period j , that is,

$$R_\gamma^o = \text{I}(\Gamma \geq \gamma \text{ and } C_r > \gamma). \quad (2.29)$$

In terms of the observed data, (A, δ) , R^o can be equivalently defined by

$$R_\gamma^o = \text{I}([A \geq \gamma \text{ and } \delta = 1] \text{ or } [A > \gamma \text{ and } \delta = 0]). \quad (2.30)$$

Noting that for $j < a_i$ that $R_{ji} = R_{ji}^o$ and that $R_{a_i i} = R_{a_i i}^o$ if $\delta_i = 1$, the likelihood can be expressed exclusively in terms of E and R^o by

$$\begin{aligned} L_i &= [P(E_{a_i} = 1 \mid R_{a_i} = 1)]^{\delta_i} \prod_{j=1}^{a_i-1} [1 - P(E_j = 1 \mid R_j = 1)] \\ &= \prod_{j \in \{r: R_{ri}^o = 1\}} P(E_j = e_{ji}). \end{aligned} \quad (2.31)$$

All of this is still done under the assumption of noninformative censoring. To summarize, the observed data, (A_i, δ_i) , as well as the corresponding likelihood, can be restated in terms of (E_i, R_i^o) without loss of information, with the following conversion:

$$E_{ji} = \begin{cases} 1 & \text{if } A_i = j \text{ and } \delta_i = 1 \\ 0 & \text{otherwise,} \end{cases} \quad (2.32)$$

and

$$R_{ji}^o = \begin{cases} 1 & \text{if } (A_i \geq j, \delta_i = 1) \text{ or } (A_i > j, \delta_i = 0) \\ 0 & \text{otherwise.} \end{cases} \quad (2.33)$$

The following observation about the likelihood given in Equation 2.31 is crucial to understanding the motivation for using the LCR framework to model discrete-time survival data: The likelihood in terms of E and R^o is identical to the likelihood for a one-class ($K = 1$) LCA model with a set of binary indicators, E_j , $j = 1, \dots, J$, with the R_j^o 's treated as response (non-missingness) indicators under the MAR assumption.⁸ Thus, the maximum likelihood estimates for the E_j 's under MAR are the MLE's for the

⁸Note that the specification of the complete data using $E = (E_{R^o}, E_{\overline{R^o}})$ and R^o , where E_{R^o} is the set of E_j 's such that $R_j^o = 1$ mirrors the Little and Rubin (1987) specification for incomplete data.

$P_h(j)$'s under noninformative right censoring.⁹ In a data analysis situation, most programs do not require that the user provide the actual response indicators for missingness; rather, missingness for given observations is denoted by some symbolic representation, such as “.”. Table 2.2 shows three example event histories for the six periods of observation. The first section of the table shows the event indicator values for each subject during each period, the second section shows the observed risk indicator values for each subject during each period, and the third section shows the “observed” event indicators with the observations for which $r_{ji}^o = 0$ marked as missing. Figure 2.3 displays the unconditional event history model using the traditional path diagram representation of the LCA models.

2.2.5 Discrete-time models with covariates

The two most common representations of “grouped-time” survival data are the ordinal and the discrete. These representations form the basis for the different model specifications in the discrete-time setting. The ordinal speci-

⁹This differs from Vermunt’s (1997) suggestion that the discrete-logit model could be expressed as a structured LCA model with dependencies across the event indicators. In this MAR formulation, no further structure on the observed indicators is needed to obtain the proper likelihood estimates.

Table 2.2: Example Data for Discrete-Time Survival Using Event and Risk

Indicators

Event indicator	e_1	e_2	e_3	e_4	e_5	e_6
Event in period 5	0	0	0	0	1	0
Censored in period 4	0	0	0	0	0	0
No event in 12 months	0	0	0	0	0	0
Risk indicator	r_1^o	r_2^o	r_3^o	r_4^o	r_5^o	r_6^o
Event in period 5	1	1	1	1	1	0
Censored in period 4	1	1	1	0	0	0
No event in 12 months	1	1	1	1	1	1
Event indicator (observed)	e_{r^o1}	e_{r^o2}	e_{r^o3}	e_{r^o4}	e_{r^o5}	e_{r^o6}
Event in period 5	0	0	0	0	1	.
Censored in period 4	0	0	0	.	.	.
No event in 12 months	0	0	0	0	0	0

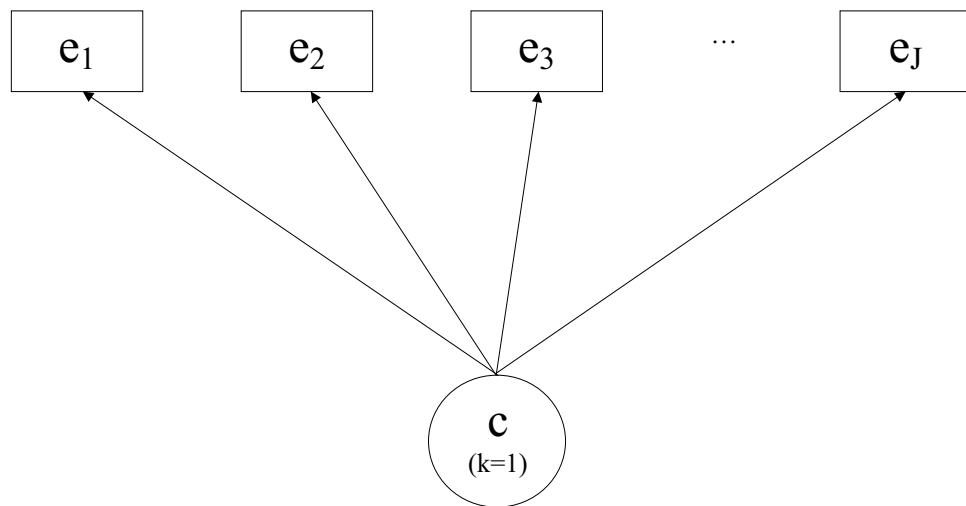


Figure 2.3: Event history LCA path diagram.

cation treats the event time period Γ_i as an ordered polytomous variable with outcomes $1, \dots, J$ and models the effects of covariates as a function of the cumulative probability. A multiplicative model using this ordinal representation is given by

$$P(\Gamma \leq \gamma \mid z) = p_0(\gamma) \cdot \psi(z). \quad (2.34)$$

In the general linear modeling approach, the probability for a categorical outcome is transformed by a *link* function and modeled as linear with respect to the observed covariates. Appropriate link functions ensure that estimated outcomes, in this case, probabilities, are in the admissible outcome space; for probabilities, that means between the values 0 and 1. Having the probability as direct linear functions of the covariates would not guarantee sensible probability estimates. The most common link function used for categorical outcomes is the *logit*, where

$$\text{logit}(p) = \log \left[\frac{p}{(1-p)} \right]. \quad (2.35)$$

The logit of the probability of an event is the log odds of that event. The multiplicative model given above with the logit link is a special case of what is known as the proportional odds model for ordered categorical data (Agresti, 1990) with censored observations. In this model, the odds of experiencing the event in a given interval j are proportional to the odds of experiencing the event *before* interval j , conditional on z .

The discrete specification models the effects of covariates on the hazard probabilities, which, as shown in the previous section, can be represented by probabilities of binary event indicators. The general multiplicative model for the discrete specification is given by

$$P_h(\gamma | z) = P_{h_0}(\gamma) \cdot \psi(z). \quad (2.36)$$

The two most common link functions use in this setting are the *complementary log-log* and the *logit*, given by

$$\log[-\log(1 - P_h(\gamma | z))] = \beta_{0\gamma} + \beta'z \quad (2.37)$$

and

$$\log\left[\frac{P_h(\gamma | z)}{1 - P_h(\gamma | z)}\right] = \beta_{0\gamma} + \beta'z, \quad (2.38)$$

respectively. The model employing the *complementary log-log* link will be referred to here as the discrete-CLL model and the model employing the *logit* link will be referred to here as the discrete-logit model. In the case of time-independent effects for all covariates, the discrete-CLL model assumes proportionality of the hazard probabilities across the time intervals; similarly, the discrete-logit assumes proportionality of the hazard odds across time intervals. As with the Cox regression model, the assumptions of proportionality for both models may be relaxed so it does not make sense to distinguish these models in name by their differing proportionality assumptions.

Since the discrete representation makes easier the incorporation of time-dependent covariates and time-dependent covariate effects as well as structured relationships between hazard probabilities across time compared to the ordinal representation, it is the basis of all further models discussed in this dissertation. Deciding to model the hazard probabilities as a function of the covariates does not, however, resolve the issue of which link function is preferable. Discussion of this very issue is given in Halford (1976); Prentice and Gloeckler (1978); Allison (1982); Hedeker et al. (in press); and Singer and Willett (2003), among others. In the case where there is an underlying continuous distribution, there is some favor in the literature for the discrete-CLL over the discrete-logit. The regression coefficients from the discrete-CLL estimate the same quantities as those in the Cox regression model—hazard ratios. Ironically, Cox himself, in his 1972 paper, gives the discrete-logit as the discrete model counterpart to his continuous-time regression model. And, as stated before, there is nothing in the nature of survival data as a whole that recommends them specifically to discrete-CLL model. Singer and Spilerman (1976) as well as Flinn and Heckman (1982) warn about the sensitivity of inferences based on the discrete-logit model to the length of the time interval compared to the discrete-CLL estimates which are interval invariant with respect to inferences about the structural parameters, i.e., baseline hazard probabilities.

Practically speaking, there is often little or no difference in the results of these two model specifications, even when the proportionality assumptions are in place, since it is well known under conditions of rare-occurring events that the odds ratio approximates the risk ratio. There is little distinction in this regard when the proportionality assumptions are relaxed and the covariate effects are allowed to be invariant. Also, by allowing the most general form of the baseline hazard in each model—estimating a separate baseline probability for each time interval—differing lengths of time intervals are automatically incorporated into the parameter estimates. However, one should be careful to note that each link function has different assumptions about the interplay between the baseline hazard probabilities and the covariates in relation to the conditional hazard probabilities—assumptions that cannot be directly verified. Given that there is no clear reason to favor one link over the other and that the *logit* link is implemented in many software applications and is more familiar and immediately accessible in understanding and application to those new to survival analysis, the discrete-logit is the specific model used for the remainder of this dissertation with the understanding that all discussions of model extensions could as easily be applied to discrete-CLL models.

2.2.6 Discrete-logit model in a latent variable framework

The primary reason for the established popularity of the discrete-logit model over the discrete-CLL in applied settings is because the full maximum likelihood estimates for the parameters of the discrete-logit model can be estimated using the regular logistic regression machinery present in most general statistical analysis software. (See Allison, 1982, and Singer and Willett, 1993, for full exposition of obtaining estimates using this framework.) Muthén & Masyn (2001) present an alternative approach for estimating the discrete-logit model using a latent variable framework.¹⁰ In modeling single-occurrence events in discrete-time with observed predictors, the latent variable framework provides an analytically equivalent model to that specified in the logistic regression framework. So, the estimation is not somehow “better” or “novel” using this framework. However, it becomes clear as the model is extended beyond single events to include unobserved population heterogeneity, recurring events, parallel and sequential longitudinal processes, and more that the latent variable framework affords much greater flexibility in modeling.

¹⁰There is nothing about this framework that requires the use of the *logit* link function. However, it is typically the link function employed in the latent variable software and is the link function currently employed in Mplus.

The most general latent variable modeling framework involves both categorical and continuous latent variables (Muthén & Shedden, 1999; Muthén, 2002) and is incorporated in the Mplus program (Muthén & Muthén, 1998-2001). Section 1.3.3 gave an overview of latent class regression and it was shown in Section 2.2.4 that the likelihood for discrete, right-censored, survival data could be equivalently expressed as the likelihood for an LCA model with missingness on the binary indicator variables. In the following section, the LCA model (under MAR) for discrete-time survival data is extended to the LCR model, allowing the hazard probabilities to be modeled as a function of observed covariates.

Discrete-logit model using LCR

Recall the representation of the discrete-time survival data in terms of E_{R^o} , $E_{\overline{R^o}}$, and R^o given in Section 2.2.4. Treating E as previously defined for survival data as a vector of categorical indicators of a latent class variable with $K = 1$ and missingness on those indicators given by R^o such that E_{R^o} represents the whole of the observed indicator data, the maximum likelihood estimates under the MAR assumption for ν_j , $j = 1, \dots, J$, are the maximum likelihood

estimates for the logit hazard probabilities.¹¹ That is,

$$\hat{P}_h(\gamma) = \hat{P}_{E_j}(E_j = 1 \mid C = 1) = \frac{1}{1 + \exp(-(\hat{\nu}_\gamma))}. \quad (2.39)$$

Using the properly formatted data, an unconditional, unstructured discrete-logit model can be fit to the violence data. The estimates for this model are given in Table 2.3. Note that the inverse logit of the ν 's match the sample probability estimates given in Table 2.1. For example,

$$\hat{P}_h(1) = \frac{1}{1 + \exp(-(-1.39))} = 0.20. \quad (2.40)$$

The maximum likelihood estimates of ν_j and β from a LCR model are the estimates of the corresponding parameters in the discrete-logit model. That is,

$$\hat{P}_h(\gamma \mid z) = \frac{1}{1 + \exp(-(\hat{\nu}_j + \hat{\beta}'z))}, \quad (2.41)$$

where $\hat{\nu}_j$ is the estimated logit of the hazard probability at time period j when $z = 0$ and $\hat{\beta}_p$ is the estimated log hazard odds ratio for a one unit increase in z_p . Alternatively, $(1 + \exp(-(\hat{\nu}_j)))^{-1}$ is the estimated baseline hazard probability

¹¹Note that the Mplus specification uses the parameter $\tau = -\nu$; this is related to the conceptualization of categorical data as originating from continuous data that has been categorized using cut-points. For a two category variable, there is one cut-point, otherwise known as threshold, and τ is the estimate for that threshold. In the discrete-logit model, the estimate for τ_γ is equal to the estimate for $-\nu_\gamma$ or $-\beta_{0\gamma}$ as given in Equation 2.38.

Table 2.3: Results for Data Example Model 1

Parameter	Est.
ν_1	-1.39
ν_2	-2.25
ν_3	-2.54
ν_4	-2.89
ν_5	-3.26
ν_6	-3.52

LL=-214.35, parameters=6

for time period j and $\exp(\beta_p)$ is the estimated hazard odds ratio for a one unit increase in z_p .

Beginning to investigate the relationship between the time-independent covariates and the hazard of domestic violence in the post-treatment periods, a series of models were fit including each time-independent covariate separately and then in combination with the other variables, all under the proportional hazard odds assumption. The two covariates that were significant predictors of the time to the first episode of domestic violence were the indicator for the behavior couples' therapy treatment (relative to both the individual-based treatment and the attention control treatment) and the indicator for wife's education not

beyond high school or GED. None of the other covariates¹² were significant in the model, including the number of pre-treatment violence episodes, and there was no evidence of interaction between any of the variables.

For a continuous variables, such as length of relationship, there is an assumption in the model, if it is entered as a continuous variable, that its relationship to the logit of the hazard probability is linear, i.e., the logit hazard changes the same amount for every one unit change in the covariate. It is possible to relax this linearity assumption by including polynomial terms for the covariates, such as $(\text{length of relationship})^2_{\text{mean-centered}}$. To explore nonlinear relationships, it can also be useful to categorize the continuous variable and represent its effect in the model with a series of dummy variables. This was done for the continuous covariates in the example and no evidence of nonlinear effects was found. The results of the model with the treatment and wife's education indicators included are given in Table 2.4. The BCT treatment has a protective effect against the onset of violence in the post-treatment period. The negative coefficient on the wife's education indicator variable suggest that subjects with wives who have no education beyond high school are less at-risk

¹²Includes husband's age (in years), length of relationship (in years), husband's race, marital status, DWI referral, alcohol dependency criteria met, household income, pre-treatment drinking, and pre-treatment violence.

for returning to violence at any given time in the post-treatment period. Perhaps these women are less likely to engage in challenges or confrontations with their husbands that might be viewed by the men as a provocation for violence.¹³ Also given in the table are the exponentiated values of the estimated coefficients which have the interpretation of the hazard odds ratio. Thus, it is estimated that subjects in the BCT treatment group have approximately half the risk of those not in the BCT treatment of returning to violence at any given period in the 12 months following treatment; and that subjects with wives with no education beyond high school have approximately half the risk of those with wives having education beyond high school or high school equivalency for returning to violence at any given period. Plots of the estimated hazard and survival probabilities for the four groups defined by BCT treatment group and wife's education level are shown in Figure 2.4.

There are two standard approaches to evaluating the statistical significance of a single covariate, say z_p , in a LCR model. One is the Likelihood

¹³At first glance, this may seem counterintuitive since in the domestic violence literature, education is an overall protective factor for women with respect to risk of spousal violence. However, one must remember that this is a sample of women who have already been subject to violence at the hands of their partners *and* have stayed in the relationships, at least through the treatment period.

Table 2.4: Results for Data Example Model 2

Covariate	Coeff. Est.	SE	Est./SE	Est. hazard OR
I(Treatment=BCT)	-0.59	0.29	-2.03	0.55
I(Wife's educ. \leq H.S.)	-0.62	0.28	-2.24	0.54
Threshold	Est.	SE	Est./SE	Est. baseline hazard
ν_1	-0.99	0.24	-4.18	0.27
ν_2	-1.82	0.31	-5.81	0.14
ν_3	-2.10	0.38	-5.54	0.11
ν_4	-2.44	0.44	-5.50	0.08
ν_5	-2.80	0.51	-5.47	0.06
ν_6	-3.07	0.60	-5.11	0.04

LL=-209.62, parameters=8

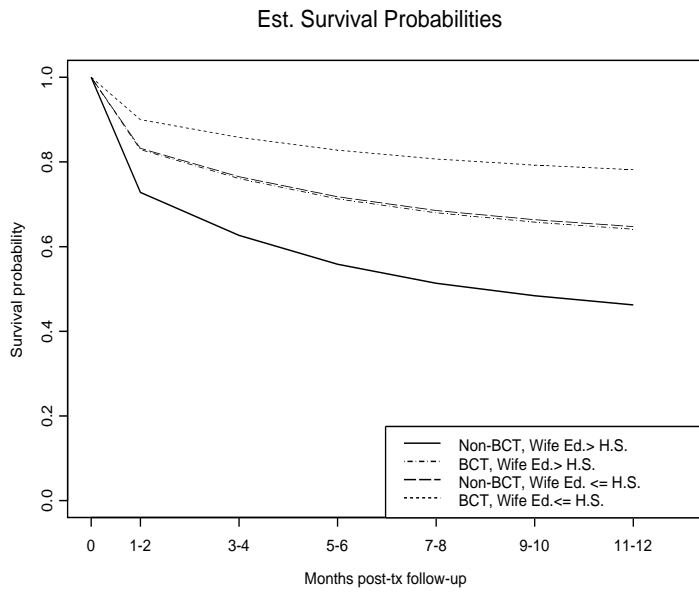
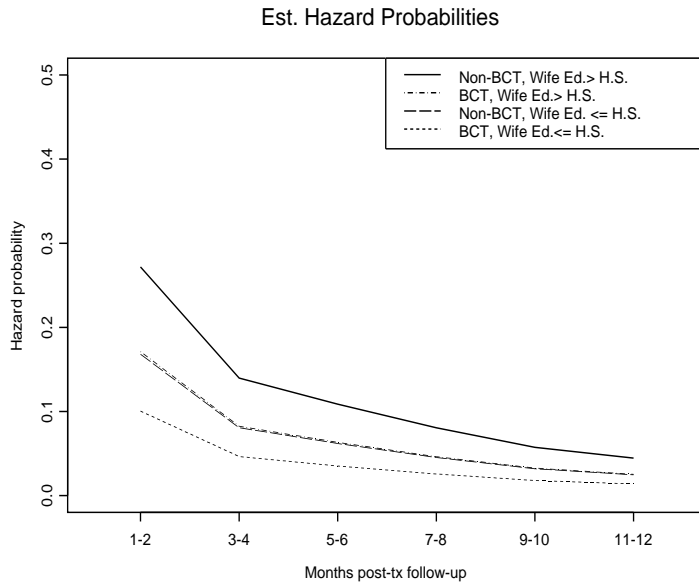


Figure 2.4: Model 2 estimated hazard and survival probabilities.

Ratio Test (LRT) and the other is the Wald Test.¹⁴ Both tests are based on asymptotics. The LRT is given by

$$LRT = -2 [LL(\text{model without } z_p) - LL(\text{model with } z_p)], \quad (2.42)$$

where the LRT is χ^2 distributed with one degree of freedom under the null hypothesis that $\beta_p = 0$. The Wald test is given by

$$W = \frac{\hat{\beta}_p}{\widehat{SE}(\beta_p)}, \quad (2.43)$$

where the Wald statistic has a standard normal distribution under the null hypothesis that $\beta_p = 0$. As cited in Hosmer and Lemeshow (2000), both Huack and Donner (1977) and Jennings (1986) have investigated the performance and adequacy of inferences based on Wald statistics and have found that the Wald test can behave in an irregular manner; the LRT is recommended over the Wald test for these models. The LRT can also be used for multivariate hypothesis testing; that is to say, the LRT can be used to evaluate the difference between

¹⁴The Score test is another method of hypothesis testing that can be used for these models but is not discussed in detail here because it is not always available in commercial software packages.

any two *nested* models. Most generally,

$$LRT = -2[LL_0 - LL_1], \quad (2.44)$$

where LL_0 is the log likelihood for the restricted model under the null hypothesis and LL_1 is the log likelihood for the larger, unrestricted model under the alternative hypothesis. The LRT statistic is χ^2 distributed with degrees of freedom equal to the difference in the number of free (estimated) parameters in the two models. That is, under the null hypothesis, $LRT \sim \chi^2(df = q_1 - q_0)$, where q_0 is the number of parameters in the null model and q_1 is the number of parameters in the alternative model. The parameters space for the null model must be contained *within* the parameter space of the alternative model. The LRT can be used for inferences about many different aspects of the hazard model including the significance of a single covariate (continuous or categorical), the time-dependency of the effects of a given covariate, and nature of the duration dependency of the baseline hazard probabilities. The use of the LRT is demonstrated in examples throughout the rest of the chapter.

The proportionality assumption

As has already been alluded to, the discrete-logit model can be made more general by relaxing the assumption of proportionality of the hazard odds—this means specifying a model that allows the effects of the covariates (in terms of

hazard odds ratios) to differ across time periods. In the regular logistic framework, modeling this would involve interaction terms between each covariate and each time period indicator. Working from the LCR parameterization of the discrete-logit model, it is a straightforward matter to extend equation (2.41) to allow for time-dependent covariate effects. This model is given by

$$P_{E_j}(e_j = 1 \mid z, c = 1) = \frac{1}{1 + \exp(-(\nu_j + \beta'_j z))}. \quad (2.45)$$

By fitting the model above and then the proportionality model with $\beta_j = \beta$ for $j = 1, \dots, J$, the model LRT can be used to assess the need for relaxation of the proportionality assumption. It may be a more structured model for the covariate effects, somewhere in between that of the full proportional hazard odds more and the model that allows different effects for each time period, could provide an optimal fit, e.g., the hazard odds ratio is the same in the early time periods but then changes in the later time periods, similar to a piece-wise regression. For example,

$$P_{E_j}(e_j = 1 \mid z, c = 1) = \frac{1}{1 + \exp(-(\nu_j + \beta'_{1j} z \cdot I(j \leq q) + \beta'_{2j} z \cdot I(j > q)))}, \quad (2.46)$$

where q represents the change-point for the covariate effects. Also, it is possible for some but not all the covariates to have time-dependent effects.

Continuing with the example, the estimated survival probabilities based on Model 2 can be plotted against the stratified sample survival probability

estimates for the four groups defined by the treatment and wife's education indicators. These four plots are displayed in Figure 2.5. These plots show that the model-based estimates generally correspond closely to the sample estimates but deviation from the sample estimates seem greatest for the middle two time periods. This suggests that there may be some improvement in fit by relaxing the proportionality assumption, allowing the effects of treatment and wife's education to vary for the first, middle, and last two time periods. The log likelihood for the model allowing for time-dependent effects was -206.89 with 12 free parameters.¹⁵ Comparing this model to the more restricted proportional hazard odds model, the LRT statistic was $-2(-209.62 - -206.89) = 5.46$; $P(\chi_4^2 > 5.46) = 0.24$. Thus, there was no significant evidence to suggest that the proportionality assumption was violated by these data.

Time-dependent covariates

Up until this point, only time-independent covariates have been considered, that is, only covariates whose values remain constant during the full period of observation. Such explanatory variables might include individual characteristics, such as race, or variables whose values are fixed at or before time zero, such as treatment status. However, it would be unusual in a longitudinal set-

¹⁵Model 3 in Appendix B.

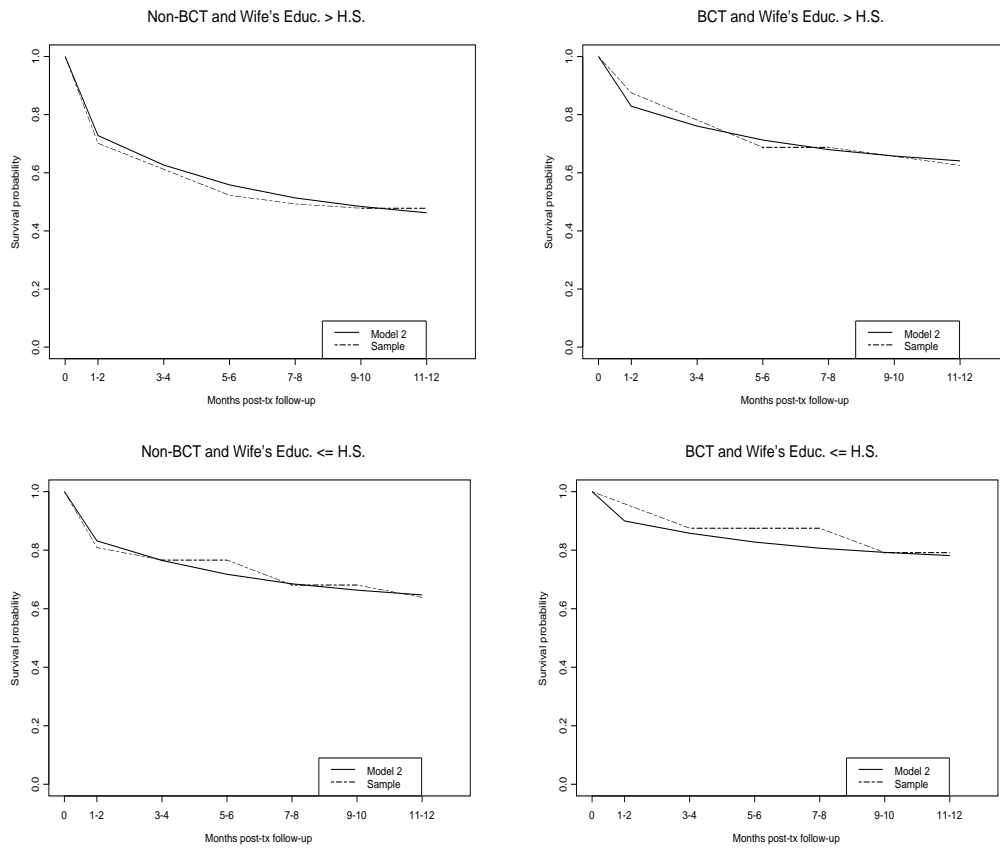


Figure 2.5: Model 2 estimated versus sample survival probabilities by treatment status and wife's education level.

ting that there would not be covariates that changed over time. It is possible to imagine a situation where a *change* in the covariate was itself predictive of survival. For example, consider how a change in marital status or a change in employment status might influence the risk of depression for an individual over time.

In terms of the LCA parameterization of the discrete-logit model, it is a straightforward matter to extend equation (2.41) to allow for time-dependent covariates. Let X be a vector of time-dependent covariates where X_j is the vector of values on those covariates at time period j . The model is given by

$$P_{E_j}(e_j = 1 \mid x_j, z, c = 1) = \frac{1}{1 + \exp(-(\nu_j + \beta'_j z + \kappa'_j x_j))}. \quad (2.47)$$

Notice that the above model allows for the possibility of time-varying effects for both the time-dependent and time-independent covariates. Figure 2.6 displays the path diagram for an event history model with both time-dependent and time-independent covariates with time-varying effects.

Although the extension in the model makes the inclusion of time-dependent covariates a simply matter, the implications for the interpretation of the model are much more complex. As in the growth modeling setting, inclusion of time-dependent covariates introduces the issue of reciprocal causation (Cox & Oakes, 1984; Singer & Willett, 2003). Essentially, when linking contemporaneous information on a predictor and outcome, the directional ar-

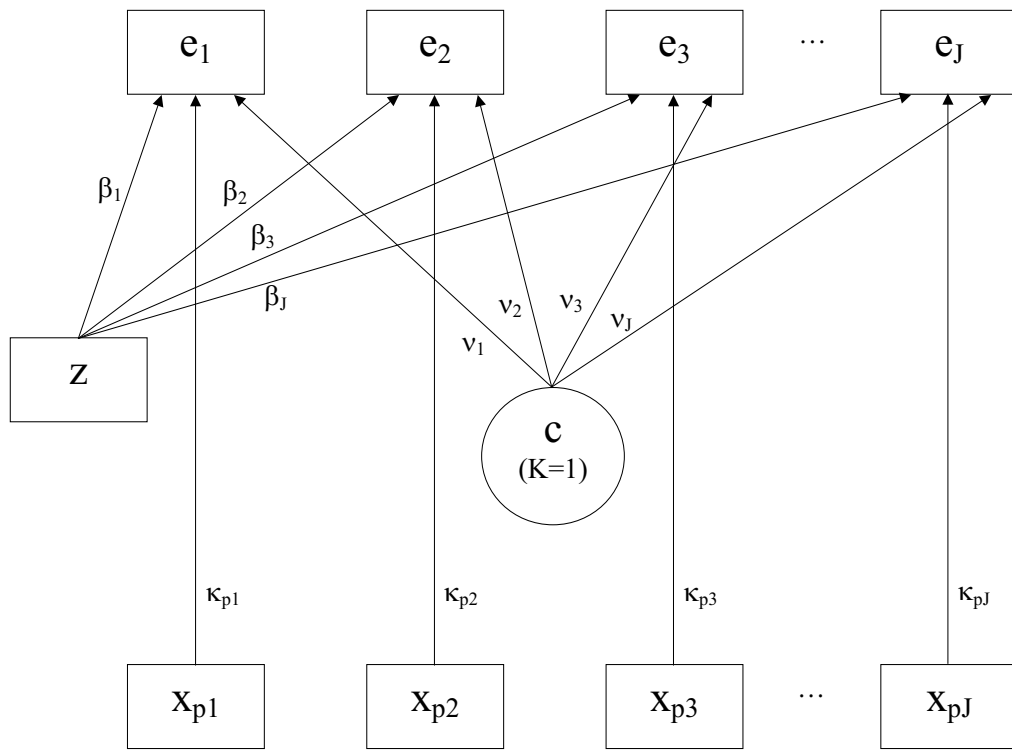


Figure 2.6: Event history LCR path diagram.

row, that is, which variables is doing the influencing, is impossible to infer from the data for certain types of covariates whose values *could* be influenced by the co-occurring outcome. In the discrete-time survival context, this issue is classified into two sub-issues: *state* and *rate* dependence. In the case of *state dependence*, the covariate value in time period j is influence by an individual's "state" in period j , that is, whether or not she experiences the event. In the case of *rate dependence*, the covariate value in time period j is influenced by the individual's hazard probability in time period j (Singer & Willett, 2003). Both cases pose problems for model-based inference and interpretation that can only be resolved by the substantive theory behind the model.

One method for side-stepping the issue of reciprocal causation is to time-lag the covariate values. That is, to have the covariate value in time period $j - 1$ predicting the outcome in time period j , or having the *change* in the covariate value from $j - 1$ to j influence the outcome in time period j . Such covariates are referred to as *lagged* predictors. Something to consider when specifying a time-lagged relationship from a time-dependent covariate to the hazard probabilities is whether an average lag the size of the time intervals that define the discrete time periods in the data is reasonable. For example, if

a subject is more likely to return to drinking within a week of losing his job, then a time lag for employment status over two month time periods would not capture the more proximal lag effect.

Another issue is the implicit assumption in the discrete-time model that the values of a time-varying covariate are essentially constant within each time period. If the data do not correspond to this assumption, there are several alternatives for including such a covariate. One option is to simply take the mean value of the covariate for each individual within each time period and use that as the assumed value for the entire period. It is also possible to incorporate a time-varying covariate that embodies something about the variability of the covariate for a given individual in each time period as well as the mean value.

Although time-dependent variables do present some issues in modeling and inference, they should not be ignored because of the central role they may play in the understanding of the outcome, i.e., survival process of interest. However, there are alternative approaches to specifying the relationship between such time-dependent variables and survival in the latent variable framework that may not be subject to some of the issues described about. For example, suppose that the relationship between the covariate and outcome may be better conceived as two co-occurring longitudinal processes where the

trajectory of the predictor over time, rather than its specific value at any given time, is most predictive of event time. This type of model is beyond the scope of the current dissertation but is a promising future extension.

Returning to the example, there are two variables related to drinking behavior that could be included as time-dependent predictors of violence. One variable is an indicator of return to drinking for each month and the other is the percent-days-drinking for each month. There is a substantial amount of literature exploring the connections between alcohol and violence (Fals-Stewart, 2003), so it could be expected that one or both of these would be significant predictors. A series of models were fit, including both concurrent and two-month lag effects of both these covariates, separately and together. There was significant evidence for the concurrent association of percent-days-drinking with the hazard of first violence post-treatment. In this example, if drinking and violence are considered to be closely linked in time (e.g., within 24 hours)—although there is still debate in the literature about the direction and causal relationship of the association—it is not surprising that there is not a significant lag effect given that the time periods are two months in length. However, as discussed earlier in this section, care must be taken when interpreting the concurrent time-dependent effects. The effect of the BCT treatment group membership became nonsignificant when including percent-days-drinking, sug-

gesting that the treatment effect on time-to-violence is mediated by drinking behavior. An indicator of household income greater than \$35,000 enters the model as a protective factor in return to violence. The effect of wife's education level remains from Model 3. There was no evidence to suggest that the proportionality restriction should be relaxed for percent-days-drinking; that is, there was no evidence of time-dependent effects for this time-dependent variable. Further details of this model are given in the next section.¹⁶

Structure for the hazard probabilities

Until now, the discrete-logit model has been presented as specified with unrestricted baseline hazard probabilities, in the form of the ν'_j s, that are allowed to vary freely across the J time intervals. There are advantages to leaving the baseline hazard unstructured including accounting for different time interval lengths and avoiding bias in covariate effect estimates due to baseline hazard misspecification. However, if the underlying hazard function does have a structural form that could be represented by a set of constraints on the baseline hazard probabilities, there can be a gain in statistical efficiency and parsimony by doing so. Like the previous section on time-dependent effects, models plac-

¹⁶Model 4 in Appendix B.

ing constraints on the baseline hazard can be compared via the model LRT, as they are nested within the unrestricted baseline hazard probabilities model.

The most restricted baseline hazard probabilities model is one where the hazard probabilities are constrained to be equal across time. That is,

$$P_{E_j}(e_j = 1 \mid z, c = 1) = \frac{1}{1 + \exp(-(\nu + \beta'_j z))}. \quad (2.48)$$

One could also imagine a piecewise constant hazard probability model, such as

$$P_{E_j}(e_j = 1 \mid z, c = 1) = \frac{1}{1 - \exp(-(\nu_1 \cdot I(j < q) + \nu_2 \cdot I(j \geq q) + \beta'_j z))}, \quad (2.49)$$

where q represents the change-point for the hazard probabilities.

