Contents

Contents ..... **ix**

Contributors ..... **xvii**

1 Introduction **1**
1.1 What is in the book? 1
1.2 Software 5
1.3 How to use this book if you are an instructor 6
1.4 What we did not do and why 7
1.5 How to cite R and associated packages 7
1.6 Our R programming style 8
1.7 Getting data into R 9

2 Limitations of linear regression applied on ecological data **11**
2.1 Data exploration 12
2.2 The linear regression model 17
2.3 Violating the assumptions; exception or rule? 19
2.3.1 Introduction 19
2.3.2 Normality 20
2.3.3 Heterogeneity 20
2.3.4 Fixed X 21
2.3.5 Independence 21
2.3.6 Example 1; Wedge clam data 22
2.3.7 Example 2; Moby's teeth 26
2.3.8 Example 3; Nereis 29
2.3.9 Example 4; Pelagic bioluminescence 30
2.4 Where to go from here 32

3 Things are not always linear; additive modelling **35**
3.1 Introduction 35
3.2 Additive modelling 36
3.2.1 GAM in gam and GAM in mgcv 37
3.2.2 GAM in gam with LOESS 38
3.2.3 GAM in mgcv with cubic regression splines 43
3.3 Technical details of GAM in mgcv* 45
3.4 GAM example 1; bioluminescent data for two stations 56

3.5 GAM example 2: Dealing with collinearity 64
3.6 Inference* 68
3.7 Summary and where to go from here? 69

4 Dealing with heterogeneity 71
4.1 Dealing with heterogeneity 72
4.1.1 Linear regression applied on squid 72
4.1.2 The fixed variance structure 74
4.1.3 The varIdent variance structure 75
4.1.4 The varPower variance structure 78
4.1.5 The varExp variance structure 80
4.1.6 The varConstPower variance structure 80
4.1.7 The varComb variance structure 81
4.1.8 Overview of all variance structures 82
4.1.9 Graphical validation of the optimal model 84
4.2 Benthic biodoversity experiment 86
4.2.1 Linear regression applied on the benthic biodiversity data 86
4.2.2 GLS applied on the benthic biodiversity data 89
4.2.3 A protocol 90
4.2.4 Application of the protocol on the benthic biodiversity data 91

5 Mixed effects modelling for nested data 101
5.1 Introduction 101
5.2 2-stage analysis method 103
5.3 The linear mixed effects model 105
5.3.1 Introduction 105
5.3.2 The random intercept model 106
5.3.3 The random intercept and slope model 109
5.3.4 Random effects model 111
5.4 Induced correlations 112
5.5 The marginal model 114
5.6 Maximum likelihood and REML estimation* 116
5.7 Model selection in (additive) mixed effects modelling 120
5.8 RIKZ data: Good versus bad model selection 122
5.8.1 The wrong approach 122
5.8.2 The good approach 127
5.9 Model validation 128
5.10 Begging behaviour of nestling barn owls 129

6 Violation of independence – Part I 143
6.1 Temporal correlation and linear regression 143
6.2 Linear regression model and multivariate time series 152
6.3 Owl sibling negotiation data 157

7 Violation of independence – Part II 161
7.1 Tools to detect violation of independence 161

7.2 Adding spatial correlation structures to the model 166
7.3 Revisiting the Hawaiian birds 171
7.4 Nitrogen isotope ratios in whales 172
7.4.1 Moby 172
7.4.2 All whales 174
7.5 Spatial correlation due to a missing covariate 177
7.6 Short godwits time series 182
7.6.1 Description of the data 182
7.6.2 Data exploration 183
7.6.3 Linear regression 185
7.6.4 Protocol time 186
7.6.5 Why all the fuss? 191

8 Meet the exponential family 193
8.1 Introduction 193
8.2 The Normal distribution 194
8.3 The Poisson distribution 196
8.4 The negative binomial distribution 199
8.5 The gamma distribution 201
8.6 The Bernoulli and binomial distributions 202
8.7 The natural exponential family 204
8.8 Zero truncated distributions for count data 207

9 GLM and GAM for count data 209
9.1 Introduction 209
9.2 Gaussian linear regression as a GLM 210
9.3 Introducing Poisson GLM with an artificial example 211
9.4 Likelihood criterion 213
9.5 Introducing the Poisson GLM with a real example 215
9.5.1 Introduction 215
9.5.2 R code and results 217
9.5.3 Deviance 218
9.5.4 Sketching the fitted values 218
9.6 Model selection in a GLM 220
9.6.1 Introduction 220
9.6.2 R code and output 221
9.6.3 Options for finding the optimal model 221
9.6.4 The drop1 command 222
9.6.5 Two ways of using the anova command 223
9.6.6 Results 224
9.7 Overdispersion 224
9.7.1 Introduction 224
9.7.2 Causes and solutions for overdispersion 225
9.7.3 Quick fix: Dealing with overdispersion in a Poisson GLM 225
9.7.4 R code and numerical output 226
9.7.5 Model selection in quasi-Poisson 227

9.8 Model validation in a Poisson GLM 229
9.9 Illustration of model validation in quasi-Poisson GLM 231
9.10 Negative binomial GLM 233
9.10.1 Introduction 233
9.10.2 Results 236
9.11 GAM 239

10 GLM and GAM for presence-absence and proportional data 245
10.1 Introduction 245
10.2 GLM for presence − absence data 246
10.2.1 Tuberculosis in wild boar 246
10.2.2 Parasites in cod 252
10.3 GLM for proportional data 255
10.4 GAM for presence-absence data 258
10.5 Where to go from here? 260

