Contents

PREFACE v
ACKNOWLEDGEMENTS v
DATASETS USED IN THIS BOOK v

1 INTRODUCTION 1
1.1 SPEAKING THE SAME LANGUAGE 1
1.2. GENERAL POINTS 2
1.3 OUTLINE OF THIS BOOK 5

2 OUTLIERS 7
2.1 WHAT IS AN OUTLIER? 7
2.2 BOXPLOT TO IDENTIFY OUTLIERS IN ONE DIMENSION 8
2.2.1 Simple boxplot 8
2.2.2 Conditional boxplot 10
2.2.3 Multi-panel boxplots from the lattice package 13
2.3 CLEVELAND DOTPLOT TO IDENTIFY OUTLIERS 15
2.3.1 Simple Cleveland dotplots 15
2.3.2 Conditional Cleveland dotplots 17
2.3.3 Multi-panel Cleveland dotplots from the lattice package 18
2.4 BOXPLOTS OR CLEVELAND DOTPLOTS? 20
2.5 CAN WE APPLY A TEST FOR OUTLIERS? 21
2.5.1 Z-score 22
2.5.2 Grubbs' test 22
2.6 OUTLIERS IN THE TWO-DIMENSIONAL SPACE 24
2.7 INFLUENTIAL OBSERVATIONS IN REGRESSION MODELS 25
2.8 WHAT TO DO IF YOU DETECT POTENTIAL OUTLIERS 27
2.9 OUTLIERS AND MULTIVARIATE DATA 31
2.10 THE PROS AND CONS OF TRANSFORMATIONS 33

3 NORMALITY AND HOMOGENEITY 37
3.1 WHAT IS NORMALITY? 37
3.2 HISTOGRAMS AND CONDITIONAL HISTOGRAMS 38
3.2.1 Multipanel histograms from the lattice package 39
3.2.2 When is normality of the raw data considered? 41
3.3 KERNEL DENSITY PLOTS 42
3.4 QUANTILE–QUANTILE PLOTS 43
3.4.1 Quantile–quantile plots from the lattice package 44
3.5 USING TESTS TO CHECK FOR NORMALITY 45
3.6 HOMOGENEITY OF VARIANCE 47
3.6.1 Conditional boxplots 47
3.6.2 Scatterplots for continuous explanatory variables 49
3.7 USING TESTS TO CHECK FOR HOMOGENEITY 50
3.7.1 The Bartlett test 50
3.7.2 The F-ratio test 50
3.7.3 Levene's test 51
3.7.4 So which test would you choose? 51

3.7.5 R code *51*
3.7.6 Using graphs? *52*

4 RELATIONSHIPS **55**
4.1 SIMPLE SCATTERPLOTS 55
4.1.1 Example: Clam data *55*
4.1.2 Example: Rabbit data *57*
4.1.3 Example: Blow fly data *58*
4.2 MULTIPANEL SCATTERPLOTS 60
4.2.1 Example: Polychaeta data *60*
4.2.2 Example: Bioluminescence data *61*
4.3 PAIRPLOTS 62
4.3.1 Bioluminescence data *63*
4.3.2 Cephalopod data *64*
4.3.3 Zoobenthos data *65*
4.4 CAN WE INCLUDE INTERACTIONS? 66
4.4.1 Irish pH data *66*
4.4.2 Godwit data *68*
4.4.3 Irish pH data revisited *70*
4.4.4 Parasite data *71*
4.5 DESIGN AND INTERACTION PLOTS 73

5 COLLINEARITY AND CONFOUNDING **77**
5.1 WHAT IS COLLINEARITY? 77
5.2 THE SAMPLE CORRELATION COEFFICIENT 77
5.3 CORRELATION AND OUTLIERS 78
5.4 CORRELATION MATRICES 79
5.5 CORRELATION AND PAIRPLOTS 80
5.6 COLLINEARITY DUE TO INTERACTIONS 82
5.7 VISUALISING COLLINEARITY WITH CONDITIONAL BOXPLOTS 83
5.8 QUANTIFYING COLLINEARITY USING VIFS 85
5.8.1 Variance inflation factors *85*
5.8.2 Geometric presentation of collinearity *86*
5.8.3 Tolerance *88*
5.8.4 What constitutes a high VIF value? *88*
5.8.5 VIFs in action *89*
5.9 GENERALISED VIF VALUES 91
5.10 VISUALISING COLLINEARITY USING PCA BIPLOT 93
5.11 CAUSES OF COLLINEARITY AND SOLUTIONS 94
5.12 BE STUBBORN AND KEEP COLLINEAR COVARIATES? 96
5.13 CONFOUNDING VARIABLES 97
5.13.1 Visualising confounding variables *99*
5.13.2 Confounding factors in time series analysis *100*

6 CASE STUDY: METHANE FLUXES **103**
6.1 INTRODUCTION 103
6.2 DATA EXPLORATION 104
6.2.1 Where in the world are the sites? *104*

6.2.2 Working with ggplot2 ..... *105*
6.2.3 Outliers *108*
6.2.4 Collinearity *111*
6.2.5 Relationships *112*
6.2.6 Interactions *114*
6.2.7 Where in the world are the sites (continued)? *115*
6.3 STATISTICAL ANALYSIS USING LINEAR REGRESSION 118
6.3.1 Model formulation *118*
6.3.2 Fitting a linear regression model *118*
6.3.3 Model validation of the linear regression model *120*
6.3.4 Interpretation of the linear regression model *125*
6.4 STATISTICAL ANALYSIS USING A MIXED EFFECTS MODEL 131
6.4.1 Model formulation *131*
6.4.2 Fitting a mixed effects model *132*
6.4.3 Model validation of the mixed effects model *132*
6.4.4 Interpretation of the linear mixed effects model *132*
6.5 CONCLUSIONS 134
6.6 WHAT TO PRESENT IN A PAPER 134

7 CASE STUDY: OYSTERCATCHER SHELL LENGTH 135
7.1 IMPORTING THE DATA 136
7.2 DATA EXPLORATION 136
7.3 APPLYING A LINEAR REGRESSION MODEL 138
7.4 UNDERSTANDING THE RESULTS 140
7.5 TROUBLE 143
7.6 CONCLUSIONS 146

8 CASE STUDY: HAWAIIAN BIRD TIME SERIES 147
8.1 IMPORTING THE DATA 147
8.2 CODING THE DATA 148
8.3 MULTI-PANEL GRAPH USING XYPLOT FROM LATTICE 148
8.3.1 Attempt 1 using xyplot *149*
8.3.2 Attempt 2 using xyplot *150*
8.3.3 Attempt 3 using xyplot *151*
8.4 MULTI-PANEL GRAPH USING GGPLOT2 153
8.5 CONCLUSIONS 154

REFERENCES 155

INDEX 159

BOOKS BY HIGHLAND STATISTICS 161