EVENT HISTORY ANALYSIS

STATISTICAL THEORY AND APPLICATION IN THE SOCIAL SCIENCES

Hans-Peter Blossfeld
Alfred Hamerle
Karl Ulrich Mayer

Psychology Press

Event History Analysis

Statistical Theory and Application in the Social Sciences

This page intentionally left blank

Hans-Peter Blossfeld
Alfred Hamerle
Karl Ulrich Mayer

Event History Analysis

Statistical Theory and Application in the Social Sciences

Psychology Press
Taylor & Francis Group
NEW YORK AND LONDON

First Published 1989 by
Lawrence Erlbaum Associates, Inc.

Published 2014 by Psychology Press
711 Third Avenue, New York, NY 10017

and by Psychology Press
27 Church Road, Hove, East Sussex, BN3 2FA

Psychology Press is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in Publication Data

Blossfeld, Hans-Peter.
Event history analysis/ Hans-Peter Blossfeld, Alfred Hamerle, Karl Ulrich Mayer.
p. cm.
Includes index.
1. Event history analysis. I. Hamerle, Alfred, 1947– .
II. Mayer, Karl Ulrich. III. Title.
H61.B49 1989
001.4'3—dc19

88-7073
CIP

ISBN 13: 978-0-805-80126-2 (hbk)

Publisher's Note
The publisher has gone to great lengths to ensure the quality of this reprint but points out that some imperfections in the original may be apparent.

Foreword

In the social sciences, especially in economics and sociology, there is an increasing interest in the analysis of event histories. Compared to traditional panel or time-series data, event histories are often better suited to the dynamic nature of empirical phenomena. For each unit of analysis event histories provide information about the exact duration until a state transition as well as the occurrence and sequence of events. Examples of event histories include the survival rates of patients in medical studies; periods of unemployment in economic studies; the "lifetime" of political systems in the field of political science; the time span in which a technical apparatus works without defect in engineering science; required learning time in psychological research; periods of stability in migration and mobility analyses; recidivism in criminological studies; the length of time in which children remain living in their parent's household in youth and family sociological studies, and so on.

The statistical theory and practical examples of event history analysis presented in this book are thus of interest to readers in a large circle of disciplines. However, the examples presented in this book are especially designed for the needs of modern economic and social science research.

The book is written for students and scientists who want to learn how to analyze event history data. It also may be used as a handbook and reference guide for users in practical research. We have tried to present the statistical foundations of event history analysis and we have especially attempted to illustrate the entire research path required in applications of event history analysis: (1) the problems of recording event oriented data; (2) specific questions of data organization; (3) the application of statistical programs; and (4) interpretation of the obtained results.

Compared with other textbooks in this field of applied statistics, it was our special intention in writing this book to provide many examples of studies in which covariates are included in semiparametric and parametric regression models. We have also sought to complement practical examples with concise explanations of the underlying statistical theory. Parameter-free methods of analysis of event history data and the possibilities for their graphical presentation are also discussed in detail. Much space is devoted to the specific problems of multistate and multiepisode models, the introduction of time-depending covariates, and the question of unobserved population heterogeneity. Detailed examples demonstrate how to check the assumptions of the models, how to test hypotheses, and how to choose the right model.

The data used in examples throughout the book are drawn from the German Life History Study (GLHS) conducted by Karl Ulrich Mayer as principal investigator at the Max Planck Institute for Human Development and Education in Berlin. The original data collection was funded by the Deutsche Forschungsgemeinschaft (DFG) within its Sonderforschungsbereich 3 "Micro-analytical Foundations of Social Policy" (Mikroanalytische Grundlagen der Gesellschaftspolitik).

Work on this textbook has been part of the research project "Life Courses and Social Change" (Lebensverläufe und gesellschaftlicher Wandel) at the Max Planck Institute for Human Development and Education in Berlin in close collaboration with the Statistics Department of the University of Constance, FRG. We wish to express a note of thanks to the Max Planck Institute for Human Development and Education for its support in the publication of both the German and English version of this book.

Our special thanks go to Doris Gampig, who in a highly professional manner and with great care and precision prepared the manuscript. Also, we wish to express our gratitude to Ulrich Kuhnert and Dieter Schmidt for the layout and preparing the figures, tables, and program examples. Our appreciation also goes to Gottfried Pfeffer, Peter Wittek, and the copy editors at Lawrence Erlbaum Associates for their editorial assistance. Furthermore, we wish to thank Gerhard Tutz from the University of Regensburg, FRG, for commenting on parts of the manuscript. Bettina Althainz, Peter Baumann, Holger Hainke, and Rolf Hackenbroch provided research assistance in the preparation of the practical examples. Finally, we wish to express our thanks to Trond Petersen (Harvard University) for his kind permission to document his BMDP subprogram P3RFUN.

The English version of this book has been translated by Michael B. Gilroy of the Department of Economics (University of Constance, FRG) and revised by Constance A. Witte (Berlin) and Jacqui Smith (Max Planck Institute for Human Development and Education, Berlin). We are grateful to Frank Schwoerer of Campus Publishing House for granting permission for the English language publication. We are especially indebted to Aage B. Sørensen (Harvard University), David L. Featherman (University of Wisconsin at Madison), and John Nesselroade (Pennsylvania State University), who took the initiative for the English publication. Last but not least we would like to express our appreciation to Aage B. Sørensen (Harvard University) and Michael T. Hannan (Cornell University) for their comments on the statistical parts of this book. Although these friends and colleagues eliminated some of our mistakes, only the authors bear the responsibility for those that remain.

