

Alain F. Zuur
Elena N. Ieno

Beginner's Guide to Zero-Inflated Models with R

Published by Highland Statistics Ltd.
Highland Statistics Ltd.
Newburgh
United Kingdom
highstat@highstat.com

ISBN: 978-0-9571741-8-4
First published in May 2016

Contents

PREFACE.....	V
ACKNOWLEDGEMENTS.....	V
DATASETS USED IN THIS BOOK	VI
COVER ART.....	VI
CONTENTS	VII
1 INTRODUCTION.....	1
1.1 HOW DO THE 2012 AND 2016 BOOKS DIFFER?	1
1.2 WHAT DO YOU NEED TO KNOW TO USE THIS BOOK?	1
1.3 OUTLINE OF THE BOOK	2
1.4 WHAT IS <i>NOT</i> IN THIS BOOK?	3
1.5 HOW TO READ THIS BOOK	4
1.6 ACCESSING THE DATA AND THE R CODE	4
1.7 A FEW FINAL COMMENTS	4
2 ESSENTIAL DISTRIBUTIONS FOR ZERO-INFLATED MODELS	5
2.1 DISTRIBUTIONS.....	5
2.2 POISSON DISTRIBUTION.....	5
2.3 NEGATIVE BINOMIAL DISTRIBUTION	7
2.4 BERNOULLI DISTRIBUTION.....	9
2.5 BINOMIAL DISTRIBUTION	9
2.6 GAMMA DISTRIBUTION	11
2.7 LOGNORMAL DISTRIBUTION	13
2.8 SUMMARY OF DISTRIBUTIONS.....	14
3 INTRODUCING ZERO-INFLATED POISSON MODELS.....	17
3.1 POISSON GLM.....	18
3.1.1 <i>Simulating Poisson distributed data</i>	18
3.1.2 <i>Applying the Poisson GLM</i>	19
3.1.3 <i>Model validation</i>	20
3.1.4 <i>Dispersion</i>	21
3.1.5 <i>Visualising the model fit</i>	23
3.1.6 <i>Using two covariates in a Poisson GLM</i>	25
3.2 BERNOULLI GLM	26
3.2.1 <i>Crocodile attack data</i>	26
3.2.2 <i>The model</i>	27
3.2.3 <i>Applying the Bernoulli GLM in R</i>	27
3.2.4 <i>Model validation</i>	28
3.2.5 <i>Visualising the model fit of the Bernoulli GLM</i>	28
3.3 CONCEPTUAL EXPLANATIONS OF ZIP MODELS	29
3.3.1 <i>Nature flips a coin</i>	30
3.3.2 <i>Nature creates the counts</i>	31
3.3.3 <i>Fitting the ZIP model in R</i>	33
3.3.4 <i>Model validation for the ZIP model</i>	34

3.3.5 Validation for the ZIP model.....	35
3.3.6 Poisson GLM applied on zero-inflated data.....	37
3.4 TRUE AND FALSE ZEROS	40
3.4.1 The origin of zeros.....	40
3.4.2 The density function of a ZIP	42
3.5 COVARIATES IN BOTH PARTS OF THE ZIP MODEL	43
3.5.1 The same covariate in both parts of the ZIP model	43
3.5.2 Using two different covariates in the ZIP model.....	48
4 ZERO-INFLATED MODELS APPLIED TO ORANGE-CROWNED	
WARBLERS	53
4.1 ORANGE-CROWNED WARBLERS	53
4.2 DATA EXPLORATION	54
4.3 POISSON GLM.....	56
4.3.1 Model formulation.....	56
4.3.2 Fitting the Poisson GLM	57
4.3.3 Simulating data from the model.....	57
4.3.4 Fitted values.....	59
4.3.5 What is next?.....	60
4.4 ZIP MODEL	60
4.4.1 Fitting the ZIP model.....	60
4.4.2 Simulating data from the ZIP model.....	61
4.5 MODEL SELECTION FOR THE ZIP MODEL USING AIC.....	64
4.5.1 What is the AIC?.....	64
4.5.2 How do you calculate the AIC?.....	64
4.5.3 AICs for 16 models.....	67
4.5.4 The optimal ZIP model	67
4.6 DISCUSSION	69
5 ZERO-INFLATED MODELS APPLIED TO SHARK ABUNDANCE	
DATA	71
5.1 SHARKS	71
5.2 POISSON GLM APPLIED TO TIGER SHARKS.....	72
5.2.1 Specifying the model.....	72
5.2.2 Fitting the Poisson GLM in R.....	74
5.2.3 Poisson GLM results for the tiger sharks	74
5.2.4 Assessing overdispersion	75
5.2.5 Model selection	75
5.2.6 Model validation.....	76
5.2.7 Model interpretation.....	76
5.3 NB GLM APPLIED TO TOTAL NUMBER OF ALL SHARK SPECIES.....	77
5.3.1 Poisson GLM results.....	77
5.3.2 Negative binomial GLM results.....	78
5.3.3 ZIP model results.....	80
5.4 ZERO-INFLATED POISSON MODEL AND SILVERTIP SHARKS	81
5.4.1 Poisson and quasi-Poisson GLM.....	81
5.4.2 NB GLM applied to silvertip shark data.....	82

