

CONTENTS

List of Tables	xii
List of Figures	xiv
List of Boxes	xvii
Preface	xviii
1 The fundamentals of survival and event history analysis	1
1.1 Introduction: what is survival and event history analysis?	1
1.2 Key concepts and terminology	2
1.3 Censoring and truncation	4
1.3.1 Right-censoring	5
1.3.2 Interval censoring	6
1.3.3 Truncation	6
1.4 Mathematical expression and relation of basic statistical functions	7
1.5 Why use survival and event history analysis?	9
1.5.1 Potential problems that might arise if censored data is ignored	9
1.5.2 What does survival analysis offer that ordinary regression models do not?	11
1.6 Overview of survival and event history models and this book	11
1.6.1 Non-, semi- and parametric models	11
1.6.2 Outline of this book	14
Exercises	17
2 An introduction to R and data exploration via descriptive statistics and graphics	18
2.1 An introduction to R and data exploration	18
2.2 Downloading R on your personal computer	20
2.3 The R base system and add-on packages	21
2.3.1 Add-on packages and how to install them	21
2.3.2 Loading an add-on package	22
2.4 Running R	22
2.4.1 Running R interactively by typing at the > prompt	22
2.4.2 Running R non-interactively using a script file	23

2.4.3	Running R using the R Commander graphical user interface	24
2.5	Determining and setting your working directory	25
2.5.1	Determining your working directory	25
2.5.2	Setting a new working directory	26
2.6	Help and documentation	27
2.7	Importing data into R	27
2.7.1	Importing Stata or SPSS data into R	28
2.7.2	Importing ASCII text or Excel data into R	30
2.8	Working with data: opening and accessing variables from a data frame	31
2.8.1	Placing the name of the data within a function	32
2.8.2	Using the \$ sign	32
2.8.3	Using (and abusing) the attach function	33
2.8.4	Using data that is part of an existing library package	34
2.8.5	Saving data	34
2.9	Saving your work and quitting R	35
2.9.1	Save to file and capture output options	35
2.9.2	Quitting R and saving your workspace	35
2.9.3	Saving your history	37
2.10	Basic descriptive statistics	37
2.10.1	The example data	37
2.10.2	Descriptive summary statistics	38
2.11	Descriptive data exploration with graphics	43
	Exercises	45
3	Survival and event history data structures	47
3.1	Introduction: why discuss data structures?	47
3.2	Sources of event history data	48
3.3	Single-episode data	49
3.4	Multi-episode data	50
3.4.1	Understanding multi-episode data	50
3.4.2	Converting single-episode to multi-episode data	51
3.5	Subject-period (discrete-time) data, episode-splitting and counting process format	53
3.5.1	Subject-period or discrete-time data	53
3.5.2	Creating a subject-period file: survSplit in survival library	54
3.5.3	Creation of a subject-period file: to Binary in eha package	55
3.5.4	Episode-splitting	56
3.5.5	Counting process style of data	58

3.6	A note on dates	58
3.6.1	Using <code>as.Date</code>	59
3.6.2	Converting date variables to a numeric format	59
3.6.3	Using <code>chron</code>	60
	Exercises	61
4	Non-parametric methods: the Kaplan–Meier estimator	62
4.1	Introduction	62
4.2	The Kaplan–Meier (KM) estimator	63
4.3	Undertaking KM estimations in R	64
4.3.1	The survival package in R	64
4.3.2	Loading <code>RcmdrPlugin.survival</code> to use in the R Commander	66
4.4	Kaplan–Meier estimation	67
4.4.1	Producing KM estimates using the R Commander	67
4.4.2	Producing KM estimates with a script file	69
4.4.3	Interpretation of KM estimates	71
4.5	Plotting the Kaplan–Meier survival curve	73
4.5.1	Plotting a univariate KM survival curve	73
4.5.2	Comparing two KM survival curves	75
4.6	Testing differences between two groups using <code>survdif</code>	79
4.6.1	The Fleming–Harrington test	80
4.6.2	The log-rank (Mantel–Haenszel) test	80
4.6.3	The Peto and Peto test	81
4.6.4	Comparing tests: which test to choose?	82
4.7	Stratifying the analysis by a covariate	83
	Exercises	85
5	The Cox proportional-hazards regression model	86
5.1	Introduction: The Cox regression model	86
5.1.1	The Cox proportional hazard model with fixed covariates	87
5.1.2	The Cox proportional hazards model with time-varying covariates	89
5.1.3	Why is the Cox model so popular?	90
5.2	Estimating and interpreting the Cox model with fixed covariates	91
5.2.1	The <code>coxph</code> object	91
5.2.2	Estimating the Cox regression model	91
5.2.3	Interpreting covariate estimates in the Cox regression model	93
5.2.4	Significance of the model	97
5.2.5	Plotting the estimated survival function	98