Berlin and Constance, April 1988

Hans-Peter Blossfeld
Alfred Hamerle
Karl Ulrich Mayer

Contents

Foreword 5

1. Aim and Structure of the Book 11

2. Domains and Rationale for the Application of Event History Analysis 14

2.1 Application Examples 15

2.2 Life Course Studies and the German Life History Study 17

2.3 Advantages of Event History Data 22

3. The Statistical Theory of Event History Analysis 26

3.1 Event History Analysis: A Special Stochastic Process 27

3.2 Fundamental Statistical Concepts (Single Spell Model) 30

3.2.1 Density, Distribution, and Survivor Functions, and the Hazard Rate 31

3.2.2 Special Probability Distributions for Durations 34

3.2.3 Life Table Method 42

3.2.4 Product-Limit Estimator (Kaplan-Meier Estimator) of the Survivor Function 44

3.2.5 Comparisons of Survivor Functions 45

3.3 Introducing Covariates: Regression Models 47

3.3.1 Quantitative and Qualitative Covariates 47

3.3.2 Parametric Regression Models 50

3.3.3 Cox's Proportional Hazards Regression Model 55

3.4 Multistate Models—Competing Risks 57

3.5	Regression Models for the Multiepisode Case	60
3.6	Estimation	64
3.6.1	The General Theory of Maximum Likelihood Estimation	65
3.6.2	Censoring	69
3.6.3	Maximum Likelihood Estimator for Parametric Regression Models	71
3.6.4	Cox Proportional Hazards Model: Partial Likelihood	73
3.6.5	Maximum Likelihood Estimation for Competing Risks Models	75
3.6.6	Maximum Likelihood Estimation of the Multiepisode Case	76
3.6.7	Least Squares Estimation	79
3.7	Hypotheses Tests and Model Choice	82
3.7.1	Residual Analysis and Model Tests	82
3.7.2	Proportional Hazards Model Tests	85
3.7.3	Tests for Regression Coefficients or Model Parts	87
3.8	Introducing Time-Dependent Covariates	89
3.9	Introducing Unobserved Population Heterogeneity	91
3.9.1	Examples of Unobserved Heterogeneity	91
3.9.2	Models and Parameter Estimation for Given Distributions of the Heterogeneity	95
3.9.3	Simultaneous Estimation of the Structural Model Parameter and the Distribution of Heterogeneity	98
3.9.4	Comparison of Some Estimators in the Presence of Unobserved Heterogeneity	99
3.9.5	Testing for Neglected Heterogeneity	101
3.10	Discrete Hazard Rate Regression Models	104
4.	**Data Organization and Descriptive Methods**	109
4.1	Managing Event Oriented Data Structures	109
4.2	Graphical Presentation of Event History Data	113
4.3	Life Table Method and Kaplan-Meier Estimator	117
5.	**Semi-Parametric Regression Models: The Cox Proportional Hazards Model**	141
5.1	Testing the Proportionality Assumption	142

5.2 Interpretation of the Estimated Results 149
5.3 Stepwise Regression Applied to the Cox Model 152
5.4 Introducing Time-Dependent Covariates 159
5.4.1 Discrete Time-Dependent Covariates 161
5.4.2 Continuous Time-Dependent Covariates 166
5.5 Modeling Multistate Models 168

6. Parametric Regression Models 175
6.1 Checking the Parametric Distribution Assumptions Graphically 176
6.2 Models Without Duration Dependency of the Hazard Rate: The Exponential Model 185
6.2.1 The Exponential Model Without Covariates 186
6.2.2 The Exponential Model With Time-Constant Covariates 189
6.2.3 Examination of the Residuals in the Exponential Model 194
6.3 Inclusion of Time-Dependent Covariates in Parametric Models 199
6.3.1 Method of Episode Splitting Given Discrete Time-Dependent Covariates 199
6.3.2 Method of Episode Splitting Given Continuous Time-Dependent Covariates 206
6.4 Models With Periodic Durations 211
6.5 Models With Duration Dependency of the Hazard Rate: The Gompertz-(Makeham), Weibull, Log-Logistic, and Lognormal Model 215
6.5.1 Gompertz-(Makeham) Model 217
6.5.2 Weibull Model 238
6.5.3 Log-Logistic Model 249
6.5.4 Lognormal Distribution 261
6.6 Models With Unobserved Heterogeneity 263

Appendices

Appendix 1: List of Variable Names Used in Examples 271
Appendix 2: Listing of the FORTRAN Program P3RFUN Written by Trond Petersen 274

Appendix 3: Listing of the FORTRAN Program for Episode Splitting Given Discrete Time-Dependent Covariates 283

Appendix 4: Listing of the FORTRAN Program for Episode Splitting Given Continuous Time-Dependent Covariates 285

Appendix 5: Listing of the GLIM Macros to Estimate the Weibull and Log-Logistic Models of Roger and Peacock 286

References 288

Index 295

Chapter 1:
Aim and Structure of the Book

This book tries to give a comprehensive overview of the most important methods of event history analysis. By "event history analysis" we mean statistical methods used to analyze time intervals between successive state transitions or events. The number of states occupied by the analyzed units are finite, but the events may occur at any point in time. Consequently, in event history analyses statistical methods for analyzing stochastic processes with discrete states and continuous time are used.

A wide range of statistical tools are available today to analyze event history data as exemplified in a variety of models, approaches, and methods. These statistical methods have, however, not yet found their place in standard statistics textbooks. There are several reasons for this. First, event history analysis applies stochastic models that are not often found in normal applications. Second, incomplete or censored data frequently occur only in very specific research designs. And third, due to the development and application of these methods in various disciplines such as medicine, demography, technology, economics, and the social sciences, the terminology is not uniform and thus the methods are not easily accessible to the user.

Consequently, the aim of this book is to show how modern statistical methods can be used to analyze event history data as well as to give some examples of event history analysis in practical research. To complement a comprehensive presentation of the statistical background we will use examples taken from current sociological research to illustrate the applications of event history analysis.

Following this overview (Chapter 1), Chapter 2 first illustrates three different ways in which in event history analysis problems are conceptualized and solved. Section 2.1 discusses the wide palette of subject areas in which event history analysis may be applied. Section 2.2 reviews the design of the German Life History Study (GLHS) which forms the empirical base of the examples used in Chapters 4 to 6. Finally, the theoretical and methodological advantages of collecting and analyzing event history data as compared to cross-sectional and traditional panel data are discussed in Section 2.3.

The statistical foundations of event history analysis are presented in Chapter 3. In addition to the classification of event history analysis within the structure of stochastic processes, Section 3.1 presents the basic concepts of event history analysis such as the hazard rate, the survivor function, cumulative hazard rates, and so on, as well as nonparametric methods of estimation including the life table method and the Kaplan-Meier estimator (Section 3.2). Of special importance for the textbook is Section 3.3 in which the inclusion of explanatory variables in semiparametric Cox models and parametric models such as the exponential, Weibull, Gompertz-(Makeham), and the log-logistic model are presented. The general theory of multiple state and multiple event cases are then given in Sections 3.4 and 3.5. Section 3.6 follows with the maximum likelihood estimation of unknown model parameters and Section 3.7 discusses methods of constructing hypotheses tests and how to choose models. The inclusion of time-dependent covariates is dealt with in Section 3.8, and models with unobserved population heterogeneity are presented in Section 3.9. Finally, the theoretically oriented Chapter 3 closes with a brief presentation of hazard rate models with discrete time.