It may also be the case that baseline hazard probabilities can be described by some sort of function of time, e.g., linearly increasing in time. Keep in mind that with the discrete-logit, it is the *logit* of the hazard probability that is being modeled. In the latent variable framework, time is not automatically included as an explicit set of variables in the model, as in the logistic regression—it is represented implicitly by the set of event indicators, one for each time period. However, time structure to the logit baseline hazard probabilities can be imposed using a latent growth modeling approach (Muthén & Masyn, 2001; Muthén, 2002). Since the discrete intervals are design-specific

rather than individual-specific, all subjects can be considered to be observed at the same times. The *logit* hazard probabilities are then expressed as a function of an intercept and slope factor, represented by latent variables, with factor loadings constraining the structure. For example, a model with a logit baseline hazard linear in time can be expressed as

$$P_{E_j}(e_j = 1 \mid z, \phi, c = 1) = \frac{1}{1 + \exp(-(\eta_0 + \lambda_j \eta_1 + \beta'z))}, \quad (2.50)$$

where η_0 is a latent intercept variable with loadings all are fixed at the value one and η_1 is the latent linear slope variable with loadings such that λ_j is the factor loading for e_j fixed at $\lambda_j = j - 1$ or some other expression of time such as $\lambda_j = \frac{1}{2}(t_j - t_{j-1})$. For identification, the thresholds, ν_j 's, are fixed at zero; at least one λ_j must be fixed at zero and at least one other fixed at a constant value.. If $\lambda_1 = 0$ then the mean of η_0 is then the logit hazard baseline probability at the first time period. The mean of η_1 is the time slope of the logit hazard probabilities. To allow nonlinear growth in polynomial form, additional growth factors are added with loadings corresponding to higher powers of time. To allow for less specific nonlinear growth, all but two of the loadings on η_1 could be freely estimated rather than fixed at given time values. This specification also allows for a slightly different representation of time-independent covariate effects since the growth parameters, as variables themselves, may be expressed as dependent upon the covariates. For example, $\eta_0 = \alpha_0 + \alpha'z$ is

equivalent to specifying a discrete-logit model with time-invariant effects for z , i.e., the proportional hazard odds model. For $\eta_1 = \alpha_0 + \alpha'z$, the effects of z on the hazard probability increase across the time periods according to the λ_j 's. For now, assume $Var(\eta | z) = 0$ —unobserved heterogeneity is discussed further in Chapter 3.

For the example, comparing Model 4 to the more restricted constant baseline hazard model¹⁷, the LRT statistic was $-2(-211.50 - -193.84) = 35.32$; $P(\chi_{9-4=5}^2 > 35.32) < 0.001$. Thus, there is strong evidence against a constant baseline hazard. As an alternative, a linear logit baseline hazard model was fit¹⁸. Such a shape is suggested when looking at the unstructured hazard probability estimates. The unstructured does not offer any significant improvement over the linear model: $LRT = -2(-194.82 - -193.84) = 1.96$; $P(\chi_{9-5=4}^2 > 1.96) = 0.74$). The linear model also offers a significant improvement over the constant hazard model: $LRT = -2(-211.50 - -194.82) = 33.36$; $P(\chi_{5-4=1}^2 > 33.36) < 0.001$. Table 2.5 gives the results for the model with the linear logit of the baseline hazard. Note the estimate for the mean of η_1 represents a decreasing logit baseline hazard with the estimated odds ratio comparing the hazard odds between time $j + 1$ and j , controlling for treat-

¹⁷Model 5a in Appendix B.

¹⁸Model 5 in Appendix B.

Table 2.5: Results for Data Example Model 5

Covariate	Coeff. Est.	SE	Est.SE	Est. hazard OR
I(Wife's educ. \leq H.S.)	-0.68	0.28	-2.43	0.51
I(Income $>$ \$35K)	-0.67	0.33	-2.04	0.51
% days drinking	2.37	0.42	5.72	10.70
$E(\eta_0)$	-1.72	0.23	-7.49	0.18
$E(\eta_1)$	-0.50	0.10	-4.93	0.61

LL=-194.82, parameters=5

ment, wife's education, and concurrent drinking. The estimated hazard odds ratio for percent-days-drinking is comparing the hazard odds between those drinking 100% of days and those drinking 0% of days. Figures 2.7–2.9 display the estimated hazard and survival probabilities for each covariate effect at the sample mean value for the other covariates.

Censored and truncated data

Until this point, the models presented have been for complete and right-censored data. This section will address how to estimate the discrete-logit

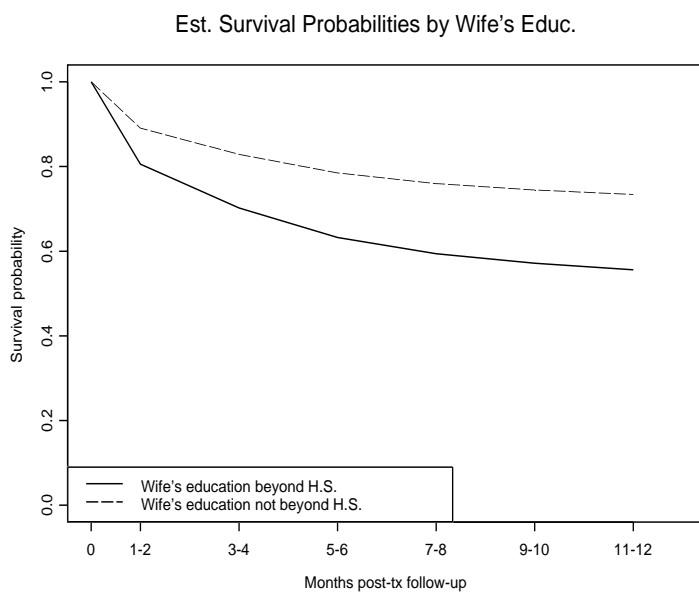
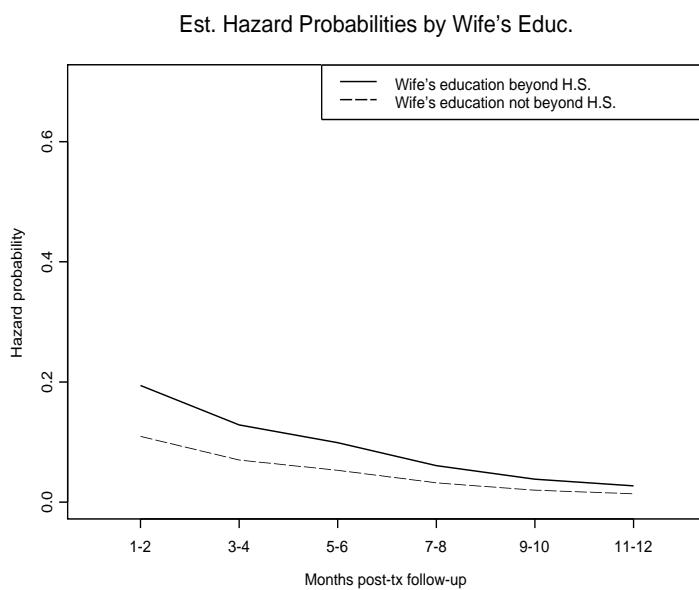


Figure 2.7: Model 5 estimated hazard and survival probabilities by wife's education level.

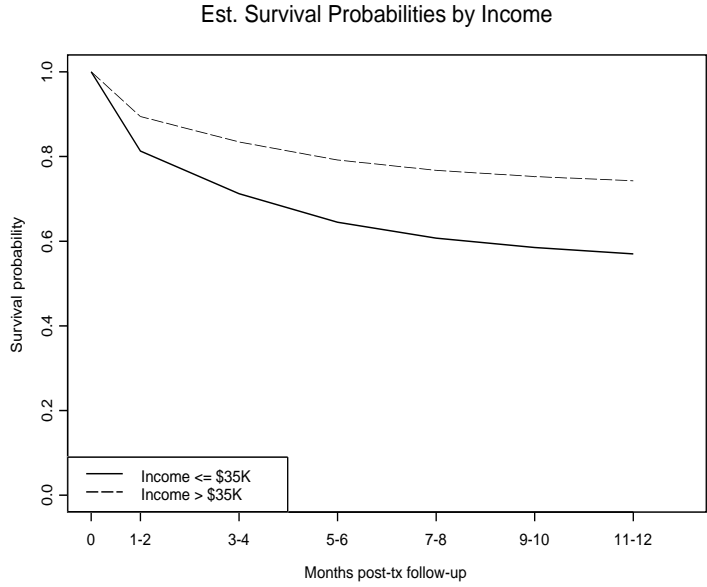
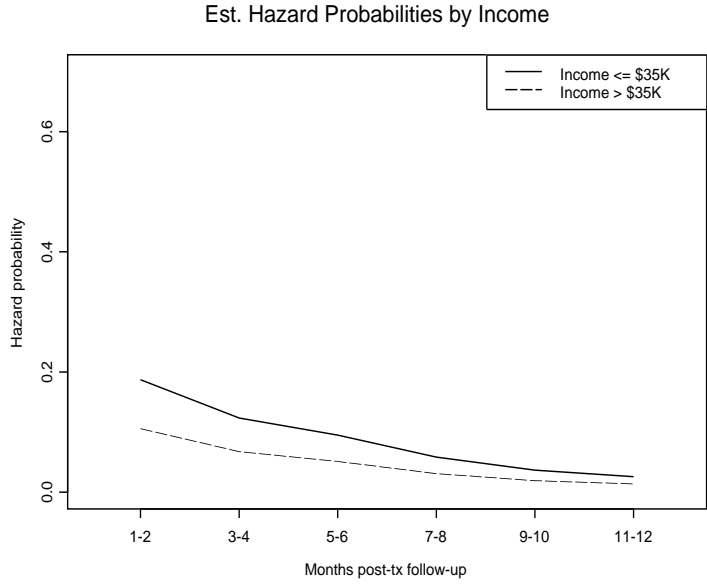


Figure 2.8: Model 5 estimated hazard and survival probabilities by household income.

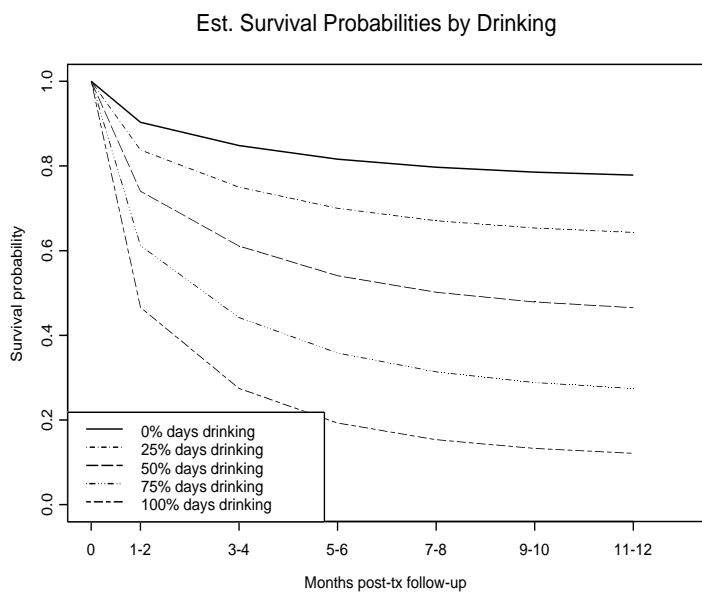
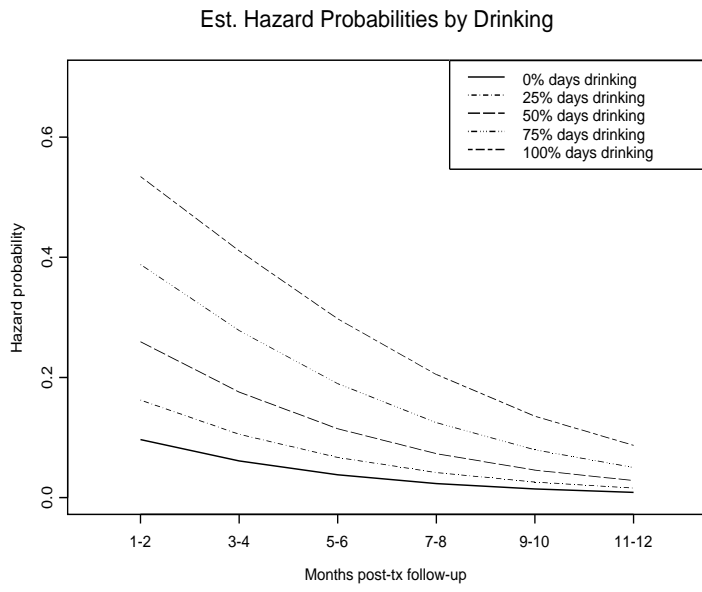


Figure 2.9: Model 5 estimated hazard and survival probabilities by % days drinking.

model with other forms of noninformative censoring and truncation. All the methods discussed in the following section can be applied to data with covariates as well.

Left-truncated data. Recall that left truncation occurs when subjects are only known to the researcher if they have experienced an intermediate event or have *not* experienced the event of interest at the initiation of the observation period. For example, consider a disease survival study that recruits subjects only 18 years of age or older and measures time to death. Any persons having died before the age of 18 will not be included in the sampling frame. Let Q_{LT_i} be the period at which subject i is left-truncated. Assuming $\Gamma \perp Q_{LT}$, the following equivalence holds:

$$P(\Gamma_i = \gamma \mid \Gamma_i \geq \gamma, \Gamma_i \geq Q_{LT_i}) = P(\Gamma_i = \gamma \mid \Gamma_i \geq \gamma), \quad (2.51)$$

for $\gamma \geq Q_{LT}$. That is, the hazard probabilities for left-truncated data are the same as the unconditional hazard probabilities. This holds for the inclusion of covariates so long as $(\Gamma|Z) \perp (Q_{LT}|Z)$. Note that obtaining the corresponding estimates for the survival probabilities does not lead to the same equality. If the left truncation precludes any observations for periods 1 to $Q_{LT} - 1$, then

the survival probability estimates will be conditional on the truncation. That is,

$$\prod_{j=Q_{LT}}^{\gamma} (1 - P_h(j)) = P(\Gamma > \gamma \mid \Gamma \geq Q_{LT}) = \frac{P_S(\gamma)}{P_S(Q_{LT} - 1)}. \quad (2.52)$$

Right-truncated data. Right truncation occurs when only subjects experiencing the event before a time period Q_{RT} are known to the researcher, that is, $\Gamma \leq Q_{RT}$ for all sampled individuals. As with left-truncated data, if event time is independent of the truncation time, then the hazard probabilities for right-truncated data are the same as the unconditional hazard probabilities but the survival probabilities are conditional on the truncation.

Double-censored data. Recall that left censoring occurs when subjects have experienced the event in a time before the onset of the study but the exact timing of the event is unknown to the researcher. It is rare that left censoring occurs in sample in the absence of right censoring. What is proposed here to deal with such censoring is an iterative algorithm, similar to the nonparametric one suggested by Turnbull (1974) for the continuous-time setting. Take, for a moment, a data set with one left-censored observation, v , all the rest being complete or right-censored. And suppose that observation is left-censored at time q_{LT} so that $\Gamma_v < q_{LT}$. This means that $\Gamma_v \in \{1, 2, \dots, q_{LT} - 1\}$. Begin by estimating the hazard probabilities for each time period based on the data with observation v excluded. Then replace observation v in the

original data set with $q_{LT} - 1$ pseudo-observations, $(v_1, v_2, \dots, v_{q_{LT}})$, such that $(A_{v_k}, \delta_{v_k}) = (k, 1)$ for $k = 1, 2, \dots, q_{LT}$. Create a weight vector such that all complete or right-censored observations are given a weight of one. The weight for the pseudo-observations are computed as follows:

$$w_{v_k} = \frac{\hat{P}_h(k)}{W_v} \quad \text{where} \quad W_v = \sum_{j=1}^{q_{LT}} \hat{P}_h(j). \quad (2.53)$$

Re-estimate the discrete-logit model with the pseudo-observations and the new weight vectors. Based on those hazard probabilities, recalculate the weights and repeat until the hazard probabilities for each time period converge.

Interval-censored data. Interval-censored observations can be treated in much the same way as described above for left-censored observations. Recall that interval-censored subjects are only known to have experienced the event during a stretch of intervals. In this case, the same iterative algorithm is applied, creating as many pseudo observations for each interval-censored subjects as the number of time period contained in the censoring interval. Compute the weights based on the sum of the hazard probabilities across those same time periods.

Discretizing continuous-time data

One of the questions that typically arises when researchers seek to utilize some of the unique features¹⁹ of the discrete-time survival framework is how continuous-time data should be discretized. This is too general a question to be able to provide specific rules-of-thumb suitable to all data. However, there are some things to consider when discretizing: 1) For the most general model with an unstructured hazard probability structure, the model estimation algorithm will not converge if there are any “empty” intervals, that is, intervals with no events within; 2) If there are time-dependent covariates, it is wise to choose interval widths within which there is little individual variability in each covariate value; 3) The fewer events in each time period, the less precise the estimates for *time-dependent* effects will be; 4) Time intervals should make substantive sense; and 5) Sensitivity to the discretizing can be assessed by comparing estimates from different time-period definitions to each other and to a comparable continuous-time model.

¹⁹The gap between extensions of continuous-time models and discrete-time models is ever narrowing, but there are still some differences, at least in available software implementations, that may make one approach more appealing than another for a given research question.

2.2.7 Model assessment

Previous sections have already discussed the comparison of nested models to assess the appropriateness of various model assumptions such as proportionality of the hazard odds for covariate effects and different structures for the baseline hazard probabilities. There is very little that has been done in the survival literature with goodness-of-fit in the discrete-time setting. Since the single event model can be specified in the logistic regression framework, corresponding goodness-of-fit tests may be applied such as the Osius and Rojek (1992) large sample normal approximation to the Pearson χ^2 and the Hosmer-Lemeshow decile of risks test (Hosmer & Lemeshow, 2000). In the LCA literature, the Pearson χ^2 is the statistic of choice. However, the Osius and Rojek version is needed when there are continuous covariates. For LCR model, the most recent goodness-of-fit test was proposed by Huang and Bandeen-Roche (in press). This G^2 statistics is described in greater detail in Chapter 3. The Osius and Rojek model has not been extended to LCR models and the G^2 formulation does not accommodate missing data. Thus, there is not goodness-of-fit measure currently available that is appropriate for the broader class of models described in remaining chapters.

With regards to issues of power and sample size, one may draw from the field of categorical data analysis. In the discrete-time survival setting,

the necessary sample size to fit specific models or the power to detect certain covariate effects is related to the rarity of the event relative to the sample size. Peduzzi, Concato, Kemper, Holford, and Feinstein (1996) demonstrate the “rule of 10”, showing that at least 10 *events* per parameter are necessary to obtain reliable estimate of logistic regression coefficients and their standard errors; to apply this rule with an event history model, one would sum up the total number of observed events across all the time periods.

Chapter 3

Unobserved Heterogeneity

In the last chapter, population heterogeneity in the survival process was assumed to be *observed*, that is to say, all variability in individual survival processes could be explained by covariates with known values for all individuals in the sample. The discrete-logit model, estimated in a latent class regression framework, was used to model the relationship between the measured covariates and the hazard probabilities. Unlike the traditional linear regression models or continuous indicator latent variable models, no random error term was included. It is not likely, however, no matter how well-designed a study may be, that all sources of population heterogeneity have been noted and accounted for. Nor is it likely that all individuals with the same observed covariate values have identical hazard probabilities. It may also be the case

that there are sources of population heterogeneity that *cannot* be directly observed. Thus, whether these covariates of survival could not be measured or simply were not measured, their absence can lead to biases in both the estimation of the hazard function as well as estimation of parameters and inferences regarding the measured covariates.¹ This chapter addresses the topic of unobserved population heterogeneity in discrete-time survival processes.

3.1 Ignoring unobserved heterogeneity

It may seem peculiar to speak of “ignoring” something that is admittedly *unobserved*. However, when one considers that spurious associations may be inferred and population time-to-event patterns misrepresented when the possibility of unobserved heterogeneity is not considered in the modeling process, it becomes clear that ignoring, or denying the possibility of interdependencies not accounted for by the measured data, could pose serious threats to the internal and well as external validity of a study. The purpose of this section is to highlight the potential problems when one assumes that there is no unobserved heterogeneity.

¹Heterogeneity could also be the results of covariates measured with error.

To begin the discussion, consider the following set of simple illustrations of the impact on various parameter estimates when unobserved heterogeneity is ignored. Suppose the population from which a sample is drawn is made up of two unobserved subpopulations, Group A and Group B, both with constant hazard functions, but one with a higher constant hazard than the other. Group A and Group B both represent 50% of the overall population. The top plot in Figure 3.1 shows the “true” hazard functions for both groups and the estimated hazard probabilities for a model based on a sample of 10,000 that does not include group membership as a covariate. Notice that the unadjusted hazard probabilities show a decrease over time even though the hazard probabilities within each group are constant. Recall that the hazard probability at each time period is a conditional probability—only those individuals surviving to the beginning of the time period are included in the risk set for that period. In the first interval, all of Group A and Group B are at-risk; thus, the estimated hazard probability is a 50/50 mixture of the two groups’ hazards. However, as time goes on, more individuals in Group A than Group B experience the event due to the higher hazard probability and thus the proportion of Group A subjects remaining in the risk set for later time periods decreases, moving the overall unadjusted hazard closer to the Group B hazard probability. The bottom plot of Figure 3.1 shows a similar scenario. In this case, there are

not two distinct groups but rather a continuous random error that is used to characterize the unobserved heterogeneity. The hazard for the population is assumed to be constant but individuals are allowed to vary around the *logit* hazard probability according to a standard normal distribution. As can be seen in the figure, the estimated hazard probabilities, unadjusted for the random variation, are biased downward as in the earlier picture for the same reason: even if initially the population shows random variation around a mean hazard, the proportion of individuals at-risk with variation above the mean, i.e., individuals with greater frailty or susceptibility, will drop across time. Similar results would be obtained with other error distributions on both the *logit* hazard probabilities and the probabilities themselves.

Figure 3.2 introduces an observed binary covariate, defined by membership in Group 1 or Group 2, along with an unobserved binary covariate, defined by membership in Group A or Group B. In the top plot, membership in Group 1 or 2 is independent of membership in Groups A or B. The population distribution for Groups 1 and 2 is 50/50 as is the distribution for Groups A and B. As before, the hazard in each group is constant. The hazard odds ratio for the outcome in Group 1 compared to Group 2 was set at 2.25. The odds ratio was defined to be constant across time with no interaction with Group A/B membership. The figure shows the estimated *logit* hazard

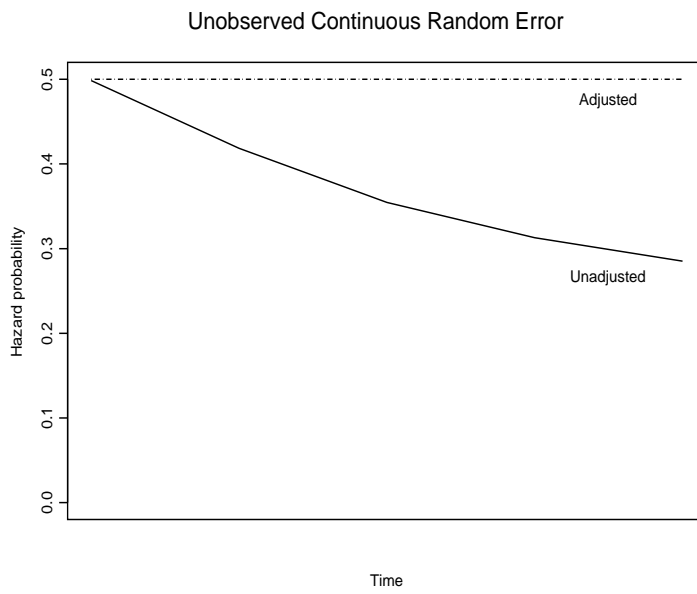
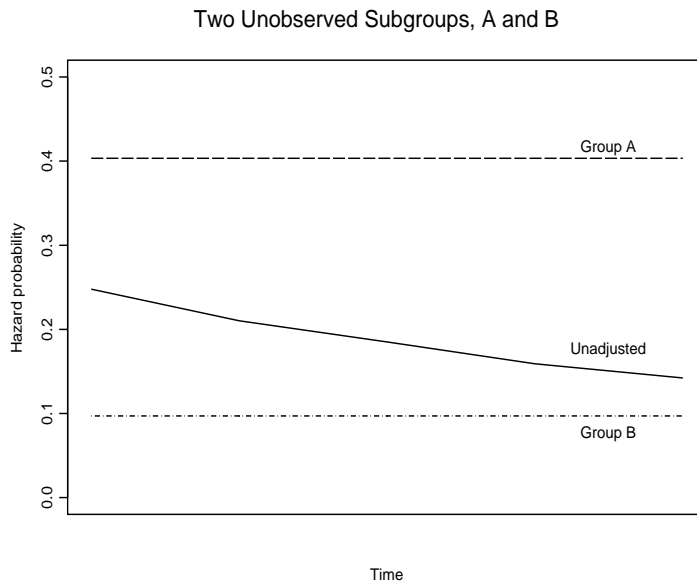


Figure 3.1: Examples of unobserved heterogeneity.

probabilities from the “unadjusted” model that does not include, i.e., ignores, Group A/B membership, and the “adjusted” model that includes an indicator of A/B membership². As can be seen, although the proportionality of the hazard odds for Group 1 versus Group 2 holds³, the value of that odds ratio is clearly underestimated in the unadjusted model. Thus, even though the Group A/B indicator is not a confounder of the Group 1/2 effect, ignoring Group A/B membership can still bias the Group 1/2 estimated effect.

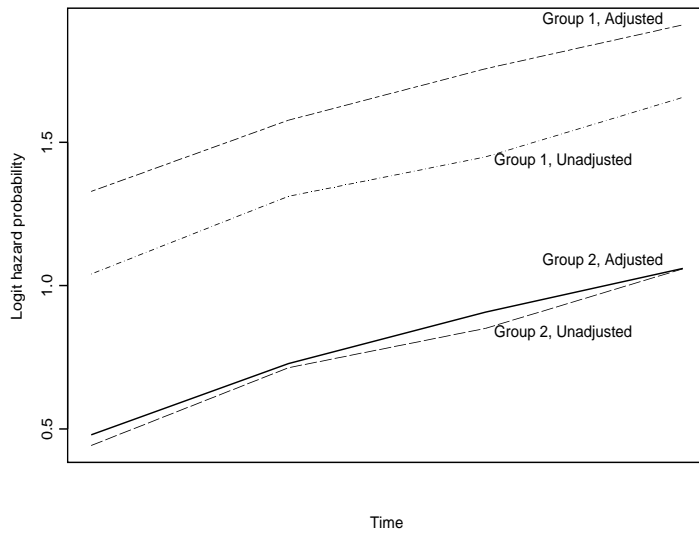
In the bottom plot of Figure 3.2, the Group A/B indicator is defined as a confounder. The population distribution for A/B is still 50/50 but membership in Group 1/2 depends on A/B membership: $P(\text{Group 1} \mid A) = 0.25$ and $P(\text{Group 1} \mid B) = 0.75$. In this case, as shown in the figure, not only is the effect of Group 1/2 membership incorrectly estimated in the unadjusted model, but there is also a spurious time-dependent effect of Group 1/2 membership that would be inferred if Group A/B membership were ignored.

The article by Vaupel, Manton, and Stallard (1979) is typically cited as the first serious treatment of unobserved heterogeneity in survival analysis.

²For the adjusted model, the points plotted for each group are at the sample average of the Group A/B indicator.

³On the *logit* scale, proportionality of the odds is seen by equal vertical distances between the group hazard probabilities over time.

Two Unobserved Subgroups Unassociated with Observed Groups, 1 and 2



Two Unobserved Subgroups Associated with Observed Groups, 1 and 2

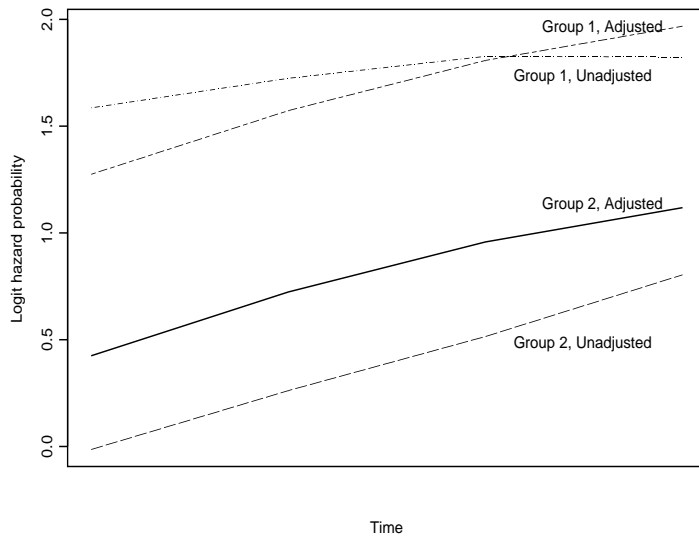


Figure 3.2: Examples of unobserved heterogeneity with an observed covariate.

They use the term *frailty* to refer to individual differences in longevity; hence, the phrase *frailty models* is now used to refer to survival models that in some way attempt to account or adjust for unobserved population heterogeneity. Heckman and Singer (1984a) proved, by application of the Cauchy-Schwartz theorem, that ignoring unobserved covariates of survival will bias the estimated hazards “towards negative duration dependence” (p. 77), meaning that an increasing hazard will appear more slowly increasing while a decreasing hazard will appear more rapidly decreasing. Consider the scenario for the top plot of Figure 3.1 to see how this proposition plays out analytically for that example. Let $P_h(\gamma | g)$ be the hazard probability, conditional on g , an indicator of Group A/B membership, and let $P_{\bar{h}}(\gamma)$ be the marginal hazard probability. What is obtained during estimation ignoring Group A/B membership is the marginal probability since, if Group A/B membership is unobserved, so too is the conditional hazard. The relationship between the conditional probability and the marginal probability is given by

$$P_{\bar{h}}(\gamma) = P_h(\gamma | g = 1) \cdot P_\gamma(g = 1) + P_h(\gamma | g = 0) \cdot P_\gamma(g = 0), \quad (3.1)$$

where $P_\gamma(g = 1)$ is the proportion in Group A at time period γ . What is evident from the figure appears in the equation above. The marginal hazard probability will be equal to the average of the hazard probabilities across the two groups at the first time period but will then move closer to the lesser

hazard probability as the proportion of subjects in the higher hazard group declines.

In summary, ignoring population heterogeneity will bias hazard probabilities (sometimes referred to as duration dependence) downward and underestimate time-independent covariate effects, even if the sources of said heterogeneity are *not* associated with the observed variables. If such sources are correlated with the observed variables, there may be spurious time-dependent effects found for the observed variables. Thus, it is clearly desirable, from an analysis standpoint, to attempt to account for unobserved heterogeneity in modeling time-to-events. The next section discusses approaches to incorporating the possibility of unobserved heterogeneity in model specification.

3.2 Modeling unobserved heterogeneity

Once the alarm had been sounded in the literature about the dangers of ignoring unobserved heterogeneity, the question became how best to incorporate such heterogeneity into models of the survival process. In keeping with the convention of modeling the hazard function in the continuous-time setting, the multiplicative hazard model with unobserved heterogeneity can be expressed as

$$h(t \mid x_t, z, \theta_t) = h_0(t) \cdot \psi(z, x_t) \cdot \varphi(\theta_t), \quad (3.2)$$

where θ_t is a vector of unobserved variables. Similarly, for the discrete-time setting, the discrete-logit model with unobserved heterogeneity can be expressed as

$$\text{logit } P_h(j \mid x_j, z, \theta_j) = \nu_j + \beta_j' z + \kappa_j' x_j + \phi(\theta_j). \quad (3.3)$$

For estimation, the marginal distribution based on the observed variables and duration data must be written in terms of the fully conditional distribution, including the unobserved variables, integrated over the distribution of the unobserved variables. In order to estimate the parameters in this expression, some form for the distribution of the unobserved variables must be specified. In the Vaupel et al. (1979) analysis, the authors assume that frailty is gamma distributed. This is a common distributional assumption in continuous-time models. Heckman and Singer (1984a, 1984b), in addition to proving the resultant biases of ignoring unobserved heterogeneity, also showed that the parameter estimates for the duration dependence and observed covariate effects in continuous-time models are very sensitive to the distributional assumption about θ . Vermunt (1997) raises the possibility that Heckman and Singer's findings regarding this sensitivity were due to misspecification of the duration dependence (for which they used a parametric model). However, Land, Nagin, and McCall (2001) showed that if the gamma distribution for the frailty was wrong, then the variance estimates, and hence inferences surrounding the

covariates, were incorrect, even with no assumption about the underlying duration dependence, i.e., using a semiparametric hazard model. Heckman and Singer suggested a nonparametric approach for modeling unobserved heterogeneity. Instead of a parametric distribution, they recommend modeling the distribution of θ as series of “mass points” or “points of support” with the number, locations, and weights of the support points to be empirically determined in order to avoid the pitfalls of misspecification. This is essentially what was introduced earlier as a *finite mixture model*. Their suggested model can be expressed by

$$\bar{h}(t) = \sum_{k=1}^K h(t) \cdot \pi_k(t) \cdot \theta_k, \quad (3.4)$$

where the number of support points is K , $\pi_k(t)$ is the weight of the support point k , and θ_k is the location of the support point.

Trussel and Richards (1985) showed that even with the Heckman-Singer procedure, results were still sensitive to misspecification of the duration dependence, i.e., the baseline hazard function. This suggests that both the duration dependence and the distribution of the unobserved covariates should modeled nonparametrically or with as few assumptions as possible. Ham and Rea (1987) approached this task in the discrete-time setting by estimating a discrete-logit mixture model with unstructured baseline hazard probabilities. More recently, Vermunt (1997) made the connection between latent class anal-

ysis, log-linear models, discrete-time Markov processes and discrete-time event models, using latent class log-linear models to account for unobserved population heterogeneity. The one limitation of Vermunt's demonstrated models is that in the traditional log-linear framework, all covariates must be categorical, although he does allude to possible extensions with continuous covariates. Land et al. (2001) specify a semiparametric mixed Poisson regression model, equivalent to a discrete-CLL model with nonparametric unobserved heterogeneity. In Vermunt's models, the support points for the distribution of the unobserved covariates are allowed to depend upon the observed covariates; in Land et al., the unobserved covariates are assumed to be independent of the observed covariates. Related to this issue, Trussel and Richards noted that model results could be sensitive to whether or not the support points of the distribution of the unobserved covariates were allowed to depend upon the observed covariates. Not specified in the multiplicative formulation is the possibility of interactions between the observed and unobserved covariates and duration dependence of the unobserved covariates; however, such relationships may exist in the process that generated the data and it is reasonable to assume that there may be similar model sensitivities to ignoring such relationships. The conventional random effects model, where there is a random "error" or "frailty" term added to the linear expression of covariates does not allow for

these more complex relationships—the concept of frailty as specified in a random effects model is an unobserved time-independent shift of the baseline hazard for each individual.

As previously noted, the approach of nonparametric specification of the distribution of unobserved covariates using support points is essentially a mixture model formulation. Equivalently, these support points could be reformulated as categories or classes of a latent class variable, where the number of support points is the number of latent classes, the weights are the class proportion in the samples, and the locations are given by the class-specific shifts in the logit baseline hazard probabilities. Allowing for more complex heterogeneity, there could be, in the mixture setting, class-specific parameter estimates for all the hazard probabilities as well as the covariate effects. By specifying the discrete-logit model in the latent class regression framework, the model readily extends to a mixture model allowing for unobserved heterogeneity by specifying $K > 1$. The observed data likelihood for a single subject i is then given by

$$\begin{aligned}
 L_i &= \sum_{k=1}^K \left(P_C(k | z_i) \cdot [P_h(a_i | C = k, x_i, z_i)]^{\delta_i} \prod_{j=1}^{a_i-1} [1 - P_h(j | C = k, x_{ij}, z_i)] \right) \\
 &= \sum_{k=1}^K \left(P_C(k | z_i) \prod_{j \in \{r: R_{ri}^o = 1\}} P(E_j = e_{ji} | C = k, x_{ij}, z_i) \right). \tag{3.5}
 \end{aligned}$$

Notice that the likelihood for a K -class mixture LCR discrete-time model can again be put in terms of the event and observed risk indicators as defined in Chapter 2. For the discrete-logit mixture model, the relationship between C and the time-independent covariates is specified as a multinomial logistic regression given by

$$P_C(k | z) = \frac{\exp(\alpha_{0k} + \alpha'_k z)}{\sum_{m=1}^K \exp(\alpha_{0m} + \alpha'_m z)}, \quad (3.6)$$

where $\alpha_{0K} = 0$ and $\alpha_K = 0$ for the reference class, K . α_{0k} is the log odds ratio for being in class k versus class K given membership in one of the two. α_{kp} is the log odds ratio for being in class k versus class K for a one unit increase in z_p , controlling for all other covariates in the regression.

As before, the relationship between the event indicators and the covariates, now including C , is specified as a logistic regression with class-specific parameters given by

$$P_{E_j}(E_j = 1 | C = k, x_j, z) = \frac{1}{1 + \exp(-(\nu_{jk} + \beta'_{jk} z + \kappa'_{jk} x_j))}. \quad (3.7)$$

ν_{jk} is the logit of the baseline hazard probability at time period j for class k ; β_{jkp} is the log hazard odds ratio for a one unit increase in z_p at time period j for class k , controlling for all other covariates included in the regression; and κ_{jkp} is the log hazard odds ratio for a one unit increase in x_p at time period j for class k , controlling for all other covariates included in the regression.

