

Statistics for Biology and Health

Philip Hougaard

Analysis of Multivariate Survival Data

Survival data or more general time-to-event data occur in many areas, including medicine, biology, engineering, economics, and demography, but previously standard methods have required that all time variables are univariate and independent. This book extends the field by allowing for multivariate times. Applications where such data appear are survival of twins, survival of married couples and families, time to failure of right and left kidneys for diabetic patients, life history data with time to outbreak of disease, complications and death, recurrent episodes of diseases, and crossover studies with time responses.

As the field is rather new, the concepts and the possible types of data are described in detail, and basic aspects of how dependence can appear in such data are discussed. Four different approaches to the analysis of such data are presented. The multistate model where a life history is described as the subject moving from state to state is the most classical approach. The Markov model makes up an important special case, but it also describes how easily more general models are set up and analyzed. Frailty models, which are random effects models for survival data, make a second approach; they extend from the most simple shared frailty models, which are considered in detail, to models with more complicated dependence structures over individuals or over time. Marginal modelling has become a popular approach to evaluate the effect of explanatory factors in the presence of dependence, but without having specified a statistical model for the dependence. Finally, the completely nonparametric approach to bivariate censored survival data is described.

This book is aimed at investigators who need to analyze multivariate survival data. It can be used as a textbook for a graduate course in multivariate survival data. It is written from an applied point of view and covers all the essential aspects of applying multivariate survival models. More theoretical evaluations, like asymptotic theory, are also described, but only to the extent useful in applications and for understanding the models. To read the book, it is useful, but not necessary, to have an understanding of univariate survival data.

Philip Hougaard is a statistician at the pharmaceutical company Novo Nordisk. He has a Ph.D. in nonlinear regression models and is Doctor of Science based on a thesis on frailty models. He is associate editor of *Biostatistics*, *Lifetime Data Analysis* and *Statistical Modelling: An International Journal* and former associate editor of *Biometrics*. He has published over 90 papers in the statistical and medical literature. He recently received an honorary title as external professor at the University of Southern Denmark.

ISBN 978-1-4612-7087-4
www.springer-ny.com

ISBN 978-1-4612-7087-4



9 781461 270874



1	Introduction	1
1.1	What is survival data?	2
1.2	Univariate data	6
1.2.1	Population data	6
1.2.2	Survival of diabetics	7
1.3	Interval censored data	7
1.3.1	Diabetic nephropathy	8
1.4	Multivariate data structures	9
1.4.1	Parallel data	9
1.4.2	Longitudinal data	10
1.5	Several individuals	11
1.5.1	Danish twins	12
1.5.2	Adoption study	13
1.5.3	Litter-matched tumorigenesis experiment	13
1.5.4	Length of leukemia remission	15
1.6	Similar organs	16
1.6.1	Diabetic retinopathy	16
1.6.2	Amalgam fillings	17
1.7	Recurrent events	17
1.7.1	Mammary tumors	18
1.7.2	Epileptic seizures	19
1.7.3	Infections in catheters for patients on dialysis	21
1.7.4	Insurance claims	21
1.8	Repeated measurements	22

1.8.1	Watermaze	23
1.8.2	Exercise test times	23
1.9	Different events	24
1.9.1	Albuminuria	25
1.9.2	Colon cancer	26
1.9.3	Survival after myocardial infarction	26
1.9.4	Stanford heart transplant data	26
1.10	Competing risks	27
1.10.1	Radiation-exposed mice	28
1.10.2	Malignant melanoma	29
1.11	Types of censoring	29
1.12	Types of truncation	30
1.13	Explanatory variables	31
1.14	Purpose of multivariate survival studies	32
1.15	Chapter summary	33
1.16	Bibliographic comments	34
1.17	Exercises	35
2	Univariate survival data	36
2.1	Special features of survival data	37
2.1.1	Censored data	38
2.1.2	Conditioning	39
2.1.3	Choice of parametric/non-parametric models	40
2.1.4	Multiple events	42
2.2	Parametric methods	43
2.2.1	The exponential distribution	45
2.2.2	The Weibull distribution	47
2.2.3	The Gompertz distribution	54
2.2.4	Piecewise constant hazards	56
2.2.5	Hidden causes of death	58
2.2.6	Truncated distributions	58
2.2.7	Mixture distributions	59
2.2.8	Comparison of parametric models	65
2.3	Simple non-parametric methods	69
2.3.1	The Kaplan-Meier estimate	70
2.3.2	The Nelson-Aalen estimate	74
2.4	Regression models	76
2.4.1	The proportional hazards probability model	77
2.4.2	The partial likelihood estimation method	79
2.4.3	Estimation of absolute risk	83
2.4.4	Truncation and time-dependent covariates	84
2.4.5	Residuals and robust variance estimates	89
2.4.6	Neglected covariates	90
2.4.7	Computational aspects	92
2.4.8	Applications	93