Chapters 4 to 6 are specifically designed for the potential users of event history analysis in research. One may, however, also use the material in these chapters as a type of workbook to introduce the empirical analysis of occupational and job trajectories within labor market research. Based on the GLHS, the strategies required for preparing and evaluating event history data are discussed in a stepwise fashion. Using concrete examples, it is then shown how the formulation of a research question may be realized at the methodological and statistical level, which available computer program packages are adequate (SPSS, BMDP, GLIM, RATE, SAS) for specific analytical aims, how the control cards must be structured, and how the results of the analyses are to be interpreted and evaluated.

Chapter 4 looks at aspects of the technical implementation of event history data structures (Section 4.1) and the various ways to present them graphically (Section 4.2). The application of the life table method and the Kaplan-Meier estimator are also discussed (Section 4.3).

Chapter 5 focuses on the application of Cox models and the partial likelihood estimation. After an examination of the proportionality assumption in Section 5.1, the interpretation of the Cox model is discussed in detail in Section 5.2. Model choice with the aid of stepwise regression is demonstrated in Section 5.3. Especially important for the application of event history analysis in economics and the social sciences are those instances in which time-dependent covariates are introduced into a Cox model (Section 5.4) and the practical application of the multiple state cases (Section 5.5).

Chapter 6 is devoted to the application of parametric models. After the graphical examination of the distribution assumptions in Section 6.1, Section 6.2 discusses in detail the exponential model, its interpretation and residual analysis. This is followed by examples of introducing time-dependent covari-

ates with the aid of episode splitting (Section 6.3) and examples of models with periodical durations (Section 6.4). Special duration models are presented in Section 6.5, whereby extensive interpretative examples and residual tests of the Gompertz-(Makeham) (Section 6.5.1), the Weibull (Section 6.5.2), the log-logistic (Section 6.5.3), and the lognormal distributions (Section 6.5.4) are given. Section 6.6 concludes with applications and examples of unobserved population heterogeneity for parametric models.

Chapter 2:
Domains and Rationale for the Application of Event History Analysis

In the fields of economics and the social sciences there are many good reasons for studying the processes and course of development. First of all, an adequate description of reality necessitates the systematic characterization of processes, change, and transitions. Naturally, this proposal is not new. However, interest in characterizing change has increased in a time that is seen as the turning point for many middle and long-term economic and social developments. Recently, it has been recognized that explanations based upon cross-sectional data are appropriate only in the relatively rare cases where there is no change in causal variables (Tuma and Hannan, 1984; Petersen, 1988). In other situations processes of change are best comprehended with the aid of longitudinal data. Furthermore, only those models of processes that capture the right causal mechanisms, and so do more than just account for certain outcomes, should be used as the basis of rational political intervention.

In the past, in the field of economics and the social sciences, the possibility to measure and formalize processes using mathematical models was rather limited. This was due, not only to the lack of available data, but also to the lack of mathematical and statistical methods. The application of differential equations requires continuously measured metric variables over time (Hannan, Blossfeld, and Schömann, 1988). These variables are sometimes available in economics as monetary units, but rarely in other social sciences. Two- and multiple-wave panel studies collect—as we show in Section 2.3—processes over time incompletely, and as a rule are distorted by externally set time points of data collection. On the other hand, time series analysis and the numerous types of econometric models require a large number of points of measurement with constant intervals.

Nowadays event histories are increasingly being collected or made available in which the exact time of transition between states of the unit analyzed are registered. Such data offers information about the exact duration until events and their sequence occur. In addition to these durations or waiting times, variables that individually or in combination influence the timing of an event are of interest. These may be time stable characteristics or attributes that vary over time.

2.1 Application Examples

In the following, some examples are presented that illustrate the specific way in which in event history analysis problems are conceptualized. It should then become quite clear, that event history analysis is amenable to a wide range of questions.

Example 1: Unemployment Studies

In labor market research, event history analysis has been applied to the study of unemployment (Heckman and Borjas, 1980; Flinn and Heckman, 1983; Heckman and Singer, 1982, 1984a; Hamerle, 1988; Sørensen, 1988; Hujer and Schneider, 1988). These studies start from the idea that in analyzing unemployment, cross sections of unemployed or the number of entrants into unemployment in a given period are only partially informative and may even be misleading. Such indicators do not permit differentiation between short and long-term unemployment, and time-dependent covariates may not be included in the analysis.

In unemployment studies, the successive phases of unemployment a worker experiences represent the "duration" variable that is included in event history analyses. Periods of unemployment might be terminated due to various reasons, for example, by beginning a new occupation, through governmental job programs, re-education or re-training, retirement or the recognition of an employment disability. Such different end states may be formulated and examined as "competing risks" or multiple state models.

Example 2: Consumer Behavior Studies

A wide range of applications for event history analysis are to be found in the area of consumer research. Various product brands are offered for sale in a market. Consumers choose and purchase one of the brand names and, at a later point in time, they may either purchase the same brand again or switch to another brand. In this example, an episode or duration is equivalent to the time a consumer sticks to a given product. The states are initiated by the various brand names.

According to the methods presented in this book, the durations of brand loyalty may be related to exogeneous influences, some of which may change over time. Such influence factors include demographic variables (e.g., age, sex, family status, household size), socio-economic characteristics (e.g., income, education, occupation, social status), geographical aspects (metropolitan area, countryside location), or psychological conditions (e.g., personal attitudes, preferences, price awareness, quality awareness, buying habits). Furthermore, the duration a product is purchased, may also depend upon previous experiences with the commodity. Data from the prior history of the consumer process can be included in models and analyses of consumer behavior with the aid of the methods presented in this book.

Example 3: Medical Studies on the Course of Illness

In recent years the methods discussed in this book have been used in analyses of the healing process and survival time in medical and epidemiological studies (see, e.g., Kalbfleisch and Prentice, 1980 and the examples discussed there). Most of these studies deal with one or more absorbing end states. Here, "absorbing" means that once a respective end state has been obtained it is no longer possible to leave it, as, for example, is the case in the death of a patient.