The maximum-likelihood estimation uses an EM algorithm (see Muthén & Shedden, 1999), where data on C are considered missing.

As with other mixture model settings, such as growth mixture modeling, there will always remain a question of whether these latent classes, or points of support, represent “true” homogeneous subgroupings within the populations or whether such a characterization is simply a reification of an analytically convenient and empirically driven specification. Trussel and Richards (1985) showed that it was, practically speaking, impossible to empirically distinguish between hazard functions that were actually decreasing and represented the whole of the population and hazard functions that were decreasing due to a mix of high- and low-risk unobserved subpopulations. Thus, it is left up to substantive researchers and theorists to validate interpretations of such mixture models and guard against reification.

The most common use of this nonparametric approach to modeling unobserved heterogeneity in the survival literature is a specialized mixture model with two points of support, $k = 1, 2$, and $\theta_2 = 0$, that conceives of these two support points as representing two characteristically different groups of individuals in the population. This model is has several names including the *mover-stayer* model, the *long-term survivor* model, or the *cure-rate* model. The heterogeneity of survival in the population is characterized by two groups:

1) the long-term survivor group that has a hazard of zero for the entirety of the observation period, and 2) the non-long-term survivors, who are at non-zero risk for the event during the period of observation. Thus, the mixture model is specified with $K = 2$ and $P_E(e | C = k, z, x) = 0$ for the LTS class. If $C = 2$ designates a long-term survivor, then the marginal probability, $P_C(C = 2)$, represents the proportion of the sample that has zero risk for the event. Not only is this model a more restricted version of the general model for hidden heterogeneity presented above, since the duration dependence and covariate effects are constrained for one class, but the value of C is observed for a portion of the sample—all subjects who experience the event during the period of observation are known not to belong to the long-term survivor class. Thus, data on C are only partially missing.⁴ This model has been applied in both the continuous-time setting (see Maller and Zhou, 1996, for an extensive discussion) and the discrete-time setting (see, for example, Steele, 2003).

One caution about interpretation is necessary for models in which the duration dependence is also unstructured: the idea of membership in a “risk-free” class should not be extrapolated beyond the period of observation. That is to say, there is no information from the data, as used in the model, to infer risk status beyond the observation period—the risk for the long-term

⁴This is handled in Mplus by the “training data” feature.

survivor class is constrained to zero for the observation period; however, there is nothing in the model specification that implies a zero risk beyond the last observation period. This is not the case with continuous-time models with parametric specification of duration dependence—the functional form of the risk is defined on a time scale that stretches to infinity; thus, a risk fixed at zero for such a model implies a risk-free status for all time, rather than simply during observation. The next section addresses the issue of *identifiability*, which is to say, under what conditions it is actually possible to include some representation of individual frailty or susceptibility in the model and still estimate the parameters of interest.

3.3 Identifiability

As with any latent variable model, it is important to address the issue of identifiability. Heuristically, the essential question is: Is there enough information in the data to uniquely estimate all the specified parameters of the model? For example, in the simple linear regression model, it is well known that you cannot estimate more regression parameters than the number of observations in the data set. More generally, the number of parameters to be estimated cannot exceed the number of pieces of unique information in the data set, related to and including the outcome of interest. In a conventional latent factor model,

the pieces of information are the unique elements are the variances covariances of the observed data. In a latent class analysis with no covariates, the pieces of information are the frequency counts for each pattern of the binary indicators. In a LCA with J indicators, there are 2^J possible response patterns. For the discrete-logit in the LCA framework, with no covariates, there are J unique pieces of information: There are $J + 1$ frequency counts for response patterns, corresponding to (1) the number of individuals experiencing the event in time period 1; (2) the number of individuals experiencing the event in time period 2; ... (J) the number of individuals experiencing the event in time period J ; and (J+1) the number of individuals not experiencing the event in time periods $1, \dots, J$; and then the restraint that the frequency counts across the $J + 1$ patterns must sum to the total sample size n yields $J + 1 - 1 = J$ *degrees of freedom*. This enumeration is useful but not complete in assessing the identifiability of a model. For example, Goodman (1974) showed that an unrestricted three-class model (14 parameters) with four binary indicators (15 degrees of freedom) is not identified. Thus, it is clear that having more degrees of freedom than parameters to be estimated is necessary but not sufficient for model identification. In the one-class unstructured discrete-logit with no observed covariates, there are J parameters and J degrees of freedom. This means that without additional information in the form of covariates or constraints placed

on the hazard probabilities, the discrete-logit mixture model is not identified. This leads to the questions of what is necessary for model identification when incorporating unobserved heterogeneity. For discrete-logit models with unobserved heterogeneity specified as latent classes, it necessary to consider conditions of identifiability for latent class regression models as well as conditions of identifiability for hazard models with unobserved heterogeneity.

To begin, take the most general definition of identifiability for mixture distributions as given by McLachlan and Peel (2000, p. 27): Let

$$f(y_i; \Psi) = \sum_{k=1}^K \pi_k f_k(y_i; \theta_k) \quad (3.8)$$

and

$$f(y_i; \Psi^*) = \sum_{k=1}^{K^*} \pi_k^* f_k(y_i; \theta_k^*) \quad (3.9)$$

be any two members of a family of mixture densities. This class of finite mixtures is said to be identifiable for $\Psi \in \Omega$ if

$$f(y_i; \Psi) \equiv f(y_i; \Psi^*) \quad (3.10)$$

if and only if $K = K^*$ and the component labels can be permuted so that

$$\pi_k = \pi_k^* \text{ and } f_k(y_i; \theta_k) = f_k(y_i; \theta_k^*) \text{ } (k = 1, \dots, K). \quad (3.11)$$

As noted by Bandeen-Roche, Miglioretti, Zeger, and Rathouz (1997) as well as Huang and Bandeen-Roche (in press), identifiability for latent class

and latent regression models has traditionally focused on *local* identifiability. The definition of local identifiability resembles the definition given above with the less global requirement that the conditions hold true for all $\Psi \in \tau$ where $\tau \subset \Omega$. McHugh (1956), Goodman (1974), and Formann (1992) all discuss that identifiability of the latent class model reduces to ensuring that the information matrix is positive definite (see Vermunt, 1997, pp. 316–318). Although this has been shown to hold true for LCR models as well, the Jacobian can have intractably large row-dimensions for verifying full column rank when continuous covariates are included. Huang and Bandeen-Roche developed a method for checking identifiability of the LCR model with covariates effects on both the latent class indicators as well as on the latent class probabilities. To summarize their results heuristically, the following conditions must hold:

1. The number of unique model parameters cannot exceed the number of independent pieces of observed information;
2. The covariate effects on the indicator probabilities are finite;
3. The covariate effects the latent class distribution are finite;
4. All covariate values are finite;
5. The probability distributions for the possible response patterns, conditional on each covariate pattern, are linearly independent; and

6. The design matrix of all covariates influencing the indicators directly and the design matrix of all covariates influencing the latent class variable distribution both have full column rank.

In the model specified by these authors, the covariate effects on the latent class indicators are constrained to be equal across the latent classes.

Although these conditions could be applied to the special case of LCR with a limited number of response patterns that corresponds to the discrete logistic model specification, it is useful to also examine the literature on identifiability of hazard models with unobserved heterogeneity. Elbers and Ridder (1982) proved three conditions which, if fulfilled, ensured the identifiability of the hazard model. Heckman and Singer (1984a) proved another condition that could be used in lieu of one the original conditions set by Elbers and Ridder. Van de Pol and Langeheine (1990) discussed identification for discrete-time mixed Markov models. Vermunt (1997) summarizes across all of these results that the parameters of a single event discrete-time hazard model with

unobserved heterogeneity are assured to be identifiable if *at least one* of the following three conditions hold:⁵

1. The model is a proportional hazard model;
2. The duration dependence is structured; or
3. The mixing distribution is parameterized. (p.201)

It is possible to combine these conditions, taking into account a wider range of model specification, including the potential for class-varying effects of covariates on the indicators and time-dependent effects of the unobserved heterogeneity, i.e., the “location” of the points of support are time-dependent. For example, assuming the conditions given by Huang and Bandeen-Roche are also met, the following conditions are necessary and sufficient for identification with $K = 2$:

- Without any observed covariates, the duration dependence must be structured. and some restrictions on the latent class effects may be required.

⁵Vermunt actually gives four conditions, the last one being that identifiability is ensured if the model is a multivariate hazard model, e.g., recurrent events, clusters observations, etc. This chapter only addresses unobserved heterogeneity for single event models with no observation clustering. For more on random effects models in clustered discrete-time settings, see for example, Hedeker, Siddiqui, and Hu (in press) or Steele (2003).

- With at least one observed covariate, z , the following models, with structured duration dependence for the hazard probabilities and $K = 2$ latent classes representing the nonparametric distribution of unobserved covariates, are identified:

1. The model with the latent class distribution dependent upon z , and no effects from z to the event indicators. The model specification is given by

$$P_C(k = 1 | z) = \frac{1}{1 + \exp(-(\alpha_0 + \alpha'z))} \text{ and}$$

$$P_{E_j}(E_j = 1 | C = k, z) = \frac{1}{1 + \exp(-(\nu_{kj}))}.$$

2. The model with the latent class distribution independent of z and the effects of z on the event indicators allowed to be time-dependent but constrained equal across the latent classes, given by

$$P_C(k = 1 | z) = \frac{1}{\exp(-\alpha_0)} \text{ and}$$

$$P_{E_j}(E_j = 1 | C = k, z) = \frac{1}{1 + \exp(-(\nu_{kj} + \beta'_j z))}.$$

3. The model with the latent class distribution dependent upon z and the effects of z on the event indicators constrained to be time-independent and equal across the latent classes, given by

$$P_C(k = 1 | z) = \frac{1}{\exp(-(\alpha_0 + \alpha'z))} \text{ and}$$

$$P_{E_j}(E_j = 1 | C = k, z) = \frac{1}{1 + \exp(-(\nu_{kj} + \beta'_j z))}.$$

4. The model with the latent class distribution dependent upon z and the effects of z on the event indicators allowed to be time-dependent and class-dependent while the baseline hazard probabilities are freely estimated across the time period but constrained to have proportional odds across the latent classes, given by

$$P_C(k = 1 | z) = \frac{1}{\exp(-(\alpha_0 + \alpha'z))} \text{ and}$$

$$P_{E_j}(E_j = 1 | C = k, z) = \frac{1}{1 + \exp(-(\nu_j + \delta_k + \beta'_{kj}z))}.$$

Model specification is not limited to the above models. The above models are identified but many other identifiable specifications are possible with the addition of other observed covariates. Identification can also be complicated by the addition of time-dependent covariates. Furthermore, it is important to note that although the conditions may “ensure” theoretic identification, the model may not be empirically identifiable. The easiest way to think of this is that information matrix may be so empirically *near* non-positive defi-

nite that whatever software is being used for estimation rejects the model as unidentified.⁶

All of these conditions presuppose a given number of classes, K . This leads to the ubiquitous issue in all mixture modeling settings which is determining the number of classes to specify. The next session address this problem of class enumeration.

3.4 Class enumeration

Analogous to choosing the number of factors in an traditional factor analysis, selecting the number of classes in a mixture analysis is also analytically challenging. Without prior knowledge or substantive theory to inform the class enumeration, it is left up to the analyst to choose the “appropriate” number of classes. The likelihood ratio test that is normally used to compare nested

⁶Related to identification is nonconvergence in estimation which may or may not signal identification problems. Mixture models are notorious for their sensitivity to starting values and estimation algorithms arriving at local rather than global optima or not converging at all because of poor starting values. Therefore, it is recommended that a series of alternate starting values are tried when fitting any given model. The forthcoming Mplus, Version 3, will have the feature of automatic random perturbations of the starting values, with the number of random starts specified by the user.

models cannot be applied to compare $K = k$ to a $K < k$ model because regularity conditions are not met, i.e., the solution is on the boundary of the parameter space. This is easier to understand by considering a model with the null hypothesis of $K = k$ and an alternative of $K = k + 1$. There are two ways to restrict the alternative model to obtain the null: 1) set one of the class probabilities to zero (a value on the boundary of the permissible parameter space for the class probability), or 2) constrain all the class-specific parameters in two of the classes to be equal across the two classes (resulting in a non-positive definite information matrix). A frequent practice for LCA modeling is simply to compare each estimated model to a saturated model (Goodman, 1974; McCutcheon, 1987; Formann, 1992) using the Pearson χ^2 and choose K to be the lowest number that yields an acceptable fit (Bandein-Roche et al., 1997). However, since it is the likely case for discrete-logit models that covariates are included to even allow identifiability of a $K > 1$ model, this technique cannot be used—the χ^2 distribution is not asymptotically valid for the Pearson χ^2 statistic usually applied for LCA goodness-of-fit for LCR models that include any continuous covariates. This is as true for regular logistic regressions with continuous covariates. To understand this better, consider that the Pearson χ^2 statistic is based on observed versus expected frequencies in each of the “cells” delimited by the categories of the covariates and the outcome variables. The

distribution of the statistic is based on “n-asymptotics”, that is, as the sample size becomes infinitely large, the expected counts within each cell also become large. This is why the Pearson χ^2 does not perform well, i.e., the asymptotics do not hold, when a small sample size results in low expected cell counts. However, with a continuous covariate, there is essentially a covariate “category” for each individual in sample and as the sample become infinitely large, so does the number of cells, meaning that the expected cell counts never approach sizes for which the asymptotics will hold. Bandeen-Roche et al. (1997) propose a method using *pseudo-classes* to assess the adequacy of the assumptions of the LCR model, thus allowing a qualitative comparison across models with different numbers of classes. However, their method does not allow for direct covariate effects to the latent class indicators. Huang and Bandeen-Roche (in press) extend this technique to allow for such effects. In addition, Huang and Bandeen-Roche offer a new test statistic for LCR goodness-of-fit with direct and indirect covariate effects. They term the goodness-of-fit statistic G^2 and prove that it converges to a χ^2 distribution. In the continuous mixture modeling literature, there have been some advances in latent class enumeration, including the empirical likelihood ratio test, present by Lo, Mendell, and Rubin (2001) based on early work by Vuong (1989). Preliminary simulation studies show this test to have promise for the continuous multivariate setting

as well as the univariate setting for which it was developed. However, this test has not been adequately explored for mixtures of categorical outcomes and will not be presented here. Entropy-based information heuristics are popular in the economic literature for mixture models but these are not clearly applicable nor have they been well-tested in the LCR setting. Because of the lack of consensus about model-building for latent class regression models, many authors simply examine the change in the likelihood as they increment the number of classes (for example, Ham and Rea, 1987). Land et al. (2001) use the change in the log likelihood as well as the Akaike Information Criterion (AIC) and the BIC to judge the “optimal” number of classes. Vermunt (1997) only cites Laird (1978) in offering a strategy for fitting latent class models. Some authors simply specify the number of classes a priori, such as using a long-term survivor model (for example, Steele, 2003; Farewell, 1982). For the purposes of this dissertation, only the AIC, BIC, and the G^2 will be presented. The next paragraph describes their calculation.

The Akaike Information Criterion (AIC), developed by Akaike (1973), is a likelihood-based criterion that essentially penalizes for the number of parameters, trying to gage the automatic improvement in the likelihood from the added parameters that come from increasing the number of classes against the loss in parsimony with the estimation of additional parameters. The AIC is

defined by

$$\text{AIC} = -2LL + 2r, \quad (3.12)$$

where LL is the log likelihood value at the conclusion of the estimation, r is the number of free model parameters. The Bayesian Information Criterion (BIC), developed by Schwartz (1978), includes a penalty that involves both the number of parameters and the sample size. The BIC is “better” than the AIC in that it is statistically consistent by including n . The BIC is defined by

$$\text{BIC} = -2LL + r \log n, \quad (3.13)$$

where n is the sample size. For both the AIC and the BIC, the smaller the number, the better the model. The calculation of the G^2 statistic is a little more involved. In a sample with J periods, there are $J + 1$ possible response patterns corresponding to

- 1) Event occurs in period 1,
- 2) Event occurs in period 2,
- ⋮
- J) Event occurs in period J ,
- $J + 1$) Event does not occur during any of the J time periods.

The formulations of the statistic does not allow for missingness on the outcome which, in this setting, means that the only censoring time permitted is at the conclusion of the observation period. Define a random vector

$Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{iJ})'$, where $Y_{ij} = I[e_{ij} = 1]$. Let $\pi_{ij} = P(Y_{ij} = 1)$. The estimated marginal probabilities of each response pattern can be calculated as follows:

$$\hat{\pi}_{ij} = \sum_{k=1}^K \left(\left[\prod_{m=1}^{j-1} (1 - \hat{P}_h(m | k, x_{mi}, z)) \right] \cdot \hat{P}_h(j | k, x_{ji}, z) \cdot \hat{P}_C(k | z) \right). \quad (3.14)$$

A vector of differences between the observed and estimate values on the Y_{ij} 's is given by

$$\hat{S}_i = Y_i - \hat{\pi}_i, \quad (3.15)$$

where $\hat{\pi}_i = (\hat{\pi}_{i1}, \dots, \hat{\pi}_{ij})'$. The estimated variance matrix for Y_i is given by

$$\widehat{V}_i = \begin{bmatrix} \hat{\pi}_{i1}(1 - \hat{\pi}_{i1}) & -\hat{\pi}_{i1}\hat{\pi}_{i2} & \cdots & -\hat{\pi}_{i1}\hat{\pi}_{iJ} \\ -\hat{\pi}_{i2}\hat{\pi}_{i1} & \hat{\pi}_{i2}(1 - \hat{\pi}_{i2}) & \cdots & -\hat{\pi}_{i2}\hat{\pi}_{iJ} \\ \vdots & \vdots & & \vdots \\ -\hat{\pi}_{iJ}\hat{\pi}_{i1} & -\hat{\pi}_{iJ}\hat{\pi}_{i2} & \cdots & \hat{\pi}_{iJ}(1 - \hat{\pi}_{iJ}) \end{bmatrix} \quad (3.16)$$

The G^2 statistic of Huang and Bandeen-Roche is then defined as

$$G^2 = \hat{S}'_N \hat{\Sigma}_N^{-1} \hat{S}_N \sim \chi^2_J, \quad (3.17)$$

where

$$\hat{S}_N = \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{S}_i, \quad (3.18)$$

and

$$\hat{\Sigma}_N = \frac{1}{n} \sum_{i=1}^n \widehat{V}_i. \quad (3.19)$$

Huang and Bandeen-Roche prove that $\widehat{S}'_N \Sigma^{-1} \widehat{S}_N \xrightarrow{\mathcal{L}} \chi^2_J$ under the conditions that the parameters and covariates are all finite, the pattern probabilities are all nonzero, Σ_N is positive definite, and that $\widehat{\pi}_i$ converges in probability to π_i . The use of G^2 follows that of the conventional goodness-of-fit test for the LCA in that the model selected is the one with the smallest number of classes that has a non-significant p-value.

The remainder of this section will examine various strategies for class enumeration in the absence of outside supporting knowledge. Assume that there is at least one measured covariate. The first matter to address is the model specification within and across classes. What sort of model should be fit when incrementing class number and comparing across models? Should one use the most general model, allowing direct and indirect effects of covariates and as much flexibility in the direct covariate effects across time periods and latent classes as possible while still maintaining identifiability? Should the baseline hazard odds across classes be constrained to be proportional? How sensitive are the information heuristics to these possible variations in model specification? This discussion begins with the matter of allowing covariates to influence the latent class distribution since this is an appealing specification in regards to achieving model identification.

Vermunt (1997) as well as Trussell and Richards (1985) warn of the sensitivity of results to whether or not covariates are allowed to influence the latent class distribution. To understand this sensitivity, consider an simple example with five time periods and two covariates, x_1 and x_2 . A sample of $n = 10,000$ was generated according the following Population A:

$$P_{E_j} = \frac{1}{1 + \exp(-(-2 + x_1 + 2x_2))}, \quad \forall j = 1, \dots, 5, \quad (3.20)$$

where $x_1 \sim N(0, 1)$ and $x_2 \sim Bernoulli(p = 0.5)$. For this example, only x_1 was considered observed; thus, x_2 is unobserved and it's distribution can be appropriately represented by a $K = 2$ latent class variable. Four different 2-class models were fit to the simulated sample data. Table 3.1 displays the specifications of the four models. Figure 3.3 displays the path diagrams for the four models.

All four model specifications are identified. Note that Model 0 does not include any effect of the observed covariate; Model 1 allows x_1 to have an effect on the distribution of the latent classes but not the event indicators (indirect); Model 2 allows x_1 to influence the event indicators (direct) but constrains the effects to be time-independent and class-independent; Model 3 allows both direct and indirect effects but constrains the direct effects to be both time-independent and class-independent. Note also that the hazard is constrained to be constant across time periods within each class for all four

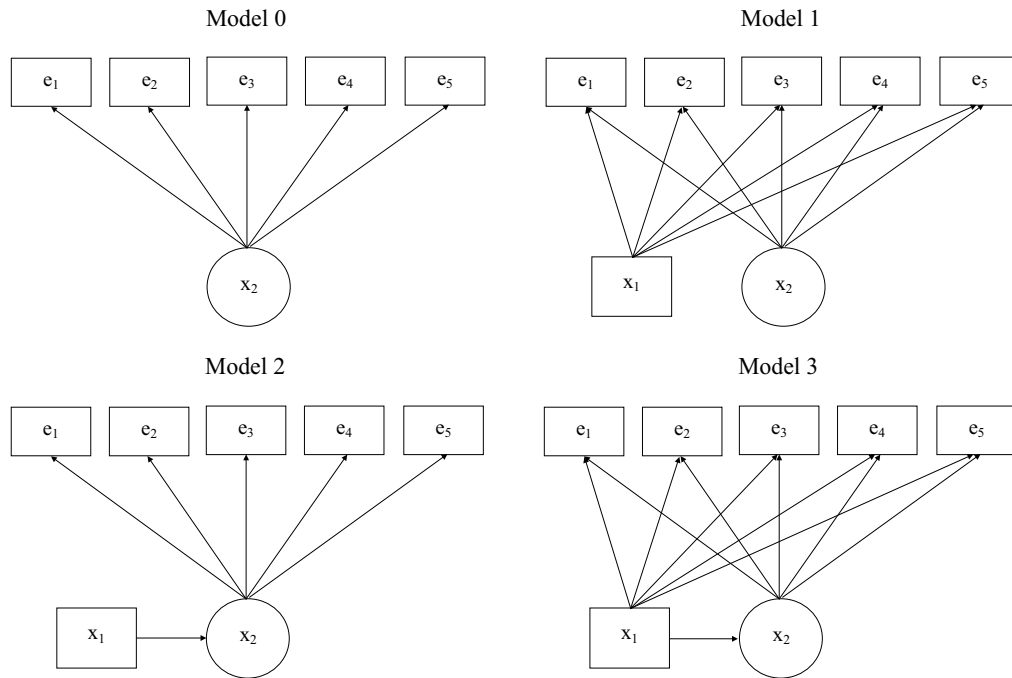


Figure 3.3: Path diagrams for class enumeration Models 0–3.

Table 3.1: Class Enumeration Models 0–3 Specification

Regression	$P_{E_j}(E_j = 1 \mid C = k, x_1)$	$P_C(k \mid x_1)$
Model 0	$\frac{1}{1+\exp(-(\nu_k))}$	$\frac{\exp(\alpha_{0k})}{\sum_{m=1}^K \exp(\alpha_{0m})}$
Model 1	$\frac{1}{1+\exp(-(\nu_k + \beta' x_1))}$	$\frac{\exp(\alpha_{0k})}{\sum_{m=1}^K \exp(\alpha_{0m})}$
Model 2	$\frac{1}{1+\exp(-(\nu_k))}$	$\frac{\exp(\alpha_{0k} + \alpha'_{1k} x_1)}{\sum_{m=1}^K \exp(\alpha_{0m} + \alpha_{1m} x_1)}$
Model 3	$\frac{1}{1+\exp(-(\nu_k + \beta' x_1))}$	$\frac{\exp(\alpha_{0k} + \alpha'_{1k} x_1)}{\sum_{m=1}^K \exp(\alpha_{0m} + \alpha_{1m} x_1)}$

model specifications. For Population A, Model 1 is the correctly specified model. The results of the four models fit to the simulated sample from population A are given in Table 3.3. For comparisons sake, consider the same four models fit to data drawn from Populations B and C given in Table 3.2. For Population B, Model 2 is the correctly specified model. The results of the four estimated models are given in Table 3.4. For Population C, Model 3 is the

Table 3.2: Class Enumeration Populations A–C Definition

Regression	$P_{E_j}(E_j = 1 \mid x_1, x_2)$	$P(X_2 = 1 \mid x_1)$
Pop ⁿ A	$\frac{1}{1+\exp(-(-2+x_1+2x_2))}$	0.5
Pop ⁿ B	$\frac{1}{1+\exp(-(-2+2x_2))}$	$\frac{1}{1+\exp(-x_1)}$
Pop ⁿ C	$\frac{1}{1+\exp(-(-2+x_1+2x_2))}$	$\frac{1}{1+\exp(-x_1)}$

$$x_1 \sim N(0, 1), j = 1, \dots, 5$$

correctly specified model. The results of the four estimated models are given in Table 3.5.

Table 3.3 demonstrates that in the Population A example, when the observed covariate, x_1 , has a direct effect on the event indicators and no relation to the unobserved variable, x_2 , but the model fit to the data is misspecified in that it does not allow a direct effect from x_1 to the covariates, as in Models 0 and 2, the estimated effect of the unobserved covariate, given in the table by $\nu_2 - \nu_1$, is overestimated. Also, in Model 2 where only an indirect effect is allowed through a regression of x_2 on x_1 , one would incorrectly infer that x_2 and x_1 were strongly associated. Using a likelihood ratio test to compare

Table 3.3: Class Enumeration Models 0–3 Results for Population A

Results	True Value	Model 0	Model 1	Model 2	Model 3
LL	–	-15,761.38	-14,635.71	-14,741.00	-14,634.34
# parameters	–	3	4	4	5
ν_1	0.00	0.43	-0.09	0.26	-0.10
ν_2	-2.00	-1.92	-2.08	-2.18	-2.13
$\nu_2 - \nu_1$	-2.00	-2.35	-1.99	-2.44	-2.03
β	1.00	@ 0	1.00*	@ 0	1.11*
α_{01}	0.00	-0.40	0.12	-0.17	0.17
α_{11}	0.00	@ 0	@ 0	1.83*	-0.21
$P_C(k = 1)$	0.50	0.40	0.53	0.47	0.54

* $p < 0.05$;
 @ = “fixed at”

Table 3.4: Class Enumeration Models 0–3 Results for Population B

Results	True Value	Model 0	Model 1	Model 2	Model 3
LL	–	-16,048.54	-15,767.68	-15,763.15	-15,763.08
# parameters	–	3	4	4	5
ν_1	0.00	0.08	0.17	0.09	0.08
ν_2	-2.00	-1.96	-1.80	-1.94	-1.93
$\nu_2 - \nu_1$	-2.00	-2.04	-1.97	-2.03	-2.01
β	0.00	@ 0	0.46*	@ 0	0.02
α_{01}	0.00	-0.06	-0.27	-0.09	-0.09
α_{11}	1.00	@ 0	@ 0	0.90*	0.87*
$P_C(k = 1)$	0.50	0.48	0.43	0.48	0.48

* $p < 0.05$;
 @ = “fixed at”

Table 3.5: Class Enumeration Models 0–3 Results for Population C

Results	True Value	Model 0	Model 1	Model 2	Model 3
LL	–	-15,034.46	-12,937.67	-13,199.94	-12,926.43
# parameters	–	3	4	4	5
ν_1	0.00	0.51	-0.25	0.37	-0.13
ν_2	-2.00	-2.31	-2.30	-2.57	-2.18
$\nu_2 - \nu_1$	-2.00	-2.82	-2.05	-2.94	-2.05
β	1.00	@ 0	1.52*	@ 0	1.03*
α_{01}	0.00	-0.14	0.46*	0.19*	0.25
α_{11}	1.00	@ 0	@ 0	2.47*	0.98*
$P_C(k = 1)$	0.50	0.46	0.61	0.52	0.55

* $p < 0.05$;
 @ = “fixed at”

Table 3.6: Class Enumeration Measures for Population A

Results	# of classes	Model 0	Model 1	Model 2	Model 3
LL	1-class	-16,163.78	-14,837.77	-16,163.78	-14,837.77
	2-class	-15,761.38	-14,635.71	-14,741.00	-14,634.34
	3-class	-15,761.07	NID	-14,650.23	-14,632.81
# parameters	1-class	1	2	1	2
	2-class	3	4	4	5
	3-class	5	6	7	8
AIC	1-class	32,329.55	29,679.54	32,329.55	29,679.54
	2-class	31,528.75	29,279.43	29,490.00	29,278.68
	3-class	31,532.15	–	29,314.46	29,281.63
BIC	1-class	32,336.76	29,693.96	32,336.76	29,693.96
	2-class	31,550.38	29,308.27	29,518.84	29,314.73
	3-class	31,568.20	–	29,364.93	29,339.31
$G_5^2, (p)$	1-class	787.07 (<0.01)	283.54 (<0.01)	787.07 (<0.01)	283.54 (<0.01)
	2-class	0.60 (0.99)	0.61 (0.98)	8.01 (0.16)	0.22 (1.00)
	3-class	0.00 (1.00)	–	0.54 (0.99)	0.43 (0.99)

Table 3.7: Class Enumeration Measures for Population B

Results	# of classes	Model 0	Model 1	Model 2	Model 3
LL	1-class	-16,330.64	-15,992.14	-16,330.64	-15,992.14
	2-class	-16,048.54	-15,767.68	-15,763.15	-15,763.08
	3-class	NID	NID	-15,761.00	-15,760.64
# parameters	1-class	1	2	1	2
	2-class	3	4	4	5
	3-class	5	6	7	8
AIC	1-class	32,663.28	31,988.27	32,663.28	31,988.27
	2-class	32,103.08	31,543.35	31,534.30	31,536.16
	3-class	–	–	31,536.00	31,537.28
BIC	1-class	32,670.49	32,002.69	32,670.49	32,002.69
	2-class	32,124.71	31,572.19	31,563.14	31,572.21
	3-class	–	–	31,586.48	31,594.96
$G_5^2, (p)$	1-class	544.83 (<0.01)	413.93 (<0.01)	544.83 (<0.01)	413.93 (<0.01)
	2-class	1.26 (0.94)	4.52 (0.48)	1.37 (0.93)	2.44 (0.79)
	3-class	–	–	2.37 (0.80)	2.55 (0.77)

Table 3.8: Class Enumeration Measures for Population C

Results	# of classes	Model 0	Model 1	Model 2	Model 3
LL	1-class	-15,779.00	-13,142.43	-15,779.00	-13,142.43
	2-class	-15,034.46	-12,937.67	-13,199.94	-12,926.43
	3-class	-15,031.83	-12,928.13	-12,946.54	-12,925.13
# parameters	1-class	1	2	1	2
	2-class	3	4	4	5
	3-class	5	6	7	8
AIC	1-class	31,560.00	26,288.86	31,560.00	26,288.86
	2-class	30,074.93	25,883.35	26,407.88	25,862.85
	3-class	30,073.66	25,868.26	25,907.09	25,866.27
BIC	1-class	31,567.21	26,303.28	31,567.21	26,303.28
	2-class	30,096.56	25,912.19	26,436.72	25,898.90
	3-class	30,109.72	25,911.52	25,957.56	25,923.95
$G^2_5, (p)$	1-class	1438.60 (<0.01)	274.74 (<0.01)	1438.60 (<0.01)	274.74 (<0.01)
	2-class	5.36 (0.37)	10.25 (0.07)	16.47 (0.01)	1.82 (0.87)
	3-class	0.09 (1.00)	8.16 (0.15)	5.40 (0.37)	0.72 (0.98)

Model 2 to another model (not shown) relaxing the constant baseline hazard within class constraint, one would incorrectly infer strong evidence *against* the null model with a constant hazard. Thus, misspecification in this case, by not allowing a direct effect when one is present, results in biased estimates of the effect of the unobserved covariate, the relationship between the observed and unobserved covariate, and the baseline hazard probabilities.

Table 3.4 demonstrates that in the Population B example, when the observed covariate, x_1 , has only an indirect effect on the event indicators through the unobserved covariate, x_2 , but the model fit to the data is misspecified in that it only allows a direct effect from x_1 to the event indicators, the mean of the unobserved covariate, given by $P_C(k = 1)$ is slightly underestimated, but the effect of x_2 on the event indicators is unaffected. By not allowing an indirect effect from x_1 to x_2 , one would incorrectly infer that x_1 has a significant association with the event indicators within each class.

Table 3.5 demonstrates that in the Population C example, when the observed covariate, x_1 , has both direct and indirect effects on the event indicators, that allowing only an indirect effect overestimates the relationship between x_1 and x_2 as well as the effect of x_2 on the hazard probabilities, while allowing only a direct effect overestimates the effect of x_1 on the hazard probabilities and overestimates the mean of x_2 . Also, only allowing an indirect

effect leads to an incorrect inference, rejecting the constant baseline hazard model.

In all three population examples, when comparing Models 1 and 2 to Model 3, which allows both direct and indirect effects for x_1 , using the likelihood ratio χ^2 test, the correct specification can be selected from among the three models, since both Models 1 and 2 are nested within Model 3. This suggests that for a given number of classes, in this example, $K = 2$, the least restrictive model, allowing both direct and indirect observed covariate effects, should be fit and then tested against more restricted models. If there are problems in terms of available data and identification, it would seem prudent to specify first a model with only direct effects for the observed covariates rather than one with only indirect effects given the resultant biases evidenced in the above examples. That is to say, ignoring indirect effects may be the “lesser of two evils” in regards to model misspecification in this survival context.

Tables 3.6–3.8 display the computed log likelihood, AIC, BIC, and G^2 values from 1-, 2-, and 3-class model runs under each of the four specifications for Populations A, B, and C.⁷ The bolded value represent the k-class model under each of the four specifications that was chosen by each selection criterion.

⁷“NID” in the tables refers to models that are not identified. This is a good example of a case of empirical non-identification that was previously mentioned.

For Population A, both the AIC and BIC incorrectly select the 3-class model under Model 2 specification. Recall the Model 2 only estimates an indirect effect for x_1 while the sample has been drawn from a population with only a direct effect from x_1 directly to the covariates. Although the G^2 is non-significant for the 2-class Model 2, the p-value is much smaller than it is for the other three models. One may suspect that with a smaller sample size, the G^2 would also select the 3-class model under this misspecification. For Population B, with only an indirect covariate effect, all criteria correctly select the 2-class model for all specifications. For Population C, with both direct and indirect covariate effects, the AIC incorrectly selects the 3-class models for all the misspecified models 1–3, the BIC selects the 3-class models for misspecified models 1 and 2, and the G^2 selects the incorrect 3-class model for Model 2 and nearly so for Model 1. As with the general discussion of model specification, the greatest trouble comes when there are direct effects of covariates present and a model is specified without those paths, leading to overestimation of the number of classes.

An issue not directly dealt with in this chapter, related to both identification as well as class enumeration is that there may be other sources from which information regarding the nature of the mixture distributions may be drawn. Although beyond the scope of this dissertation, the latent variable

framework does allow extensions that could include concurrent traditional LCA or LCR models, concurrent or preceding growth mixture models, etc. that could inform or drive the estimation of the survival mixture components. (See, for example, Muthén and Masyn, 2001; or Larsen, in press.)

3.4.1 Long-term survivors

One special case of the survival mixture models perviously introduced is the long-term survivor (LTS) model, also known as the cure rate model or the mover/stayer model. Since it can be difficult to confidently enumerate the latent classes empirically, there is much appeal in the long-term survivor model that specifies not only the number of classes, $K = 2$, but the hazard probabilities within one of the classes, all set to zero. There is also appeal in that such a specification simplifies the model to be estimated. And it can be quite defensible in many substantive contexts to imagine a class of individuals who are never actually at-risk for the event in question. However, given the sensitivity of the survival model estimates to the specification of the distribution of the unobserved heterogeneity as well as the duration dependence of the hazard probabilities, this particular model should be reexamined.