5.4.3 ZIP model.....	83
5.5 POSSIBLE EXTENSIONS.....	86
6 HURDLE MODELS FOR RIPARIAN SPIDER COUNTS.....	87
6.1 RIPARIAN SPIDERS	87
6.1.1 <i>Introducing the data</i>	87
6.1.2 <i>Data exploration</i>	88
6.2 POISSON GLM RESULTS	94
6.3 EXPLANATION OF HURDLE MODELS.....	96
6.3.1 <i>Nature flips a coin</i>	96
6.3.2 <i>Nature creates count data</i>	97
6.3.3 <i>Crossing a hurdle</i>	100
6.3.4 <i>Density function of the ZAP model</i>	101
6.4 ZAP MODEL FOR THE SPIDER DATA.....	102
6.4.1 <i>Model formulation</i>	102
6.4.2 <i>Running the hurdle model in R</i>	102
6.5 DOING IT MANUALLY IN TWO STEPS	109
6.6 SIMULATING FROM THE MODEL	110
7 MODELS FOR ZERO-INFLATED CONTINUOUS DATA APPLIED TO CHINESE TALLOW TREES	113
7.1 CHINESE TALLOW TREES	113
7.1.1 <i>Introduction to the data</i>	113
7.1.2 <i>Data exploration</i>	114
7.2 MULTIPLE LINEAR REGRESSION APPLIED TO THE CTT DATA	116
7.3 GAMMA GLM	117
7.3.1 <i>Simulating gamma distributed data</i>	117
7.3.2 <i>Applying the gamma GLM</i>	118
7.3.3 <i>Visualising the model fit</i>	119
7.4 EXPLANATION OF HURDLE MODELS.....	120
7.4.1 <i>Nature flips a coin</i>	120
7.4.2 <i>Nature creates continuous data</i>	121
7.4.3 <i>ZAG model</i>	122
7.4.4 <i>Density function of the hurdle GLM</i>	126
7.5 THE HURDLE MODEL APPLIED TO THE CTT DATA	126
7.5.1 <i>Hurdle model formulation for the CTT data</i>	126
7.5.2 <i>Applying the gamma GLM to the CTT data</i>	127
7.5.3 <i>Bernoulli GLM applied to the CTT data</i>	128
7.5.4 <i>Gluing together the two components</i>	129
7.6 DISCUSSION	131
7.7 WHAT TO PUT IN A PAPER.....	132
8 LINEAR MIXED EFFECTS MODELS	133
8.1 LILIES AND BEAVERS	133
8.1.1 <i>Data exploration</i>	133
8.1.2 <i>Model formulation</i>	135
8.1.3 <i>Fitting the model using <code>lmer</code></i>	138