There are few medical multiepisode models in the empirical literature although they are often appropriate. For example, the course of an illness is usually a succession of various stages characterized by events such as remission or death. Hamerle (1985b), for example, studied in female patients, the periods of nonillness following a breast cancer operation. One of the interesting points with regard to this example is the finding that the time period of nonillness appeared to be an especially good predictor of the final survival rate. In this example, separate examinations of the respective phases with single episode models were not adequate descriptors of the problem because they did not take into consideration the inherent dependence of the events and their temporal occurrence. In this book, we suggest methods to deal with special cases like these.

Example 4: Learning Experiments in Psychology and Instruction Research

In the psychology of learning, event history analysis may be used to obtain information regarding the temporal process of learning. The durations being modeled and analyzed here are simply the time spells required for learning some specific fact or task. Here it is possible to observe the speed of learning in relationship to personal and environmental factors.

In practical research on instruction, for example, event history analysis based upon video recordings has been applied to evaluate the concentration spells of pupils. One would ask, for example, whether instructional groups within classes or level of achievement influenced the concentration levels of pupils (Felmlee and Eder, 1983).

Example 5: Insurance and Accident Studies

Application of event history analysis methods may be used to research accidents, especially the conditional risk of an accident occurring based on time dependency as well as numerous other risk factors. For example, the duration might simply be characterized by the length of time a driver has driven without having an accident. Possible explanatory variables could include age, number of miles driven, traffic context, type of automobile, and so on.

Example 6: Studies of Migration

When analyzing residential mobility and migration between regions, event history analysis also proves to be especially useful. The states are represented here by living in a given apartment or house, city or region and the episodes by the respective durations of residence. The rate of migration could be related to various factors and motives such as earning opportunities, the housing market, access to services, tenure status, or the recreational value of a place. Important in this regard, is the question whether migration is influenced by varying resources of persons or through other life events such as job change, marriage or childbirth (Courgeau, 1984, 1985; Mayer and Wagner, 1986; Sandefur and Scott, 1981; Wagner, 1987a, 1987b).

Example 7: Analysis of Family Formation and Fertility

Event history analysis is especially suited to the study of marriage, fertility, and divorce behavior. Although work on population research up till now has mostly used the simple life table method and has mostly not moved beyond aggregate data to individual life histories, a growing number of applications of event history analysis have been published in recent years (Michael and Tuma, 1985; Diekmann, 1987; Huinink, 1987; Papastefanou, 1987; Sørensen and Sørensen, 1986; Wu, 1988). Of interest is how an individual's age or length of prior marriage is related to fertility or divorce, for example, or how the "risks" for marriage, divorce, or pregnancy are distributed over time. For example, the risk of divorce is minimal directly after marriage, increases, and then decreases monotonically after several years of marriage. In all of these examples, the problem arises as to the choice of an appropriate functional relationship path. How such parametric models may be chosen and whether it is better to apply methods that leave the temporal development of risks unspecified is discussed in detail.

The divorce example is also illustrative of the problem of "heterogeneous populations." Consider the case in which, due to religious convictions, certain groups within a population may not face the risk of getting divorced at all (Diekmann and Mitter, 1984).

Example 8: Criminology Studies and Legal Research

In criminology, event history analysis may be used to study the inclination of a criminal recidivism amongst ex-prisoners. Duration here is defined as a spell of time between prison release and commitment of new criminal acts and may be related to certain resocialization and rehabilitation measures or to income and living conditions. Also, the number and duration of previous prison sentences may be included in an explanation (Diekmann, 1980). Another area of application could be legal research. Here, the length of time that passes until the conclusion of civil or criminal court cases may be analyzed dependent on court procedures, characteristics of the judge, or in relation to changes in legal statutes.

Example 9: Organization and Management Research

The survival time of political regimes, of firms, working groups, and similar institutions may also be effectively analyzed using event history analysis. Especially interesting in this new area of

research is the so-called "organization ecology" (Carroll and Huo, 1985, 1986; Hannan and Freeman, 1977; Freeman, Carroll, and Hannan, 1983; Carroll, 1984; Carroll and Delacroix, 1982). So far, event history analysis has been used to examine the "births" and "deaths" of newspapers, restaurants, and local worker union organizations of the 19th century.

These illustrative examples of the application of event history analysis in various empirical research areas naturally do not exhaust the potential uses. Many other areas of application can be imagined, especially in technology. In industrial reliability studies (where simpler survival models are well established), the determination of the influence of time-dependent covariates on the life span of a technical apparatus, particularly tests under extreme conditions of stress or general "accelerated life tests" appear very promising. With regard to the study of employment trajectories and occupational careers, we present in Chapters 4 to 6 further examples of practical applications of the method of event history analysis.

2.2 Life Course Studies and the German Life History Study

The recent rapidly increasing demand for longitudinal studies in the field of economics and the social sciences is closely linked to the general rise in interest in the study of the life course.

By the term *life course research* we would like to designate an interdisciplinary paradigm that has been emerging over the last decade or so. Its main objective is the representation of societal processes and the explanation of individual life events and life trajectories within a common formal, conceptual, and empirical frame of reference. The unit of analysis is the individual life course as an institutionalized sequence of activities and events in various life domains.

The observation plan involves mapping the flow of successive cohorts through institutionally defined events (such as leaving home, marriage, birth of children, job entry and exit, or retirement) and states or role incumbencies (such as class membership, marital or employment status, household membership, or schooling).

The life course paradigm is innovative not least in the sense that it is breaking down century-old barriers between scientific disciplines and schools of theory and is transcending long-held distinctions between micro- and macroanalysis. What in the past was separated into the fields of microeconomics, aggregate demography, migration theory, sociology of the family, social mobility, and status attainment research is being brought into a common and therefore—in regard to explanatory claims—competitive framework.

This development is documented by numerous research projects that have led to or are leading to many comprehensive event oriented data sets.