Consider a group of individuals in a sample who do not experience the event of interest during the observation period. Although these subjects may

have all “survived” the event, some may still come to experience the event beyond the observation period. Others, however, may never experience the event at all. In the formulation of the LTS model, there are two latent classes; the LTS-class is defined as having a hazard probability of zero for all times with all direct covariate effects fixed at zero, equivalent to never being at-risk for the event. Of course, there are those who are at-risk but do not experience the event by chance—these are not long-term survivors. This creates a special case of mixture modeling where membership in the latent classes is partially observed: none of the subjects experiencing the event during the observation period are in the LTS-class; this is known. The class membership of those subjects not experiencing the event while under observation is unknown, i.e., latent. Conceptually, one could ask whether there is a reasonable distinction to be made between a subpopulation of individuals at very low risk versus a subpopulation of people not at-risk, especially in regards to latent subpopulations in the context of social research. Many behavioral mechanisms studied in social research are considered probabilistic not deterministic. Is there a group of “reformed” alcoholics that are never at-risk for taking a drink or is their risk simply very low? Is there a group of former inmates that are never at-risk for rearrest or does their risk just become negligible over time? Also, there is an important distinction to be made between *never at-risk* and not at-risk

during the time of observation. Since the more general discrete-time models do not explicitly model the duration dependence of the hazard in terms of time, there is nothing in the model specification setting the hazard probabilities in the LTS-class to zero that allow the interpretation of membership in the LTS-class to mean *never* at-risk; the LTS-class in a discrete-time survival model can only be interpreted as a subpopulation not at-risk during the time periods defined by the study.⁸

In essence, there are three concerns that should be addressed when considering application of the LTS model: 1) Specifying a two-class mixture model that is not empirically supported by the data; 2) Misspecifying a restricted LTS model when the sample has been drawn from a more general two-class mixture; 3) Interpreting the long-term survivor class a class of indefinite zero risk. The final concern is not related to model specification as it is a caution in model interpretation and has already been discussed in the previous paragraph. For the first concern, consider that the LTS-model is often selected a priori and procedures for class enumeration are not typically followed. That is, the restricted two-class structure is taken as given. It would be advisable to follow class enumeration procedures to explore the empirical support in the data for a mixture model and then test the LTS model against

⁸This is not the case for all continuous-time models.

Table 3.9: Long-term Survivor Populations A–E Definition

Regression	$P_{E_j}(E_j = 1 \mid x_1, C = 1)$	$P_{E_j}(E_j = 1 \mid x_1, C = 2)$	$P(C = 1 \mid x_1)$
Pop ⁿ A	0	$\frac{1}{1+\exp(-(-1+x_1))}$	0.5
Pop ⁿ B	0	$\frac{1}{1+\exp(-(-1+x_1))}$	$\frac{1}{1+\exp(-x_1)}$
Pop ⁿ C	$\frac{1}{1+\exp(4)}$	$\frac{1}{1+\exp(-(-1+x_1))}$	0.5
Pop ⁿ D	$\frac{1}{1+\exp(-(-4+x_1))}$	$\frac{1}{1+\exp(-(-1+x_1))}$	0.5
Pop ⁿ E	$\frac{1}{1+\exp(-(-4+x_1))}$	$\frac{1}{1+\exp(-(-1+x_1))}$	$\frac{1}{1+\exp(-x_1)}$

$$x_1 \sim N(0, 1), j = 1, \dots, 5$$

a less restricted two class model. This leads to the second concern regarding imposing an LTS-class restriction on a model for a sample drawn from a more general two-class mixture (or in the presence of nonspecific heterogeneity). To understand this concern, consider an simple example with five time periods and one covariates, x_1 . A sample of $n = 10,000$ was generated according the 2-class populations given in Table 3.9.

Note that Populations A and B both defined to have a fraction of long-term survivors. For A, the observed covariates does not influence the likelihood

Table 3.10: Long-term Survivor Models 1–4 Specification

Regression	$P_{E_j}(E_j = 1 C = 1, x_1)$	$P_{E_j}(E_j = 1 C = 2, x_1)$	$P_C(C = 1 x_1)$
Model 1	0	$\frac{1}{1+\exp(-(\nu_2+\beta_2x_1))}$	$\frac{1}{1+\exp(-\alpha_0)}$
Model 2	0	$\frac{1}{1+\exp(-(\nu_2+\beta_2x_1))}$	$\frac{1}{1+\exp(-\alpha_0)}$
Model 3	$\frac{1}{1+\exp(-(\nu_1+\beta_1x_1))}$	$\frac{1}{1+\exp(-(\nu_2+\beta_2x_1))}$	$\frac{1}{1+\exp(-\alpha_0)}$
Model 4	$\frac{1}{1+\exp(-(\nu_1+\beta_1x_1))}$	$\frac{1}{1+\exp(-(\nu_2+\beta_2x_1))}$	$\frac{1}{1+\exp(-(\alpha_0+\alpha_1x_1))}$

of being in the LTS-class while for Population B, it does. Populations C,D, and E are all a mixture of two non-zero hazard classes. For C, x_1 only influences the hazard in the higher risk class. For D, x_1 influences the hazard probabilities in both classes. Population E is like D but with an influence of x_1 on the class probabilities as well. Four models were fit to each of these five samples drawn from the five specified populations. The models are given in Table 3.10. The results are given in Tables 3.11–3.15.

In Tables 3.11 and 3.12, where the data are drawn from Populations A and B, respectively, that have a long-term survivor class, Models 3 and 4, allowing the parameters in both classes to be freely estimated offer no sta-

Table 3.11: Long-term Survivor Models 1–4 Results for Population A

Parameter	True Value	Model 1	Model 2	Model 3	Model 4
LL	–	-11,448.48	-11,448.37	-11,448.19	11,448.07
# parameters	–	3	4	5	6
ν_1	$-\infty$	@ -15 [†]	@ -15	-15.10	-15.15
ν_2	-1.00	-1.01	-1.03	-1.01	-1.03
β_1	0.00	@ 0	@ 0	-3.30*	-3.35*
β_2	1.00	1.00*	1.01*	1.00*	1.01*
α_{01}	0.00	-0.03	-0.04	-0.03	-0.05
α_{11}	0.00	@ 0	0.02	@ 0	0.02
$P_C(k = 1)$	0.50	0.49	0.49	0.49	0.49

* $p < 0.05$;

@ = “fixed at”

[†]Since the indicator probability is specified in terms of ν , the probability itself is not fixed to zero; rather, ν must be fixed at a large negative number, making the probability very close to or essentially zero. A ν of -15 corresponds to a probability of $3.06E - 7$.

Table 3.12: Long-term Survivor Models 1–4 Results for Population B

Parameter	True Value	Model 1	Model 2	Model 3	Model 4
LL	–	-11,089.90	-10,801.12	-10,952.73	10,799.65
# parameters	–	3	4	5	6
ν_1	$-\infty$	@ -15	@ -15	-3.21	-9.78
ν_2	-1.00	-0.69	-1.05	-0.33	-1.05
β_1	0.00	@ 0	@ 0	-0.36*	1.55*
β_2	1.00	0.53*	1.03*	1.26*	1.05*
α_{01}	0.00	0.51*	-0.06	1.34*	-0.06
α_{11}	1.00	@ 0	1.08*	@ 0	1.11*
$P_C(k = 1)$	0.50	0.63	0.49	0.79	0.49

* $p < 0.05$;

@ = “fixed at”

Table 3.13: Long-term Survivor Models 1–4 Results for Population C

Parameter	True Value	Model 1	Model 2	Model 3	Model 4
LL	–	-12,235.16	-12,233.66	-12,220.63	12,220.63
# parameters	–	3	4	5	6
ν_1	-4.00	@ -15	@ -15	-3.51	-3.50
ν_2	-1.00	-1.11	-1.16	-0.99	-0.99
β_1	0.00	@ 0	@ 0	0.09	0.09
β_2	1.00	1.16*	0.94*	1.00*	1.16*
α_{01}	0.00	-0.26	-0.34	0.18	-0.01
α_{11}	0.00	@ 0	0.09	@ 0	0.18
$P_C(k = 1)$	0.50	0.44	0.42	0.54	0.55

* $p < 0.05$;

@ = “fixed at”

Table 3.14: Long-term Survivor Models 1–4 Results for Population D

Parameter	True Value	Model 1	Model 2	Model 3	Model 4
LL	–	-12,560.94	-12,532.76	-12,504.47	12,504.41
# parameters	–	3	4	5	6
ν_1	-4.00	@ -15	@ -15	-4.06	-4.11
ν_2	-1.00	-1.40	-1.09	-1.03	-1.05
β_1	1.00	@ 0	@ 0	0.98*	1.02*
β_2	1.00	0.90*	0.69*	1.06*	1.07*
α_{01}	0.00	-0.74*	-0.27*	-0.08	-0.11
α_{11}	0.00	@ 0	-0.42*	@ 0	0.03
$P_C(k = 1)$	0.50	0.32	0.43	0.48	0.47

* $p < 0.05$;

@ = “fixed at”

Table 3.15: Long-term Survivor Models 1–4 Results for Population E

Parameter	True Value	Model 1	Model 2	Model 3	Model 4
LL	–	-12,871.50	-12,869.87	-12,860.57	12,822.99
# parameters	–	3	4	5	6
ν_1	-4.00	@ -15	@ -15	-8.79	-3.96
ν_2	-1.00	-1.21	-1.16	-1.14	-0.99
β_1	1.00	@ 0	@ 0	2.46*	0.96*
β_2	1.00	0.35*	0.28*	0.35*	0.96*
α_{01}	0.00	-0.25*	-0.18*	-0.16*	0.02
α_{11}	1.00	@ 0	-0.10	@ 0	0.92*
$P_C(k = 1)$	0.50	0.44	0.46	0.46	0.51

* $p < 0.05$;

@ = “fixed at”

tistical improvement over the LTS Models 1 and 2. In other words, when there is a long-term survivor class, it is possible to distinguish empirically (given a sufficient sample size) a zero-risk class. In Tables 3.13–3.15, the story changes. These data are drawn from Populations C, D, and E that have two classes, one with a low, but non-zero risk. By fitting an LTS model to these data, the estimated class proportions and covariate effects in the non-LTS class present a much different picture than what are the actual relationships between the classes and between the hazard probabilities and the observed covariate. Again, as with the previous model specification and class enumeration examples, the lesson here seems clear: the best analysis strategy, even if substantive theory support the existence of a LTS class, is to fit an unrestricted two class model with both direct and indirect covariate effects and then test this model against the LTS model.

Returning to the example begun in Chapter 2, evidence for unobserved heterogeneity is evaluated. The G^2 statistic for the last model (5b) fit in Chapter 2 was 1.62 (df=6) which has a corresponding p-value of 0.95, suggesting the 1-class model fit the data very well and that is unlikely any significant improvement can be made by 2-class model. Table 3.16 gives the class enumeration criteria results for Model 5b and the unrestricted 2-class version in Model 6. The BIC and G^2 both favor the one class over the two-class as

Table 3.16: 1- and 2-Class Model Comparisons

Results	LL	# parameters	AIC	BIC	G^2
1-class (Model 5)	-194.82	5	399.64	415.32	1.62 (p=0.95)
2-class (Model 6)	-184.72	13	395.44	436.21	0.96 (p=0.99)

expected.

Although not supported by the data, it is still instructive to examine the results of Model 6. The parameter estimates are given in Table 3.17. Figures 3.4–3.8 display the estimated survival probabilities for baseline, at the sample mean covariate values, and for each covariate effect at the sample mean of the other covariates based on the Model 6. Figure 3.4 shows that the two classes have very similar baseline survival. The proportion of the sample in Class 1 is estimated at 30% and for Class 2, 70%. Figure 3.5 shows that Class 1 is, at the average covariate values, at much higher risk and, consequently, has a notably lower survival rate during the first year. The plots for Figures 3.6–3.8 show that the reason is not an elevated baseline rate but rather an elevated effect of the wife’s education and percent-days-drinking. Although drinking is a significant risk factor for both Class 1 and Class 2, the magnitude of the effect

is estimated to be much higher in class one. Wife's education not beyond high school is a significant *risk* factor and has no significant effect in Class 2. This is opposite what was found in the one-class model. However, wife's education is also predictive of class membership with men having wives with no education beyond high school being six time less likely to be in Class 1. Household income is not a significant predictor for Class 1 but having an income greater than \$35K is protective against return to violence for Class 2 men. Income is not significantly associated with class membership. Although there may not be two subpopulations exactly resembling the two classes characterized by this model, it does suggest that there is more heterogeneity with regards to frailty or susceptibility to risk or protective factors related to violence than heterogeneity in underlying baseline risk of returning to violence, i.e., some men who are more likely to return to violence when drinking than others, even though men who are drinking are all more likely to return to violence than those who are not. It also suggests the that relationship between wife's education and husband's time-to-violence is much more complex than the one-class model may reveal.

Table 3.17: Results for Data Example Model 6

Class 1 = 30%

Class 2 = 70%

Parameter	Class 1 Est.	SE	Est./SE	Class 2 Est.	SE	Est./SE
I(Wife's educ. \leq H.S.)	4.31*	1.60	2.70	-0.69	0.46	-1.51
I(Income $>$ \$35K)	1.82	1.11	1.64	-1.84	0.73	-2.52
% days drinking	8.80*	2.59	3.40	2.33	0.75	3.11
E(η_0)	-3.04*	1.17	-2.60	-2.45	0.64	-3.86
E(η_1)	-0.71*	0.34	-2.06	-0.19	0.16	-1.18
α_{01}	-0.34	0.87	-0.39	@0	-	-
$\alpha_{I(\text{Wife's educ.} \leq \text{H.S.}),1}$	-1.74*	0.76	-2.31	@0	-	-
$\alpha_{I(\text{Income} > \$35K),1}$	0.15	0.84	0.18	@0	-	-

* $p < 0.05$;

@ = "fixed at"

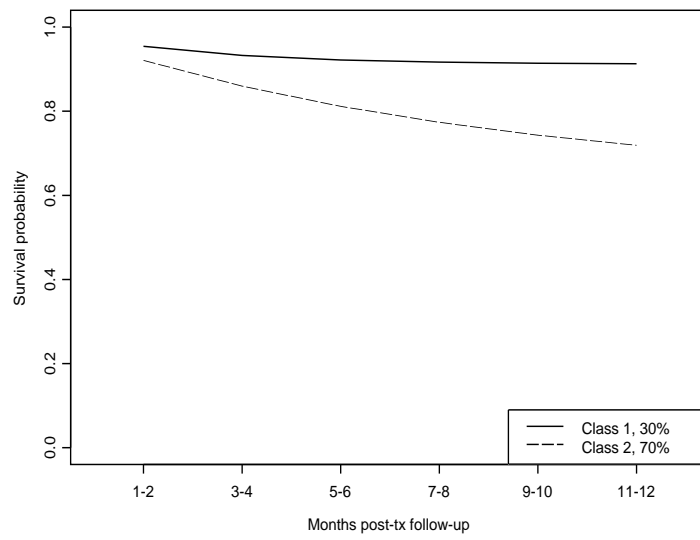


Figure 3.4: Model 6 estimated survival probabilities at baseline.

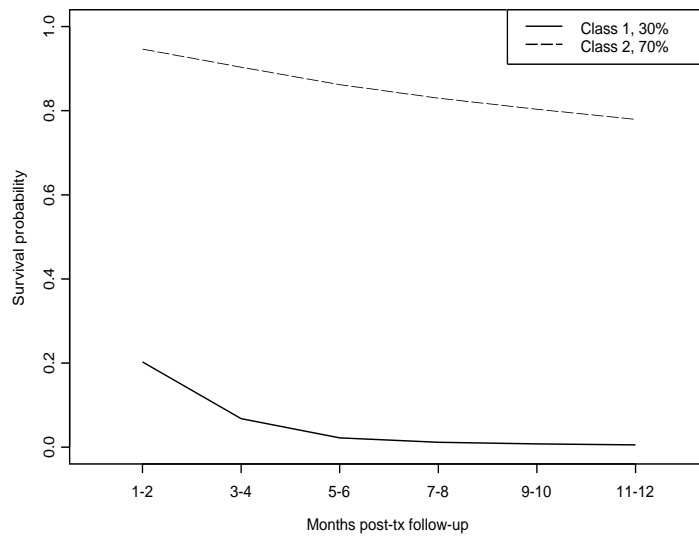


Figure 3.5: Model 6 estimated survival probabilities at sample average wife's education, household income, and % days drinking.

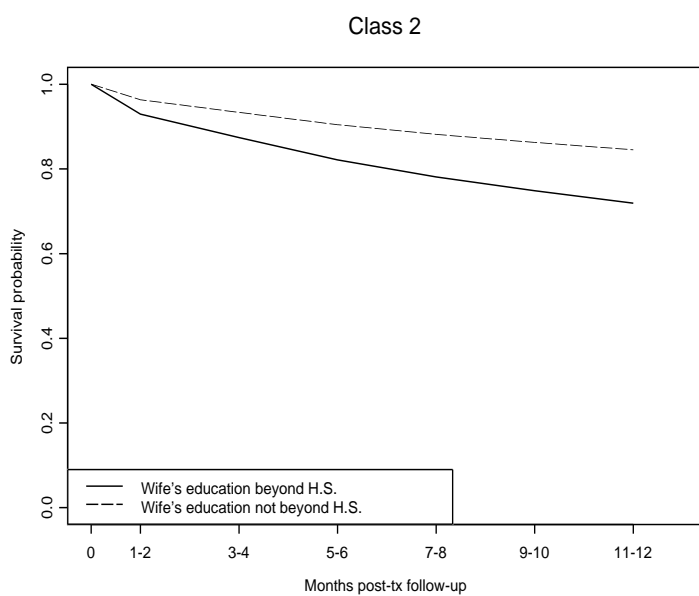
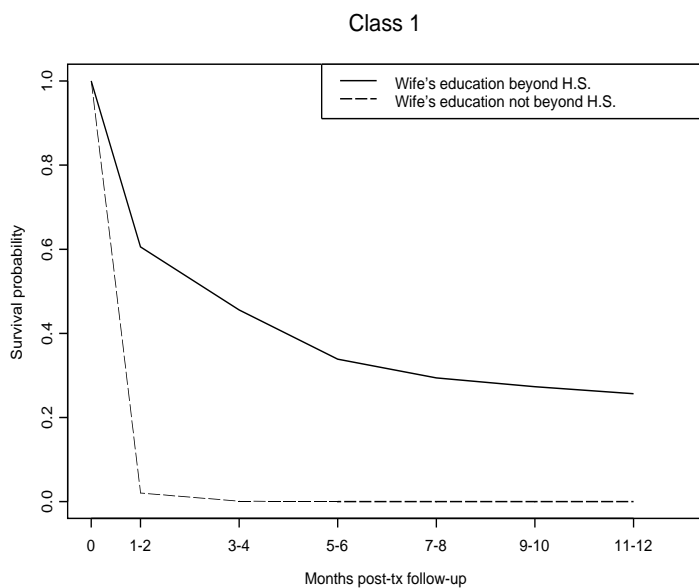


Figure 3.6: Model 6 estimated survival probabilities for Classes 1 and 2 by wife's education at sample average household income and % days drinking.

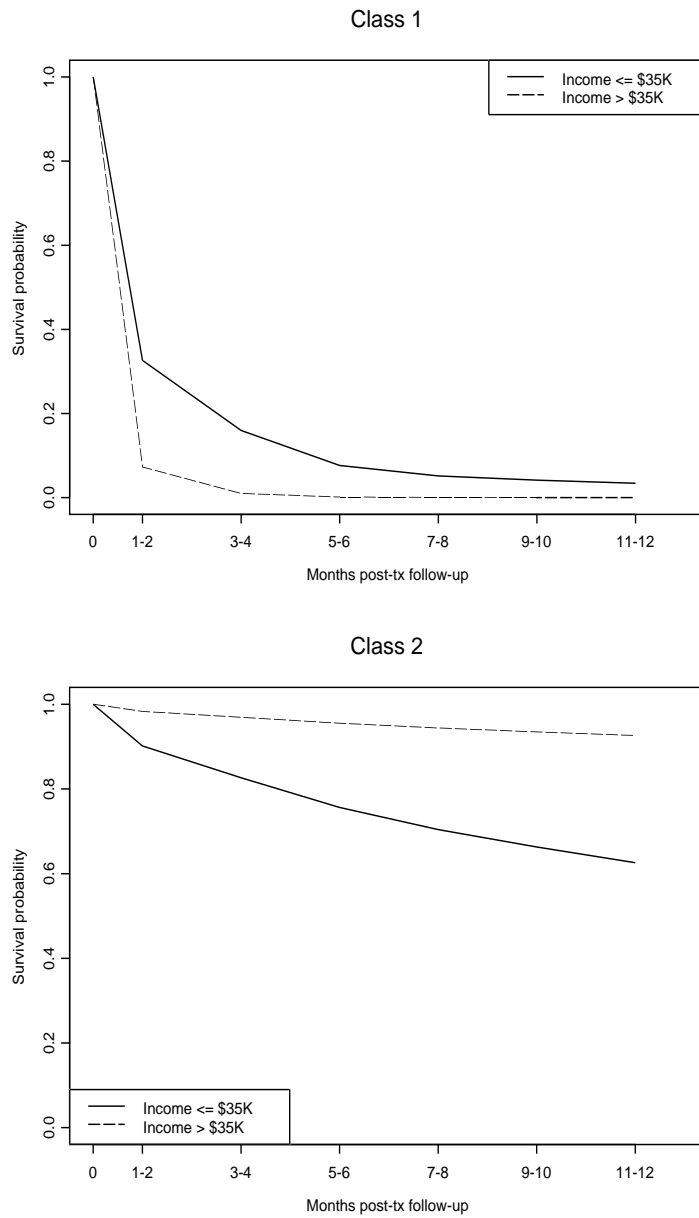


Figure 3.7: Model 6 estimated survival probabilities for Classes 1 and 2 by household income at sample average wife's education and % days drinking.

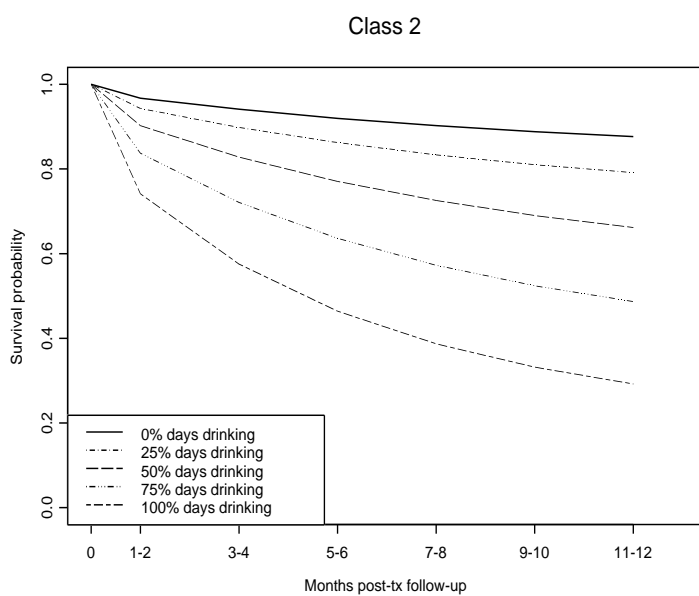
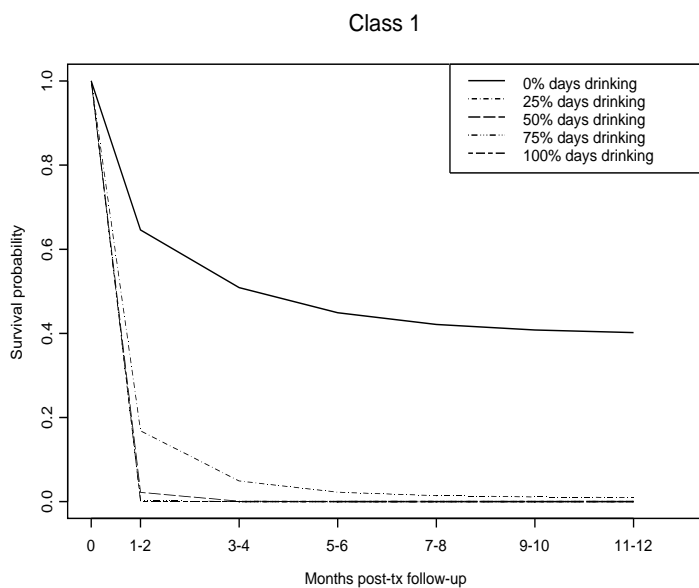


Figure 3.8: Model 6 estimated survival probabilities for Classes 1 and 2 by % days drinking at sample average wife's education and household income.

Chapter 4

Recurrent Events

Up until this point, the only types of events that have been considered are single, non-repeatable events. If an event is considered a transition from one state to another, such as living to dead, then the end state is, in the language of Markov models, *absorbing*; once an individual has transitioned into that state, there is no further movement or “risk” to the individual. Given the historical development of survival models in the area of life table analysis, it is not surprising that the main focus for methods development has been around single, terminating events. However, there are many time-to-event processes, particularly in fields of social research, that do not fit the single event model. Most generally, such data can be referred to as *multivariate*

survival or event history data. This chapter focuses on one particular type of multivariate survival process: recurrent events.

4.1 Multivariate event histories

In the survival analysis literature, multivariate event process are typically categorized as either multiple event histories or recurrent event histories.¹ Multiple event processes are also termed competing risks models. Essentially, rather than a single terminating event, there are several possible events that could terminate a subject's risk. Multiple causes of death is the most common example—instead of modeling time-to-death, one may want to model time-to-death from opportunistic infection versus time-to-death from cancer in advanced AIDS patients. Another example of multiple event data is juvenile arrest. Rather than modeling time-to-arrest, one may want to model time-to-arrest for crimes against property versus time-to-arrest for crimes against persons. In these models, a subject is at-risk simultaneously for all events but the risk for any of the events is terminated at the occurrence of one of the possible events. A distinction can be made between competing risks, where

¹Certainly, in more complex model extensions, a combination of recurrent and multiple events may be considered—multiple events and recurrent events are by no means mutually exclusive in social or behavioral processes.

the occurrence of any one of the possible events precludes the occurrence of the others, and a multiple process scenario, where several events may occur to a given individual without any necessary ordering or sequence. Analysis of either type of multivariate survival data, multiple event or multiple process, is beyond the scope of this dissertation.

Recurrent event processes are also termed repeatable events or multiple spell models. Essentially, the event of interest is the same but it is not terminating, that is, after an individual's first occurrence of the given event she returns to an "at-risk" status for a subsequent occurrence. Examples of recurrent events include pregnancies, school suspensions, and suicide attempts. Hougaard (2002) makes the usefully distinction between recurrent event processes with high and low enumerations. That is, some recurrent event processes have a small number of maximum events that are ever observed for a single subject, such as pregnancies. Other recurrent event processes may have such a large number of recurrences for some subjects, such as epileptic seizures, that enumerating them becomes cumbersome. In all cases, it is possible to define

some time, $t = 0$, that marks the beginning of risk for any event for a given individual.² However, in the case with frequent occurring events, the subject may experience an unknown number prior to the beginning of observation period, resulting in multiple left-censored episode times. In the case for low enumerative events, it is more likely that a full event history from $t = 0$ may be known to the researcher. For the purposes of this dissertation, only single, recurrent events of the low enumerative kind will be considered.³

Although some authors reformulate recurrent event data as multiple event data, the following discussion highlights the importance of considering recurrent event processes as distinct from other multivariate event processes. There are three key features of recurrent event data—the first two distinguishes them from other multiple event data and the third feature is common to all multivariate processes: 1) an individual may only be at-risk for one event at a time; 2) an individual may not be at-risk for the m^{th} event until she has experienced the $(m - 1)^{th}$ event, i.e., the events are ordered; and 3) the presence of

²In the recurrent events setting, if there is not any clear $t = 0$, it may be possible to instead begin the process modeling with $t = 0$ defined at the timing of the first event, i.e., modeling from the time a subject is at-risk for recurrence rather than at-risk for first occurrence.

³Hougaard (2002) and Vermunt (1997) conceive of these types of processes in terms of multi-state Markov models in continuous-time and discrete-time, respectively.

within-subject event time correlation. Features 1 and 2 are what make recurrent event data markedly different from competing risk or multiple event data where an individual is considered to be at-risk for all events simultaneously. Feature 3 falls under the topic of unobserved heterogeneity. Just as in Chapter 3, the assumption of independent observations may not be applicable; it is unlikely that the assumption of independent observations *within* individuals, let alone between individuals, would hold true for most recurrent event data. That is, it is unlikely that each subject's individual susceptibility (or frailty) for a sequence of events is entirely captured in values of the observed covariates on the individual. These are the features that must be kept central when considering the model specification for recurrent event data. An additional feature that may also apply to a variety of multivariate survival processes, including recurrent events, is event-specific processes. For recurrent events, that would mean the hazard for the m^{th} event may have a different duration dependence and behave differently with respect to covariates than the hazards for preceding events.

Kelly and Lim (2000) wrote a comprehensive review and comparison of the most recent models used to handle recurrent event data in continuous time. They present a systematic framework used to evaluate the similarities and differences across these models, as well as the appropriateness of those

models in handling recurrent events data.⁴ Currently, there is no comparable review of recurrent event models for discrete-time. The purpose of this chapter is to present a similar framework for understanding the different formulations of recurrent event processes in discrete-time and to develop corresponding specifications of each formulation in terms of latent variables, extending the single event models presented in Chapters 2 and 3. The next section focuses on the definition of risk and risk periods, leaving aside the issue of correlated event times (Feature 3) for a later section.

4.2 Defining risk for recurrent event histories

In order to understand and model a recurrent event process, one must answer three basic questions: “Who?”, “What?”, and “When?”. Who is at-risk? For what are they at-risk, e.g., first event, second event, etc.? When are they at-risk? The first two questions can be answered in a joint manner. At $t = 0$, all subjects are at-risk. They are at-risk for the first event. They are at-risk for the first event until the first event occurs or they are censored. This answers

⁴Kelly and Lim’s framework is based on what they call the four *key* model components: definition of the risk interval; definition of the risk set; choice of a common versus event-specific baseline hazard (as noted by Kelly and Lim, risk set definition incorporates the choice of baseline hazard); and handling of within-subject correlation (p.14).

the “when” question for the first event. This time period of risk for the first event may also be referred to as the first “spell”. A spell is ended by either the occurrence of the event or by cessation of observation by censoring or study conclusion. For the first spell, time zero ($t = 0$), the left end point of the first time period, is defined as it would be in a single event model, e.g., birth, start of school, end of treatment, etc. As soon as a subject experiences the first event, she then becomes at-risk for a second event, not before. All those having experienced a first event but not yet a second are at-risk for a second event. The second spell is the time period of risk from the occurrence of the first event to the occurrence of the second event or censoring. Here is an issue in discrete-time regarding spell definition that does not come up in the continuous-time framework: defining the beginning of spells for $m > 1$. By Feature 1, a subject may not be at-risk for event m until event $m - 1$ has occurred. In continuous time, it is the instant right after the occurrence of event $m - 1$ that the risk of m begins. In discrete-time, event $m - 1$ may happen any time during a certain interval, say j . There are two difficulties if risk for event m is defined to start at period $j + 1$: 1) There is some portion of interval j during which the subject is at-risk for event m that is not counted; and 2) It must be assumed that only one event is possible per interval. Allison (1982) offers three solutions to this issue: 1) Chose intervals such that no subject experiences more than one event

in a single period, 2) Simply treat any number of events great than one in any given interval as only one event, or 3) Model the number of events within each time interval with a Poisson regression. The first solution can be applied only if the continuous-time data is available and the analyst is able to select the period length for discretizing the event times. However, this is a non-option if the data only exist in discrete-time form. Also, even if possible, creating smaller intervals could also result in a higher number of periods containing no observed events, especially for $m > 1$, which could cause problems with the model estimation. The second solution is not a reasonable option if one wants to accommodate Feature 2, allowing the baseline hazard to be event-specific. The event enumeration by number of periods in which at least one event occurred can lead to very different results than enumeration by the number of events themselves if subjects are experiencing multiple recurrences within each time period. The third solution, like the second, does not allow for the estimation of event-specific hazards. Land, Nagin, and McCall (2001) use the Poisson regression approach.⁵

Here, a more suitable solution to the problem of multiple recurrence in a single time period is proposed, keeping with the idea of an underlying

⁵Although their model does not allow event-specific hazards, it does allow inclusion of unobserved heterogeneity in the form of finite mixture.

continuous-time process. Essentially, risk for event m is defined to start in the *same* period as the occurrence of event $m - 1$. Allowing this overlap does not violate Feature 1 of recurrent events data. Since event $m - 1$ occurring in period j implies that the event time for $m - 1$ is between t_{j-1} and t_j , the risk for event m begins in that same interval. This definition of spell beginning for event $m > 1$ is applied for the time formulations described below. The primary problem with this definition of risk, and most likely the reason that it has not been utilized in discrete-time for recurrent events, is the bias present in the hazard duration dependence by allowing such an overlap if not accounting for reduced risk duration. To understand this issue, consider a subject who experiences a first and second event in the same time period, j . For the first event in j , the underlying hazard rate on the continuous-time scale is assumed to be constant for the interval. However, for the second event, the underlying hazard rate must be smaller in the beginning of the interval since the first event in that period occurs before the second. Another way to look at this issue is to return to the definition of the hazard probability in terms of the underlying continuous-time process. The hazard probability for a first event in period j can be defined as $P(t_{j-1} \leq T_1 < t_j \mid T_1 \geq t_{j-1})$. The hazard probability

for a second event in period j is defined as $P(t_1 \leq T_2 < t_j \mid T_2 \geq t_1)$.⁶ As can be seen, the average risk duration is *not* the same for two consecutive events in the same interval. The bias in risk durations can increase the more events occurring in a single time period for each individual. Note that in the case of discretely occurring events, multiple recurrence in a single time period is not an issue. The beginning of risk for event m should be defined as the period immediately following the period of occurrence for event $m - 1$. This chapter allows the overlap in risk intervals for consecutive events (assuming an underlying continuous-time process) and proposes a possible bias correction in the section on estimation.⁷

The above discussion helped better define the “when” question for second and later events but there is another dimension to the “when” question with regards to the time scale that is chosen. To understand this issue better, consider the three example cases depicted in Figure 4.1. The beginning of period 1 is assumed to be $t = 0$ and the same for all three cases. Subject A

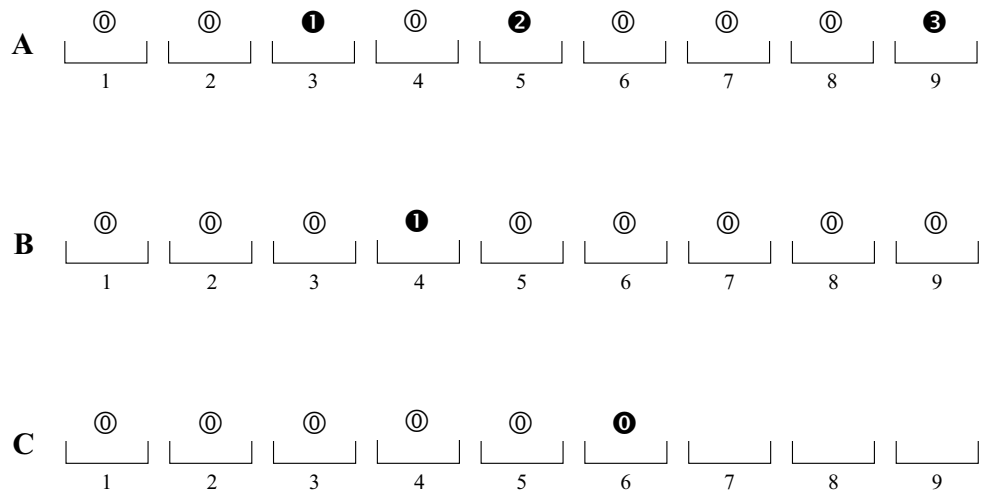
⁶There are different definitions of the hazard probability depending of the time formulations discussed in subsequent sections.

⁷Allowing the overlap in risk intervals for adjacent events and the proposed bias correction are novel with respect to the existing literature and will need to be more carefully examined analytically and empirically before they are accepted into current analysis practices.

experienced a first event in period 3. Thus, according to the reasoning above, Subject A is at-risk for a second event beginning in period 3. Subject B experiences a first event in period 4 and is thus at-risk for a second event beginning in period 4. There are two ways to look at the timing of risk for the second event. One way is with respect to the occurrence of the first event. That is, considering period 3 for Subject A and period 4 for Subject B as equivalent periods in that both periods represent for each subject the first period at-risk for the second event. In this time scale, Subject A is at-risk for a second event during periods 0–2 following the first event and Subject B is at-risk for a second event in periods 0–5 following the first event, where period zero represents the same period of the prior event occurrence. This is referred to in the literature as *Gap Time*, or GT, formulation.