2.4.9	Summary of the Cox model	98
2.5	Accelerated failure times	99
2.5.1	Comparison of accelerated failure time models and proportional hazards models	100
2.5.2	Neglected covariates	101
2.5.3	Applications	103
2.6	Counting processes	104
2.7	Chapter summary	105
2.8	Bibliographic comments	106
2.9	Exercises	107
3	Dependence structures	112
3.1	Probability mechanisms	113
3.1.1	Common events	114
3.1.2	Common risks models	114
3.1.3	Event-related dependence	116
3.1.4	Relevance of the mechanisms	117
3.2	Dependence time frame	117
3.2.1	Early/late dependence	119
3.2.2	Instantaneous dependence	120
3.2.3	Short-term dependence	120
3.2.4	Long-term dependence	121
3.3	Examples	122
3.3.1	Birth histories	122
3.3.2	Married couples	123
3.3.3	Twin dependence	123
3.3.4	Adoptive children	123
3.3.5	Traffic deaths	123
3.3.6	Epileptic seizures	124
3.3.7	Bleeding patterns	124
3.3.8	Amalgam fillings	125
3.3.9	Competing risks	125
3.4	Chapter summary	125
3.5	Bibliographic comments	126
3.6	Exercises	126
4	Bivariate dependence measures	128
4.1	Correlation coefficients	129
4.2	Kendall's coefficient of concordance	131
4.2.1	Applications	133
4.3	Spearman's correlation coefficient	134
4.4	Median concordance	135
4.5	Integrated hazard correlation	136
4.6	Local dependence measures	136
4.7	Chapter summary	137

4.8	Bibliographic comments	137
4.9	Exercises	138
5	Probability aspects of multi-state models	139
5.1	Transition probabilities	144
5.2	Observations and censoring patterns	145
5.3	Selection of state structure	145
5.3.1	The fertility model	146
5.3.2	The disability model	151
5.3.3	Bivariate parallel data	152
5.4	Progressive models	155
5.4.1	Competing risks	156
5.4.2	Bivariate parallel data	157
5.4.3	The disability model	157
5.4.4	Recurrent events	159
5.4.5	General progressive models	159
5.5	Homogeneous Markov processes	159
5.5.1	The disability model	161
5.5.2	Recurrent events	162
5.5.3	The Freund model	163
5.5.4	The Marshall-Olkin model	165
5.5.5	The combined model	166
5.5.6	The alternating state model	167
5.6	General Markov processes	167
5.7	Markov extension models	168
5.8	General models	169
5.9	Conditionally simple processes	170
5.10	Hidden states	170
5.11	Chapter summary	173
5.12	Bibliographic comments	173
5.13	Exercises	173
6	Statistical inference for multi-state models	177
6.1	Statistical models	178
6.1.1	The disability model	179
6.1.2	Bivariate parallel data	179
6.1.3	Recurrent events	180
6.1.4	Alternating states	180
6.1.5	General models	180
6.2	Likelihood evaluation	181
6.2.1	Markov likelihood functions	181
6.2.2	General likelihood functions	182
6.3	Estimation of transition hazards	183
6.3.1	Separate and common constant hazards	183
6.3.2	Parametric models	184