In the Federal Republic of Germany these include:

- the German Life History Study (GLHS) conducted at the Max Planck Institute for Human Development and Education in Berlin (see Mayer et al., 1988);
- the socio-economic panel, carried out within the framework of the DFG-Sonderforschungsbereich 3 and based at the German Institute for Economic Research (DIW), Berlin (see Hanefeld, 1987; Krupp and Hanefeld, 1987);
- the follow-up survey of former Gymnasium (upper secondary school) pupils (Gymnasiasten-Wiederholungsbefragung), conducted by Meulemann and Wiese at the Central Survey Archive in Cologne (see Meulemann et al., 1984);
- and the project "Generative behavior in Nordrhein-Westfalen" (see Strohmeier, Schultz, and Kaufmann, 1985) as well as the project "Labor market dynamics, family development, and generative behavior" conducted at the Institute for Population Research and Social Policy of the University of Bielefeld (see Birg et al., 1985).

Whereas most of these projects are at a relatively early stage (see Mayer and Tuma, 1987, 1988), a number of other similar studies have already been concluded in other countries, especially in connection with American, Norwegian, French, and Israeli life history studies (Featherman and Sørensen, 1983; Matras, 1983; Courgeau, 1984; Michael and Tuma, 1985).

A common approach in all these studies is the examination of educational and occupational histories from a dynamic perspective based on certain historical time periods; the changes observed are not restricted to the field of education and work, but also include other spheres of life (such as social origin, family, spatial mobility, etc.); finally the educational and occupational histories are recorded retrospectively, prospectively, or on the basis of process-induced data.

Similarly, it is the aim of the GLHS (Mayer et al., 1988) to demonstrate and reconstruct the German social history since the end of the Second World War using quantitative life histories. This study also examines the effects of social institutions, especially the educational system, the employment system, and the family on the individual life course. The study attempts to answer the following questions: What do the processes of family formation look like and to what extent have the life courses of women changed? Is the social system in the Federal Republic of Germany characterized by age norm processes and what effect do they have on individual life courses? How has the relationship between the educational and the employment system evolved and what transitional effects can be observed in the choice of careers in different birth cohorts? What quantitative importance do certain migration paths play and what is the degree of spatial mobility within life courses?

Because, however, the main goal of this book is to demonstrate, in a stepwise fashion, how to apply and use event history analysis given an event

history database, the examples presented concentrate primarily on life events and trajectories—rather than their macrosocial implications—and specifically on selected aspects of labor market processes.

The GLHS is useful for this purpose since it provides detailed data about the life histories of 2,171 women and men from the birth cohorts 1929–31, 1939–41, and 1949–51, collected in the years 1981–1983. These birth cohorts were chosen so that the respondents' phase of transitions from school to work fell in particularly significant periods in history: for the 1930 cohort, this transition phase lies in the immediate postwar period; those born around 1940 left school in a time of large-scale economic growth, and the cohort 1949–51 entered the labor market during a phase marked by the expansion of the welfare state. The underlying hypothesis is that these specific historic conditions at the point of transition had a substantial impact on the respondents' subsequent careers.

The educational and occupational histories of the GLHS were recorded retrospectively in accordance with the event oriented observation plan. This method is demonstrated by an abstract from the questionnaire where respondents were interviewed about their work careers (Figure 2.1). It is characteristic that apart from collecting theoretically interesting information about the area of employment, number of working hours, income, and so on, the exact beginning and end of each job were recorded on a monthly basis. When this information about the sequence of job episodes is combined with records of periods of training and interruption, the educational and occupational history of an individual can be completely reconstructed. Such an event oriented observation plan provides detailed information on the states of a given respondent's career at any point in the period of observation.

A study conducted prior to the actual drawing of the GLHS sample demonstrated that the reliability of retrospectively recorded data about objective life histories is not systematically affected by a lack of ability to answer questions or deficient memory capacity (Papastefanou, 1980; Tölke, 1980). This study indicated that while the possibility of recall errors was minimal, the form and precision of the survey instrument proved to be of key importance with respect to the quality of the responses. In particular, it was important to divide the life history interview into different spheres of life (education/training, employment, residence, etc.). Lengthy and extensive data editing, data checks, and cross-comparisons also vouched for the quality of the collected information (Brückner et al., 1984). Finally, an examination of the representative quality of the life history data on the basis of census and microcensus surveys shows that the GLHS data provide a reliable picture of sociostructural cross-sections of the past (Blossfeld, 1987a).

Because the data cover not only educational and occupational histories, but also provide information on the whole spectrum of the various spheres of life (i.e., information on social background, family history, the spouse's history, residence history, etc.), it is possible to study the effects of events in

Figure 2.1: Example of an Event Oriented Observation Plan to Record Work Careers

400 Now I want to ask you about your occupation and employment. I shall proceed as I did for the other questions and go through all occupational activities, e.g., including part-time employment or temporary jobs you may have had. Any changes should be recorded as exactly as possible.
INT: If respondent was never employed—go on to Q414, p. 32.

401 Let's begin with your first job.
What occupation did you hold in your first job?
INT: Note exact job title in column 1, continue with Q402

401a What about your next job?
What was your occupation then?
INT: Continue with Q402

402 What was your exact activity at the beginning of this job?
INT: Note below and go on to Q403

403 How did your activity change during this job?—I'm also referring to e.g. changes between full-time and part-time jobs
INT: Let respondent describe the activities and note them down. For each activity go to the next box below. When all activities per page are filled in, go on to Q404

404 In what month and year did the job begin and in what month and year did the job end?

405 *INT: 1st job: Q405a, all subsequent jobs: Q405b*

405a Was this job in the firm in which you did your apprenticeship/ vocational training?
INT: Only ask for 1st job

405b Was this the same firm/place of employment as your previous job?

Occupation	Activity at the beginning and changes of activity		M.	Y.	Training establishment / 405b
(KA 3)		fr.			yes.....1
		to			no 2
(KA 4)		fr.			same firm 1
		to			other firm 2
(KA 5)		fr.			same firm 1
		to			other firm 2
(KA 6)		fr.			same firm 1
		to			other firm 2
(KA 7)		fr.			same firm 1
		to			other firm 2
(KA 8)		fr.			same firm 1
		to			other firm 2
(KA 9)		fr.			same firm 1
		to			other firm 2
(KA 10)		fr.			same firm 1
		to			other firm 2