Another way to look at risk timing is on the original observation timeline for which Subject A is at-risk for a second event in the period *before* Subject B is at-risk for a second event. Subject A is at-risk during periods 3–5 for a second event (at period 5 the second event occurs) and Subject B is at-risk during periods 4–9 (after period 9, observation is ended and B is censored with respect to the time-to-second-event process⁸. This is referred to in the

⁸As depicted in the figure, censoring of Subject B does not occur until *after* the end of the ninth time period



Darkened circles represent events, with the event number within the circle. A darkened circle with a "0" represents a censoring event; the unfilled circles represent observations of "no event" made on the subject. Time periods without any circle represent periods during which the subject ceased to be under the observation of the researcher.

Figure 4.1: Three subject example of recurrent event observations.

literature as *Counting Process*, or CP, formulation. Both these formulations are described in more detail below.

In the GT formulation, the first period of risk for event m is the period of occurrence for event $m-1$. In this formulation, the clock essentially "resets" after each event. The occurrence of all events after the first are modeled on a time scale relative to the prior event and not relative to the actual timeline of observation. In GT formulation, the $j = 0$ time period for the m event is

associated with the *same* time period during which event $m - 1$ occurred. An event m occurring in the GT period $j = 1$ implies that event m occurred in the period immediately following the period during which event $m - 1$ occurred. Consider again the three example cases depicted in Figure 4.1. In the GT formulation, Subject A is at-risk for a first event in periods 1–3, at-risk for a second event in periods 0–2, and at-risk for a third event in periods 0–4. All three of the spells for Subject A terminate with an event. Subject B is at-risk for a first event in periods 1–4 and at-risk for a second event in periods 0–5. Subject B is never observed to be at-risk for the third event because the second event does not occur during the period of observation. Similarly, Subject C is not observed to be at-risk for any event but the first. Table 4.1 gives the risk periods (spells) for each subject by each event.

A distinction here is to be made between what will be termed full-GT and partial-GT formulations. Consider the three hypothetical subjects. Period 1 for the first spell of Subject A is the first interval of time from some *beginning*, $t = 0$, while period 1 for the second and the third spells are the time intervals directly before which a prior event occurred. Thus, the spell of the first event is on a different time-scale than the gap time formulation for all subsequent events. However, if the first spell itself is begun by the occurrence of an event, i.e., onset of the recurrent event process, then the first spell is actually the

Table 4.1: Risk Periods Defined

Formulation		Gap time (GT)	Counting process (CP)	Total time (TT)
Subject A	1 st event	1-3	1-3	1-3
	2 nd event	0-2	3-5	1-5
	3 rd event	0-4	5-9	1-9
Subject B	1 st event	1-4	1-4	1-4
	2 nd event	0-5	4-9	1-9
	3 rd event	N/A	N/A	(1-9)
Subject C	1 st event	1-5	1-5	1-5
	2 nd event	N/A	N/A	(1-9)
	3 rd event	N/A	N/A	(1-9)

spell of the second event and time period 1 would be in the same gap time formulation as the subsequent spells. In the first case, where the first spell is marked by onset of risk, the gap time formulation is partial, only applying to events after the first occurrence. In the second case, where the first spell is marked by a first event, e.g., onset of IV drug use, the gap time formulation is full-GT.

The CP formulation uses the same time scale for all events, referenced to a fixed point in time, but does not allow the overlap in risk periods across events for a given subject. That is, Subject A cannot be observed to be at-risk for a second event until after A has experienced the first event. For example, in Table 4.1, Subject A has observed risk periods 3-5 for the second event;

the second spell is three periods long, as in the GT formulation, but the spell terminates at period 5, on the original time scale. Both the GT and CP formulations are conditional in that a subject's risk status for the m^{th} event is conditional upon the occurrence of $(m - 1)$ earlier events. As explained by Allison (1995), it is possible to imagine some processes where the hazard for an event depends on the time since an individual first became at-risk, regardless of the number and timing of prior events, e.g., time in the labor force for risk on unemployment. The fact is, in many processes, the hazard may depend on *both* time since the last event *and* time since overall risk onset. It is possible to account for this dual duration dependence in both the CP and GT formulations as is demonstrated in a later section.

There is a third way of defining risk that was not mentioned in the above discussion. The *total time*, or TT, formulation defines the risk periods for event m (m^{th} spell) as beginning at a select point on the observation time scale. In this case, compared to the GT formulation, the clock does not reset for each event—the beginning of each spell is at the same point in the observation timeline. And, unlike the CP formulation, risk periods for different events for the same subject overlap. The TT formulation is essentially marginal—the risk status for each event is *not* conditional on the occurrence of prior events. Rather than determining risk at each time period based on the event history

up to that period, the TT formulation looks at risk $t = 0$ and forward, e.g., subjects could experience a first, second, and third event during the first time period and are considered at risk for all three events. Risk periods for each event end either by occurrence of the corresponding event, at which point the subject continues to remain at-risk for all subsequent events, or by censoring, at which point the subject ceases to be observed at-risk for *all* remaining events. As shown in Table 4.1, the first spell for Subject A is periods 1–3, periods 1–5 for the second spell, and periods 1–9 for the third spell. For Subject B, the first spell is periods 1–4 and the second spell is periods 1–9. The periods for the third spell are in parentheses. There is a variation on the TT formulation that defines spells in TT but conditional upon the occurrence of a prior event during the observation period. In this version, only subjects who experiencing a first event during the observation period are considered to be at-risk for a second event from period 1. This formulation makes little sense since it violates the desired feature of not allowing overlapping risk periods and does not provide a marginal risk determination as does the first TT formulation described. Only the first marginal formulation is presented in the remainder of this chapter. Although the marginal definition of risk may provide some information about the recurrent events process, allowing subjects to be at risk for more than one event at a time is really more appropriate for multiple events or competing

risks for which it was originally developed. Wei, Linn, and Weissfeld (1989), who developed the model with this TT formulation recommended its use for recurrent events, even though it includes those risk periods in parentheses. Guo and Lin (1994) reformulate the model for discrete-time, again recommending it for recurrent events, as well as competing risks. Ironically, despite this model's acknowledged shortcomings in the literature with respect to its application to recurrent events data, particularly related to correct estimation of covariate effects, it enjoys wide use (Kelly & Lim, 2000). The differences between using the GT, CP, and TT formulations will be further discussed throughout this chapter, including full illustration of each with the real data example.

For all three formulations, it is possible to allow the baseline hazard and covariate effects to be event-specific, allowing for testing of common effects or duration dependence across events.⁹ Specifying a common hazard in the GT formulation, for example, would suggest the risk of an event during the first period following the first event was the same as the risk of an event during the first period following the second event. Specifying a common hazard in the CP formulation would suggest that the risk of a second event in the time

⁹Some recurrent event model specifications *presuppose* common baseline hazards and/or common covariate effects, failing to accommodate the desired Feature 2 previously mentioned.

period j was the same as the risk of a third event in during period j . A common hazard in the TT formulation is not a reasonable model for recurrent events. This is because specifying a common hazard is the same as equating the marginal event probabilities for all events in each time period, ignoring the ordered nature of the events which makes it highly unlikely that initial events and later event would have the same hazard in the earlier time periods of risk.

4.3 Basic notation and likelihood

Let the given sample consist of n independent individuals i , with $i = 1, \dots, n$. Let T_{mi} be the survival time for individual i corresponding to event m . Let M be the maximum number of possible events. In the discrete-time setting, events are only observed to fall within J grouped time intervals, $[t_j, t_{j+1})$, where $j = 0, \dots, J - 1$ and $t_j = \infty$. Let Γ_{mi} represent the time interval into which T_{mi} falls, so that Γ_{mi} is a discrete random variable with the event set $\{1, \dots, J\}$. Then $\Gamma_{mi} = \gamma_m$ if $t_{\gamma_m-1} \leq T_{mi} < t_{\gamma_m}$. Also, the following ordering is imposed: $\Gamma_{1i} \leq \Gamma_{2i} \leq \dots \leq \Gamma_{Mi}$. In all formulations, the hazard and survival probabilities for the first event is the same as it is in the single event case:

$$\begin{aligned}
P_{h_1}(\Gamma_1 = \gamma) &= P(\Gamma_1 = \gamma \mid \Gamma_1 \geq \gamma) \\
&= P(t_{\gamma-1} \leq T_1 < t_\gamma \mid T_1 \geq t_{\gamma-1}), \tag{4.1}
\end{aligned}$$

and

$$\begin{aligned}
P_{S_1}(\gamma) &= P(\Gamma_1 > \gamma) \\
&= P(\Gamma_1 \neq 1 \mid \Gamma_1 \geq 1) \cdot P(\Gamma_1 \neq 2 \mid \Gamma_1 \geq 2) \cdots P(\Gamma_1 \neq \gamma \mid \Gamma_1 \geq \gamma) \\
&= \prod_{j=1}^{\gamma} (1 - P_{h_1}(j)). \tag{4.2}
\end{aligned}$$

Let the observed data for subject i corresponding to the first event be represented by $\{A_{1i}, \delta_{1i}\}$. Similar to the single event framework, A_{1i} represents the last time period during which subject i is observed to be at-risk for the first event and δ_{1i} is the indicator of whether an event ($\delta_{1i} = 1$) or censoring ($\delta_{1i} = 0$) occurred during that final period.¹⁰ The likelihood contribution of subject i relative to the first event is given by

$$L_{1i} = [P_{h_1}(a_{1i})]^{\delta_{1i}} \cdot \prod_{j=1}^{a_{1i}-1} [1 - P_{h_1}(j)]. \tag{4.3}$$

As in Chapter 2, the likelihood can be restated in terms of the event and observed risk indicators. Let E_{1j} be an event indicator for the first event

¹⁰Only noninformative right-censoring is considered in this dissertation.

such that

$$E_{1\gamma} = I(\Gamma_1 = \gamma). \quad (4.4)$$

Let R_{1j}^o be an indicator of observed risk for the first event in period j , that is,

$$R_{1\gamma}^o = I([A_1 \geq \gamma \text{ and } \delta_1 = 1] \text{ or } [A_1 > \gamma \text{ and } \delta_1 = 0]). \quad (4.5)$$

The observed data likelihood for the first event¹¹ for subject i can be restated in terms of R_{1i}^o and E_{1i} by

$$L_{1i} = \prod_{j \in \{r: R_{1r}^o = 1\}} P(E_{1j} = e_{1ji}). \quad (4.6)$$

To summarize, the observed data for the first event (or first spell), (A_{1i}, δ_{1i}) , as well as the corresponding likelihood, can be restated in terms of (E_{1i}, R_{1i}^o) without loss of information, with the following conversion:

$$E_{1ji} = \begin{cases} 1 & \text{if } A_{1i} = j \text{ and } \delta_{1i} = 1 \\ 0 & \text{otherwise,} \end{cases} \quad (4.7)$$

and

$$R_{1ji}^o = \begin{cases} 1 & \text{if } (A_{1i} \geq j, \delta_{1i} = 1) \text{ or } (A_{1i} > j, \delta_{1i} = 0) \\ 0 & \text{otherwise.} \end{cases} \quad (4.8)$$

Table 4.2 gives the values for R_{1i}^o and E_{R^o1} for the hypothetical Subjects, A, B, and C, from the example introduced in the preceding section.

¹¹As before, MAR corresponds to the assumption of noninformative censoring.

Table 4.2: Example Data for First Event

Risk indicator	$r_{1,1}^o$	$r_{1,2}^o$	$r_{1,3}^o$	$r_{1,4}^o$	$r_{1,5}^o$	$r_{1,6}^o$	$r_{1,7}^o$	$r_{1,8}^o$	$r_{1,9}^o$
Subject A	1	1	1	0	0	0	0	0	0
Subject B	1	1	1	1	0	0	0	0	0
Subject C	1	1	1	1	1	0	0	0	0
Event indicator	$e_{r^o1,1}$	$e_{r^o1,2}$	$e_{r^o1,3}$	$e_{r^o1,4}$	$e_{r^o1,5}$	$e_{r^o1,6}$	$e_{r^o1,7}$	$e_{r^o1,8}$	$e_{r^o1,9}$
Subject A	0	0	1
Subject B	0	0	0	1
Subject C	0	0	0	0	0

4.3.1 Gap time

In the GT formulation, the time periods of all event after the first are on a time scale relative to the occurrence of the first event. Thus, define Δ_{mi} as the time period into which T_{mi} falls, relative to the first event, such that $\Delta_{mi} = \Gamma_{mi} - \Gamma_{(m-1)i}$. Note that $\Delta_{mi} = 0$ corresponds to event m occurring in the same time period as event $m - 1$. The hazard probability for event m , for $m > 1$, is then given by

$$\begin{aligned}
 P_{gt(h_m)}(d) &= P(\Delta_m = d \mid \Delta_m \geq 0) \\
 &= P(\Gamma_m - \Gamma_{m-1} = d \mid \Gamma_m - \Gamma_{m-1} \geq d). \tag{4.9}
 \end{aligned}$$

So, $P_{gt(h_m)}(j)$ is the probability that the m^{th} event occurs in j^{th} period after the $(m - 1)^{th}$ event given that it does not occur before that. A conditional

survival probability can then be defined, similar to the hazard for each event m conditional on the occurrence of event $m - 1$.

$$\begin{aligned} P_{gt(S_m)}(d) &= P(\Delta_m > d) \\ &= \prod_{j=1}^d (1 - P_{gt(h_m)}(j)) \end{aligned} \quad (4.10)$$

Let the observed data for subject i corresponding to the second event be represented by $\{D_{gt(2)i}, \delta_{2i}\}$ where $D_{gt(2)i}$ represents the last time period during which subject i is observed to be at-risk for the second event (in gap time, $D_{gt(2)} = A_2 - A_1$) and δ_{2i} is the indicator of whether an event ($\delta_{2i} = 1$) or censoring ($\delta_{2i} = 0$) occurred during that final period. The likelihood contribution of subject i relative to the second event is given by

$$L_{gt(2)i} = \left([P_{gt(h_2)}(d_{gt(2)i})]^{\delta_{2i}} \cdot \prod_{j=0}^{d_{gt(2)i}-1} [1 - P_{gt(h_2)}(j)] \right)^{\delta_{1i}}. \quad (4.11)$$

More generally, for any event m , $m > 1$, the likelihood contribution of subject i relative to the m^{th} event is

$$L_{gt(m)i} = \left([P_{gt(h_m)}(d_{gt(m)i})]^{\delta_{mi}} \cdot \prod_{j=0}^{d_{gt(m)i}-1} [1 - P_{gt(h_m)}(j)] \right)^{\delta_{(m-1)i}}. \quad (4.12)$$

The full likelihood for the gap time formulation is then given by

$$L_{gt} = \prod_{i=1}^n \left[L_{1i} \prod_{m=2}^M L_{gt(m)i} \right]. \quad (4.13)$$

As for the first event, the likelihood can be restated in terms of the event and observed risk indicators. Let $E_{gt(m)j}$ be the GT event indicator for

the m^{th} event such that

$$E_{gt(m)\gamma} = I(\Delta_m = \gamma). \quad (4.14)$$

Let $R_{gt(m)j}^o$ be the GT indicator of observed risk for the m^{th} event in period j , that is,

$$R_{gt(m)\gamma}^o = I([D_m \geq \gamma \text{ and } \delta_m = 1] \text{ or } [D_m > \gamma \text{ and } \delta_m = 0]). \quad (4.15)$$

The observed data likelihood for the m^{th} event for subject i can be restated in terms of $R_{gt(m)i}^o$ and $E_{gt(m)i}$ by

$$L_{mi} = \prod_{j \in \{r: R_{gt(m)ri}^o = 1\}} P(E_{gt(m)j} = e_{mji}). \quad (4.16)$$

To summarize, the observed data for the m^{th} event (or m^{th} spell), (D_{mi}, δ_{mi}) , as well as the corresponding likelihood, can be restated in terms of $(E_{gt(m)i}, R_{gt(m)i}^o)$ without loss of information, with the following conversion:

$$E_{gt(m)ji} = \begin{cases} 1 & \text{if } D_{mi} = j \text{ and } \delta_{mi} = 1 \\ 0 & \text{otherwise,} \end{cases} \quad (4.17)$$

and

$$R_{gt(m)ji}^o = \begin{cases} 1 & \text{if } (D_{mi} \geq j, \delta_{mi} = 1) \text{ or } (D_{mi} > j, \delta_{mi} = 0) \\ 0 & \text{otherwise.} \end{cases} \quad (4.18)$$

Table 4.3 gives the values for $R_{gt(m)}$ and $E_{gt(m)}$ for example subjects, A, B, and C, for the second and third event.

Table 4.3: Example Data for Second and Third Event in GT

$r_{gt(2),*}^o$	1	2	3	4	5	6	7	8	9
Subject A	1	1	1	0	0	0	0	0	0
Subject B	1	1	1	1	1	1	0	0	0
Subject C	0	0	0	0	0	0	0	0	0

$e_{r_{gt(2),*}^o}$	1	2	3	4	5	6	7	8	9
Subject A	0	0	1
Subject B	0	0	0	0	0	0	.	.	.
Subject C

$r_{gt(3),*}^o$	1	2	3	4	5	6	7	8	9
Subject A	1	1	1	1	1	0	0	0	0
Subject B	0	0	0	0	0	0	0	0	0
Subject C	0	0	0	0	0	0	0	0	0

$e_{r_{gt(3),*}^o}$	1	2	3	4	5	6	7	8	9
Subject A	0	0	0	0	1
Subject B
Subject C

4.3.2 Counting process

In the CP formulation, the time periods of all events after the first are on the same time scale as the first event. Thus, the hazard probability for event m , given $m > 1$, is defined by

$$P_{cp(h_m)}(\gamma) = P(\Gamma_m = \gamma \mid \Gamma_m \geq \gamma, \Gamma_{m-1} \leq \gamma). \quad (4.19)$$

Note the additional conditioning event so that the hazard probability of event m occurring in time period γ is conditional upon not only that event m did not occur before time period γ but also that event $m - 1$ occurred before or during time period γ . This condition is implicit in the gap time formulation in the expression $\Delta_m \geq 0$. It is possible to also define a conditional survival probability, such that

$$\begin{aligned} P_{cp(S_m)}(\gamma) &= P(\Gamma_m > \gamma \mid \Gamma_{m-1} \leq \gamma) \\ &= \prod_{j=\Gamma_{m-1}}^{\gamma} (1 - P_{cp(h_m)}(j)). \end{aligned} \quad (4.20)$$

Although admittedly it is more intuitive to think about survival probabilities than hazard probabilities when conceiving of time-to-event processes, the meaning of survival becomes much more complicated and less intuitive in the recurrent events setting under the CP formulation. The equation given above defines the survival probability for event m conditional on the random time

period for event $m - 1$. It would also be possible to compute the mean conditional survival probability for event m at time period γ by computing a weighted average over all possible time periods for the $m - 1$ event.

Let the observed data for subject i corresponding to the second event be represented by $\{A_{cp(2)i}, \delta_{2i}\}$ where $A_{cp(2)i}$ represents the last time period during which subject i is observed to be at-risk for the second event (in CP time, $A_{cp(2)} = A_2$) and δ_{2i} is the indicator of whether an event ($\delta_{2i} = 1$) or censoring ($\delta_{2i} = 0$) occurred during that final period. The likelihood contribution of subject i relative to the second event is given by

$$L_{cp(2)i} = \left([P_{cp(h_2)}(a_{cp(2)i})]^{\delta_{2i}} \cdot \prod_{j=0}^{a_{cp(2)i}-1} [1 - P_{cp(h_2)}(j)] \right)^{\delta_{1i}}. \quad (4.21)$$

More generally, for any event m , $m > 1$, the likelihood contribution of subject i relative to the m^{th} event is

$$L_{cp(m)i} = \left([P_{cp(h_m)}(a_{cp(m)i})]^{\delta_{mi}} \cdot \prod_{j=0}^{a_{cp(m)i}-1} [1 - P_{cp(h_m)}(j)] \right)^{\delta_{(m-1)i}}. \quad (4.22)$$

The full likelihood for the gap time formulation is then given by

$$L_{cp} = \prod_{i=1}^n \left[L_{1i} \prod_{m=2}^M L_{cp(m)i} \right]. \quad (4.23)$$

As in gap time, the likelihood can be restated in terms of the event and observed risk indicators. Let $E_{cp(m)j}$ be the CP event indicator for the m^{th} event such that

$$E_{cp(m)\gamma} = \mathbf{I}(A_{cp(m)} = \gamma). \quad (4.24)$$

Let $R_{cp(m)j}^o$ be the CP indicator of observed risk for the m^{th} event in period j , that is,

$$R_{cp(m)\gamma}^o = \mathbb{I} \left(\begin{array}{l} [A_{cp(m-1)} \leq \gamma \leq A_{cp(m)} \quad \text{and} \quad \delta_m = 1] \quad \text{or} \\ [A_{cp(m-1)} \leq \gamma < A_{cp(m)} \quad \text{and} \quad \delta_m = 0] \end{array} \right). \quad (4.25)$$

The observed data likelihood for the m^{th} event for subject i can be restated in terms of $R_{cp(m)i}^o$ and $E_{cp(m)i}$ by

$$L_{mi} = \prod_{j \in \{r: R_{cp(m)r}^o = 1\}} P(E_{cp(m)j} = e_{mji}). \quad (4.26)$$

To summarize, the observed data for the m^{th} event (or m^{th} spell), $(A_{cp(m)i}, \delta_{mi})$, as well as the corresponding likelihood, can be restated in terms of $(E_{cp(m)i}, R_{cp(m)i}^o)$ without loss of information, with the following conversion:

$$E_{cp(m)ji} = \begin{cases} 1 & \text{if } A_{cp(m)i} = j \text{ and } \delta_{mi} = 1 \\ 0 & \text{otherwise,} \end{cases} \quad (4.27)$$

and

$$R_{cp(m)ji}^o = \begin{cases} 1 & \text{if } (A_{cp(m-1)i} \leq j \leq A_{cp(m)i}, \delta_{mi} = 1) \text{ or} \\ & (A_{cp(m-1)i} \leq j < A_{cp(m)i}, \delta_{mi} = 0) \\ 0 & \text{otherwise.} \end{cases} \quad (4.28)$$

Table 4.4 gives the values for $R_{cp(m)}$ and $E_{cp(m)}$ for example subjects, A, B, and C, for the second and third event.

Table 4.4: Example Data for Second and Third Event in CP

$r_{cp(2),*}^o$	1	2	3	4	5	6	7	8	9
Subject A	0	0	1	1	1	0	0	0	0
Subject B	0	0	0	1	1	1	1	1	1
Subject C	0	0	0	0	0	0	0	0	0

$e_{r^o cp(2),*}$	1	2	3	4	5	6	7	8	9
Subject A	.	.	0	0	1
Subject B	.	.	.	0	0	0	0	0	0
Subject C

$r_{cp(3),*}^o$	1	2	3	4	5	6	7	8	9
Subject A	0	0	0	0	1	1	1	1	1
Subject B	0	0	0	0	0	0	0	0	0
Subject C	0	0	0	0	0	0	0	0	0

$e_{r^o cp(3),*}$	1	2	3	4	5	6	7	8	9
Subject A	0	0	0	0	1
Subject B
Subject C

4.3.3 Total time

In the TT formulation, the time periods of all events after the first are on the same time scale as the second event; however, the hazard is not conditional upon the time of the prior event. Thus, the hazard probability for event m , given $m > 1$, is given by

$$P_{tt(h_m)}(\gamma) = P(\Gamma_m = \gamma \mid \Gamma_m \geq \gamma). \quad (4.29)$$

So, $P_{tt(h_m)}(j)$ is the probability that the m^{th} event occurs in period j given that it does not occur before period j . The survival probability based on the TT hazard yields the marginal survival probability for each event:

$$\begin{aligned} P_{tt(S_m)}(\gamma) &= P(\Gamma_m > \gamma) \\ &= \prod_{j=1}^{\gamma} (1 - P_{tt(h_m)}(j)). \end{aligned} \quad (4.30)$$

Let the observed data for subject i corresponding to the second event be represented by $\{A_{tt(2)i}, \delta_{2i}\}$ where $A_{tt(2)i}$ represents the last time period during which subject i is observed to be at-risk for the second event (in TT time, $A_{tt(2)} = A_2$) and δ_{2i} is the indicator of whether an event ($\delta_{2i} = 1$) or censoring ($\delta_{2i} = 0$) occurred during that final period. The likelihood contribution of subject i relative to the second event is given by

$$L_{tt(2)i} = \left([P_{tt(h_2)}(a_{tt(2)i})]^{\delta_{2i}} \cdot \prod_{j=0}^{a_{tt(2)i}-1} [1 - P_{tt(h_2)}(j)] \right). \quad (4.31)$$

More generally, for any event m , $m > 1$, the likelihood contribution of subject i relative to the m^{th} event is

$$L_{tt(m)i} = \left([P_{tt(h_m)}(a_{tt(m)i})]^{\delta_{mi}} \cdot \prod_{j=0}^{a_{tt(m)i}-1} [1 - P_{tt(h_m)}(j)] \right). \quad (4.32)$$

The full likelihood for the gap time formulation is then given by

$$L_{tt} = \prod_{i=1}^n \left[L_{1i} \prod_{m=2}^M L_{tt(m)i} \right]. \quad (4.33)$$

As in GT and CP, the likelihood can be restated in terms of the event and observed risk indicators. Let $E_{tt(m)j}$ be the TT event indicator for the m^{th} event such that

$$E_{tt(m)\gamma} = I(A_{tt(m)} = \gamma). \quad (4.34)$$

Let $R_{tt(m)j}^o$ be the TT indicator of observed risk for the m^{th} event in period j , that is,

$$R_{tt(m)\gamma}^o = I([A_{tt(m)} \geq \gamma \text{ and } \delta_m = 1] \text{ or } [A_{tt(m)} > \gamma \text{ and } \delta_m = 0]). \quad (4.35)$$

The observed data likelihood for the m^{th} event for subject i can be restated in terms of $R_{tt(m)i}^o$ and $E_{tt(m)i}$ by

$$L_{mi} = \prod_{j \in \{r: R_{tt(m)ri}^o = 1\}} P(E_{tt(m)j} = e_{mji}). \quad (4.36)$$

To summarize, the observed data for the m^{th} event (or m^{th} spell), $(A_{tt(m)i}, \delta_{mi})$, as well as the corresponding likelihood, can be restated in terms of

$(E_{tt(m)i}, R_{tt(m)i}^o)$ without loss of information, with the following conversion:

$$E_{tt(m)ji} = \begin{cases} 1 & \text{if } A_{tt(m)i} = j \text{ and } \delta_{mi} = 1 \\ 0 & \text{otherwise,} \end{cases} \quad (4.37)$$

and

$$R_{tt(m)ji}^o = \begin{cases} 1 & \text{if } (A_{tt(m)i} \geq j, \delta_{mi} = 1) \text{ or } (A_{tt(m)i} > j, \delta_{mi} = 0) \\ 0 & \text{otherwise.} \end{cases} \quad (4.38)$$

Table 4.5 gives the values for $R_{tt(m)}$ and $E_{tt(m)}$ for example subjects, A, B, and C, for the second and third event.

4.4 Estimation

Similar to what was shown in Chapter 2, the specification of the observed data likelihood in terms of the hazard probabilities is identical to the specification of the observed data likelihood in terms of the event indicators vectors, E_m , e.g., the maximum likelihood estimates for $P(E_{gt(m)j})$, under MAR with $R_{gt(m)j}^o$ as the missingness indicator, are the same as the estimates for $P_{gt(h_m)}(j)$. Using the logistic link as before, each hazard probability can be expressed as a function of observed covariates. Again, as in Chapter 2, using the latent class regression analysis, treating each event indicator as a latent class indicator, the MLE's for the parameters of the LCR and the parameters of discrete-logit

Table 4.5: Example Data for Second and Third Event in TT

$r_{tt(2),*}^o$	1	2	3	4	5	6	7	8	9
Subject A	1	1	1	1	1	0	0	0	0
Subject B	1	1	1	1	1	1	1	1	1
Subject C	1	1	1	1	1	0	0	0	0
$e_{r^o tt(2),*}$	1	2	3	4	5	6	7	8	9
Subject A	0	0	0	0	1
Subject B	0	0	0	0	0	0	0	0	0
Subject C	0	0	0	0	0
$r_{tt(3),*}^o$	1	2	3	4	5	6	7	8	9
Subject A	1	1	1	1	1	1	1	1	1
Subject B	1	1	1	1	1	1	1	1	1
Subject C	1	1	1	1	1	0	0	0	0
$e_{r^o tt(3),*}$	1	2	3	4	5	6	7	8	9
Subject A	0	0	0	0	0	0	0	0	1
Subject B	0	0	0	0	0	0	0	0	0
Subject C	0	0	0	0	0

hazard probability model are equal. Figure 4.2 displays the path diagram for a recurrent event history model with three possible event occurrences and only time-independent covariates. Allowing for event-specific baseline hazard probabilities, which correspond to event-specific thresholds in the regression, and event-specific covariate and covariate effects as well as time-varying covariates and covariate effects, the full and unrestricted regression model for each event indicator is given, using the gap time formulation as an example, by

$$\text{logit}(P(E_{gt(m)j} \mid x_{gt(m)j}, z_m)) = \nu_{gt(m)j} + \beta'_{gt(m)j} z_m + \kappa'_{gt(m)j} x_{gt(m)j}. \quad (4.39)$$

The intercept, $\nu_{gt(m)j}$ is the logit of the baseline GT hazard probability for period j ; that is, $\nu_{gt(m)j}$ is the logit probability for $z_m = 0$ and $x_{gt(m)j} = 0$ that the m^{th} event happens j periods after the $(m-1)^{\text{th}}$ event given that it has not happened before that. $\beta_{gt(m)jp}$ is the log GT hazard odds ratio in period j for a one unit increase in z_{mp} . $\kappa_{gt(m)jp}$ is the log GT hazard odds ratio in period j for a one unit increase in x_{mjp} . The time varying covariates, x , are given in terms of their value on the GT time scale, since the first period of spell m in gap time may be at different locations on the original time scale for different subjects. Notice that allowing event-specific baseline hazard probabilities with event-specific covariate sets and event-specific effects is equivalent to doing a separate analysis for each successive event. An event-stratified analysis is one of the conventional approaches to recurrent events but can be tedious and

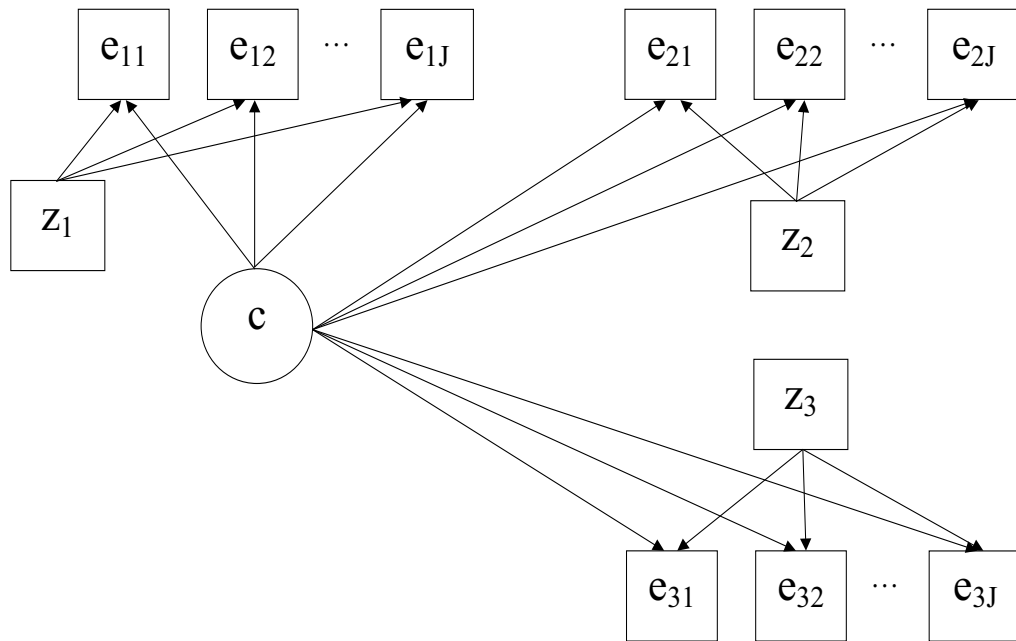


Figure 4.2: Recurrent event history LCR path diagram.

statistically inefficient if the event processes have any parameters in common (Allison, 1984). Also, it does not allow statistical testing of parameter equality across events. The restriction of $\nu_{mj} = \nu_j$ for all m constrains the baseline hazard probabilities to be the same at time period j regardless of the event number¹². Notice that interpretation of ν_j depends on whether the observed data has a GT, TT, or CP formulation. For the full-GT formulation, the restriction to a common baseline hazard would be a reasonable specification if the data did not call for event-specific baseline hazard probabilities. However, for the partial-GT formulation, the restriction of $\nu_{mj} = \nu_j$ for all $m > 1$ is the more reasonable restriction since the first spell is not on the same time scale as the subsequent spells.

Although the specification allows event-specific hazards for all possible M recurrences, there may be a very small number of subjects observed at-risk

¹²The TT formulation with common hazard for continuous-time was proposed by Lee, Wei, and Amato (1992), the LWA model. Both the GT and CP formulations with event-specific hazards for continuous-time were proposed by Prentice, Williams, and Peterson (1981), the PWP-GT and PWP-CP models. The CP formulation with common hazard was proposed by Andersen and Gill (1982), the AG model. Kelly and Lim (2000) also consider the GT formulation with a common hazard, the GT-UR model, and the CP formulation with event-specific hazards, the TT-R model. The GT formulation with event-specific hazard for discrete-time was presented by Willett and Singer (1995).

for and/or experiencing those higher numbered events. In such cases, there are two options for practical analysis to avoid unstable and unreliable event-specific hazards for the larger m : 1) restrict the data to p events where $p < M$, and all those subject who experience p events cease to be at-risk for any event after p ; or 2) constrain the baseline hazard probabilities and covariate effects to be the same across some p to M events, where $p < M$. To be conservative, apply the same guidelines as for single events to each event-specific model.

It is also possible to impose a structure on each event-specific or common hazard across time periods using the η factors as shown in Chapter 2. The restriction of $\beta_{mj} = \beta_j$ for all m constrains the effects of the time-independent covariates to be the same for all events at time period j . Recall that the restriction of $\beta_j = \beta$ for all j constrains the effects of the time-independent covariates to be the same for all time periods—the proportional hazard odds model.

4.4.1 Correcting bias in duration dependence

As discussed in Section 4.2, it is necessary to make an adjustment in the model to account for the average shortened risk duration that is present when allowing risk intervals for consecutive events to overlap. In the GT formulation, there is already an adjustment implicit in the model for two events occurring

in the same interval. That is, allowing for events in time period zero and allowing the baseline probability to be unrestricted by event number, already absorbs the shortened risk duration into the baseline hazard probability estimate. However, there is not an automatic adjustment if more than one prior event occurred in the same time period. Such an adjustment can be made to the baseline hazard probability of event m for $m > 2$ at time period zero by including an indicator variable for all events $1, \dots, (m - 2)$ in the regression model as given below.

$$\begin{aligned} \text{logit}(P(E_{gt(m)0} \mid x_{gt(m)0}, z_m)) &= \nu_{gt(m)0} + \beta'_{gt(m)0} z_m + \kappa'_{gt(m)j} x_{gt(m)0} \\ &+ \sum_{p=1}^{m-2} (\omega_{gt(0)p} \cdot \mathbb{I}[A_p = A_m]). \end{aligned} \quad (4.40)$$

$\omega_{gt(0)p}$ represents the adjustment to the baseline hazard for event p , $p < m - 1$ having occurred in the same interval as event m . $\nu_{gt(m)0}$ is now the logit probability for $z_m = 0$ and $x_{gt(m)j} = 0$ that the m^{th} event happens in the same period as event $m - 1$ given that event $m - 2$ occurred in a prior time period. Because gap time is conditional on each event occurrence, the only adjustment needed is for period zero for each event spell after the second. This is not the case in the CP formulation, where an adjustment must be made to each baseline hazard probability for $m > 1$ at each time period as given below.

$$\begin{aligned} \text{logit}(P(E_{cp(m)j} \mid x_{cp(m)j}, z_m)) &= \nu_{cp(m)j} + \beta'_{cp(m)j} z_m + \kappa'_{cp(m)j} x_{gt(m)j} \\ &+ \sum_{p=1}^{m-1} (\omega_{cp(j)p} \cdot \mathbb{I}[A_p = j]). \end{aligned} \quad (4.41)$$

$\omega_{cp(j)p}$ represents the adjustment to the baseline hazard for event p , $p < m$, having occurred in the same time period, j . Note that including a set of time-dependent indicator variables for the time periods of all prior events makes the inclusion of the ω terms redundant. No such adjustment is necessary for the TT models since risk for event m is not conditional upon the occurrence of the $(m - 1)^{th}$ event—all risk intervals begin at the same time, $t = 0$.