5.2.6	Plotting the estimated survival function by a covariate	99
5.3	The Cox regression model with time-varying covariates	100
5.3.1	Creating a subject–period file to accommodate time-varying covariates	100
5.3.2	Modelling time-varying covariates using person-period data	103
5.3.3	Creating a subject–period file with lagged variables to reduce problems of causal ordering	106
5.3.4	Lagged time-varying covariates to reduce problems of causal ordering	107
5.3.5	Interactions with time as time-dependent covariates: episode-splitting at time intervals	108
	Exercises	113
6	Parametric models	114
6.1	Introduction	114
6.2	Relationship of the probability density, hazard and survival function	115
6.3	Proportional hazards (PH) versus accelerated failure time (AFT) models	116
6.4	Specification of parametric models	117
6.4.1	Summary of selected parametric survival distributions	117
6.4.2	The exponential model	118
6.4.3	Piecewise constant exponential model	121
6.4.4	The Weibull model	121
6.4.5	Log-logistic and log-normal models	124
6.4.6	Other parametric models	125
6.5	Estimating parametric survival models using the survival and eha packages	125
6.5.1	Estimating parametric models using the survreg function in the survival library	125
6.5.2	Estimating parametric models using the phreg and aftreg functions in the eha library	126
6.6	Estimation and interpretation of parametric models	126
6.6.1	Exponential model: PH parameterization	126
6.6.2	Exponential model: AFT parameterization	128
6.6.3	Piecewise exponential model: PH and AFT parameterization	133
6.6.4	Weibull model: PH parameterization	136
6.6.5	Weibull model: AFT parameterization	136
6.6.6	Log-logistic and log-normal models: AFT parameterization	137

6.7	What happens if a parametric model is specified incorrectly?	139
	Exercises	140
7	Model-building and diagnostics	141
7.1	Introduction	141
7.2	Model-building and selection of covariates and a model	142
7.2.1	Purposeful selection of covariates	142
7.2.2	The decision path to choosing an appropriate model	144
7.3	Assessing the overall goodness of fit of your model	146
7.3.1	The log-likelihood and likelihood ratio tests	146
7.3.2	Akaike information criterion (AIC) and evaluation of standard errors	148
7.4	Testing overall model adequacy: Cox–Snell residuals	149
7.5	Testing the proportional hazards assumption: Schoenfeld residuals	151
7.5.1	Understanding and estimating Schoenfeld residuals	151
7.5.2	Dealing with non-proportional hazards: introducing an interaction effect	154
7.5.3	Dealing with non-proportional hazards: stratifying the data	155
7.6	Checking for influential observations: score residuals	157
7.6.1	What should be done if influential observations are identified?	160
7.7	Assessing nonlinearity: martingale residual and component-plus-residual plots	160
	Exercises	163
8	Frailty and recurrent event models	164
8.1	Introduction	164
8.2	Shared frailty: modelling recurrent events and clustering in groups	166
8.2.1	Recurrent events	166
8.2.2	Shared clustering in groups	167
8.3	Additional frailty models: unshared, nested, joint and additive models	169
8.3.1	Individual (unshared) frailty models	169
8.3.2	Nested frailty models	170
8.3.3	Joint and additive frailty models	170
8.4	Estimating frailty models in R	171
8.4.1	Using the frailty function	171
8.4.2	The frailtypack and survrec library in R	171
8.5	Frailty model estimation and interpretation	172
8.5.1	Description of the data	172

8.5.2	Frailty model with a gamma distribution	173
8.5.3	Frailty model with a Gaussian distribution	177
	Exercises	178
9	Discrete-time models	179
9.1	Introduction	179
9.2	Discrete-time models	181
9.2.1	Specification of the hazard, survival and cumulative probability density functions	181
9.2.2	Models to estimate discrete-time data: logit, probit and complementary log-log functions	182
9.3	Restructuring data for discrete-time modelling	184
9.4	Estimation and interpretation of discrete-time models	184
9.4.1	Estimation of logit, probit and cloglog discrete-time models	184
9.4.2	Interpretation and comparison of estimates	187
9.5	Advantages and disadvantages of discrete-time models	189
	Exercises	189
10	Competing risk and multi-state models	190
10.1	Introduction	190
10.2	Competing risk models	191
10.2.1	Three central techniques to model competing risks	192
10.2.2	The latent or cause-specific approach	192
10.2.3	The cumulative incidence curve (CIC)	193
10.3	Estimating competing risks using the latent versus CIC approach	195
10.3.1	Data preparation and restructuring	195
10.3.2	Estimating CIC estimates and their standard errors	197
10.4	Regression analysis with competing risks	198
10.5	Multi-state models	202
10.5.1	A brief introduction to multi-state models and their applications	202
10.5.2	Markov, semi-Markov and extended Markov model properties	203
10.6	Estimation of multi-state models	204
10.6.1	Preparation of data for multi-state models using the <code>mstate</code> package	204
10.6.2	Estimation of Markov model with stratified hazards	207
10.6.3	Estimation of Markov model with proportional hazards	210

10.6.4	Estimation of state arrival extended Markov proportional hazards model	211
10.6.5	Further predictions and estimation of multi-state models with the cumulative incidence function	212
	Exercises	212
11	Sequence analysis	213
11.1	Introduction: sequence analysis	213
11.1.1	A brief introduction to sequence analysis	213
11.1.2	Optimal-matching techniques	215
11.2	Sequence analysis data and estimation using the TraMineR package	215
11.2.1	Sequence data	216
11.2.2	The transition from school to work using the mvad data	216
11.3	Describing and visualizing sequence datasets	217
11.3.1	Exploring the data, sequence frequency and state distribution plots	217
11.3.2	Calculating entropy and turbulence	219
11.4	Measuring similarities and distances between sequences	221
11.5	Producing typologies of trajectories: cluster analysis	221
11.6	Event sequence analysis	223
11.7	Criticisms of the OM approach and the dynamic future of sequence analysis	224
	Exercises	225
	Appendix 1: Description of the data used in this book	227
	Appendix 2: Survival and event history analysis using stata	232
	Glossary	255
	References	261
	Index	273