406 What sector does (did) this firm belong to?
INT: Present blue list 6.

407 How many people are/were employed in the firm you work(ed) for?

408 Did your place of employment belong to the public sector?

409 What was your occupational status at the time?
Which of the following applies?
INT: Present white card C.

410 How many hours was your average working week in this job?

411 What about your working hours? Did you work regular hours or, e.g., shiftwork, nightwork, Sundays, etc.?

412 What was your net monthly salary at the beginning and end of your activity as *(INT: name activity)*?

413 Why did your activity then change? / Why did you change jobs?

Sector	Size of firm	yes/no	Occup. status	Hours	Work. hours	Net salary	
		yes ......1 no 2			norm. ..1 oth. 2	at start DM at end DM	
		yes ......1 no 2			norm. ..1 oth. 2	at start DM at end DM	
		yes ......1 no 2			norm. ..1 oth. 2	at start DM at end DM	
		yes ......1 no 2			norm. ..1 oth. 2	at start DM at end DM	
		yes ......1 no 2			norm. ..1 oth. 2	at start DM at end DM	
		yes ......1 no 2			norm. ..1 oth. 2	at start DM at end DM	
		yes ......1 no 2			norm. ..1 oth. 2	at start DM at end DM	
		yes ......1 no 2			norm. ..1 oth. 2	at start DM at end DM	

other parallel processes (e.g., in the case of family history, the event "marriage") have on occupational careers (e.g., "stability" of occupational trajectories). Similarly, prior history can be analyzed to examine the extent to which the subsequent career has been predetermined and channelled in certain directions. This database thus can be used as a good example for illustrating the various stages and potentials of event history analysis.

2.3 Advantages of Event History Data

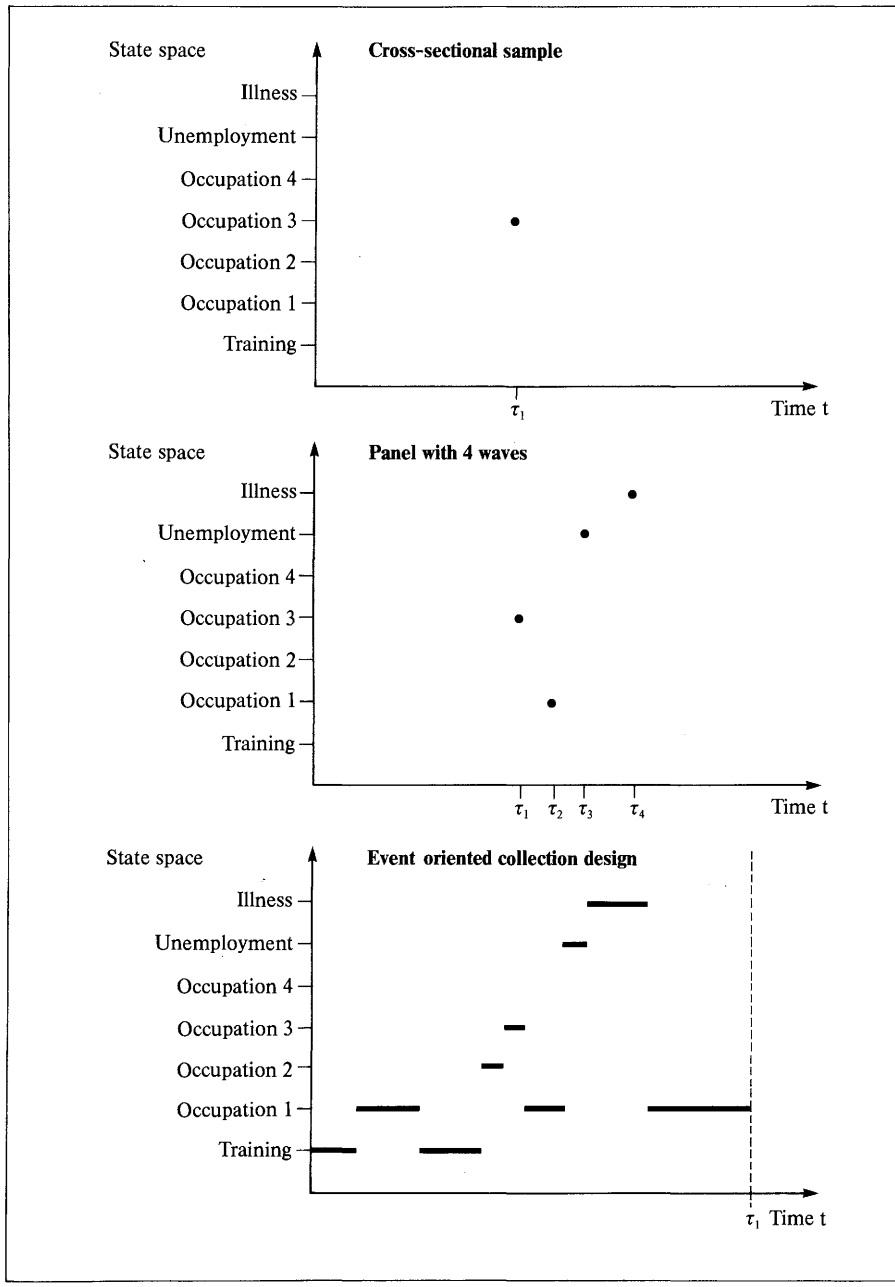
What are the advantages of the event oriented observation plan that make it so attractive to modern economic and social science research? In order to answer this question let us look at a simple example in which data has been collected for an individual with regard to education and occupation with the aid of a cross-sectional sample, a panel, and an event oriented sample design (Figure 2.2). The individual's career path is differentiated into seven states (training, occupation 1, occupation 2, occupation 3, occupation 4, unemployment, and illness) which the individual may occupy.

First, looking at Figure 2.2 one observes that in a cross-sectional survey the educational and occupational history of a person is only represented by a single point, that being the state at the time of the interview. Somewhat more information is obtained by the four-wave panel in which the circumstances of the respondent can be observed at four different points in time. However, the career between the four waves of the panel remains unclear. It is only the *event oriented collection design* in which changes in states and their precise times are explored. Such a design allows the educational and occupational career to be reconstructed in detail in its various phases and at any point in time.