Within the GT or CP formulations, the need for an unrestricted baseline hazard can be evaluated by comparing the event-specific and common baseline hazard models using the LRT.¹³ In addition, equivalent models can be obtained across the GT and CP formulation by including covariates in each model corresponding to the event history in the alternate time scale.¹⁴ For example, in the CP formulation, a time-varying covariate representing the

¹³Care should be taken in interpreting the meaning of common baseline hazard probabilities when the bias adjustment given above has been included in the model.

¹⁴Such equivalence does *not* exist in the continuous-time setting using partial likelihood methods because the baseline hazard, on whatever time scale is chosen, is not explicitly estimated.

number of time periods from the prior event can be included to account for the possible hazard dependence on the time since last occurrence. In the GT formulation, a time-independent covariate representing the time period of the prior event in terms of the original time scale can be included to account for the possible hazard dependence on the time since the onset of risk for the very first event. Thus, the choice between the GT and CP formulations is a matter of preference and is analogous, for example, to the decision of how and whether to center covariates in a multilevel linear model—the likelihood will be the same but the parameter estimates and their corresponding interpretations will change. In applying these models, a researcher may want to fit models to more than one time-scale specification, careful to note that the parameters, both baseline hazard probabilities and covariate effects, have different interpretations depending on the time scale.

The TT model is not comparable to the GT and CP models in that it provides marginal rather than conditional estimates of event likelihood across time. If one uses a TT formulation, it is important to recognize how the covariate effects may present in such a model. Suppose that there is a covariate that only predicts time to the first event but not the time to the second event or beyond conditional on the occurrence of the first event. As demonstrated clearly in the simulation examples by Kelly and Lim (2000), when there is a

covariate effect for only the first event, the TT models display what they term a “carry-over” effect. This should be expected given the marginal nature of the TT model. If there is a positive covariate effect on the first event, it makes all subsequent events, marginally speaking, more likely as well. Even when the covariate effect is constant across events and the spells are independent within subjects, the TT models can overestimate the treatment effect.

4.5 Unobserved heterogeneity

One of the important features of recurrent events data named at the beginning of this chapter was within-subject correlation (Feature 3). Up until this point, all of the model specifications given assumed that event times across individuals in the sample as well as within individuals were independent. In Chapter 3, the problems of ignoring unobserved heterogeneity in survival analysis were demonstrated. Those issues are just as relevant, if not more so, for analysis of recurrent events data; that is because, as previously mentioned, it is unlikely that the correlation between spell durations for a given individual are completely explained by observed covariates. Not only will the standard errors on parameters be underestimated if all spells are treated as distinct and independent observations, but the hazard probabilities and covariate effects

will themselves be biased¹⁵. Allison (1995) mentions that using the time from or time of the prior occurrence as a way to detect dependence between spells within subjects in models that treat each spell as a distinct observation¹⁶ but warns that this approach does not necessarily correct the resultant biases. In the marginal approach proposed by Wei, Lin, and Weissfeld (1989), events within individuals are treated as independent and then a robust variance estimate is obtained using a “sandwich estimator”. Allison (1995) notes that this approach does not correct for biases in the actual hazard or coefficient estimates. Also, Kelly and Lim (2000) demonstrated through simulations that the robust variance estimate did not adequately account for the within-subject correlations.

Another approach to dealing with the within-subject correlation is random effects models, also called frailty models, similar to what was introduced in Chapter 3. Because there are multiple events observed for each individual in the sample, identification of models that explicitly contain a random effect in

¹⁵Kelly and Lim (2000) showed in their simulation studies that for data generated from populations with a constant treatment effect across events and correlated event times within subjects that the estimated treatment effect was attenuated in models that assumed spell independence

¹⁶Dependence between observations within subjects is ignored in the Willett and Singer (1995) recurrent event model for discrete-time.

the model to account for the unobserved heterogeneity may not require model constraints or covariate information as was necessary in the single event case. The challenges of model identification depend, in large part, on the constraints imposed in the model. For example, since allowing event-specific baseline hazard and covariate effects is equivalent to fitting separate, single event models for each event number, the identification issues are the *same* as in the single event case. If constraints such as a common baseline hazard across event numbers are imposed, identification of the unobserved heterogeneity model is simplified. Steele (2003) has taken the hierarchical nonlinear modeling approach, as available through the MLWiN software application, treating spells as clustered within individuals, i.e., spells are the level 1 units and subjects are the level 2 units (similar to the hierarchical modeling approach to repeated measures data). This specification assumes a constant baseline hazard across events with a mean shift possible by including the number of prior events as a covariate. This framework allows for random coefficients on the covariates as well as random effect. All random terms are assumed to be normally distributed. Hence, this approach is susceptible to the same shortcoming as parametric models of unobserved heterogeneity in the single event setting: parameters estimates are sensitive to misspecification of the frailty distribution. Also, this approach makes the assumption that a subject's frailty is the *same*

for each event, i.e., if a subject is at increased susceptibility for the first event, she has the same increased susceptibility for the second event, and so on. Using the LCR framework allows not only nonparametric specification for the frailty distribution, it allows event-specific hazard, covariate effects, *and* frailties through intercepts and coefficients specified by event number and latent class.

Surprisingly, little can be found in the literature extending the Heckman and Singer (1984a) concept of nonparametric modeling of unobserved heterogeneity to multivariate survival data. Vermunt (1997) does offer some discussion of latent classes as a way to account, nonparametrically, for both between and within subject time dependencies but does not focus on recurrent events as much as the more general multiple event models. Given that specification of all three recurrent event time formulations can be fit into the LCR framework, as shown in the previous section, it is possible to specify any of those models with $K > 1$. As in Chapter 3, it is recommended that the nature of the classes be data driven rather than determined a priori, e.g., not presupposing a long-term survivor class, members of which would have a zero hazard probability for all events and all time periods. Extending to a LCR model with $K > 1$, the regression for E on the covariates and C is then given,

using the GT formulation for example, by

$$\text{logit}(P(E_{gt(m)}[j] | C = k, x_{gt(m)j}, z_m)) = \nu_{k,gt(m),j} + \beta'_{k,gt(m),j} z_m + \kappa'_{k,gt(m),j} x_{gt(m)j}, \quad (4.42)$$

and, as before, the regression C is given by

$$P_C(k | z) = \frac{\exp(\alpha_{0k} + \alpha'_k z)}{\sum_{m=1}^K \exp(\alpha_{0m} + \alpha'_m z)}. \quad (4.43)$$

4.6 Example

Consider again the example begun in Chapter 2 regarding domestic violence. In the previous chapters, the time to first domestic violence was modeled as a single, nonrecurrent event.

In the sample, 69 out of the 170 men have at least one violent episode during the 12 month post treatment period. Forty-three have a second episode and 26 have a third episode. The maximum number of observed episodes during the observation period was seven but since only 26 are at-risk for a fourth episode or beyond, only the hazards for the first three episodes will be modeled here. Tables 4.6–4.8 display the number at-risk for each event during each of the six time periods as well as the number of occurrence in each period and the ratio of events to number at-risk. These proportions correspond to the sample hazard probability estimates in each time formulation.

Table 4.6: Sample Frequencies and Proportions for First, Second, and Third Violence Episodes in GT Formulation

	Months	1-2	3-4	5-6	7-8	9-10	11-12
First episode	# at-risk	170	136	123	114	108	104
	# of events	34	13	9	6	4	3
	Hazard	0.20	0.10	0.07	0.05	0.04	0.03
	Period	1	2	3	4	5	6
Second episode	# at-risk	69	51	31	24	21	14
	# of events	16	18	5	2	1	1
	Hazard	0.23	0.35	0.16	0.08	0.05	0.07
Third episode	# at-risk	43	30	22	17	12	5
	# of events	12	8	3	1	2	0
	Hazard	0.28	0.27	0.14	0.06	0.17	0.00

Table 4.7: Sample Frequencies and Proportions for First, Second, and Third Violence Episodes in CP Formulation

	Months	1-2	3-4	5-6	7-8	9-10	11-12
First episode	# at-risk	170	136	123	114	108	104
	# of events	34	13	9	6	4	3
	Hazard	0.20	0.10	0.07	0.05	0.04	0.03
Second episode	# at-risk	34	39	41	35	32	29
	# of events	8	7	12	7	6	3
	Hazard	0.24	0.18	0.29	0.20	0.19	0.10
Third episode	# at-risk	8	14	24	28	29	26
	# of events	1	2	3	5	6	9
	Hazard	0.13	0.14	0.13	0.18	0.21	0.35

Table 4.8: Sample Frequencies and Proportions for First, Second, and Third Violence Episodes in TT Formulation

	Months	1-2	3-4	5-6	7-8	9-10	11-12
First episode	# at-risk	170	136	123	114	108	104
	# of events	34	13	9	6	4	3
	Hazard	0.20	0.10	0.07	0.05	0.04	0.03
Second episode	# at-risk	170	162	155	143	136	130
	# of events	8	7	12	7	6	3
	Hazard	0.05	0.04	0.08	0.05	0.04	0.02
Third episode	# at-risk	170	169	167	164	159	153
	# of events	1	2	3	5	6	9
	Hazard	0.01	0.01	0.02	0.03	0.04	0.06

The analysis of the recurrent events was done for each formulation and then the results compared. Recall that each formulation deals with a different angle or conception of the process and should be compared not with the goal of choosing one over the others but with the goal of assimilating complementary information from different models. The general analysis strategy was the same across the three formulations and parallels analysis strategy for multiple group analyses: 1) Fit models for each episode separately, investigating covariate effects and hazard structure; 2) Combine models for each episode into a single model, investigating equality of parameters across episodes; and 3) Fit a series of mixture models, investigating the presence and influence of unobserved heterogeneity, including dependence between spell times within subjects.

4.6.1 Gap time analysis

Tables 4.9 and 4.10 gives the results of the separate models for the second and the third episode in the gap time formulation.¹⁷ For the second episode, the only predictor that had a significant effect was the length of the relationship. There was a nonlinear effect modeled by including a mean-centered squared

¹⁷Only three subjects experienced the second and the third episodes in the same time period so no bias adjustment was included in the GT model for the third event. By similar reasoning, no adjustment was included in the CP models for the second and third events.

Table 4.9: Results for Data Example Model 8a: Second Episode, GT Formulation

Parameter	Est.	SE	Est./SE
Length of relationship	-0.03	0.02	-1.06
(Length of relationship) ²	-0.006*	0.003	-2.01
E(η_0)	-0.23	0.33	-0.71
E(η_1)	-0.33*	0.12	-2.68

LL=-98.55, parameters=4

* $p < 0.05$

term in the model. Figure 4.3 shows the contribution to the logit hazard probabilities by length of relationship. From the coefficients on the linear and quadratic term, length of relationship is protective but becomes less so from one to nine years in relationship length and then becomes increasing protective from 9 years upward. A length of relationship of one year and 17 years have a hazard odds ratio of one. Figure 4.4 shows the estimated hazard and survival probabilities for select values of relationship length. There was no evidence of time-dependent effects for length of relationship. A linear structure adequately modeled the baseline hazard.

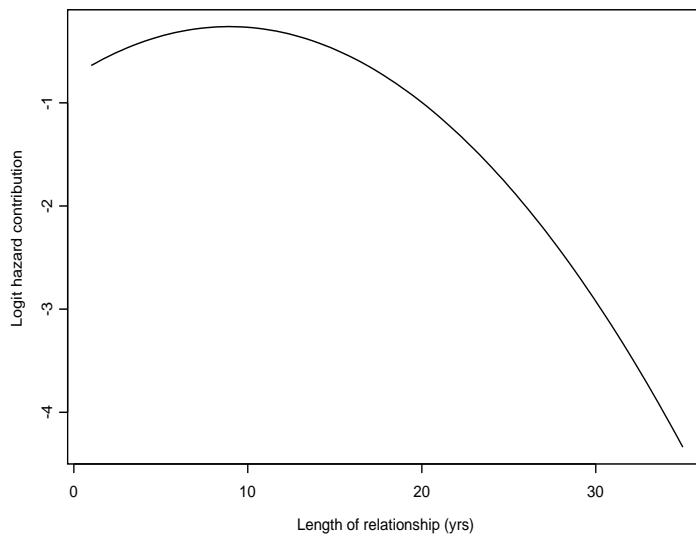


Figure 4.3: Estimated contribution of length of relationship (in years) to the logit hazard probabilities of second violent episode.

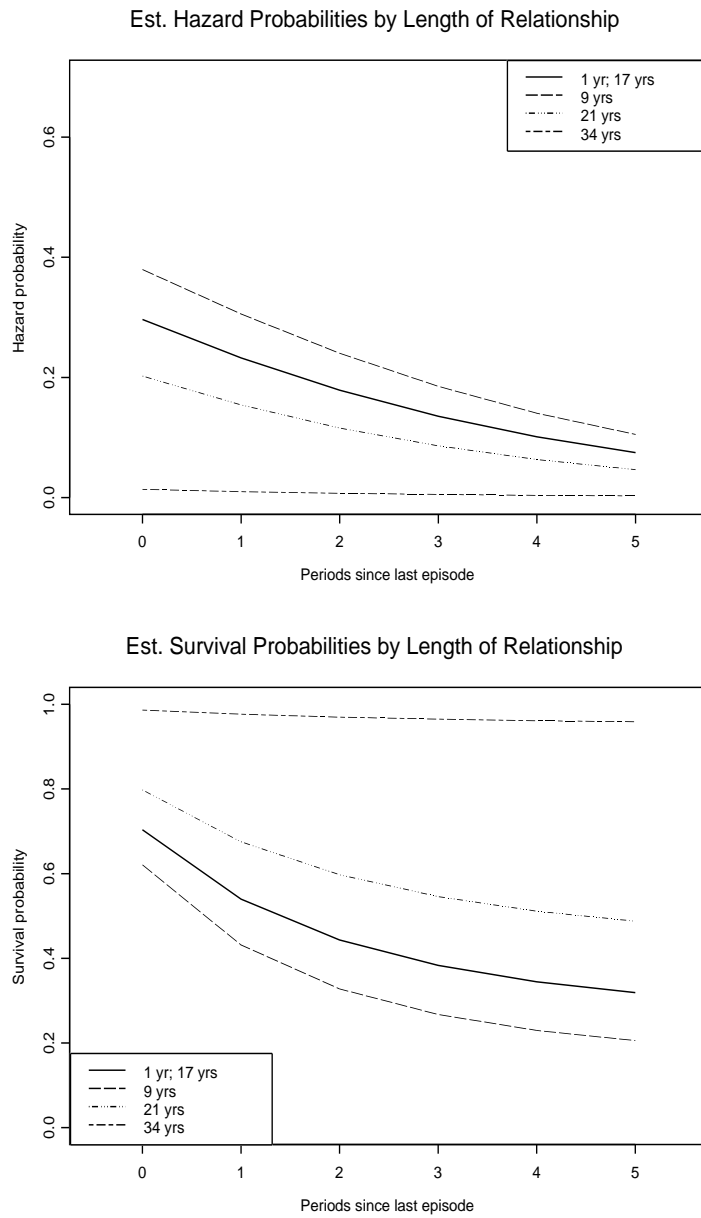


Figure 4.4: Model 8a estimated hazard and survival probabilities for second episode by length of relationship.

For the third episode, the pre-treatment level of violence and the actual timing of the second episode, relative to the end of treatment, were significant predictors of time-to-violence. Pre-treatment violence was found to have a nonlinear effect, with those having more than four reported episodes being at increased risk at all time periods relative to those reporting four or less episodes in the three months preceding treatment. The actual time between the first and the second episode was *not* predictive of the time to the third episode but the time from end of treatment to the second episode was, with the hazard increasing as the time from treatment increased. Figure 4.5 displays the estimated hazard and survival probabilities for pre-treatment violence levels and for all six values of the second episode timing. Only values of the hazard and survival in the observed range were plotted, i.e., for a second episode occurring in third (5-6 months) post-treatment period, only a maximum of four gap time periods of risk would have been observed. There was no evidence of time-dependent effects for either pre-treatment violence or timing of second episode. The baseline hazard was adequately modeled by a constant value.

Models for the first, second, and third episode were combined into one (see Model 9 in Appendix B). However, since there are no shared covariates across the episode and the structure of the baseline hazard for the second and the third episode are different, there were no equality constraints to be

reasonably tested. There is clearly a different process determining time to first episode of violence in the post-treatment period than the process determining the time between the first and second episode or the process determining the time between the second the third episode. It is important to remember that the risk sets of men for the second and third episode are conditional on the occurrence of a prior event. Drinking levels are not predictive of the gap times for the second and third event. That does not mean these men are not drinking. Those who drink are more likely to end up in the risk set for the second episode because they are more likely to return to violence. It suggests that once these men return to violence and to drinking, drinking is *not* predictive of when they will offend again. For the third episode, the fact that a higher level of pre-treatment violence increases the hazard suggests that these men may begin to fall back into their pattern of more frequent violent behavior once they have committed at least two offenses within the first year. The increased hazard for the third episode due to the time since treatment for the second episode suggests that there may be an overall impact of treatment (of any kind) for these men. Certainly, the overall rate of violence in the sample is much lower than pre-treatment considering that all of the subject had at least one reported episode of violence in the three months preceding treatment but only 41% return to violence in the 12 months following treatment. Among those who

Table 4.10: Results for Data Example Model 8b: Third Episode, GT Formulation

Parameter	Est.	SE	Est./SE
I(Pre-tx violence > 4)	2.10*	0.68	3.09
Time of 2 nd episode	1.00*	0.24	4.12
ν	-4.86	1.03	-4.74

LL=-49.06, parameters=3

* $p < 0.05$

do return to violence, this relationship to the time since treatment may imply that treatment may temporarily diffuse the frequency of violence among those who do return to that behavior but that effect dissipates as they move further away from treatment in time. None of these differential covariate relationships could be explored in a recurrent events model that assumed a common hazard and/or common covariate effects across the first three occurrences.

4.6.2 Counting process analysis

Tables 4.11 and 4.12 give the results for the separate models for the second and third episode in the counting process formulation. The most striking

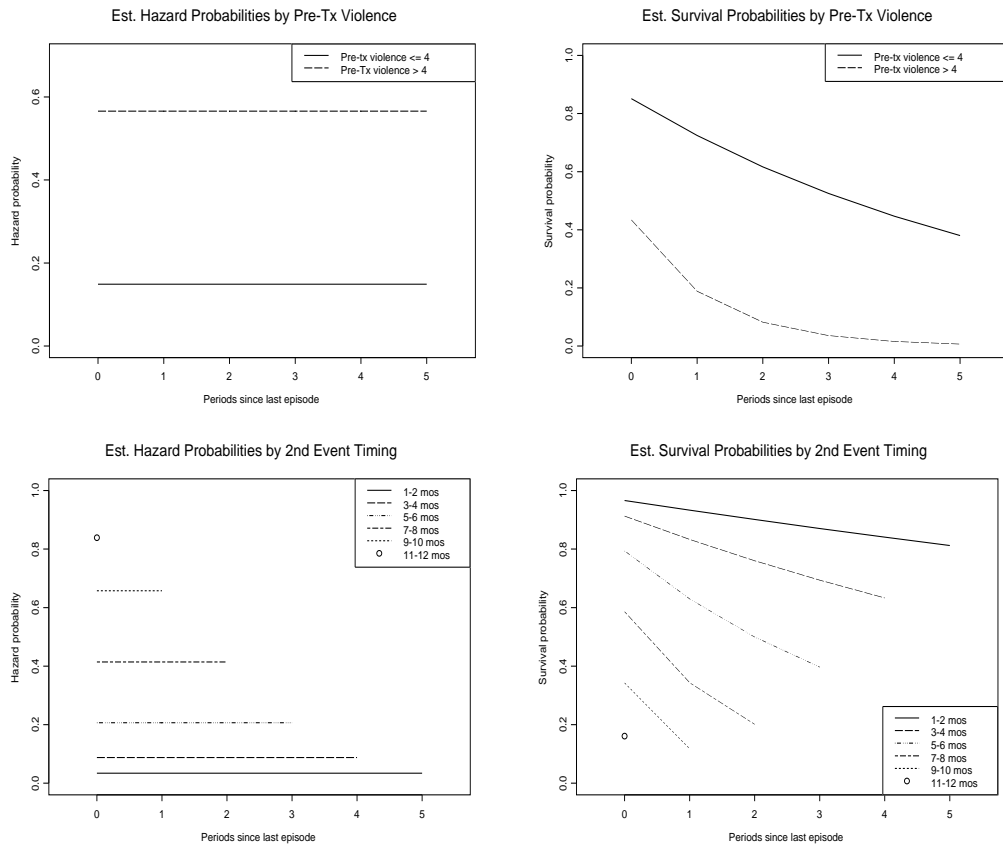


Figure 4.5: Model 8b estimated hazard and survival probabilities for third episode by pre-treatment violence and timing of second episode.

thing to observe is that the model results for the second episode in the CP version are essentially identical to the results for the second episode in the GT formulation (Model 8a). Note that the likelihood and number of free parameters are the same. The quadratic relationship between time-to-violence and length of relationship was found. Including a time-varying covariate of time since first episode captures the linear duration dependence estimated for the baseline hazard in the gap time model. This illustrates the point that the CP and GT models can model the same time dependence processes, gap time with an observation-timeline covariate and counting process with a time-varying time-since-last-event covariate. A similar relationship is seen for the results of the third episode. The increasing linear trend in the baseline hazard of the counting process model mirrors the linear effect of the timing of the second episode found in the gap time model. As with the GT combined model, there were no reasonable equality constraints to be made across the hazards for the three episodes. Figure 4.6 displays the estimated hazard and survival probabilities at the mean covariate values. Since the hazard probabilities for the second and third episode are also in terms of the original observation timeline, they can be plotted on the same scale as the first episode (unlike the partial gap time model). The plots show the hazard for the second episode greater than the first across time with the hazard for the third beginning

Table 4.11: Results for Data Example Model 10a: Second Episode, CP Formulation

Parameter	Est.	SE	Est./SE
Length of relationship	-0.03	0.02	-1.06
(Length of relationship) ²	-0.006*	0.003	-2.01
Time since 1 st episode	-0.33*	0.12	-2.68
ν	-0.23	0.33	-0.71

LL=-98.55, parameters=4

* $p < 0.05$

lower than those for the two prior episode but then increasing dramatically over time. Recall that these are conditional plots—the low hazard for the third episode in the initial time periods indicate that it is unlikely that subject with a first and second episode in those early period have a third in so near a time to treatment. However, as time from treatment increases, the likelihood of a third episode in the later periods given first and second offense in any of the prior periods is quite high.

Table 4.12: Results for Data Example Model 10b: Third Episode, CP Formulation

Parameter	Est.	SE	Est./SE
I(Pre-tx violence > 4)	2.15*	0.72	2.99
Time since 2 nd episode	-0.99*	0.26	-3.82
E(η_0)	-4.18	1.10	-3.79
E(η_1)	1.08	0.30	3.61

LL=-49.34, parameters=4

* $p < 0.05$

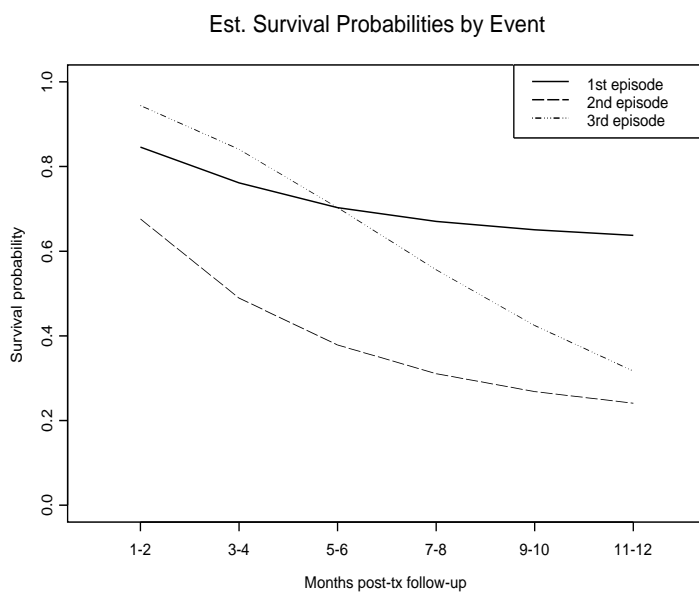
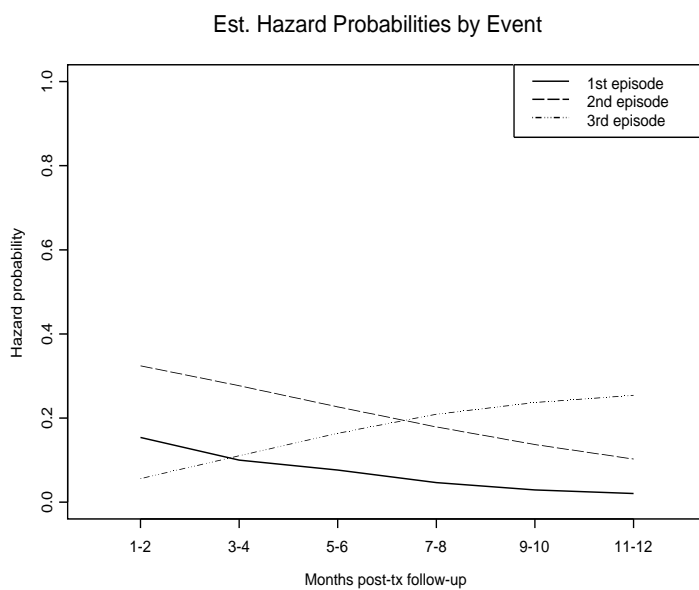


Figure 4.6: Model 10 estimated hazard and survival probabilities for episodes 1–3.

4.6.3 Total time analysis

Table 4.13 gives the results for the combined model in the TT formulation. Recall that unlike the conditional GT and CP models, the TT model is a marginal model and different results were expected. The separate models for the first, second, and third episode were combined (Model 5, 12a, and 12b, respectively) and several equality constraints were made for parameters across the three events. For the second and third episode, percent-days-drinking was predictive of time-to-violence and the effects was found to be essentially the same for episodes 1–3. This is the expected “carry-over” effect mentioned previously. Because drinking is predictive of the time to first episode, it is predictive of the time to all subsequent episodes, marginally speaking. There was also a carry-over effect for household income for the first to the second episode but not the third. There was no carry-over effect for wife’s education. The time-varying covariate of time since the previous episode was found to have a significant effect on both the second and third episode and that effect was equivalent. A positive linear effect of pre-treatment violence was found for the second and the third episode. The baseline hazard for the first and second episode were adequately fit with a linear structure—the intercepts for the two events were different but the slopes were found to be equivalent and decreasing; the baseline hazard for the third episode was modeled as a constant.

Figure 4.7 displays the hazard and survival probabilities at the mean covariate levels. These plots look quite different than the conditional plots from the GT and CP models. Since these are marginal probabilities, the hazards for the second and the third episodes are lower than those for the first episode across all time periods but are close to each other. The survival probabilities plot illustrates that almost 60% of the sample does not return to violence in the first year. A much smaller percent fails to survive having two or three episodes of violence but the survival curves for the second the third episode are close together, suggesting that many of those who have a second episode have a third episode, more so than first to second episode.

4.6.4 Mixture model analysis

The modeling of multiple latent classes was conducted with only the gap time formulation; the counting process model has already been shown to be equivalent to gap time model and the total time includes carry-over effects that could confound the detection of unobserved heterogeneity. Table 4.14 gives the results of a two class model, allowing each episode's baseline hazard as well as covariate effects to differ across classes. For the quadratic term for length of relationship and timing of the second episode, allowing class varying effects resulted in model non-convergence. Wife's education level was the only

Table 4.13: Results for Data Example Model 14: Combined Model, TT Formulation

First episode parameters	Est.	SE	Est./SE
I(Wife's educ. \leq H.S.)	-0.67*	0.27	-2.45
I(Income $>$ \$35K)	-0.69*	0.30	-2.32
% days drinking	2.19*	0.34	6.50
E(η_0)	-1.69*	0.22	-7.72
E(η_1)	-0.47*	0.09	-5.37
Second episode parameters	Est.	SE	Est./SE
I(Income $>$ \$35K)	-0.69*	0.30	-2.32
Length of relationship	-0.03	0.03	-1.14
(Length of relationship) ²	-0.003	0.003	-1.02
Pre-tx violence	0.05*	0.02	2.08
% days drinking	2.19*	0.34	6.50
Time since 1 st episode	0.65*	0.07	9.34
E(η_0)	-2.51	0.42	-5.93
E(η_1)	-0.47	0.09	-5.37
Third episode parameters	Est.	SE	Est./SE
Pre-tx violence	0.05*	0.02	2.08
% days drinking	2.19*	0.34	6.50
Time since 2 nd episode	0.65*	0.07	9.34
ν	-4.67	0.27	-17.05

LL=-430.31, parameters=11

* $p < 0.05$

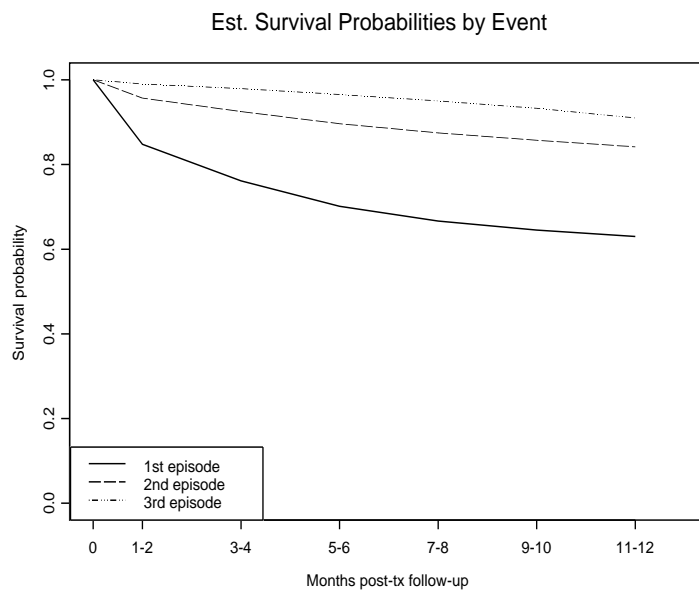
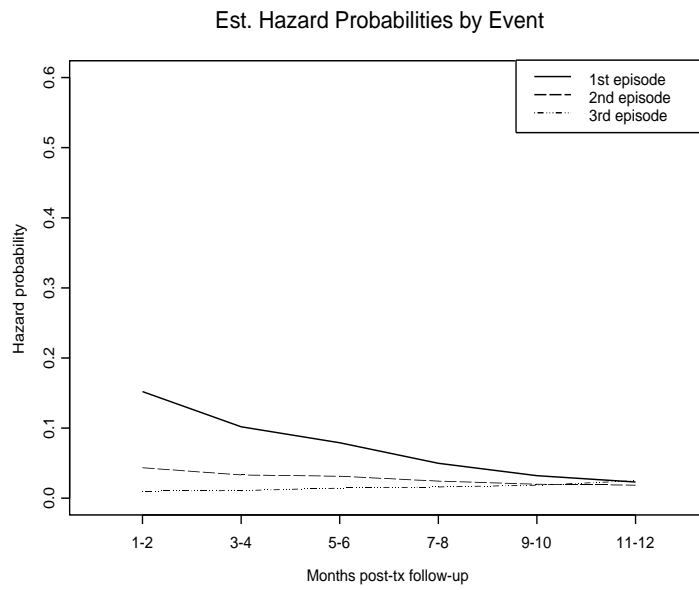


Figure 4.7: Model 14 estimated hazard and survival probabilities for episodes 1–3.

predictor of class membership. The class distribution and parameter estimates for the first episode suggest that the estimation of heterogeneity in the form of latent classes is most driven by the first episode which is not surprising considering there are no shared parameters across the episodes and the first episode model contains the greatest number of observations. Also, as in the mixture model for the first episode, the data does not evidence a two class model. However, examining the results in the table as well as the class plots for each episode in Figure 4.8 suggest that there is heterogeneity in underlying susceptibility to violence over time as well as heterogeneity in the effect of the protective and risk factors for time-to-violence across episodes.

Table 4.14: Results for Data Example Model 15

Class 1 = 32%

Class 2 = 68%

1 st episode parameters	Class 1			Class 2		
	Est.	SE	Est./SE	Est.	SE	Est./SE
I(Wife's educ. ≤ H.S.)	2.78	1.69	1.65	-1.47*	0.61	-2.42
I(Income > \$35K)	1.58	0.80	1.98	-3.17*	1.01	-3.12
% days drinking	3.00	2.06	1.46	3.71*	1.13	3.29
E(η_0)	-1.91*	0.70	-2.75	-2.51*	0.46	-5.40
E(η_1)	-0.90*	0.29	-3.14	-0.12	0.14	-0.90

2 nd episode parameters	Class 1			Class 2		
	Est.	SE	Est./SE	Est.	SE	Est./SE
Length of relationship	-0.01	0.06	-0.19	-0.03	0.06	-0.58
(Length of relationship) ²	-0.01	0.005	-1.90	-0.01	0.005	-1.90
E(η_0)	-0.47	0.78	-0.60	0.05	0.54	0.10
E(η_1)	-0.71*	0.29	-2.43	0.42	0.27	1.55

3 rd episode parameters	Class 1			Class 2		
	Est.	SE	Est./SE	Est.	SE	Est./SE
I(Pre-tx violence > 4)	1.56	1.50	1.05	2.72*	0.99	2.76
Time of 2 nd episode	1.24*	0.31	4.03	1.24*	0.31	4.03
ν	-4.58	1.10	-4.18	-6.13	1.40	-4.39

Class regression parameters	Class 1			Class 2		
	Est.	SE	Est./SE	Est.	SE	Est./SE
α_{01}	-0.21	0.53	-0.40	@0	–	–
$\alpha_{I(\text{Wife's educ.} \leq \text{H.S.}),1}$	-1.58*	0.76	-2.08	@0	–	–

LL=-329.17

parameters=24

* $p < 0.05$;

@ = "fixed at"

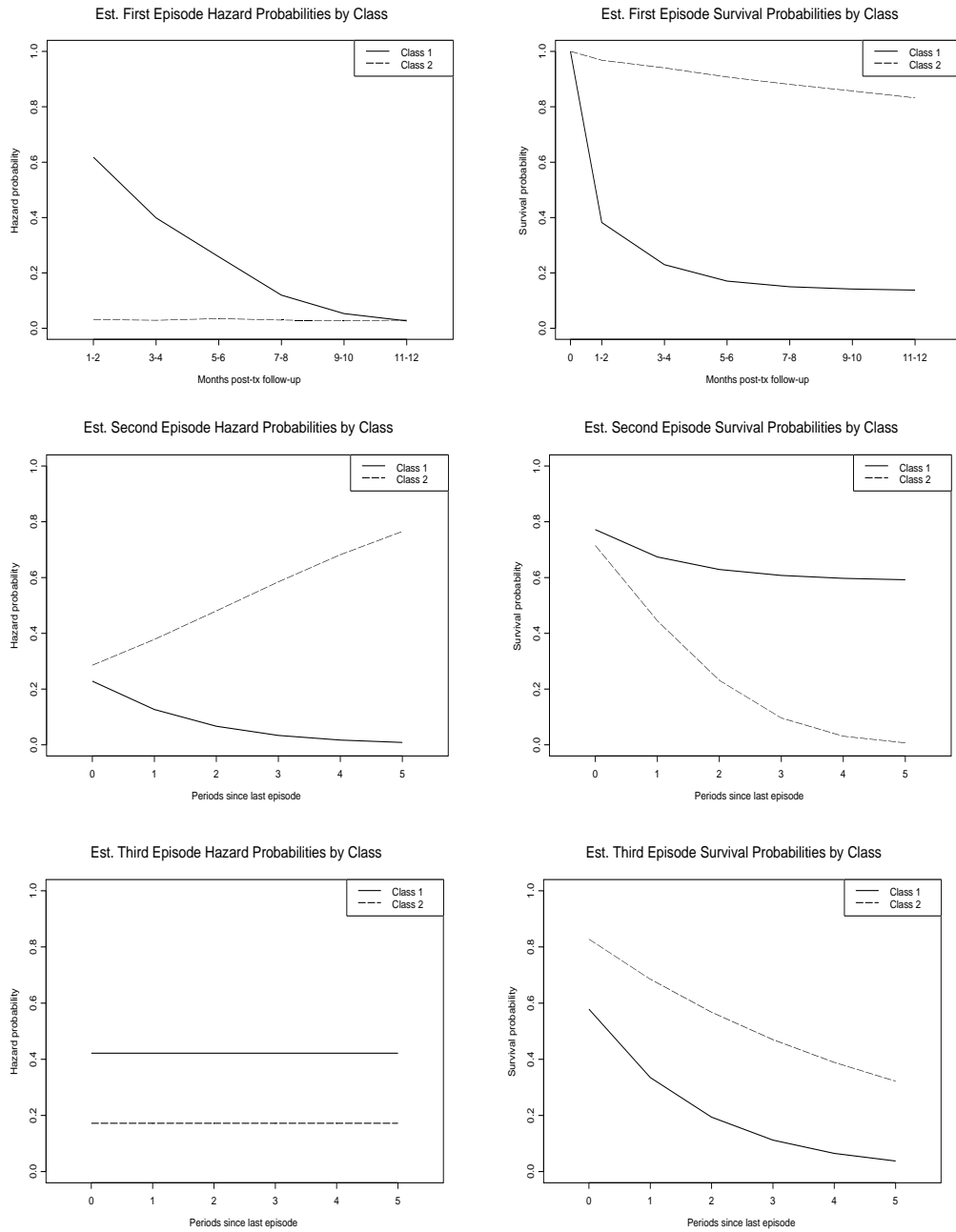


Figure 4.8: Model 15 estimated hazard and survival probabilities for first, second, and third episode by latent class.