8.1.4 Model validation.....	139
8.1.5 Sketching the fitted values.....	140
8.2 CHACMA BABOONS	143
8.2.1 Dependency structure.....	143
8.2.2 Data exploration	146
8.2.3 Model formulation.....	147
8.2.4 Fitting the model using <i>lmer</i>	147
8.2.5 Model validation.....	148
8.2.6 Sketching the fitted values.....	149
8.3 DISCUSSION	150
8.3.1 Lilies and beaver dataset.....	150
8.3.2 Baboon dataset.....	151
9 ZERO-ALTERED MODELS WITH TWO-WAY NESTED AND CROSSED RANDOM EFFECTS	153
9.1 CLIMATE CHANGE AND GRASSLAND SPECIES.....	153
9.2 DATA EXPLORATION.....	155
9.3 POISSON GLMM.....	156
9.3.1 Model formulation.....	156
9.3.2 Fitting the Poisson GLMM using <i>lme4</i>	157
9.3.3 Model validation.....	158
9.3.4 How to continue: Zero inflation or GAMM?.....	162
9.3.5 Model fit of the Poisson GLMM.....	163
9.4 ZERO-ALTERED MODELS WITH RANDOM EFFECTS	164
9.4.1 Bernoulli GLMM for absence and presence data.....	165
9.4.2 Once <i>T. montanum</i> is present	167
9.4.3 Combining both models.....	171
9.5 DISCUSSION	176
10 INTRODUCTION TO BAYESIAN STATISTICS	177
10.1 GENERAL PROBABILITY RULES	177
10.2 BIVARIATE LINEAR REGRESSION APPLIED TO OSPREY DATA.....	180
10.2.1 Ospreys.....	180
10.2.2 Ordinary least squares.....	181
10.2.3 The frequentist interpretation	182
10.2.4 Changing the notation	185
10.2.5 Likelihood estimation as an alternative to OLS.....	187
10.3 WHY GO BAYESIAN?.....	188
10.4 A CARTOON EXPLANATION OF MCMC	190
10.5 CONCEPT OF MCMC	192
10.5.1 Starting values	193
10.5.2 Parameters to save.....	193
10.5.3 Burn-in.....	193
10.5.4 The Metropolis-Hastings algorithm.....	193

10.6 DO-IT-YOURSELF MCMC IN R	195
10.6.1 <i>Standardisation of continuous covariates</i>	195
10.6.2 <i>Steps 1 and 2 of the MCMC algorithm in R</i>	196
10.6.3 <i>Step 3 of the MCMC algorithm in R</i>	197
10.6.4 <i>Mixing</i>	200
10.6.5 <i>Posterior mean values and posterior distributions</i>	204
10.7 MCMC APPLIED TO OSPREY DATA IN JAGS	204
10.7.1 <i>Installing JAGS and R2jags</i>	204
10.7.2 <i>Flowchart for running a model in JAGS</i>	204
10.7.3 <i>Preparing the data for JAGS</i>	206
10.7.4 <i>JAGS code</i>	206
10.7.5 <i>Initial values and parameters to save</i>	209
10.7.6 <i>Running JAGS</i>	210
10.7.7 <i>Accessing numerical output from JAGS</i>	211
10.7.8 <i>Assess mixing</i>	212
10.7.9 <i>Posterior information</i>	213
10.8 WHAT TO REMEMBER FROM THIS CHAPTER.....	214
10.9 RECOMMENDED LITERATURE	214
10.10 WHAT'S NEXT?	214
10.11 EXERCISES WITH VIDEO SOLUTION FILES	214
10.11.1 <i>Exercise 1: Irish pH data</i>	214
10.11.2 <i>Exercise 2: Crayfish data</i>	215
11 BAYESIAN ANALYSIS FOR POISSON, NB, ZIP AND BERNOULLI	
MODELS	217
11.1 FITTING A POISSON GLM IN A BAYESIAN CONTEXT	217
11.1.1 <i>Poisson GLM for tiger shark data</i>	217
11.1.2 <i>Preparing the data for JAGS</i>	218
11.1.3 <i>JAGS code for a Poisson GLM</i>	219
11.1.4 <i>Initial values and parameters to save</i>	221
11.1.5 <i>Running JAGS</i>	221
11.1.6 <i>Results from JAGS for the Poisson GLM</i>	222
11.1.7 <i>Assess overdispersion</i>	223
11.1.8 <i>Model validation</i>	226
11.2 NEGATIVE BINOMIAL GLM IN JAGS.....	228
11.3 ZERO-INFLATED MODELS IN JAGS FOR SILVERTIP SHARKS	231
11.3.1 <i>Data for JAGS</i>	232
11.3.2 <i>JAGS code for a ZIP model</i>	232
11.3.3 <i>Initial values and parameters to save</i>	233
11.3.4 <i>Running JAGS</i>	233
11.3.5 <i>Mixing of chains</i>	235
11.3.6 <i>Calculating Pearson residuals retrospectively</i>	236
11.4 BAYESIAN BERNOULLI GLM.....	237
11.4.1 <i>Data for JAGS</i>	238
11.4.2 <i>JAGS code for the Bernoulli GLM</i>	238
11.4.3 <i>Initial values and parameters to save</i>	239
11.4.4 <i>Running JAGS</i>	239

