

Alain F. Zuur · Elena N. Ieno · Neil J. Walker ·
Anatoly A. Saveliev · Graham M. Smith

Mixed Effects Models and Extensions in Ecology with R

 Springer

Contents

1	Introduction	1
1.1	What Is in the Book?	1
1.1.1	To Include or Not to Include GLM and GAM	3
1.1.2	Case Studies	4
1.1.3	Flowchart of the Content	4
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	6
1.5	How to Cite R and Associated Packages	7
1.6	Our R Programming Style	8
1.7	Getting Data into R	9
1.7.1	Data in a Package	10
2	Limitations of Linear Regression Applied on Ecological Data	11
2.1	Data Exploration	12
2.1.1	Cleveland Dotplots	12
2.1.2	Pairplots	14
2.1.3	Boxplots	15
2.1.4	xyplot from the Lattice Package	15
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	19
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	28
2.3.9	Example 4; Pelagic Bioluminescence	30
2.4	Where to Go from Here	31

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	42
3.3	Technical Details of GAM in mgcv	44
3.3.1	A (Little) Bit More Technical Information on Regression Splines	47
3.3.2	Smoothing Splines Alias Penalised Splines	49
3.3.3	Cross-Validation	51
3.3.4	Additive Models with Multiple Explanatory Variables	53
3.3.5	Two More Things	53
3.4	GAM Example 1; Bioluminescent Data for Two Stations	55
3.4.1	Interaction Between a Continuous and Nominal Variable	59
3.5	GAM Example 2: Dealing with Collinearity	63
3.6	Inference	66
3.7	Summary and Where to Go from Here?	67
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 Biodiversity 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	92
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
5.4.1	Intraclass Correlation Coefficient	114
5.5	The Marginal Model	114
5.6	Maximum Likelihood and REML Estimation	116
5.6.1	Illustration of Difference Between ML and REML	119
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
5.10.1	Step 1 of the Protocol: Linear Regression	130
5.10.2	Step 2 of the Protocol: Fit the Model with GLS	132
5.10.3	Step 3 of the Protocol: Choose a Variance Structure	132
5.10.4	Step 4: Fit the Model	133
5.10.5	Step 5 of the Protocol: Compare New Model with Old Model	133
5.10.6	Step 6 of the Protocol: Everything Ok?	134
5.10.7	Steps 7 and 8 of the Protocol: The Optimal Fixed Structure	135
5.10.8	Step 9 of the Protocol: Refit with REML and Validate the Model	137
5.10.9	Step 10 of the Protocol	139
5.10.10	Sorry, We are Not Done Yet	139
6	Violation of Independence – Part I	143
6.1	Temporal Correlation and Linear Regression	143
6.1.1	ARMA Error Structures	150
6.2	Linear Regression Model and Multivariate Time Series	152
6.3	Owl Sibling Negotiation Data	158
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	184
7.6.4	Protocol Time	186
7.6.5	Why All the Fuss?	190

8	Meet the Exponential Family	193
8.1	Introduction	193
8.2	The Normal Distribution	194
8.3	The Poisson Distribution	196
8.3.1	Preparation for the Offset in GLM	198
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
8.7.1	Which Distribution to Select?	205
8.8	Zero Truncated Distributions for Count Data	206
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	216
9.5.3	Deviance	217
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	220
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	223
9.7	Overdispersion	224
9.7.1	Introduction	224
9.7.2	Causes and Solutions for Overdispersion	224
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	228
9.8.1	Pearson Residuals	229
9.8.2	Deviance Residuals	229
9.8.3	Which One to Use?	230
9.8.4	What to Plot?	230
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	238
9.11.1	Distribution of larval Sea Lice Around Scottish Fish Farms	239
10	GLM and GAM for Absence–Presence and Proportional Data	245
10.1	Introduction	245
10.2	GLM for Absence–Presence Data	246
10.2.1	Tuberculosis in Wild Boar	246
10.2.2	Parasites in Cod	252
10.3	GLM for Proportional Data	254
10.4	GAM for Absence–Presence Data	258
10.5	Where to Go from Here?	259
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	265
11.3	Too Many Zeros	269
11.3.1	Sources of Zeros	270
11.3.2	Sources of Zeros for the Cod Parasite Data	271
11.3.3	Two-Part Models Versus Mixture Models, and Hippos	271
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	286
11.5.1	Mathematics of the ZAP and ZANB	287
11.5.2	Example of ZAP and ZANB	288
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	299
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	304
12.2.4	Step 3 of the GEE: The Association Structure	304
12.3	Why All the Fuss?	309
12.3.1	A Bit of Maths	310

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	323
13.1	Setting the Scene for Binomial GLMM	324
13.2	GLMM and GAMM for Binomial and Poisson Data	327
13.2.1	Deer Data	327
13.2.2	The Owl Data Revisited	333
13.2.3	A Word of Warning	339
13.3	The Underlying Mathematics in GLMM	339
14	Estimating Trends for Antarctic Birds in Relation to Climate Change	343
	A.F. Zuur, C. Barbraud, E.N. Ieno, H. Weimerskirch, G.M. Smith, and N.J. Walker	
14.1	Introduction	343
14.1.1	Explanatory Variables	344
14.2	Data Exploration	345
14.3	Trends and Auto-correlation	350
14.4	Using Ice Extent as an Explanatory Variable	352
14.5	SOI and Differences Between Arrival and Laying Dates	354
14.6	Discussion	360
14.7	What to Report in a Paper	361
15	Large-Scale Impacts of Land-Use Change in a Scottish Farming Catchment	363
	A.F. Zuur, D. Raffaelli, A.A. Saveliev, N.J. Walker, E.N. Ieno, and G.M. Smith	
15.1	Introduction	363
15.2	Data Exploration	365
15.3	Estimation of Trends for the Bird Data	367
15.3.1	Model Validation	368
15.3.2	Failed Approach 1	372
15.3.3	Failed Approach 2	373
15.3.4	Assume Homogeneity?	374
15.4	Dealing with Independence	374
15.5	To Transform or Not to Transform	378
15.6	Birds and Explanatory Variables	378
15.7	Conclusions	380
15.8	What to Write in a Paper	381

