

Contents

1	Introducing R	1
1.1	Statistical packages and statistical modelling	1
1.2	Getting started in R	1
1.3	Reading data into R	3
1.4	Assignment and data generation	6
1.5	Displaying data	8
1.6	Data structures and the workspace	9
1.7	Transformations and data modification	11
1.8	Functions and suffixing	12
1.8.1	Structure functions	13
1.8.2	Mathematical functions	13
1.8.3	Logical operators	14
1.8.4	Control functions	14
1.8.5	Statistical functions	14
1.8.6	Random numbers	15
1.8.7	Suffixes in expressions	16
1.8.8	Extracting subsets of data	17
1.8.9	Recoding variates and factors into new factors	17
1.9	Graphical facilities	18
1.10	Text functions	21
1.11	Writing your own functions	22
1.12	Sorting and tabulation	23
1.13	Editing R code	26
1.14	Installing and using packages	27
2	Statistical modelling and inference	28
2.1	Statistical models	28
2.2	Types of variables	30
2.3	Population models	31
2.4	Random sampling	44
2.5	The likelihood function	44

2.6	Inference for single parameter models	46
2.6.1	Comparing two simple hypotheses	47
2.6.2	Information about a single parameter	49
2.6.3	Comparing a simple null hypothesis and a composite alternative	54
2.7	Inference with nuisance parameters	58
2.7.1	Profile likelihoods	59
2.7.2	Marginal likelihood for the variance	63
2.7.3	Likelihood normalizing transformations	66
2.7.4	Alternative test procedures	68
2.7.5	Bayes inference	71
2.7.6	Binomial model	74
2.7.7	Hypergeometric sampling from finite populations	80
2.8	The effect of the sample design on inference	81
2.9	The exponential family	82
2.9.1	Mean and variance	83
2.9.2	Generalized linear models	83
2.9.3	Maximum likelihood fitting of the GLM	84
2.9.4	Model comparisons through maximized likelihoods	87
2.10	Likelihood inference without models	89
2.10.1	Likelihoods for percentiles	89
2.10.2	Empirical likelihood	92
3	Regression and analysis of variance	97
3.1	An example	97
3.2	Strategies for model simplification	107
3.3	Stratified, weighted and clustered samples	111
3.4	Model criticism	114
3.4.1	Mis-specification of the probability distribution	116
3.4.2	Mis-specification of the link function	119
3.4.3	The occurrence of aberrant and influential observations	119
3.4.4	Mis-specification of the systematic part of the model	123
3.5	The Box–Cox transformation family	123
3.6	Modelling and background information	126
3.7	Link functions and transformations	136
3.8	Regression models for prediction	138
3.9	Model choice and mean square prediction error	140
3.10	Model selection through cross-validation	141
3.11	Reduction of complex regression models	144
3.12	Sensitivity of the Box–Cox transformation	153
3.13	The use of regression models for calibration	156

3.14	Measurement error in the explanatory variables	159
3.15	Factorial designs	161
3.16	Unbalanced cross-classifications	168
3.16.1	The Bennett hostility data	168
3.16.2	ANOVA of the cross-classification	170
3.16.3	Regression analysis of the cross-classification	174
3.16.4	Statistical package treatments of cross-classifications	176
3.17	Missing data	178
3.18	Approximate methods for missing data	180
3.19	Modelling of variance heterogeneity	180
3.19.1	Poisson example	184
3.19.2	Tree example	191
4	Binary response data	195
4.1	Binary responses	195
4.2	Transformations and link functions	197
4.2.1	Profile likelihoods for functions of parameters	202
4.3	Model criticism	207
4.3.1	Mis-specification of the probability distribution	207
4.3.2	Mis-specification of the link function	207
4.3.3	The occurrence of aberrant and influential observations	207
4.4	Binary data with continuous covariates	208
4.5	Contingency table construction from binary data	223
4.6	The prediction of binary outcomes	235
4.7	Profile and conditional likelihoods in 2×2 tables	242
4.8	Three-dimensional contingency tables with a binary response	246
4.8.1	Prenatal care and infant mortality	246
4.8.2	Coronary heart disease	248
4.9	Multidimensional contingency tables with a binary response	255
5	Multinomial and Poisson response data	269
5.1	The Poisson distribution	269
5.2	Cross-classified counts	271
5.3	Multicategory responses	279
5.4	Multinomial logit model	285
5.5	The Poisson-multinomial relation	287
5.6	Fitting the multinomial logit model	293
5.7	Ordered response categories	298
5.7.1	Common slopes for the regressions	299
5.7.2	Linear trend over response categories	301