This example illustrates the following:

- As a rule, cross-sectional analysis presupposes a steady state (i.e., the distribution at any given point in time is only informative if the underlying process remains relatively stable over time). In cases of major fluctuations and changes, the "snapshot" of a cross-section will not be a good picture of the situation because the analysis will depend upon the specific conditions prevailing at the time of survey. In contrast, panel and event oriented data explicitly take into account change and the dynamics of empirical phenomena. Accordingly, any research survey related to economic and social policy should be backed up by information based on longitudinal data on the level of the units of analysis.
- Even if empirical conditions are predominantly stable, panel and event history data are more informative than cross-sections. Cross-sectional data can be regarded as a special case of panel and event history data because cross-sections can be reconstructed from the latter. Moreover, in cases of empirical application, only the recording of panel or event history data can demonstrate whether stability really exists over time. Finally, unlike cross-

Figure 2.2: Recording of a Person's Educational and Occupational Career on the Basis of a Cross-Sectional Sample, a Panel and an Event Oriented Collection Design



sections, panel and event history data provide information on prior history which can help to improve the explanatory and prognostic capacity of statistical models.

- Whereas in the panel method the course of events between the individual survey points remains unknown, the event oriented observation plan permits the reconstruction of the continuous process. The panel method may also be suitable to determine the course of events if the changes take place at clearly defined points in time coinciding with the survey intervals (e. g., the determination of yearly income on a yearly basis) or if a continuous variable (e. g., a person's weight) can only be appropriately observed on the basis of time discrete surveys. Yet, all other changes in qualitative variables that may occur at any point in time can only be fully reconstructed if the states and time of their changes are exactly registered. Therefore, the event oriented observation plan proves to be a necessary precondition for the adequate reconstruction of change in many fields of research.
- Finally, if one considers the dynamic analysis of *complex feedback processes*, the continuous survey of qualitative variables would seem to be the only adequate method to assess empirical change. This is particularly true if the events of parallel processes occur not only at arbitrary points in time, but also have an interactive effect at a later stage.

The major advantage of an event oriented observation plan that concurs with the growing interest in the analysis of change is the fact that it permits an adequate representation of changes in qualitative variables which may occur at any point in time. The question nevertheless remains: Why has the event oriented observation plan thus far only seldom been used in economic and social science research?

One explanation can certainly be found in the extensive and costly observation procedures necessary to record event histories. One way of doing this is to observe the process and follow the development of the characteristics of individuals with the survey instrument over a lengthy period of time. However, it then often takes a long time before the data is finally available and theory has sometimes developed in a different direction. Event history data is therefore often collected retrospectively. As was also the case in the GLHS, the history of events is thus reconstructed over a long period of time. This type of data collection is sometimes the only way of obtaining event oriented information (e. g., today, the educational and occupational careers of the 1929–31, 1939–41, and 1949–51 birth cohorts can only be recorded on a retrospective basis). However, such data is often criticized as being unreliable, in particular when the events to be recalled took place in the distant past. Retrospective recording of event history data therefore requires a greater degree of care and control and this can generally only be achieved by extensive data checking and time consuming data editing. Moreover, if the data is retrospectively recorded on only one occasion or for only one birth cohort,

there is a considerable risk of the database becoming obsolete relatively quickly.

This is why in the case of the socio-economic panel (Hanefeld, 1987), the advantages of the traditional panel are combined with the retrospective recording of event history data. Thus each new panel wave provides not only up-to-date information for the time of survey, but by retrospective questions one also records the most important changes and their precise point of time between these waves (for comparison of panel and retrospective studies see Featherman, 1979-80).

Regardless which of the described procedures to record event history data is selected, it is always an *extensive and costly exercise.* However, this does not seem to be the main reason for the lack of usage of event oriented data. Another reason is certainly the fact that many economic and social science researchers simply do *not know how to use methods of dynamic analysis.* The structure of the data often is regarded as being too complex, the stochastic models that are part of event history analysis are also not well known, and the statistical programs required for samples with censored data are seldom used. This situation is changing rapidly as researchers recognize that in many cases it is unavoidable to base explanatory, causal, and dynamic inferences on event history data and corresponding stochastic models. Thus there exists a potentially strong demand for dynamic analysis of processes and courses in the fields of economics and the social sciences. This growing demand should lead to an increased supply of event history oriented data structures in these fields in the future.

Chapter 3:
The Statistical Theory of Event History Analysis

In this chapter we present the statistical fundamentals of event history analysis. After discussing the classification of event history analysis within the framework of stochastic processes in Section 3.1, in Section 3.2 we thoroughly discuss the fundamental concepts of event history analysis. The basic concepts presented are the hazard function (frequently referred to in the literature as the failure rate, the instantaneous death rate, or the force of mortality), the survivor function, the cumulative hazard rate, as well as some important classes of distributions that are relevant for describing the episode (spell, duration, lifetime), which is the period of time between successive events. These concepts are then defined for the one-episode case, although many of these statistical concepts are applicable to more complex situations exhibiting recurrent episodes or competing risks. Finally, this section concludes with a presentation of nonparametric estimation methods such as the life table technique and the Kaplan-Meier estimate (commonly referred to as the product-limit estimate) of the survivor function as well as comparative tests of survivor functions.

Of central importance in this chapter is Section 3.3 in which the inclusion of covariates in the regression approaches is presented. Parametric models, such as the exponential, Weibull, Gompertz or the log-logistic regression models, and the semi-parametric Cox model are discussed in depth. The following sections of this chapter deal with the inclusion of covariates or prognostic factors.

In Sections 3.4 and 3.5 general multistate and multiepisode regression models are analyzed. A general theory for the presentation of the models is developed in which a central concept (element) is the episode and state specific hazard function.

The maximum likelihood estimation of unknown model parameters is the subject of Section 3.6. After a brief introduction on the general principles of the maximum likelihood estimation procedure, the censoring problem that may arise when analyzing event history data is dealt with. This is then

followed by application of the maximum likelihood estimation method. For the Cox model, a modified approach is necessary.

In Section 3.7 we present methods of constructing tests of hypotheses and model choice, followed by the possibilities for the inclusion of time-dependent and stochastic covariates as exemplified in Section 3.8. Section 3.9 then discusses various methods, solutions, and potential problems of incorporating unobserved population heterogeneity in the analysis including individual specific disturbance terms.

Finally, we conclude this chapter with a brief presentation of regression methods for discrete hazard functions given "grouped" durations.