11 Zero truncated and zero inflated models for count data 261
11.1 Introduction 261
11.2 Zero truncated data 263
11.2.1 The underlying mathematics for truncated models 263
11.2.2 Illustration of Poisson and NB truncated models 266
11.3 Too many zeros 269
11.4 ZIP and ZINB models 274
11.4.1 Mathematics of the ZIP and ZINB 274
11.4.2 Example of ZIP and ZINB models 278
11.5 ZAP and ZANB models, alias hurdle models 287
11.5.1 Mathematics of the ZAP and ZANB 288
11.5.2 Example of ZAP and ZANB 289
11.6 Comparing Poisson, quasi-Poisson, NB, ZIP, ZINB, ZAP and ZANB GLMs 291
11.7 Flowchart and where to go from here 293

12 Generalised Estimation Equations 295
12.1 GLM: Ignoring the dependence structure 295
12.1.1 The California bird data 295
12.1.2 The owl data 298
12.1.3 The deer data 300
12.2 Specifying the GEE 302
12.2.1 Introduction 302
12.2.2 Step 1 of the GEE: Systematic component and link function 303
12.2.3 Step 2 of the GEE: The variance 303
12.2.4 Step 3 of the GEE: The association structure 304
12.3 Why all the fuss? 309
12.4 Association for binary data 313
12.5 Examples of GEE 314
12.5.1 A GEE for the California birds 314

12.5.2 A GEE for the owls 316
12.5.3 A GEE for the deer data 319
12.6 Concluding remarks 320

13 GLMM and GAMM 321
13.1 Setting the scene for binomial GLMM 322
13.2 GLMM and GAMM for binomial and Poisson data 325
13.2.1 Deer data 325
13.2.2 The owl data revisited 331
13.2.3 A word of warning 337
13.3 The underlying mathematics in GLMM 338

14 Estimating trends for Antarctic birds in relation to climate change 341
14.1 Introduction 341
14.2 Data exploration 343
14.3 Trends and auto-correlation 349
14.4 Using ice extent as an explanatory variable 350
14.5 SOI and differences between arrival and laying dates 353
14.6 Discussion 358
14.7 What to report in a paper 360

15 Large-scale impacts of land-use change in a Scottish farming catchment 361
15.1 Introduction 361
15.2 Data exploration 363
15.3 Estimation of trends for the bird data 365
15.4 Dealing with independence 372
15.5 To transform or not to transform 376
15.6 Birds and explanatory variables 377
15.7 Conclusions 378
15.8 What to write in a paper 379

16 Negative binomial GAM and GAMM to analyse amphibian roadkills ... 381
16.1 Introduction 381
16.2 Data exploration 383
16.3 GAM 388
16.4 Understanding what the negative binomial is doing 392
16.5 GAMM: Adding spatial correlation 395
16.6 Discussion 396
16.7 What to write in a paper 396

17 Additive mixed modelling applied on deep-sea pelagic bioluminescent organisms 397
17.1 Biological introduction 397
17.2 The data and underlying questions 399
17.3 Construction of multi-panel plots for grouped data 400
17.4 Estimating common patterns using additive mixed modelling 408

17.5 Choosing the best model 418
17.6 Discussion 418
17.7 What to write in a paper 420

18 Additive mixed modelling applied on phytoplankton time series data 421
18.1 Introduction 421
18.2 Data exploration 425
18.3 A statistical data analysis strategy for DIN 428
18.4 Results for temperature 438
18.5 Results for DIAT1 440
18.6 Comparing phytoplankton and environmental trends 443
18.7 Conclusions 446
18.8 What to write in a paper 446

19 Mixed effects modelling applied on American Foulbrood affecting honey bees larvae 447
19.1 Introduction 447
19.2 Data exploration 448
19.3 Analysis of the data 450
19.4 Discussion 458
19.5 What to write in a paper 458

20 Three-way nested data for age determination techniques applied to cetaceans 459
20.1 Introduction 459
20.2 Data exploration 460
20.3 Data analysis 462
20.4 Discussion 467
20.5 What to write in a paper 468

21 GLMM applied on the spatial distribution of koalas in a fragmented landscape 469
21.1 Introduction 469
21.2 The data 471
21.3 Data exploration and preliminary analysis 473
Collinearity 474
21.4 Generalised linear mixed-effects modelling 481
21.5 Discussion 490
21.6 What to write in a paper 492

22 A comparison of GLM, GEE and GLMM applied on Badger activity data 493
22.1 Introduction 493
22.2 Data exploration 495
22.3 GLM results assuming independence 497
22.4 GEE results 499
22.5 GLMM results 500

22.6 Discussion 501
22.7 What to write in a paper 502

23 Incorporating temporal correlation in seal abundance data with MCMC 503
23.1 Introduction 503
23.2 Preliminary results 505
23.3 GLM 508
23.4 What is Bayesian statistics? 510
23.5 Fitting the Poisson model in BRugs 513
23.6 Poisson model with random effects 520
23.7 Poisson model with random effects and auto-correlation 523
23.8 Negative binomial distribution with auto-correlated random effects 525
23.9 Conclusions 529

A. Required pre-knowledge: A linear regression and additive modelling example 531
A.1 The data 531
A.2 Data exploration 532
A.3 Linear regression 536
A.4 Additive modelling 545
A.5 Further extensions 549
A.6 Information Theory and Multi-Model Inference 550
A.7 Maximum likelihood estimation in linear regression context 551

References 553