11.4.5 <i>Mixing of chains</i>	240
11.4.6 <i>Numerical results</i>	240
11.4.7 <i>Sketching the results</i>	242
11.5 REFERENCES.....	247
12 BAYESIAN ANALYSIS FOR LINEAR MIXED EFFECTS MODELS – BEAVER AND LILIES	249
12.1 LILIES AND BEAVER DATA	249
12.2 DATA FOR JAGS.....	249
12.3 JAGS CODE.....	251
12.4 INITIAL VALUES AND PARAMETERS TO SAVE	252
12.5 RUNNING JAGS.....	252
12.6 ASSESS MIXING	253
12.7 NUMERICAL RESULTS.....	254
12.8 MODEL VALIDATION.....	254
12.9 VISUALISING THE RESULTS.....	257
12.10 MISSING VALUES	260
13 THE ZERO TRICK TO FIT ANY DISTRIBUTION IN A BAYESIAN ANALYSIS	265
13.1 UNDERLYING MATHEMATICS	265
13.2 ZERO TRICK FOR A POISSON GLM	267
13.2.1 <i>Preparing the data for JAGS</i>	267
13.2.2 <i>JAGS code</i>	268
13.2.3 <i>Initial values and parameters to save</i>	269
13.2.4 <i>Running JAGS</i>	269
13.2.5 <i>Numerical output</i>	269
13.3 ZERO TRICK FOR THE ZIP MODEL	270
13.4 ZERO TRICK FOR A ZAP MODEL.....	271
13.5 APPLYING A BAYESIAN GAMMA GLM TO THE CTT DATA.....	272
13.6 THE HURDLE MODEL APPLIED TO THE CTT DATA.....	275
13.6.1 <i>Data for JAGS</i>	275
13.6.2 <i>JAGS code for the hurdle model</i>	276
13.6.3 <i>Initial values and parameters to save</i>	277
13.6.4 <i>Running JAGS</i>	277
13.6.5 <i>Mixing of chains</i>	278
13.6.6 <i>Numerical results</i>	278
13.6.7 <i>Model validation</i>	280
13.6.8 <i>Sketching the fitted values</i>	282
13.7 LOGNORMAL REGRESSION APPLIED TO THE CTT DATA	284
13.8 DISCUSSION	286
14 BAYESIAN MODEL SELECTION TECHNIQUES	287
14.1 CRITICAL NOTES ON MODEL SELECTION.....	288
14.1.1 <i>Dropping covariates based on p-values?</i>	288
14.1.2 <i>Using the AIC</i>	289
14.1.3 <i>Information theoretic approach</i>	289
14.1.4 <i>Other approaches</i>	289