16 Negative Binomial GAM and GAMM to Analyse Amphibian Roadkills 383
A.F. Zuur, A. Mira, F. Carvalho, E.N. Ieno, A.A. Saveliev, G.M. Smith, and N.J. Walker

16.1 Introduction 383
16.1.1 Roadkills 383

16.2 Data Exploration 385

16.3 GAM 389

16.4 Understanding What the Negative Binomial is Doing 394

16.5 GAMM: Adding Spatial Correlation 396

16.6 Discussion 397

16.7 What to Write in a Paper 397

17 Additive Mixed Modelling Applied on Deep-Sea Pelagic Bioluminescent Organisms 399
A.F. Zuur, I.G. Priede, E.N. Ieno, G.M. Smith, A.A. Saveliev, and N.J. Walker

17.1 Biological Introduction 399

17.2 The Data and Underlying Questions 401

17.3 Construction of Multi-panel Plots for Grouped Data 402

17.3.1 Approach 1 402

17.3.2 Approach 2 407

17.3.3 Approach 3 408

17.4 Estimating Common Patterns Using Additive Mixed Modelling ... 410

17.4.1 One Smoothing Curve for All Stations 410

17.4.2 Four Smoothers; One for Each Month 414

17.4.3 Smoothing Curves for Groups Based on Geographical Distances 417

17.4.4 Smoothing Curves for Groups Based on Source Correlations 418

17.5 Choosing the Best Model 419

17.6 Discussion 420

17.7 What to Write in a Paper 421

18 Additive Mixed Modelling Applied on Phytoplankton Time Series Data 423
A.F. Zuur, M.J. Latuhihin, E.N. Ieno, J.G. Baretta-Bekker, G.M. Smith, and N.J. Walker

18.1 Introduction 423

18.1.1 Biological Background of the Project 424

18.2 Data Exploration 427

18.3 A Statistical Data Analysis Strategy for DIN 429

18.4 Results for Temperature 439

18.5 Results for DIAT1 441

18.6 Comparing Phytoplankton and Environmental Trends 443

18.7	Conclusions	445
18.8	What to Write in a Paper	446
19	Mixed Effects Modelling Applied on American Foulbrood Affecting Honey Bees Larvae	447
	A.F. Zuur, L.B. Gende, E.N. Ieno, N.J. Fernández, M.J. Eguaras, R. Fritz, N.J. Walker, A.A. Saveliev, and G.M. Smith	
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
	E.N. Ieno, P.L. Luque, G.J. Pierce, A.F. Zuur, M.B. Santos, N.J. Walker, A.A. Saveliev, and G.M. Smith	
20.1	Introduction	459
20.2	Data Exploration	460
20.3	Data Analysis	462
20.3.1	Intraclass Correlations	466
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
	J.R. Rhodes, C.A. McAlpine, A.F. Zuur, G.M. Smith, and E.N. Ieno	
21.1	Introduction	469
21.2	The Data	471
21.3	Data Exploration and Preliminary Analysis	473
21.3.1	Collinearity	473
21.3.2	Spatial Auto-correlation	479
21.4	Generalised Linear Mixed Effects Modelling	481
21.4.1	Model Selection	483
21.4.2	Model Adequacy	487
21.5	Discussion	490
21.6	What to Write in a Paper	492
22	A Comparison of GLM, GEE, and GLMM Applied to Badger Activity Data	493
	N.J. Walker, A.F. Zuur, A. Ward, A.A. Saveliev, E.N. Ieno, and G.M. Smith	
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
	A.A. Saveliev, M. Cronin, A.F. Zuur, E.N. Ieno, N.J. Walker, and G.M. Smith	
23.1	Introduction	503
23.2	Preliminary Results	504
23.3	GLM	507
23.3.1	Validation	509
23.4	What Is Bayesian Statistics?	510
23.4.1	Theory Behind Bayesian Statistics	510
23.4.2	Markov Chain Monte Carlo Techniques	511
23.5	Fitting the Poisson Model in BRugs	513
23.5.1	Code in R	513
23.5.2	Model Code	514
23.5.3	Initialising the Chains	515
23.5.4	Summarising the Posterior Distributions	517
23.5.5	Inference	518
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
23.8.1	Comparison of Models	528
23.9	Conclusions	528
A	Required Pre-knowledge: A Linear Regression and Additive Modelling Example	531
A.1	The Data	531
A.2	Data Exploration	532
A.2.1	Step 1: Outliers	532
A.2.2	Step 2: Collinearity	533
A.2.3	Relationships	536
A.3	Linear Regression	536
A.3.1	Model Selection	540
A.3.2	Model Validation	542
A.3.3	Model Interpretation	543
A.4	Additive Modelling	546
A.5	Further Extensions	550
A.6	Information Theory and Multi-model Inference	550
A.7	Maximum Likelihood Estimation in Linear Regression Context	552
	References	553
	Index	563