6.3.3	Non-parametric models for the disability model	185
6.3.4	Non-parametric models for bivariate parallel data	186
6.4	Estimation of transition probabilities	188
6.5	Censoring patterns and asymptotics	190
6.5.1	Estimation of variance	191
6.6	Model checking	192
6.7	Applications	192
6.7.1	Heart transplant data	193
6.7.2	Danish twins	194
6.7.3	Myocardial infarctions	201
6.7.4	Albuminuria progression	202
6.7.5	Other applications	206
6.8	Chapter summary	206
6.9	Bibliographic comments	207
6.10	Exercises	207
7	Shared frailty models	215
7.1	Unspecified frailty	218
7.2	General probability results	220
7.2.1	The conditional parametrization	221
7.2.2	The marginal parametrization	223
7.2.3	Updating	224
7.2.4	Weibull conditional distributions	227
7.2.5	Quantification of dependence	228
7.2.6	Derived quantities	230
7.2.7	Regression models	231
7.3	Gamma frailty models	233
7.3.1	Updating	234
7.3.2	Weibull models	235
7.3.3	Extension to negative dependence	236
7.4	Positive stable frailty distributions	237
7.4.1	Updating	238
7.4.2	Regression models	238
7.4.3	The stable-Weibull model	238
7.4.4	Weibull regression models	240
7.5	PVF frailty distributions	241
7.5.1	Weibull and regression models	243
7.5.2	Updating	243
7.6	Lognormal frailty distributions	244
7.7	Other distributions	245
7.8	Comparison of frailty distributions	246
7.8.1	Theoretical comparison	247
7.8.2	Comparison of fit	247
7.8.3	Quantification of dependence	252
7.9	Chapter summary	256

7.10	Bibliographic comments	257
7.11	Exercises	258
8	Statistical inference for shared frailty models	263
8.1	Parametric models	264
8.2	Simple estimates	265
8.3	The EM-algorithm	265
8.4	The three-stage approach	267
8.5	The full marginal estimate	268
8.6	The full conditional estimate	269
	8.6.1 The gamma model	270
	8.6.2 The PVF model	272
8.7	The penalized likelihood approach	274
8.8	Other approaches	275
8.9	Software	275
8.10	Goodness-of-fit	275
8.11	Asymptotic evaluations	277
	8.11.1 The gamma-Weibull model	278
	8.11.2 The stable-Weibull model	279
	8.11.3 The non-parametric gamma model	282
	8.11.4 The two-stage approach	283
8.12	Applications	283
	8.12.1 Tumorigenesis data	283
	8.12.2 Exercise data	285
	8.12.3 Catheter infections	288
	8.12.4 Amalgam fillings	290
	8.12.5 Danish twin data	291
	8.12.6 Other applications	308
8.13	Chapter summary	309
8.14	Bibliographic comments	309
8.15	Exercises	310
9	Shared frailty models for recurrent events	312
9.1	Notation and observation plan	314
9.2	Poisson processes	316
	9.2.1 Evaluation in SAS	317
9.3	Constant frailty overdispersion models	317
	9.3.1 Common observation period	318
	9.3.2 Negative binomial models for period count data	319
	9.3.3 Negative binomial models	321
	9.3.4 PVF frailty models for period count data	322
	9.3.5 PVF frailty models	328
	9.3.6 Positive stable frailty models	328
9.4	Dependence measures	329
9.5	Regression models	329

9.5.1	Period count data	330
9.5.2	Interval count data	330
9.5.3	Gap times	331
9.6	Estimation	331
9.6.1	Data of exact times	332
9.6.2	Period count data	332
9.6.3	Interval counts for several intervals	333
9.7	Asymptotics	333
9.8	Applications	334
9.8.1	Period count data	335
9.8.2	Time data for a fixed period	339
9.8.3	Interval count data	339
9.8.4	Other applications	340
9.9	Chapter summary	340
9.10	Bibliographic comments	341
9.11	Exercises	341
10	Multivariate frailty models	345
10.1	Comparison of dependence	348
10.1.1	Estimation	349
10.1.2	Application to twin data	350
10.2	Differential effect of a shared frailty	352
10.3	Shared frailty models for bivariate marginals	353
10.3.1	Applications	353
10.4	The multiplicative stable model	354
10.4.1	The trivariate model	355
10.4.2	Combined model for monozygotic and dizygotic twins	357
10.4.3	Treatment by center interaction	359
10.4.4	Goodness-of-fit	359
10.4.5	Estimation	359
10.4.6	Applications	360
10.5	Additive models	362
10.5.1	Bivariate additive gamma models	364
10.5.2	Compound symmetry model	369
10.5.3	Combined model for monozygotic and dizygotic twins	369
10.5.4	Treatment by center interaction	370
10.5.5	Father-mother-child model	370
10.5.6	General models	371
10.5.7	Estimation	373
10.5.8	Asymptotic evaluation	373
10.5.9	Applications	374
10.6	Multivariate lognormal frailty	379
10.7	Negative dependence models	380