14.2 BAYESIAN ANALYSIS OF ORANGE-CROWNED WARBLERS.....	290
14.2.1 <i>Fitting the Poisson GLM</i>	290
14.2.2 <i>Fitting a zero inflated Poisson model</i>	293
14.3 COMPARING THE POISSON AND ZIP MODELS.....	295
14.4 BAYESIAN MODEL SELECTION: AIC AND DIC.....	297
14.4.1 <i>Using the AIC</i>	297
14.4.2 <i>Using the DIC</i>	299
14.5 BAYESIAN MODEL SELECTION: LASSO	302
14.5.1 <i>Simulation study for LASSO</i>	303
14.5.2 <i>Bayesian LASSO</i>	305
14.5.3 <i>LASSO and ZIP model?</i>	307
14.6 BAYESIAN MODEL SELECTION WITH INDICATOR FUNCTIONS.....	307
14.6.1 <i>Kuo and Mallick</i>	308
14.6.2 <i>Gibbs variable selection</i>	314
14.7 BAYESIAN MODEL SELECTION: MODEL PROBABILITIES.....	318
14.7.1 <i>Non-mathematical introduction</i>	318
14.7.2 <i>A little bit of mathematics</i>	319
14.8 DISCUSSION	322
15 BAYESIAN MODEL SELECTION APPLIED TO ZERO-INFLATED BUTTERFLY DATA	325
15.1 BUTTERFLIES.....	325
15.2 DATA EXPLORATION.....	326
15.3 POISSON GLMM.....	329
15.3.1 <i>Data for JAGS</i>	330
15.3.2 <i>JAGS code</i>	331
15.3.3 <i>Initial values and parameters to save</i>	332
15.3.4 <i>Running JAGS</i>	332
15.3.5 <i>Mixing</i>	332
15.3.6 <i>Dispersion</i>	333
15.3.7 <i>Model validation</i>	333
15.4 ZIP, ZAP OR THE NB GLM?	335
15.4.1 <i>Model probabilities</i>	336
15.4.2 <i>Out of sample prediction</i>	339
15.5 ZAP MODEL WITH RANDOM EFFECTS.....	341
15.5.1 <i>Data for JAGS</i>	342
15.5.2 <i>JAGS code</i>	342
15.5.3 <i>Initial values and parameters to save</i>	344
15.5.4 <i>Running JAGS</i>	344
15.5.5 <i>Mixing</i>	344
15.5.6 <i>Numerical output</i>	344
15.6 ZAP MIXED MODELS AND BAYESIAN MODEL SELECTION	345
15.6.1 <i>Data for JAGS</i>	346
15.6.2 <i>JAGS modelling code</i>	347
15.6.3 <i>Initial values and parameters to save</i>	349
15.6.4 <i>Running JAGS</i>	350
15.6.5 <i>Mixing</i>	350

15.6.6 Numerical information.....	350
15.7 VISUALISING THE OPTIMAL ZAP MODEL	351
16 ZERO-INFLATED SEAGRASS COVERAGE DATA.....	353
16.1 SEAGRASS.....	353
16.2 BRAINSTORMING.....	354
16.2.1 Which covariates?.....	354
16.2.2 Adding dependency	354
16.2.3 Anticipated problems	355
16.2.4 Subsetting the data to shorten computing time.....	356
16.3 DATA EXPLORATION.....	356
16.4 THE ZERO-INFLATED BETA MODEL.....	359
16.4.1 The beta distribution	359
16.4.2 The beta model.....	360
16.4.3 Zero-inflated beta model.....	362
16.5 SEAGRASS DATA AND THE ZERO-ALTERED BETA MODEL.....	364
16.5.1 Frequentist approach.....	364
16.5.2 Bayesian approach.....	369
16.6 VISUALISATION OF RESULTS AND BAYESIAN POST-HOC TEST	372
16.7 HALF-CAUCHY DISTRIBUTION	376
17 OTHER DISTRIBUTIONS.....	379
17.1 ZERO-INFLATED BINOMIAL DATA.....	379
17.2 GENERALISED POISSON MODEL.....	380
17.3 ZERO-INFLATED NEGATIVE BINOMIAL MODELS.....	386
18 MULTIVARIATE GLMM.....	387
18.1 PSEUDO-REPLICATION	387
18.2 MULTIVARIATE GLMM FOR PLANT POLLEN.....	388
18.2.1 Pollen data	388
18.2.2 Univariate Poisson GLMM	389
18.2.3 Specification of a multivariate GLMM	390
18.2.4 Multivariate GLMM in JAGS.....	391
18.2.5 Multivariate GLMM results for the pollen data.....	395
18.2.6 Differences between univariate and multivariate GLMMs	397
18.2.7 Discussion	397
REFERENCES	399
INDEX	407
BOOKS BY HIGHLAND STATISTICS.....	411