Chapter 5

Conclusion

This chapter summarizes the material presented in this dissertation and discusses the limitations of the methods herein. It also provides a roadmap for future methodology development in the area of discrete-time survival analysis using a latent variable framework.

5.1 Single events

This dissertation built on the work of Muthén and Masyn (2001), demonstrating the specification and estimation of a single event discrete-time survival model using latent class regression. It was shown in the case of complete and noninformative right-censored data that the maximum likelihood estimates for the probability of the event indicators (treated as binary latent class indicators

with $K = 1$) under the assumption of Missing-at-Random (MAR), with non-missingness in each time period corresponding to observed risk, were equal to the MLE's of the hazard probabilities for each time period. It was also shown how to model the relationship between the hazard probabilities for each time period and both time-independent and time-dependent covariates using a logit link function.¹ Testing of certain modeling constraints, such as the proportionality of the hazard odds ratio and the time-independence of covariate effects, was explained, as was the testing of different structures that can be imposed on the baseline hazard probabilities.

Discrete-time survival analysis allows researchers with event history data to model the time-to-event process, investigating how risk for an event may change over time and how risk over time may be influenced by both time-independent and time-dependent covariates. As stated in Chapter 2, the LCR framework for discrete-time survival models does not offer anything new, per se, with regards to specification or estimation—the LCR for single events

¹There was also a discussion about the choice of link functions with the note that nothing about the LCR modeling framework favored one link function over another except in the practical sense, in that the logit link is the one utilized by the Mplus software. Also, the logit model specifically may be sensitive to interval length, but it was posited that this sensitivity may have little practical impact on model estimation; that supposition, however, needs further substantiation.

is equivalent to the discrete-time model specified in the logistic regression framework. However, the LCR framework does readily allow for many more complex model extensions than does the traditional logistic regression. Some of these extensions were explored in Chapters 3 and 4.

The model for single event discrete-time survival analysis explicated in Chapter 2 has the same limitations as any of the alternate discrete formulations. If one believes that there is an underlying continuous-time process, there is an untestable assumption about the constancy of the hazard rate within each time interval. Also, the model does not explicitly allow for changes in a time-dependent covariate *within* a given time period. The issue of reciprocal causation in modeling and interpreting the effects of time-dependent covariates, present in all longitudinal models, is complicated in the discrete-time setting because of the grouped-time structure of the data. In the area of model estimation, a discrete-time model with zero observed events in a given time period is not identified unless the time period is combined with an adjacent period, essentially constraining the baseline hazard probabilities and covariate effects to be the same in those periods. There has been very little work done on model assessment in terms of goodness-of-fit tests and power calculations for discrete-time models. This dissertation refers to the existent literature, primarily from the area of categorical data analysis and logistic regression,

but does not offer anything beyond these current conventions. Finally, although other forms of censoring and truncation are addressed, the assumption of noninformative right-censoring is applied throughout the dissertation in the discussion of model specification and estimation. An iterative algorithm for model estimation with double-censored data is proposed but its properties, such as self-consistency, are not established nor is the implementation of the algorithm demonstrated.

5.2 Unobserved heterogeneity

The literature on unobserved heterogeneity and unmeasured covariates in continuous time, often found under the topic of *frailty* models, was reviewed. In the survival data setting, ignoring unobserved heterogeneity can lead to biases in the estimated baseline hazard probabilities, covariate effects, and even spurious covariate time-dependent effects depending on the relationship of the unobserved measures to the survival process and the observed covariates. Several simulated examples were presented in Chapter 3 to illustrate the potential dangers of failing to account for unobserved heterogeneity. The recommendation of modeling unobserved heterogeneity using finite mixture models in the continuous-time setting to counteract model sensitivity to misspecification of the distribution of the unmeasured covariates, was implemented in the discrete-

time LCR framework by allowing the number of latent classes to be greater than one. Attention was given to the issue of identification which can prove challenging with single-event data. A discrete-time survival mixture model with unstructured baseline hazard probabilities is not identified without at least one measured covariate. There was also a discussion on the matter of class enumeration, that is, assessing the empirical evidence in differential support of mixture models with increasing numbers of classes. The AIC, BIC, and G^2 were compared for a set of simulated examples to which correctly specified and misspecified models were fit. It was shown that misspecifying a model, particularly by allowing only indirect effects of a measured covariate on the hazard probabilities through the latent class variable, could lead to incorrect model selection with respect to the class enumeration as well as biases in the estimated covariate effects. As a related issue, it was also shown that presupposing a class number and structure, such as is done with long-term survivor models, can lead to biases in the model estimation. From the simulated examples presented, it seems advisable that a long-term survival model is tested against an unrestricted two-class model rather than assuming it to be correct.

The dangers of ignoring unobserved heterogeneity are clear. One of the primary advantages to specifying discrete-time survival models in the LCR framework is the straightforward extension that can be made to a multi-class

mixture model. By allowing more than one latent class, the model can accommodate the possibility of individual variability in not only baseline hazard probabilities, through class-specific thresholds, but also variability in the effects of the measured covariates, through class-specific regression coefficients. The model also allows measured covariates to influence the distribution of individual frailty, through the regression of the latent class variable on observed time-independent covariates. Such a flexible model with no distributional assumptions about the unobserved heterogeneity guards against problems of misspecification. The discrete-time survival mixture model allows researchers to explore not only the mean survival process but to also better understand how overall event susceptibility as well as susceptibility to various risk and protective factors may vary within the population. And such differences in susceptibility may also depend on a set of measured covariates. This is a clear improvement over models that require multilevel data (e.g., individuals clusters in observed groups such as classrooms or families) or a restricted specification of the unobserved heterogeneity (e.g., long-term survivor models) to account for unobserved heterogeneity.

The discrete-time survival mixture model for non-clustered single event data is limited by the requirements for identification. Depending on the number of time periods, number of observed events, total sample size, number of

measured covariates, and the nature of the unobserved heterogeneity, the LCR framework allows a more unrestricted model than may be theoretically or empirically identified for a given sample. One could argue that if a particular mixture model is not empirically identified that the bias present in a more restricted model, even if misspecified, may be negligible, practically speaking. However, this matter needs to be more fully explored. Also, there is a limitation in the methods currently available for class enumeration. The use of the AIC, BIC, and G^2 was demonstrated in Chapter 3 but the performance of these indices and statistical test along with others available for mixture modeling in different settings, such as the Lo, Mendell, and Rubin LRT, was not systematically evaluated in the discrete-time setting. Until such a time as there is a reliable test for the number of classes, substantive knowledge must be used in combination with empirical evaluation to inform the mixture model specification. The need for more methods related to models assessment and power calculations in the non-mixture case applies to the mixture models as well. This dissertation, although detailing the motivation and specification of discrete-time mixture models does not establish a complete set of “best practices” for model building and assessment.

5.3 Recurrent events

Although there is some literature about recurrent event models in discrete-time, there is no thorough review of the different approaches in modeling recurrent event processes as exists for continuous time. This dissertation presented three different formulations of time and risk for recurrent events, focusing on processes with low frequencies of recurrences. It was demonstrated how the gap time, counting process, and total time formulations could all be specified in the same LCR framework used for single events. Similar to the single event models, the LCR framework specification allows a direct extension to multiple latent classes, using finite mixtures to account for the likely within subject correlation across event times as well as other sources of unobserved heterogeneity. It was proposed that risk intervals for adjacent spells be allowed to overlap at the end period for one spell and the beginning of the next. The gap time and counting process formulations were shown to be equivalent in discrete-time, when including the appropriate time-dependent covariates; however, the gap time formulation requires less explicit adjustment for the possible bias resulting from overlapping risk intervals. The choice between the two formulations should be driven by the substantive research questions and the time scale that best represents those questions. The total time formula-

tion, resulting in a marginal model, was not found to be as informative as the other two formulations in understanding the recurrent event process.

The recurrent events model in the LCR framework offers several appealing features for the applied researcher. The model allows for event-specific baseline hazard probabilities as well as event-specific covariate effects. This means that researchers may investigate the differences in duration dependence and covariate effects for different events. For example, are the significant risk factors for time to first occurrence also significant risk factors for the time between recurrences? Are those subjects who have already experienced one event at increased risk of experiencing a second event? In addition, the effects of unobserved heterogeneity (including within subject correlation) are permitted to influence not only the baseline hazard probabilities but also event-specific covariate effects. That is, the frailty or susceptibility for an individual is permitted to be different for differently numbered events. None of the current models for recurrent events in discrete-time present this degree of flexibility. The traditional logistic regression formulation allows for event-specific hazards but does not account for within or between subject correlation beyond the measured covariates. The Poisson mixture model allows for nonparametric modeling of the unobserved heterogeneity but does not allow for event-specific hazard probabilities or covariate effects. The multilevel discrete-time logistic

model also does not allow for event-specific hazard probabilities (although the number of prior events may be included as a covariate) and imposes a normal distribution on the variability of the logit baseline hazard probabilities and covariate effects.

The treatment of recurrent events is limited in this dissertation so that processes with high frequencies of recurrence are not addressed. Also, the proposed correction for the possible bias resulting from allowing the end and beginning of adjacent risk intervals to overlap was not formally proven to be a sufficient adjustment. There is no measure of goodness-of-fit presented although there is a possibility of extending the G^2 statistic. With respect to unobserved heterogeneity, there is no satisfactory work on the sensitivity to misspecification of the distribution in the recurrent events setting nor is there a clear practical strategy for mixture model specification and assessment.

5.4 Future research

The limitations of the current dissertation point to some of the many directions in which future research in the area of discrete-time survival using latent variables may go. For the single event models, more work is needed on the topics of model assessment and power calculations, particularly for the mixture models. The G^2 statistic is a promising goodness-of-fit measure but it needs to

be modified to accommodate right-censoring prior to the end of the observation period. The behavior of this statistic also needs to be examined through simulations to understand the degree of sensitivity it may have to sample size and trivial model misspecification. Also, the G^2 and the Lo, Mendell, and Rubin LRT, along with commonly employed information criteria, such as the AIC and BIC, should be evaluated with regards to their performance in the area of latent class enumeration for discrete-time models. It would also be of great value to the applied researcher to evaluate the power for detecting time-dependent and time-independent covariate effects as well as direct and indirect covariate effects in the mixture models as a function of the number of time periods, the width of the time periods, the baseline hazard probabilities, sample size, and effect sizes. The algorithms proposed for double-censored and interval-censored data should be shown to be self-consistent and implemented in a real data analysis setting. All of the work mentioned here should also be extended to recurrent event models with the additional consideration of how associations between the different event processes could improve stability of parameter estimates, power, accuracy of class enumeration, etc.

Moving beyond the scope of this dissertation is the challenge of adapting this LCR framework to accommodate other multivariate survival data. For example, the recurrent event modeling framework could be modified for use

with data from high frequency event recurrence processes. These data do not allow for event-specific baseline hazard probabilities or event-specific covariate effects and often contain double-censored observations. Another kind of multivariate survival data for which the LCR framework could be modified are the competing risks processes briefly described at the beginning of Chapter 4. For these data, occurrence of one event precludes the occurrence of any of the other possible events. A model for competing risks using this framework might resemble something of a hybrid between the single and recurrent event models. And all these extensions are still limited in scope to survival data—they do not realize the full potential of the LCR framework for multivariate longitudinal events. Using a general latent variable framework, which includes LCR, the richness and flexibility of this formulation for discrete-time event history is currently unmatched. Muthén and Masyn (2001) illustrated a model that combined a growth mixture model followed by a discrete-time survival process, in which the heterogeneity in the growth trajectories predicted heterogeneity in the time-to-event process. It would also be possible in this framework to specify a discrete-time survival process followed by a growth model, e.g., modeling time to returning to drinking and then modeling the pattern of drinking behavior once it had begun. Two consecutive survival processes could be modeled as well as concurrent survival processes or other concurrent longitudinal

processes. It would be possible to incorporate covariates measured with error (i.e., latent variables, continuous or categorical, as covariates of survival) and it may even be possible to allow the event itself to be measured with error, e.g., depression as an event where depression is not diagnosed with absolute precision.

There are certainly other modeling possibilities not mentioned here and many that have yet to be conceived. This dissertation establishes a strong foundation that will allow future exploration into the many methodology extensions that will provide researchers with full and flexible models that best represent the complexity of behavioral processes over time.

Appendix A: Splus Code

`nsize` is the sample size.

`emviopj` is an `nsize`-element vector where `emviopj[i]=1` if the m^{th} event occurred for subject i in time period j and 0 otherwise.

The value 999 indicates missingness.

Creating observed event indicators for first violence episode

`eviomj` is an `nsize`-element vector corresponding to the observed event indicators for the m^{th} event in the j^{th} time period as described in Chapter 2.

```
evio11_elviop1
evio12_elviop2
evio13_elviop3
evio14_elviop4
evio15_elviop5
evio16_elviop6

for (i in 1:nsize)
{
if (evio11[i]==1)
  {evio12[i]_999; evio13[i]_999;
  evio14[i]_999; evio15[i]_999; evio16[i]_999}

if (evio12[i]==1)
  {evio13[i]_999; evio14[i]_999;
  evio15[i]_999; evio16[i]_999}

if (evio13[i]==1)
  {evio14[i]_999; evio15[i]_999; evio16[i]_999}

if (evio14[i]==1)
  {evio15[i]_999; evio16[i]_999}

if (evio15[i]==1)
  {evio16[i]_999}
}
```

Creating observed event indicators for second and third violence episode in gap time formulation

`eviomjgt` is an `nsize`-element vector corresponding to the observed event indicators for the m^{th} event in the j^{th} time period in gap time formulation as described in Chapter 4.

```
evio20gt_rep(999,nsize)
evio21gt_rep(999,nsize)
evio22gt_rep(999,nsize)
evio23gt_rep(999,nsize)
evio24gt_rep(999,nsize)
evio25gt_rep(999,nsize)

for (i in 1:nsize)
{
if (elviop1[i]==1)
  {evio20gt[i]_0; evio21gt[i]_0; evio22gt[i]_0;
  evio23gt[i]_0; evio24gt[i]_0; evio25gt[i]_0}

if (elviop2[i]==1)
  {evio20gt[i]_0; evio21gt[i]_0; evio22gt[i]_0;
  evio23gt[i]_0; evio24gt[i]_0}

if (elviop3[i]==1)
  {evio20gt[i]_0; evio21gt[i]_0; evio22gt[i]_0;
  evio23gt[i]_0}

if (elviop4[i]==1)
  {evio20gt[i]_0; evio21gt[i]_0; evio22gt[i]_0}

if (elviop5[i]==1)
  {evio20gt[i]_0; evio21gt[i]_0}

if (elviop6[i]==1)
  {evio20gt[i]_0}
}
```

```

for (i in 1:nsize)
{
if (e2viop1[i]==1)
    {evio20gt[i]_1; evio21gt[i]_999; evio22gt[i]_999;
    evio23gt[i]_999; evio24gt[i]_999; evio25gt[i]_999}

if (e2viop2[i]==1 & elviop2[i]==1)
    {evio20gt[i]_1; evio21gt[i]_999; evio22gt[i]_999;
    evio23gt[i]_999; evio24gt[i]_999; evio25gt[i]_999}

if (e2viop2[i]==1 & elviop1[i]==1)
    {evio21gt[i]_1; evio22gt[i]_999; evio23gt[i]_999;
    evio24gt[i]_999; evio25gt[i]_999}

if (e2viop3[i]==1 & elviop3[i]==1)
    {evio20gt[i]_1; evio21gt[i]_999; evio22gt[i]_999;
    evio23gt[i]_999; evio24gt[i]_999; evio25gt[i]_999}

if (e2viop3[i]==1 & elviop2[i]==1)
    {evio21gt[i]_1; evio22gt[i]_999; evio23gt[i]_999;
    evio24gt[i]_999; evio25gt[i]_999}

if (e2viop3[i]==1 & elviop1[i]==1)
    {evio22gt[i]_1; evio23gt[i]_999; evio24gt[i]_999;
    evio25gt[i]_999}

if (e2viop4[i]==1 & elviop4[i]==1)
    {evio20gt[i]_1; evio21gt[i]_999; evio22gt[i]_999;
    evio23gt[i]_999; evio24gt[i]_999; evio25gt[i]_999}

if (e2viop4[i]==1 & elviop3[i]==1)
    {evio21gt[i]_1; evio22gt[i]_999; evio23gt[i]_999;
    evio24gt[i]_999; evio25gt[i]_999}

if (e2viop4[i]==1 & elviop2[i]==1)
    {evio22gt[i]_1; evio23gt[i]_999; evio24gt[i]_999;
    evio25gt[i]_999}

if (e2viop4[i]==1 & elviop1[i]==1)
    {evio23gt[i]_1; evio24gt[i]_999; evio25gt[i]_999}

if (e2viop5[i]==1 & elviop5[i]==1)
    {evio20gt[i]_1; evio21gt[i]_999; evio22gt[i]_999;
    evio23gt[i]_999; evio24gt[i]_999; evio25gt[i]_999}

```

```

if (e2viop5[i]==1 & elviop4[i]==1)
    {evio21gt[i]_1; evio22gt[i]_999; evio23gt[i]_999;
    evio24gt[i]_999; evio25gt[i]_999}

if (e2viop5[i]==1 & elviop3[i]==1)
    {evio22gt[i]_1; evio23gt[i]_999; evio24gt[i]_999;
    evio25gt[i]_999}

if (e2viop5[i]==1 & elviop2[i]==1)
    {evio23gt[i]_1; evio24gt[i]_999; evio25gt[i]_999}

if (e2viop5[i]==1 & elviop1[i]==1)
    {evio24gt[i]_1; evio25gt[i]_999}

if (e2viop6[i]==1 & elviop6[i]==1)
    {evio20gt[i]_1; evio21gt[i]_999; evio22gt[i]_999;
    evio23gt[i]_999; evio24gt[i]_999; evio25gt[i]_999}

if (e2viop6[i]==1 & elviop5[i]==1)
    {evio21gt[i]_1; evio22gt[i]_999; evio23gt[i]_999;
    evio24gt[i]_999; evio25gt[i]_999}

if (e2viop6[i]==1 & elviop4[i]==1)
    {evio22gt[i]_1; evio23gt[i]_999; evio24gt[i]_999;
    evio25gt[i]_999}

if (e2viop6[i]==1 & elviop3[i]==1)
    {evio23gt[i]_1; evio24gt[i]_999; evio25gt[i]_999}

if (e2viop6[i]==1 & elviop2[i]==1)
    {evio24gt[i]_1; evio25gt[i]_999}

if (e2viop6[i]==1 & elviop1[i]==1)
    {evio25gt[i]_1}
}

evio30gt_rep(999, nsize)
evio31gt_rep(999, nsize)
evio32gt_rep(999, nsize)
evio33gt_rep(999, nsize)
evio34gt_rep(999, nsize)
evio35gt_rep(999, nsize)

```

```

for (i in 1:nsize)
{
if (e2viop1[i]==1)
    {evio30gt[i]_0; evio31gt[i]_0; evio32gt[i]_0;
    evio33gt[i]_0; evio34gt[i]_0; evio35gt[i]_0}

if (e2viop2[i]==1)
    {evio30gt[i]_0; evio31gt[i]_0; evio32gt[i]_0;
    evio33gt[i]_0; evio34gt[i]_0}

if (e2viop3[i]==1)
    {evio30gt[i]_0; evio31gt[i]_0; evio32gt[i]_0;
    evio33gt[i]_0}

if (e2viop4[i]==1)
    {evio30gt[i]_0; evio31gt[i]_0; evio32gt[i]_0}

if (e2viop5[i]==1)
    {evio30gt[i]_0; evio31gt[i]_0}

if (e2viop6[i]==1)
    {evio30gt[i]_0}
}

for (i in 1:nsize)
{
if (e3viop1[i]==1)
    {evio30gt[i]_1; evio31gt[i]_999; evio32gt[i]_999;
    evio33gt[i]_999; evio34gt[i]_999; evio35gt[i]_999}

if (e3viop2[i]==1 & e2viop2[i]==1)
    {evio30gt[i]_1; evio31gt[i]_999; evio32gt[i]_999;
    evio33gt[i]_999; evio34gt[i]_999; evio35gt[i]_999}

if (e3viop2[i]==1 & e2viop1[i]==1)
    {evio31gt[i]_1; evio32gt[i]_999; evio33gt[i]_999;
    evio34gt[i]_999; evio35gt[i]_999}

if (e3viop3[i]==1 & e2viop3[i]==1)
    {evio30gt[i]_1; evio31gt[i]_999; evio32gt[i]_999;
    evio33gt[i]_999; evio34gt[i]_999; evio35gt[i]_999}
}

```



```

if (e3viop3[i]==1 & e2viop2[i]==1)
    {evio31gt[i]_1; evio32gt[i]_999; evio33gt[i]_999;
    evio34gt[i]_999; evio35gt[i]_999}

if (e3viop3[i]==1 & e2viop1[i]==1)
    {evio32gt[i]_1; evio33gt[i]_999; evio34gt[i]_999;
    evio35gt[i]_999}

if (e3viop4[i]==1 & e2viop4[i]==1)
    {evio30gt[i]_1; evio31gt[i]_999; evio32gt[i]_999;
    evio33gt[i]_999; evio34gt[i]_999; evio35gt[i]_999}

if (e3viop4[i]==1 & e2viop3[i]==1)
    {evio31gt[i]_1; evio32gt[i]_999; evio33gt[i]_999;
    evio34gt[i]_999; evio35gt[i]_999}

if (e3viop4[i]==1 & e2viop2[i]==1)
    {evio32gt[i]_1; evio33gt[i]_999; evio34gt[i]_999;
    evio35gt[i]_999}

if (e3viop4[i]==1 & e2viop1[i]==1)
    {evio33gt[i]_1; evio34gt[i]_999; evio35gt[i]_999}

if (e3viop5[i]==1 & e2viop5[i]==1)
    {evio30gt[i]_1; evio31gt[i]_999; evio32gt[i]_999;
    evio33gt[i]_999; evio34gt[i]_999; evio35gt[i]_999}

if (e3viop5[i]==1 & e2viop4[i]==1)
    {evio31gt[i]_1; evio32gt[i]_999; evio33gt[i]_999;
    evio34gt[i]_999; evio35gt[i]_999}

if (e3viop5[i]==1 & e2viop3[i]==1)
    {evio32gt[i]_1; evio33gt[i]_999; evio34gt[i]_999;
    evio35gt[i]_999}

if (e3viop5[i]==1 & e2viop2[i]==1)
    {evio33gt[i]_1; evio34gt[i]_999; evio35gt[i]_999}

if (e3viop5[i]==1 & e2viop1[i]==1)
    {evio34gt[i]_1; evio35gt[i]_999}

if (e3viop6[i]==1 & e2viop6[i]==1)
    {evio30gt[i]_1; evio31gt[i]_999; evio32gt[i]_999;
    evio33gt[i]_999; evio34gt[i]_999; evio35gt[i]_999}

```

```

if (e3viop6[i]==1 & e2viop5[i]==1)
    {evio31gt[i]_1; evio32gt[i]_999; evio33gt[i]_999;
    evio34gt[i]_999; evio35gt[i]_999}

if (e3viop6[i]==1 & e2viop4[i]==1)
    {evio32gt[i]_1; evio33gt[i]_999; evio34gt[i]_999;
    evio35gt[i]_999}

if (e3viop6[i]==1 & e2viop3[i]==1)
    {evio33gt[i]_1; evio34gt[i]_999; evio35gt[i]_999}

if (e3viop6[i]==1 & e2viop2[i]==1)
    {evio34gt[i]_1; evio35gt[i]_999}

if (e3viop6[i]==1 & e2viop1[i]==1)
    {evio35gt[i]_1}
}

```

Creating observed event indicators for second and third violence episode in counting process formulation

eviomjcp is an nsize-element vector corresponding to the observed event indicators for the m^{th} event in the j^{th} time period in counting process formulation as described in Chapter 4.

```

evio21cp_rep(999,nsize)
evio22cp_rep(999,nsize)
evio23cp_rep(999,nsize)
evio24cp_rep(999,nsize)
evio25cp_rep(999,nsize)
evio26cp_rep(999,nsize)

for (i in 1:nsize)
{
if (e1viop1[i]==1)
    {evio21cp[i]_0; evio22cp[i]_0; evio23cp[i]_0;
    evio24cp[i]_0; evio25cp[i]_0; evio26cp[i]_0}
}

```

```

if (elviop2[i]==1)
    {evio22cp[i]_0; evio23cp[i]_0; evio24cp[i]_0;
    evio25cp[i]_0; evio26cp[i]_0}

if (elviop3[i]==1)
    {evio23cp[i]_0; evio24cp[i]_0; evio25cp[i]_0;
    evio26cp[i]_0}

if (elviop4[i]==1)
    {evio24cp[i]_0; evio25cp[i]_0; evio26cp[i]_0}

if (elviop5[i]==1)
    {evio25cp[i]_0; evio26cp[i]_0}

if (elviop6[i]==1)
    {evio26cp[i]_0}
}

for (i in 1:nsize)
{
if (e2viop1[i]==1)
    {evio21cp[i]_1; evio22cp[i]_999; evio23cp[i]_999;
    evio24cp[i]_999; evio25cp[i]_999; evio26cp[i]_999}

if (e2viop2[i]==1)
    {evio22cp[i]_1; evio23cp[i]_999; evio24cp[i]_999;
    evio25cp[i]_999; evio26cp[i]_999}

if (e2viop3[i]==1)
    {evio23cp[i]_1; evio24cp[i]_999; evio25cp[i]_999;
    evio26cp[i]_999}

if (e2viop4[i]==1)
    {evio24cp[i]_1; evio25cp[i]_999; evio26cp[i]_999}

if (e2viop5[i]==1)
    {evio25cp[i]_1; evio26cp[i]_999}

if (e2viop6[i]==1)
    {evio26cp[i]_1}
}

```

```

evio31cp_rep(999, nsize)
evio32cp_rep(999, nsize)
evio33cp_rep(999, nsize)
evio34cp_rep(999, nsize)
evio35cp_rep(999, nsize)
evio36cp_rep(999, nsize)

for (i in 1:nsize)
{
if (e2viop1[i]==1)
    {evio31cp[i]_0; evio32cp[i]_0; evio33cp[i]_0;
    evio34cp[i]_0; evio35cp[i]_0; evio36cp[i]_0}

if (e2viop2[i]==1)
    {evio32cp[i]_0; evio33cp[i]_0; evio34cp[i]_0;
    evio35cp[i]_0; evio36cp[i]_0}

if (e2viop3[i]==1)
    {evio33cp[i]_0; evio34cp[i]_0; evio35cp[i]_0;
    evio36cp[i]_0}

if (e2viop4[i]==1)
    {evio34cp[i]_0; evio35cp[i]_0; evio36cp[i]_0}

if (e2viop5[i]==1)
    {evio35cp[i]_0; evio36cp[i]_0}

if (e2viop6[i]==1)
    {evio36cp[i]_0}
}

for (i in 1:nsize)
{
if (e3viop1[i]==1)
    {evio31cp[i]_1; evio32cp[i]_999; evio33cp[i]_999;
    evio34cp[i]_999; evio35cp[i]_999; evio36cp[i]_999}

if (e3viop2[i]==1)
    {evio32cp[i]_1; evio33cp[i]_999; evio34cp[i]_999;
    evio35cp[i]_999; evio36cp[i]_999}
}

```

```

if (e3viop3[i]==1)
    {evio33cp[i]_1; evio34cp[i]_999; evio35cp[i]_999;
    evio36cp[i]_999}

if (e3viop4[i]==1)
    {evio34cp[i]_1; evio35cp[i]_999; evio36cp[i]_999}

if (e3viop5[i]==1)
    {evio35cp[i]_1; evio36cp[i]_999}

if (e3viop6[i]==1)
    {evio36cp[i]_1}
}

```

Creating observed event indicators for second and third violence episode in total time formulation

eviomjtt is an nsize-element vector corresponding to the observed event indicators for the m^{th} event in the j^{th} time period in counting process formulation as described in Chapter 4.

```

evio21tt_rep(0,nsize)
evio22tt_rep(0,nsize)
evio23tt_rep(0,nsize)
evio24tt_rep(0,nsize)
evio25tt_rep(0,nsize)
evio26tt_rep(0,nsize)

for (i in 1:nsize)
{
if (e2viop1[i]==1)
    {evio21tt[i]_1; evio22tt[i]_999; evio23tt[i]_999;
    evio24tt[i]_999; evio25tt[i]_999; evio26tt[i]_999}
if (e2viop2[i]==1)
    {evio22tt[i]_1; evio23tt[i]_999; evio24tt[i]_999;
    evio25tt[i]_999; evio26tt[i]_999}

if (e2viop3[i]==1)
    {evio23tt[i]_1; evio24tt[i]_999; evio25tt[i]_999;
    evio26tt[i]_999}
}

```

```

if (e2viop4[i]==1)
    {evio24tt[i]_1; evio25tt[i]_999; evio26tt[i]_999}

if (e2viop5[i]==1)
    {evio25tt[i]_1; evio26tt[i]_999}

if (e2viop6[i]==1)
    {evio26tt[i]_1}
}

evio31tt_rep(0,nsiz)
evio32tt_rep(0,nsiz)
evio33tt_rep(0,nsiz)
evio34tt_rep(0,nsiz)
evio35tt_rep(0,nsiz)
evio36tt_rep(0,nsiz)

for (i in 1:nsiz)
{
if (e3viop1[i]==1)
    {evio31tt[i]_1; evio32tt[i]_999; evio33tt[i]_999;
    evio34tt[i]_999; evio35tt[i]_999; evio36tt[i]_999}

if (e3viop2[i]==1)
    {evio32tt[i]_1; evio33tt[i]_999; evio34tt[i]_999;
    evio35tt[i]_999; evio36tt[i]_999}

if (e3viop3[i]==1)
    {evio33tt[i]_1; evio34tt[i]_999; evio35tt[i]_999;
    evio36tt[i]_999}

if (e3viop4[i]==1)
    {evio34tt[i]_1; evio35tt[i]_999; evio36tt[i]_999}

if (e3viop5[i]==1)
    {evio35tt[i]_1; evio36tt[i]_999}

if (e3viop6[i]==1)
    {evio36tt[i]_1}
}

```

Calculation of G^2 for Model 5

```
y1_as.numeric(evio11==1)
y2_as.numeric(evio12==1)
y3_as.numeric(evio13==1)
y4_as.numeric(evio14==1)
y5_as.numeric(evio15==1)
y6_as.numeric(evio16==1)
```

```
eta0_-1.718
eta1_-0.498
```

```
lambda_c(0,1,2,3,4,5)
```

```
beta1_-0.675
beta2_-0.668
beta3_2.373
```

```
y1hat_1/(1+exp(-(eta0+eta1*lambda[1]+
beta1*edw+beta2*inc+beta3*pddp1)))
```

```
y2hat_(1-1/(1+exp(-(eta0+eta1*lambda[1]+
beta1*edw+beta2+inc+beta3*pddp1))))*
(1/(1+exp(-(eta0+eta1*lambda[2]-
beta1*edw+beta2+inc+beta3*pddp2))))
```

```
y3hat_(1-1/(1+exp(-(eta0+eta1*lambda[1]-
beta1*edw+beta2+inc+beta3*pddp1))))*
(1-1/(1+exp(-(eta0+eta1*lambda[2]-
beta1*edw+beta2+inc+beta3*pddp2))))*
(1/(1+exp(-(eta0+eta1*lambda[3]-
beta1*edw+beta2+inc+beta3*pddp3))))
```

```
y4hat_(1-1/(1+exp(-(eta0+eta1*lambda[1]-
beta1*edw+beta2+inc+beta3*pddp1))))*
(1-1/(1+exp(-(eta0+eta1*lambda[2]-
beta1*edw+beta2+inc+beta3*pddp2))))*
(1-1/(1+exp(-(eta0+eta1*lambda[3]-
beta1*edw+beta2+inc+beta3*pddp3))))*
(1/(1+exp(-(eta0+eta1*lambda[4]-
beta1*edw+beta2+inc+beta3*pddp4))))
```

$$\begin{aligned}
y5hat_ & (1-1/(1+\exp(-(\eta_0+\eta_1*\lambda[1]- \\
& \beta_1*edw+\beta_2+inc+\beta_3*pddp1)))) * \\
& (1-1/(1+\exp(-(\eta_0+\eta_1*\lambda[2]- \\
& \beta_1*edw+\beta_2+inc+\beta_3*pddp2)))) * \\
& (1-1/(1+\exp(-(\eta_0+\eta_1*\lambda[3]- \\
& \beta_1*edw+\beta_2+inc+\beta_3*pddp3)))) * \\
& (1-1/(1+\exp(-(\eta_0+\eta_1*\lambda[4]- \\
& \beta_1*edw+\beta_2+inc+\beta_3*pddp4)))) * \\
& (1/(1+\exp(-(\eta_0+\eta_1*\lambda[5]- \\
& \beta_1*edw+\beta_2+inc+\beta_3*pddp5))))
\end{aligned}$$

$$\begin{aligned}
y6hat_ & (1-1/(1+\exp(-(\eta_0+\eta_1*\lambda[1]- \\
& \beta_1*edw+\beta_2+inc+\beta_3*pddp1)))) * \\
& (1-1/(1+\exp(-(\eta_0+\eta_1*\lambda[2]- \\
& \beta_1*edw+\beta_2+inc+\beta_3*pddp2)))) * \\
& (1-1/(1+\exp(-(\eta_0+\eta_1*\lambda[3]- \\
& \beta_1*edw+\beta_2+inc+\beta_3*pddp3)))) * \\
& (1-1/(1+\exp(-(\eta_0+\eta_1*\lambda[4]- \\
& \beta_1*edw+\beta_2+inc+\beta_3*pddp4)))) * \\
& (1-1/(1+\exp(-(\eta_0+\eta_1*\lambda[5]- \\
& \beta_1*edw+\beta_2+inc+\beta_3*pddp5)))) * \\
& (1/(1+\exp(-(\eta_0+\eta_1*\lambda[6]- \\
& \beta_1*edw+\beta_2+inc+\beta_3*pddp6))))
\end{aligned}$$

$$\begin{aligned}
vc1_ & (1/nsize) * c(\text{sum}(y1hat*(1-y1hat)), \\
& \text{sum}(y1hat*-y2hat), \text{sum}(y1hat*-y3hat), \\
& \text{sum}(y1hat*-y4hat), \text{sum}(y1hat*-y5hat), \\
& \text{sum}(y1hat*-y6hat))
\end{aligned}$$

$$\begin{aligned}
vc2_ & (1/nsize) * c(\text{sum}(y2hat*-y1hat), \\
& \text{sum}(y2hat*(1-y2hat)), \text{sum}(y2hat*-y3hat), \\
& \text{sum}(y2hat*-y4hat), \text{sum}(y2hat*-y5hat), \\
& \text{sum}(y2hat*-y6hat))
\end{aligned}$$

$$\begin{aligned}
vc3_ & (1/nsize) * c(\text{sum}(y3hat*-y1hat), \\
& \text{sum}(y3hat*-y2hat), \text{sum}(y3hat*(1-y3hat)), \\
& \text{sum}(y3hat*-y4hat), \text{sum}(y3hat*-y5hat), \\
& \text{sum}(y3hat*-y6hat))
\end{aligned}$$

$$\begin{aligned}
vc4_ & (1/nsize) * c(\text{sum}(y4hat*-y1hat), \\
& \text{sum}(y4hat*-y2hat), \text{sum}(y4hat*-y3hat), \\
& \text{sum}(y4hat*(1-y4hat)), \text{sum}(y4hat*-y5hat), \\
& \text{sum}(y4hat*-y6hat))
\end{aligned}$$


```

vc5_(1/nsize)*c(sum(y5hat*-y1hat),
  sum(y5hat*-y2hat),sum(y5hat*-y3hat),
  sum(y5hat*-y4hat),sum(y5hat*(1-y5hat)),
  sum(y5hat*-y6hat))

vc6_(1/nsize)*c(sum(y6hat*-y1hat),
  sum(y6hat*-y2hat),sum(y6hat*-y3hat),
  sum(y6hat*-y4hat),sum(y6hat*-y5hat),
  sum(y6hat*(1-y6hat)))

vy_cbind(vc1,vc2,vc3,vc4,vc5,vc6)

sn1_y1-y1hat
sn2_y2-y2hat
sn3_y3-y3hat
sn4_y4-y4hat
sn5_y5-y5hat
sn6_y6-y6hat

sn_(1/sqrt(nsize))*c(sum(sn1),sum(sn2),
  sum(sn3),sum(sn4),sum(sn5),sum(sn6))

g2_t(sn)%%solve(vy)%%t(t(sn))

pg2_1-pchisq(g2,6)