3.1 Event History Analysis: A Special Stochastic Process

The basic statistical model of an event history analysis examines the length of time intervals between consecutive changes of state defined by some qualitative variable within some observation period. Events are thus changes in the set of all distinct values that the chosen qualitative variable may take on within the state space. For the observation period, the points of time at which changes of state occur, or equivalently the occurrence of the sequence of events, is given. If the length of the time intervals or, the durations of the episodes can be measured exactly, we have a stochastic process with a continuous time parameter. Time is a continuous variable since changes of state may occur at any point in time. The state variable, on the other hand, possesses only a finite number of values. In the statistical model, points of time at which transitions occur are represented by a series of nonnegative random variables $0 = T_0 \leq T_1 \leq T_2 \leq \ldots$ and the state variable is characterized by the set of random variables with a finite state space $\{Y_k : k = 0, 1, 2, \ldots\}$. The corresponding stochastic process $(Y, T) = \{(Y_k, T_k) : k = 1, 2, \ldots\}$ may be described as

$$Z = \{Z(t) : t \geq 0\}$$

with $$Z(t) = Y_{k-1} \text{ for } T_{k-1} \leq t < T_k,\ k = 1, 2, \ldots$$

which is a continuous time, discrete state stochastic process.

Although from a theoretical point of view there may be a countable number of states, practical application normally requires only observation of a finite number of states. For example, in examining aspects of unemployment, a division of the state space into "employed," "not in the labor force," or "unemployed" is meaningful (see Figure 3.1). Occasionally, it is a matter of observing the times of events that occur repeatedly. Examples of such occurrences are the intervals of time between successive childbirths in demographic studies or functionability of some technical apparatus until the appearance of some defect in industrial reliability studies. In such cases, the process Y_k is said to be a *degenerate* process, which simply means that the state space consists of

only one element. The emphasis of the analysis is then upon the times T_k, k = 1, 2, ..., of the repeated occurrences of the event in question.

In any case, the important theoretical preliminary decision concerning the state space must be made in accordance with the nature of the substantive problem. The choice of the state space significantly affects the statistical model structure as well as the interpretation of the obtained results. With the exception of the above mentioned special case, the term "event" always corresponds to changes in Z(t), that is, with a transition from one state to another.

The term *episode* or *spell* designates the period of time between successive events. Of special interest are the duration intervals

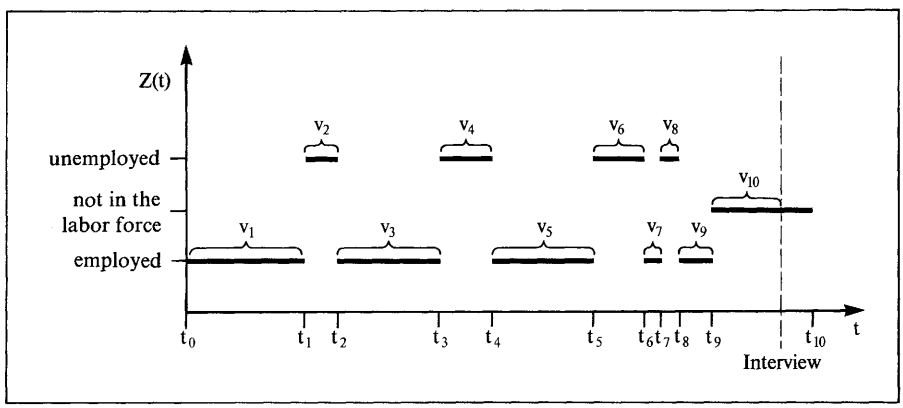
$$V_k = T_k - T_{k-1} \quad , k = 1, 2, \ldots$$

which are commonly referred to as the *waiting times.*

Example:

In research on durations of repeated episodes of unemployment, the states "employed," "not in the labor force," and "unemployed" may be distinguished. The complete history of state occupancies and times of changes, the *sample path* of an individual, is presented in Figure 3.1 below. The horizontal axis in Figure 3.1 represents time and the vertical axis states the person's status at time point t with regard to the three possible states.

Figure 3.1: Hypothetical Employment History of a Person



A very important special case that is often relevant in research studies is one in which processes exhibit only *one* episode, *one* initial state, and *one* destination state. Methods of dealing with this sort of data have been developed mainly in "survival analysis" in which research focuses on the distribution of lifetimes. If multiple outcomes on destination states exist, the method of

analysis needed may be characterized as a *multistate model* as commonly found in biometric literature under the catchword: competing risk models. Finally, *multiepisode models* are characterized by repeated transitions from one state to another or when a specific event may occur repeatedly.

Since the termination of the entire observation period is, as a rule, exogenous (e. g., due simply to the time point of the retrospective collection of event history data), the endpoint of the last episode of an individual or subject may not be observed. In such situations, event history data are said to be *right censored.* For example, it is conceivable that event history data for individual employment histories might have been collected on the first day of an individual's initial act of unemployment registration and thereafter the event history may be followed over a specific time span until some target date. The length of time between successive episodes in which the observed persons are unemployed will thus be measured. Under such a scenario it is possible that the end of the last unemployment episode in the final target date is not observable. For a detailed description of possible censoring mechanisms, the reader is referred to Kalbfleisch and Prentice (1980, Chapter 5), or Lawless (1982). The statistical modeling of the main censoring mechanisms will be presented below in Section 3.6.2.

Alternatively, there is the possibility that event history data are *left censored,* such that the length of time that an individual or subject has already resided in state y_0 is unknown. Left censoring is more difficult to handle than right censoring, since it is usually not possible to calculate the effects of the unknown event history data upon future events. For sake of simplicity, the following always assumes that either the starting point in time and the original state are given (without loss of generality that $t_0 = 0$), or that the previous history of state occupancies is irrelevant for the future event history process. Since all of the survey samples of the cohorts 1929–31, 1939–41, and 1949–51 of life histories applied for illustrative purposes in Chapters 4–6 include the individual event histories in their entirety, the problem of left censored data does not arise. For the incorporation of left censored observation analyzing hazard rate models, see Hamerle (1988a).

Often, not only the waiting times but also various covariates or prognostic factors that may affect the waiting time singly or in combination are included in the analysis. One important goal of the statistical analysis of event histories is, therefore, the quantitative determination of the impact of such exogenous or endogenous variables with suitable regression models. In such analyses the covariates may be either quantitative or qualitative. Categorical covariates can be coded in the regression model by appropriate dummy variables. This is discussed further below in Section 3.3.1. Some of the covariates may also be time dependent and stochastic.

In the simplest case, a time dependent covariate is a fixed exogenous function of time, such as age. Some covariates may, however, be stochastic processes that are active parallel to the main process being observed. As such,