5.7.3	Proportional slopes	304
5.7.4	The continuation ratio model	304
5.7.5	Other models	308
5.8	An Example	310
5.8.1	Multinomial logit model	313
5.8.2	Continuation ratio model	320
5.9	Structured multinomial responses	330
5.9.1	Independent outcomes	331
5.9.2	Correlated outcomes	339
6	Survival data	347
6.1	Introduction	347
6.2	The exponential distribution	347
6.3	Fitting the exponential distribution	349
6.4	Model criticism	354
6.5	Comparison with the normal family	361
6.6	Censoring	364
6.7	Likelihood function for censored observations	365
6.8	Probability plotting with censored data: the Kaplan–Meier estimator	368
6.9	The gamma distribution	377
6.9.1	Maximum likelihood with uncensored data	379
6.9.2	Maximum likelihood with censored data	382
6.9.3	Double modelling	384
6.10	The Weibull distribution	388
6.11	Maximum likelihood fitting of the Weibull distribution	390
6.12	The extreme value distribution	394
6.13	The reversed extreme value distribution	397
6.14	Survivor function plotting for the Weibull and extreme value distributions	398
6.15	The Cox proportional hazards model and the piecewise exponential distribution	400
6.16	Maximum likelihood fitting of the piecewise exponential distribution	403
6.17	Examples	404
6.18	The logistic and log-logistic distributions	407
6.19	The normal and lognormal distributions	411
6.20	Evaluating the proportional hazard assumption	414
6.21	Competing risks	420
6.22	Time-dependent explanatory variables	427
6.23	Discrete time models	427

7	Finite mixture models	433
7.1	Introduction	433
7.2	Example – girl birthweights	434
7.3	Finite mixtures of distributions	434
7.4	Maximum likelihood in finite mixtures	435
7.5	Standard errors	437
7.6	Testing for the number of components	440
7.6.1	Example	443
7.7	Likelihood ‘spikes’	448
7.8	Galaxy data	450
7.9	Kernel density estimates	458
8	Random effect models	461
8.1	Overdispersion	461
8.1.1	Testing for overdispersion	464
8.2	Conjugate random effects	466
8.2.1	Normal kernel: the t -distribution	466
8.2.2	Poisson kernel: the negative binomial distribution	472
8.2.3	Binomial kernel: beta-binomial distribution	477
8.2.4	Gamma kernel	478
8.2.5	Difficulties with the conjugate approach	478
8.3	Normal random effects	479
8.3.1	Predicting from the normal random effect model	481
8.4	Gaussian quadrature examples	481
8.4.1	Overdispersion model fitting	481
8.4.2	Poisson – the fabric fault data	482
8.4.3	Binomial – the toxoplasmosis data	484
8.5	Other specified random effect distributions	487
8.6	Arbitrary random effects	487
8.7	Examples	489
8.7.1	The fabric fault data	489
8.7.2	The toxoplasmosis data	492
8.7.3	Leukaemia remission data	493
8.7.4	The Brownlee stack-loss data	493
8.8	Random coefficient regression models	496
8.8.1	Example – the fabric fault data	498
8.9	Algorithms for mixture fitting	499
8.9.1	The trypanosome data	499
8.10	Modelling the mixing probabilities	503
8.11	Mixtures of mixtures	504

9 Variance component models	508
9.1 Models with shared random effects	508
9.2 The normal/normal model	508
9.3 Exponential family two-level models	511
9.4 Other approaches	513
9.5 NPML estimation of the masses and mass-points	514
9.6 Random coefficient models	514
9.7 Variance component model fitting	515
9.7.1 Children's height development	516
9.7.2 Multi-centre trial of beta-blockers	524
9.7.3 Longitudinal study of obesity	530
9.8 Autoregressive random effect models	537
9.9 Latent variable models	543
9.9.1 The normal factor model	543
9.10 IRT models	544
9.10.1 The Rasch model	544
9.10.2 The two-parameter model	545
9.10.3 The three-parameter logit (3PL) model	547
9.10.4 Example – The Law School Aptitude Test (LSAT)	547
9.11 Spatial dependence	551
9.12 Multivariate correlated responses	552
9.13 Discreteness of the NPML estimate	552
Bibliography	554
R function and constant index	567
Dataset index	570
Subject index	571