10.8	Other approaches	380
10.9	Chapter summary	381
10.10	Bibliographic comments	382
10.11	Exercises	383
11	Instantaneous and short-term frailty models	385
11.1	Independent increments frailty model	387
11.1.1	Bivariate parallel data	387
11.1.2	Recurrent events	388
11.1.3	Estimation for recurrent events	390
11.1.4	Estimation for parallel data	391
11.1.5	Asymptotics	391
11.1.6	Application	392
11.2	Polya and independent increments additive model	392
11.2.1	Application	394
11.3	Piecewise gamma model	395
11.3.1	Bivariate parallel data	397
11.3.2	Recurrent events	398
11.3.3	Applications	398
11.4	Moving average model	400
11.5	Hidden cause of death model	401
11.6	The Woodbury-Manton model	402
11.7	Chapter summary	403
11.8	Bibliographic comments	403
11.9	Exercises	404
12	Competing risks models	406
12.1	Effects of changing cause-specific hazards	408
12.2	The multivariate parallel approach	409
12.3	The multi-state approach	410
12.4	The frailty approach	411
12.5	What can be estimated?	412
12.6	Competing risks for twin data	412
12.7	Applications	414
12.8	Chapter summary	417
12.9	Bibliographic comments	417
12.10	Exercises	417
13	Marginal and copula modeling	419
13.1	The coordinate-wise approach	421
13.1.1	Recurrent events	422
13.1.2	Computational aspects	423
13.1.3	Asymptotic results	425
13.1.4	Applications	425
13.2	The independence working model	427

13.2.1	Asymptotic evaluation	428
13.2.2	Software	429
13.2.3	Applications	429
13.3	Discussion of marginal modeling	431
13.4	Recurrent events	433
13.5	Copulas	435
13.6	The combined approach	436
13.6.1	Estimation	436
13.6.2	Asymptotics	437
13.6.3	Applications	437
13.7	Chapter summary	439
13.8	Bibliographic comments	439
13.9	Exercises	440
14	Multivariate non-parametric estimates	442
14.1	Distribution concepts and observations	446
14.2	The Hanley and Parnes estimate	447
14.2.1	Homogeneous censoring	448
14.2.2	Heterogeneous censoring	451
14.3	The Dabrowska estimate	452
14.4	The Pruitt estimate	453
14.5	Comparison	454
14.6	Other estimates and extensions	456
14.7	Asymptotics	457
14.8	Graphical illustrations	457
14.8.1	Cuts	457
14.8.2	Conditional on coordinate survival	458
14.8.3	Conditional on coordinate interval survival	458
14.8.4	Three-dimensional plot	459
14.8.5	Contour curves of the survivor function	459
14.8.6	Contour curves after uniform transformation	459
14.8.7	Contour curves of the bivariate density	459
14.8.8	Bivariate density after uniform transformation	459
14.8.9	Markov assumption check	460
14.9	Quantitative evaluation of dependence	461
14.9.1	Degree of dependence	461
14.9.2	Early/late dependence	461
14.9.3	Short-term dependence	462
14.10	Model checking	462
14.11	Applications	463
14.11.1	Danish twins	463
14.11.2	Kidney catheter data	477
14.11.3	Exercise data	478
14.11.4	Heart transplant data	478
14.12	Chapter summary	479

14.13	Bibliographic comments	480
14.14	Exercises	480
15	Summary	483
15.1	Summary of the theory	483
15.2	Course of analysis	486
15.3	Summary of the applications	491
15.3.1	Twins	491
15.3.2	Length of leukemia remission	492
15.3.3	Exercise data	493
15.3.4	Kidney catheter infections	493
15.3.5	Epileptic seizures	493
15.3.6	Amalgam data	494
15.3.7	Albuminuria	494
15.3.8	Malignant melanoma	494
15.4	Future development	494
15.5	Exercises	495
A	Mathematical results	497
A.1	Laplace transforms	497
A.2	Exponential families	498
A.3	Distribution theory	500
A.3.1	Gamma distributions	500
A.3.2	Inverse Gaussian distributions	501
A.3.3	Positive stable distributions	502
A.3.4	Power variance function distributions	504
A.3.5	Lognormal distributions	506
A.3.6	Generalized inverse Gaussian distributions	508
A.3.7	Non-central gamma distributions	508
A.4	Mathematical functions	509
A.4.1	The gamma function and related functions	509
A.4.2	Bessel functions	510
A.4.3	Hypergeometric functions	511
A.5	Bibliography	511
A.6	Exercises	512
B	Iterative solutions	513
B.1	Newton-Raphson iteration	514
B.2	Extensions and modifications	514
B.3	Standard error evaluation	516
	References	517
	Index	531