```

Class enumeration data simulations: Population A

```

caseid_1:10000

x1_rnorm(10000,0,1)

x2_as.numeric(runif(10000,0,1)<.5)

u1_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-2+x1+2*x2)))))
u2_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-2+x1+2*x2)))))
u3_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-2+x1+2*x2)))))

```

```

u4_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-2+x1+2*x2)))))
u5_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-2+x1+2*x2)))))

for (i in 1:10000)
{
if (u1[i]==1)
  {u2[i]_999; u3[i]_999; u4[i]_999; u5[i]_999}

if (u2[i]==1)
  {u3[i]_999; u4[i]_999; u5[i]_999}

if (u3[i]==1)
  {u4[i]_999; u5[i]_999}

if (u4[i]==1)
  {u5[i]_999}
}

```

Class enumeration data simulations: Population B

```

caseid_1:10000

x1_rnorm(10000,0,1)

x2_as.numeric(runif(10000,0,1)<(1/(1+exp(-x1))))

u1_as.numeric(runif(10000,0,1)<(1/(1+exp(-(-2+2*x2)))))
u2_as.numeric(runif(10000,0,1)<(1/(1+exp(-(-2+2*x2)))))
u3_as.numeric(runif(10000,0,1)<(1/(1+exp(-(-2+2*x2)))))
u4_as.numeric(runif(10000,0,1)<(1/(1+exp(-(-2+2*x2)))))
u5_as.numeric(runif(10000,0,1)<(1/(1+exp(-(-2+2*x2)))))

```

```

for (i in 1:10000)
{
if (u1[i]==1)
    {u2[i]_999; u3[i]_999; u4[i]_999; u5[i]_999}

if (u2[i]==1)
    {u3[i]_999; u4[i]_999; u5[i]_999}

if (u3[i]==1) {u4[i]_999; u5[i]_999}

if (u4[i]==1)
    {u5[i]_999}
}

```

Class enumeration data simulations: Population C

```

caseid_1:10000

x1_rnorm(10000,0,1)

x2_as.numeric(runif(10000,0,1)<(1/(1+exp(-x1))))

u1_as.numeric(runif(10000,0,1)<
    (1/(1+exp(-(-2+x1+2*x2)))))
u2_as.numeric(runif(10000,0,1)<
    (1/(1+exp(-(-2+x1+2*x2)))))
u3_as.numeric(runif(10000,0,1)<
    (1/(1+exp(-(-2+x1+2*x2)))))
u4_as.numeric(runif(10000,0,1)<
    (1/(1+exp(-(-2+x1+2*x2)))))
u5_as.numeric(runif(10000,0,1)<
    (1/(1+exp(-(-2+x1+2*x2)))))

for (i in 1:10000)
{
if (u1[i]==1)
    {u2[i]_999; u3[i]_999; u4[i]_999; u5[i]_999}

if (u2[i]==1)
    {u3[i]_999; u4[i]_999; u5[i]_999}
}

```

```

if (u3[i]==1)
  {u4[i]_999; u5[i]_999}

if (u4[i]==1)
  {u5[i]_999}
}

```

Long-term survivor data simulations: Population A

```

caseid_1:10000

x1_rnorm(10000,0,1)

x2_as.numeric(runif(10000,0,1)<.5)

#x2==1 for LTS

u1_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1)))))*(x2==0)
u2_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1)))))*(x2==0)
u3_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1)))))*(x2==0)
u4_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1)))))*(x2==0)
u5_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1)))))*(x2==0)

for (i in 1:10000)
{
if (u1[i]==1)
  {u2[i]_999; u3[i]_999; u4[i]_999; u5[i]_999}

if (u2[i]==1)
  {u3[i]_999; u4[i]_999; u5[i]_999}

if (u3[i]==1)
  {u4[i]_999; u5[i]_999}

if (u4[i]==1)
  {u5[i]_999}
}
}

```

Long-term survivor data simulations: Population B

```
caseid_1:10000

x1_rnorm(10000,0,1)

x2_as.numeric(runif(10000,0,1)<.5)

u1_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-4))))*(x2==1)
u2_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-4))))*(x2==1)
u3_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-4))))*(x2==1)
u4_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-4))))*(x2==1)
u5_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-4))))*(x2==1)

for (i in 1:10000)
{
if (u1[i]==1)
  {u2[i]_999; u3[i]_999; u4[i]_999; u5[i]_999}

if (u2[i]==1)
  {u3[i]_999; u4[i]_999; u5[i]_999}

if (u3[i]==1)
  {u4[i]_999; u5[i]_999}

if (u4[i]==1)
  {u5[i]_999}
}
```

Long-term survivor data simulations: Population C

```
caseid_1:10000

x1_rnorm(10000,0,1)

x2_as.numeric(runif(10000,0,1)<.5)

u1_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-4+x1))))*(x2==1)
u2_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-4+x1))))*(x2==1)
u3_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-4+x1))))*(x2==1)
u4_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-4+x1))))*(x2==1)
u5_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-4+x1))))*(x2==1)

for (i in 1:10000)
{
if (u1[i]==1)
  {u2[i]_999; u3[i]_999; u4[i]_999; u5[i]_999}

if (u2[i]==1)
  {u3[i]_999; u4[i]_999; u5[i]_999}

if (u3[i]==1)
  {u4[i]_999; u5[i]_999}

if (u4[i]==1)
  {u5[i]_999}
}
```

Long-term survivor data simulations: Population D

```
caseid_1:10000

x1_rnorm(10000,0,1)

x2_as.numeric(runif(10000,0,1)<1/(1+exp(-x1)))

u1_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
    (1/(1+exp(-(-4+x1))))*(x2==1)
  )
)
u2_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
    (1/(1+exp(-(-4+x1))))*(x2==1)
  )
)
u3_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
    (1/(1+exp(-(-4+x1))))*(x2==1)
  )
)
u4_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
    (1/(1+exp(-(-4+x1))))*(x2==1)
  )
)
u5_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0) +
  as.numeric(runif(10000,0,1)<
    (1/(1+exp(-(-4+x1))))*(x2==1)
  )
)

for (i in 1:10000)
{
  if (u1[i]==1)
    {u2[i]_999; u3[i]_999; u4[i]_999; u5[i]_999}

  if (u2[i]==1)
    {u3[i]_999; u4[i]_999; u5[i]_999}

  if (u3[i]==1)
    {u4[i]_999; u5[i]_999}

  if (u4[i]==1)
    {u5[i]_999}
}
```

Long-term survivor data simulations: Population E

```
caseid_1:10000

x1_rnorm(10000,0,1)

x2_as.numeric(runif(10000,0,1)<1/(1+exp(-x1)))

u1_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0))
u2_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0))
u3_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0))
u4_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0))
u5_as.numeric(runif(10000,0,1)<
  (1/(1+exp(-(-1+x1))))*(x2==0))

for (i in 1:10000)
{
if (u1[i]==1)
  {u2[i]_999; u3[i]_999; u4[i]_999; u5[i]_999}

if (u2[i]==1)
  {u3[i]_999; u4[i]_999; u5[i]_999}

if (u3[i]==1)
  {u4[i]_999; u5[i]_999}

if (u4[i]==1)
  {u5[i]_999}
}
```


Appendix B: Mplus input

RIA Data Example Variable List

evio11-*evio16*: Binary event indicators for the first episode of violence in the six post-treatment periods.

tx1: Binary indicator of BCT treatment.

edw1: Binary indicator of wife's education not beyond high school.

inc5: Binary indicator of household income \$35,001 - \$40,000.

inc6: Binary indicator of household income greater than \$40,000.

pddp1-*pddp6*: Percent-days-drinking for the six post-treatment periods.

evio20gt-*evio25gt*: Binary event indicators for the second event on the gap time scale.

evio30gt-*evio34gt*: Binary event indicators for the third event on the gap time scale.

lor: Length of relationship in years.

lorsq: Square of mean-centered length of relationship.

viopre5: Binary indicator of 5-10 violent episodes in the three month pre-treatment period.

viopre6: Binary indicator of more than 10 violent episodes in the three month pre-treatment period.

tte2: Time period during which the second episode occurred on the original time scale.

evio21cp-*evio26vp*: Binary event indicators for the second event on the counting process time scale.

evio31cp-*evio36vp*: Binary event indicators for the third event on the counting process time scale.

gtel1p1-*gtel1p6*: Binary indicators for the first event occurring in period 1-6.

evio21tt-*evio26tt*: Binary event indicators for the second event on the total time scale.

evio31tt-*evio36tt*: Binary event indicators for the third event on the total time scale.

Model 1

DATA:

File is RIAdat.dat;

VARIABLE:

Names are case,whiteh,ms,dwi,dep,...;

Usevar are evio11-evio16;

Missing are all(999);

Categorical are evio11-evio16;

Classes = c(1);

ANALYSIS:

Type=Mixture missing;

MODEL:

%overall%

%c#1%

[evio11\$1*0 evio12\$1*0 evio13\$1*0 evio14\$1*0];

[evio15\$1*0 evio16\$1*0];

Model 2

DATA:

File is RIAdat.dat;

VARIABLE:

Names are case,whiteh,ms,dwi,dep,...;

Usevar are eviol1-eviol6 tx1 edw1;

Missing are all(999);

Categorical are eviol1-eviol6;

Classes = c(1);

ANALYSIS:

Type=Mixture missing;

MODEL:

%overall%

eviol1-eviol6 on tx1 (1);

eviol1-eviol6 on edw1 (2);

%c#1%

[eviol1\$1*0 eviol2\$1*0 eviol3\$1*0 eviol4\$1*0];

[eviol5\$1*0 eviol6\$1*0];

Model 3

DATA:

File is RIAdat.dat;

VARIABLE:

Names are case,whiteh,ms,dwi,dep,...;

Usevar are eviol1-eviol6 tx1 edw1;

Missing are all(999);

Categorical are eviol1-eviol6;

Classes = c(1);

ANALYSIS:

Type=Mixture missing;

MODEL:

%overall%

eviol1-eviol2 on tx1 (1);

eviol3-eviol4 on tx1 (2);

eviol5-eviol6 on tx1 (3);

eviol1-eviol2 on edw1 (4);

eviol3-eviol4 on edw1 (5);

eviol5-eviol6 on edw1 (6);

%c#1%

[eviol1\$1*0 eviol2\$1*0 eviol3\$1*0 eviol4\$1*0];

[eviol5\$1*0 eviol6\$1*0];

Model 4

DATA:

File is RIAdata.dat;

VARIABLE:

Names are case,whiteh,ms,dwi,dep,...;

Usevar are eviol1-eviol6 edw1 inc5 inc6
pddp1 pddp2 pddp3 pddp4 pddp5 pddp6;

Missing are all(999);

Categorical are eviol1-eviol6;

Classes = c(1);

ANALYSIS:

Type=Mixture missing;

MODEL:

%overall%

eviol1-eviol6 on edw1 (1);

eviol1-eviol6 on inc5 (2);

eviol1-eviol6 on inc6 (2);

eviol1 on pddp1 (3);

eviol2 on pddp2 (3);

eviol3 on pddp3 (3);

eviol4 on pddp4 (3);

eviol5 on pddp5 (3);

eviol6 on pddp6 (3);

%c#1%

[eviol1\$1*0 eviol2\$1*0 eviol3\$1*0 eviol4\$1*0];

[eviol5\$1*0 eviol6\$1*0];

Model 5a

DATA:

File is RIAdat.dat;

VARIABLE:

Names are case,whiteh,ms,dwi,dep,...;

Usevar are eviol1-eviol6 edw1 inc5 inc6
pddp1 pddp2 pddp3 pddp4 pddp5 pddp6;

Missing are all(999);

Categorical are eviol1-eviol6;

Classes = c(1);

ANALYSIS:

Type=Mixture missing;

MODEL:

%overall%

eviol1-eviol6 on edw1 (1);

eviol1-eviol6 on inc5 (2);

eviol1-eviol6 on inc6 (2);

eviol1 on pddp1 (3);

eviol2 on pddp2 (3);

eviol3 on pddp3 (3);

eviol4 on pddp4 (3);

eviol5 on pddp5 (3);

eviol6 on pddp6 (3);

%c#1%

[eviol1\$1*0 eviol2\$1*0 eviol3\$1*0 eviol4\$1*0] (1);

[eviol5\$1*0 eviol6\$1*0] (1);

Model 5

DATA:

File is RIAdat.dat;

VARIABLE:

Names are case,whiteh,ms,dwi,dep,...;

Usevar are eviol1-eviol6 edw1 inc5 inc6
pddp1 pddp2 pddp3 pddp4 pddp5 pddp6;

Missing are all(999);

Categorical are eviol1-eviol6;

Classes = c(1);

ANALYSIS:

Type=Mixture missing;

MODEL:

%overall%

eviol1-eviol6 on edw1 (1);

eviol1-eviol6 on inc5 (2);

eviol1-eviol6 on inc6 (2);

eviol1 on pddp1 (3);

eviol2 on pddp2 (3);

eviol3 on pddp3 (3);

eviol4 on pddp4 (3);

eviol5 on pddp5 (3);

eviol6 on pddp6 (3);

eta0 by eviol1-eviol6@1;

etal by eviol1@0 eviol2@1 eviol3@2

eviol4@3 eviol5@4 eviol6@5;

[eta0* etal*];

%c#1%

[eviol1\$1@0 eviol2\$1@0 eviol3\$1@0 eviol4\$1@0] (4);

[eviol5\$1@0 eviol6\$1@0] (4);

Model 6

DATA:

File is RIAdat.dat;

VARIABLE:

Names are case,whiteh,ms,dwi,dep,...;

Usevar are eviol1-eviol6 edw1 inc5 inc6
pddp1 pddp2 pddp3 pddp4 pddp5 pddp6

Missing are all(999);

Categorical are eviol1-eviol6;

Classes = c(2);

ANALYSIS:

Type=Mixture missing;

MODEL:

%overall%

eviol1-eviol6 on edw1 (1);

eviol1-eviol6 on inc5 (2);

eviol1-eviol6 on inc6 (2);

eviol1 on pddp1 (3);

eviol2 on pddp2 (3);

eviol3 on pddp3 (3);

eviol4 on pddp4 (3);

eviol5 on pddp5 (3);

eviol6 on pddp6 (3);

eta0 by eviol1-eviol6@1;

eta1 by eviol1@0 eviol2@1 eviol3@2

eviol4@3 eviol5@4 eviol6@5;

[eta0* eta1*];

```

c#1 on edw1;

c#1 on inc5 inc6 (4);

%c#1%

[evio11$1@0 evio12$1@0 evio13$1@0 evio14$1@0] (14);
[evio15$1@0 evio16$1@0] (14);

[eta0*5 etal*-1];

evio11-evio16 on edw1 (11);

evio11-evio16 on inc5 (12);
evio11-evio16 on inc6 (12);

evio11 on pddp1 (13);
evio12 on pddp2 (13);
evio13 on pddp3 (13);
evio14 on pddp4 (13);
evio15 on pddp5 (13);
evio16 on pddp6 (13);

%c#2%

[evio11$1@0 evio12$1@0 evio13$1@0 evio14$1@0] (24);
[evio15$1@0 evio16$1@0] (24);

[eta0*0 etal*0];

evio11-evio16 on edw1 (21);

evio11-evio16 on inc5 (22);
evio11-evio16 on inc6 (22);

evio11 on pddp1 (23);
evio12 on pddp2 (23);
evio13 on pddp3 (23);
evio14 on pddp4 (23);
evio15 on pddp5 (23);
evio16 on pddp6 (23);

```

Model 8a

```
DATA:
    File is RIAdat.dat;

VARIABLE:
    Names are case,whiteh,ms,dwi,dep,...;

    Usevar are evio20gt-evio25gt lor lorsq;

    Missing are all(999);

    Categorical are evio20gt-evio25gt;

    Classes = c(1);

ANALYSIS:
    Type=Mixture missing;

MODEL:
%overall%

    evio20gt-evio25gt on lor (1);
    evio20gt-evio25gt on lorsq (2);

    eta02 by evio20gt-evio25gt@1;

    eta12 by uvio20gt@0 evio21gt@1 evio22gt@2
            evio23gt@3 evio24gt@4 evio25gt@5;

    [eta02* eta12*];

%c#1%

    [evio20gt$1@0 evio21gt$1@0 evio22gt$1@0 evio23gt$1@0];
    [evio24gt$1@0 evio25gt$1@0];
```

Model 8b

DATA:

File is RIAdat.dat;

VARIABLE:

Names are case,whiteh,ms,dwi,dep,...;

Usevar are evio30gt-evio34gt viopre5 viopre6 tte2;

Missing are all(999);

Categorical are evio30gt-evio34gt;

Classes = c(1);

ANALYSIS:

Type=Mixture missing;

MODEL:

%overall%

evio30gt-evio34gt on viopre5 (1);

evio30gt-evio34gt on viopre6 (1);

evio30gt-evio34gt on tte2 (2);

%c#1%

[evio30gt\$1*0 evio31gt\$1*0 evio32gt\$1*0] (3);

[evio33gt\$1*0 evio34gt\$1*0] (3);

Model 10a

DATA:

File is RIAdata.dat;

VARIABLE:

Names are case,whiteh,ms,dwi,dep,...;

Usevar are evio21cp-evio26cp lor lorsq
gtelp1 gtelp2 gtelp3 gtelp4 gtelp5 gtelp6,;

Missing are all(999);

Categorical are evio21cp-evio26cpgt;

Classes = c(1);

ANALYSIS:

Type=Mixture missing;

MODEL:

%overall%

evio21cp-evio26cp on lor (1);
evio21cp-evio26cp on lorsq (2);

evio21cp on gtelp1 (3);
evio22cp on gtelp2 (3);
evio23cp on gtelp3 (3);
evio24cp on gtelp4 (3);
evio25cp on gtelp5 (3);
evio26cp on gtelp6 (3);

%c#1%

[evio21cp\$1@0 evio22cp\$1@0 evio23cp\$1@0] (4);
[evio24cp\$1@0 evio25cp\$1@0 evio26cp\$1@0] (4);

Model 10b

DATA:

File is RIAdat.dat;

VARIABLE:

Names are case,whiteh,ms,dwi,dep,...;

Usevar are evio31cp-evio36cp viopre5 viopre6
gte2p1 gte2p2 gte2p3 gte2p4 gte2p5 gte2p6;

Missing are all(999);

Categorical are evio31cp-evio36cp;

Classes = c(1);

ANALYSIS:

Type=Mixture missing;

MODEL:

%overall%

evio31cp-evio36cp on viopre5 (1);
evio31cp-evio36cp on viopre6 (1);

evio31cp on gte2p1 (2);
evio32cp on gte2p2 (2);
evio33cp on gte2p3 (2);
evio34cp on gte2p4 (2);
evio35cp on gte2p5 (2);
evio36cp on gte2p6 (2);

eta03 by uvio31cp-uvio36cp@1;

eta13 by uvio31cp@0 uvio32cp@1 uvio33cp@2
uvio34cp@3 uvio35cp@4 uvio36cp@5;

[eta03* eta13*];

%c#1%

[evio31cp\$1@0 evio32cp\$1@0 evio33cp\$1@0] (3);
[evio34cp\$1@0 evio35cp\$1@0 evio36cp\$1@0] (3);

Model 14

DATA:

File is RIAdata.dat;

VARIABLE:

Names are case,whiteh,ms,dwi,dep,...;

Usevar are evio11-evio16 evio21tt-evio36tt
edw1 inc5 inc6 lor lorsq viopre
pddp1 pddp2 pddp3 pddp4 pddp5 pddp6
gte1p1 gte1p2 gte1p3 gte1p4 gte1p5 gte1p6,
gte2p1 gte2p2 gte2p3 gte2p4 gte2p5 gte2p6

Missing are all(999);

Categorical are evio11-evio16 evio21tt-evio36tt;

Classes = c(1);

ANALYSIS:

Type=Mixture missing;

MODEL:

%overall%

!First Event

evio11-evio16 on edw1 (11);

evio11-evio16 on inc5 (12);

evio11-evio16 on inc6 (12);

evio11 on pddp1 (13);

evio12 on pddp2 (13);

evio13 on pddp3 (13);

evio14 on pddp4 (13);

evio15 on pddp5 (13);

evio16 on pddp6 (13);

eta01 by evio11-evio16@1;

eta11 by evio11@0 evio12@1 evio13@2

evio14@3 evio15@4 evio16@5;

[eta01*];
[eta11*] (5);

!Second Event

evio21tt-evio26tt on inc5 (12);
evio21tt-evio26tt on inc6 (12);

evio21tt-evio26tt on lor (22);
evio21tt-evio26tt on lorsq (23);

evio21tt-evio26tt on viopre (24);

evio21tt on pddp1 (13);
evio22tt on pddp2 (13);
evio23tt on pddp3 (13);
evio24tt on pddp4 (13);
evio25tt on pddp5 (13);
evio26tt on pddp6 (13);

evio21tt on gtelp1 (26);
evio22tt on gtelp2 (26);
evio23tt on gtelp3 (26);
evio24tt on gtelp4 (26);
evio25tt on gtelp5 (26);
evio26tt on gtelp6 (26);

eta02 by evio21tt-evio26tt@1;

eta12 by evio21tt@0 evio22tt@1 evio23tt@2
evio24tt@3 evio25tt@4 evio26tt@5;

[eta02*];
[eta12*] (5);

!Third Event

evio31tt-evio36tt on viopre (24);

evio31tt on pddp1 (13);
evio32tt on pddp2 (13);
evio33tt on pddp3 (13);
evio34tt on pddp4 (13);
evio35tt on pddp5 (13);
evio36tt on pddp6 (13);


```
evio31tt on gte2p1 (26);
evio32tt on gte2p2 (26);
evio33tt on gte2p3 (26);
evio34tt on gte2p4 (26);
evio35tt on gte2p5 (26);
evio36tt on gte2p6 (26);
```

```
%c#1%
```

```
!First Event
```

```
[evio11$1@0 evio12$1@0 evio13$1@0] (1);
[evio14$1@0 evio15$1@0 evio16$1@0] (1);
```

```
!Second Event
```

```
[evio21tt$1@0 evio22tt$1@0 evio23tt$1@0] (1);
[evio24tt$1@0 evio25tt$1@0 evio26tt$1@0] (1);
```

```
!Third Event
```

```
[evio31tt$1*0 evio32tt$1*0 evio33tt$1*0] (3);
[evio34tt$1*0 evio35tt$1*0 evio36tt$1*0] (3);
```

Model 15

DATA:

File is RIAdata.dat;

VARIABLE:

Names are case,whiteh,ms,dwi,dep,...;

Usevar are evio11-evio16
evio20gt-evio25gt evio30gt-evio34gt
edw1 inc5 inc6 lor lorsq
viopre5 viopre6 tte2
pddp1 pddp2 pddp3 pddp4 pddp5 pddp6;

Missing are all(999);

Categorical are evio11-evio16
evio20gt-evio25gt evio30gt-evio34gt;

Classes = c(2);

ANALYSIS:

Type=Mixture missing;

MODEL:

%overall%

!First Event

evio11-evio16 on edw1 (11);

evio11-evio16 on inc5 (12);
evio11-evio16 on inc6 (12);

evio11 on pddp1 (13);
evio12 on pddp2 (13);
evio13 on pddp3 (13);
evio14 on pddp4 (13);
evio15 on pddp5 (13);
evio16 on pddp6 (13);

```

eta01 by evio11-evio16@1;

eta11 by evio11@0 evio12@1 evio13@2
      evio14@3 evio15@4 evio16@5;

[eta01* eta11*];

!Second Event

evio20gt-evio25gt on lor (21);
evio20gt-evio25gt on lorsq (22);

eta02 by evio20gt-evio25gt@1;

eta12 by evio20gt@0 evio21gt@1 evio22gt@2
      evio23gt@3 evio24gt@4 evio25gt@5;

[eta02* eta12*];

!Third Event

evio30gt-evio34gt on viopre5 (31);
evio30gt-evio34gt on viopre6 (31);

evio30gt-evio34gt on tte2 (32);

c#1 on edw1;

%c#1%

!First Event

[evio11$1@0 evio12$1@0 evio13$1@0] (1);
[evio14$1@0 evio15$1@0 evio16$1@0] (1);

!Second Event

[evio20gt$1@0 evio21gt$1@0 evio22gt$1@0] (2);
[evio23gt$1@0 evio24gt$1@0 evio25gt$1@0] (2);

!Third Event

[evio30gt$1*0 evio31gt$1*0 evio32gt$1*0] (3);
[evio33gt$1*0 evio34gt$1*0] (3);

```

!First Event

evio11-uvio16 on edw1 (11);

evio11-uvio16 on inc5 (12);

evio11-uvio16 on inc6 (12);

evio11 on pddp1 (13);

evio12 on pddp2 (13);

evio13 on pddp3 (13);

evio14 on pddp4 (13);

evio15 on pddp5 (13);

evio16 on pddp6 (13);

eta01 by evio11-evio16@1;

eta11 by evio11@0 evio12@1 evio13@2

evio14@3 evio15@4 evio16@5;

[eta01*5];

[eta11*-1];

!Second Event

evio20gt-evio25gt on lor (21);

evio20gt-evio25gt on lorsq (22);

eta02 by evio20gt-evio25gt@1;

eta12 by evio20gt@0 evio21gt@1 evio22gt@2

evio23gt@3 evio24gt@4 evio25gt@5;

[eta02*] (23);

[eta12*] (24);

!Third Event

evio30gt-evio34gt on viopre5 (31);

evio30gt-evio34gt on viopre6 (31);

evio30gt-evio34gt on tte2 (32);

%c#2%

!First Event

[evio11\$1@0 evio12\$1@0 evio13\$1@0] (201);
[evio14\$1@0 evio15\$1@0 evio16\$1@0] (201);

!Second Event

[evio20gt\$1@0 evio21gt\$1@0 evio22gt\$1@0] (202);
[evio23gt\$1@0 evio24gt\$1@0 evio25gt\$1@0] (202);

!Third Event

[evio30gt\$1*0 evio31gt\$1*0 evio32gt\$1*0] (203);
[evio33gt\$1*0 evio34gt\$1*0] (203);

!First Event

evio11-evio16 on edw1 (110);

evio11-evio16 on inc5 (120);
evio11-evio16 on inc6 (120);

evio11 on pddp1 (130);
evio12 on pddp2 (130);
evio13 on pddp3 (130);
evio14 on pddp4 (130);
evio15 on pddp5 (130);
evio16 on pddp6 (130);

eta01 by evio11-evio16@1;

eta11 by evio11@0 evio12@1 evio13@2
evio14@3 evio15@4 evio16@5;

[eta01*0];
[eta11*0];

!Second Event

evio20gt-evio25gt on lor (210);
evio20gt-evio25gt on lorsq (22);

eta02 by uvio20gt-uvio25gt@1;

eta12 by uvio20gt@0 uvio21gt@1 uvio22gt@2
uvio23gt@3 uvio24gt@4 uvio25gt@5;

[eta02*] (230);

[eta12*] (240);

!Third Event

evio30gt-evio34gt on viopre5 (310);

evio30gt-evio34gt on viopre6 (310);

evio30gt-evio34gt on tte2 (32);

Bibliography

- (2001). *Publication Manual of the American Psychological Association*. American Psychological Association, Washington, DC, 5th edition.
- Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6:701–726.
- Agresti, A. (1990). *Categorical Data Analysis*. John Wiley & Sons, Inc., New York.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akadémiai Kiadó.
- Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, 13:61–98.

- Allison, P. D. (1984). *Event History Analysis: Regression for Longitudinal Event Data*. Sage Publications, Inc., Newbury Park, CA.
- Allison, P. D. (1995). *Survival Analysis Using the SAS System: A Practical Guide*. SAS Institute, Inc., Cary, NC.
- Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, 10:1100–1120.
- Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92(440):1375–1386.
- Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Prentice-Hall, Inc., Englewood Cliffs, NJ.
- Clogg, C. C. (1977). Some latent trait structure models for the analysis of Likert-type data. *Social Science Research*, 8:287–301.
- Collet, D. (1994). *Modelling Survival Data in Medical Research*. Chapman & Hall, London.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, 34(2):187–220.

- Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39:1–38.
- Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.
- Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association*, 83(402):414–425.
- Elbers, C. and Ridder, G. (1982). True and spurious duration dependence: The identifiability of the proportional hazard model. *Review of Economic Studies*, 49(3):403–409.
- Fals-Stewart, W. (2003). The occurrence of partner physical aggression on days of alcohol consumption: A longitudinal diary study. *Journal of Consulting and Clinical Psychology*, 71(1):41–52.
- Farewell, V. T. (1982). The uses of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38(4):1041–1046.

- Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival*. Wiley, New York.
- Flinn, C. J. and Heckman, J. J. (1982). New methods for analyzing individual event histories. *Sociological Methodology*, 13:99–140.
- Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, 87(418):476–486.
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:215–231.
- Greenwood, M. (1926). The nature of cancer duration. In *Reports on Public Health and Medical Subjects 33*, pages 1–26. His Majesty's Stationery Office, London.
- Guo, S. W. and Lin, D. Y. (1994). Regression analysis of multivariate grouped survival data. *Biometrics*, 50:632–639.
- Gupta, P. L. and Gupta, R. C. (1996). Ageing characteristics of the weibull mixtures. *Probability in the Engineering and Informational Sciences*, 10:591–600.
- Hagenaars, J. A. and McCutcheon, A., editors (2002). *Applied Latent Class Analysis*. Cambridge University Press, Cambridge, England.

- Ham, J. C. and Rea, Jr., S. A. (1987). Unemployment insurance and male unemployment duration in canada. *Journal of Labor Economics*, 5(3):325–353.
- Heckman, J. J. and Singer, B. (1984a). Econometric duration analysis. *Journal of Econometrics*, 24(84):63–132.
- Heckman, J. J. and Singer, B. (1984b). The identifiability of the proportional hazard model. *The Review of Economic Studies*, 51(2):231–241.
- Hedeker, D., Siddiqui, O., and Hu, F. B. (2003). Random-effects regression analysis of correlated group-time survival data. In press.
- Hogg, R. V. and Craig, A. T. (1978). *Introduction to Mathematical Statistics*. Macmillan Publishing Co., Inc., New York.
- Holford, T. R. (1976). Life tables with concomitant information. *Biometrics*, 32(3):587–597.
- Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley & Sons, Inc., New York, 2nd edition.
- Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag, Inc., New York.

- Huang, G.-H. and Bandeen-Roche, K. (2003). Building latent class regressions with covariate effects on underlying and measured variables. In press.
- Jöreskog, K. G. and Sörbom, D. (1979). *Advances in Factor Analysis and Structural Equation Models*. Abt Books, Cambridge, MA.
- Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481.
- Kelly, P. J. and Lim, L. L.-Y. (2000). Survival analysis for recurrent event data: An application to childhood infectious diseases. *Statistics in Medicine*, 19:13–33.
- Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, Inc., New York.
- Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811.
- Laird, N. and Oliver, D. (1981). Covariance analysis of censored survival data

using log-linear analysis techniques. *Journal of the American Statistical Association*, 76(374):231–240.

Lamport, L. (1994). *L^AT_EX: User's Guide and Reference Manual*. Addison-Wesley Publishing Company, Reading, MA, 2nd edition.

Land, K. C., Nagin, D. S., and McCall, P. L. (2001). Discrete-time hazard regression models with hidden heterogeneity: The semiparametric mixed poisson regression approach. *Sociological Methods and Research*, 29(3):342–373.

Larsen, K. (2003). Joint analysis of time-to-event and multiple categorical indications of latent classes. In press.

Larsen, R. J. and Marx, M. L. (1986). *An Introduction to Mathematical Statistics and Its Applications*. Prentice-Hall, Inc., Englewood Cliffs, NJ.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.

Lee, E. W., Wei, L. J., and Amato, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In Klein, J. P. and Goel, P. K., editors, *Survival Analysis: State of the Art*, pages 237–247. Kluwer Academic Publisher, Dordrecht.

- Lemis, L. M. (1995). *Reliability: Probabilistic Models and Statistical Methods*.
Prentice-Hall, Inc., Englewood Cliffs, NJ.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing
Data*. John Wiley & Sons, Inc., New York, 2nd edition.
- Lo, Y., Mendell, N. R., and Rubin, D. B. (2001). Testing the number of
components in a normal mixture model. *Biometrika*, 88(3):767–778.
- Maller, R. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*.
John Wiley & Sons, Inc., Chichester, England.
- McCutcheon, A. L. (1987). *Latent Class Analysis*. Sage Publications, Inc.,
Newbury Park, CA.
- McHugh, R. B. (1956). Efficient estimation and local identification in latent
class analysis. *Psychometrika*, 21:331–347.
- McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley &
Sons, Inc., New York.
- Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Be-
haviormetrika*, 29(1):81–117.
- Muthén, B. O. and Masyn, K. (2001). Discrete-time survival mixture analysis.

- Muthén, B. O. and Shedden, K. (1999). Finite mixture modelling with mixture outcomes using the EM algorithm. *Biometrics*, 55:463–469.
- Muthén, L. K. and Muthén, B. O. (2001). *Mplus User's Guide*. Muthén & Muthén, Los Angeles, CA.
- Nagin, D. S. and Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed poisson model. *Criminology*, 31:327–362.
- Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14:945–965.
- Osius, G. and Rojek, D. (1992). Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association*, 87:1145–1152.
- Pearson, K. (1894). Contributions to the theory of mathematical evolution. *Philosophical Transactions of the Royal Society of London A*, 185:71–110.
- Peduzzi, P. N., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. (1996). A simulation study of the number of events per variable in logistic regression. *Journal of Clinical Epidemiology*, 99:1373–1379.
- Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped

survival data with application to breast cancer data. *Biometrics*, 34(1):57–67.

Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68:373–379.

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA.

Schwartz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, 6:461–464.

Singer, B. and Spilerman, S. (1976). Some methodological issues in the analysis of longitudinal surveys. *Annals of Economic And Social Measurement*, 5:447–474.

Singer, J. D. and Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of event. *Journal of Educational Statistics*, 18(2):155–195.

Singer, J. D. and Willett, J. B. (2003). *Applied Lingitudinal Data Analysis: Modeling Change and Event Occurence*. Oxford University Press, New York.

- Steele, F. (2003). A multilevel mixture model for event history data with long-term survivors: An application to an analysis of contraceptive sterilisation in Bangladesh. *Lifetime Data Analysis*, 9:155–174.
- Tan, W. Y. and Chang, W. C. (1972). Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. *Journal of the American Statistical Association*, 67:702–708.
- Thompson, Jr., W. A. (1977). On the treatment of grouped observations in life studies. *Biometrics*, 33(3):463–470.
- Trussel, J. and Richards, T. (1985). Correcting for unmeasured heterogeneity in hazard models using the Heckman-Singer procedure. *Sociological Methodology*, pages 242–276.
- Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association*, 69:169–173.
- Van de Pol, F. and Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology*.
- Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of hetero-

geneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

Vermunt, J. K. (1997). *Log-Linear Models for Event Histories*. Sage Publications, Inc., Thousand Oaks, CA.

Vuong, Q. H. (1989). Likelihood ratio test for model selection and non-nested hypotheses. *Econometrica*, 57:307–333.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84:1065–1073.

Willett, J. B. and Singer, J. D. (1993). Investigating onset, cessation, relapse, and recovery: Why you should, and how you can, use discrete-time survival analysis. *Journal of Consulting and Clinical Psychology*, 61(6):952–965.

Willett, J. B. and Singer, J. D. (1995). It's déjà vu all over again: Using multiple-spell discrete-time survival analysis. *Journal of Educational and Behavioral Statistics*, 20(1):41–67.

Yamaguchi, K. (1991). *Event History Analysis*. Sage Publications, Inc., Newbury Park, CA.

