Generalized Linear Models, with Applications in Fisheries Research

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in the

DEPARTMENT OF STATISTICS

of

RHODES UNIVERSITY

by

Bonelwa Sidumo

November 2017

Abstract

Gambusia affinis (*G. affinis*) is an invasive fish species found in the Sundays River Valley of the Eastern Cape, South Africa. The relative abundance and population dynamics of *G. affinis* were quantified in five interconnected impoundments within the Sundays River Valley. This study utilised a *G. affinis* data set to demonstrate various, classical ANOVA models. Generalized linear models were used to standardize catch per unit effort (CPUE) estimates and to determine environmental variables which influenced the CPUE. Based on the generalized linear model results dam age, mean temperature, *Oreochromis mossambicus* abundance and *Glossogobius callidus* abundance had a significant effect on the *G. affinis* CPUE.

The Albany Angling Association collected data during fishing tag and release events. These data were utilized to demonstrate repeated measures designs. Mixed-effects models provided a powerful and flexible tool for analyzing clustered data such as repeated measures data and nested data, hence it has become tremendously popular as a framework for the analysis of bio-behavioral experiments. The results show that the mixed-effects methods proposed in this study are more efficient than those based on generalized linear models. These data were better modeled with mixed-effects models due to their flexibility in handling missing data.

Keywords: *Gambusia affinis*, catch per unit effort, ANOVA, generalized linear models, repeated measures designs, mixed-effects models, generalized mixed effects models.

Contents

List of Tables

List of Figures

Acknowledgments

I would like to express my special appreciation and thanks to my supervisor Mr Jeremy Baxter, you have been a tremendous mentor. I would also like to thank National Research Foundation (NRF) for their financial support during the course of this study. This work would not have been possible without NRF financial support.

I would also like to acknowledge South African Weather Service (SAWS) for providing me with their climate data. Thanks are also due to Dr Woodford from University of the Witwatersrand, South Africa and Albany Angling Association of Grahamstown, Eastern Cape for providing me with their fisheries data.

A special thanks to my family. Words cannot express how grateful I am to my mother, father, siblings for all of the sacrifices that you have made on my behalf. Your prayer for me was what sustained me thus far. Thank you for supporting me for everything and especially I cannot thank you enough for encouraging me throughout this experience. To my beloved son Oyisa Sidumo, I would like to express my thanks for being such a good boy always cheering me up.

Finally I thank my God, my good father, for letting me through all the difficulties. I have experienced your guidance day by day and I will keep on trusting you for my future. Thank you, Lord.

Chapter 1

Overview

This thesis describes various classical ANOVA models and their underlying assumptions (Chapter 2). These methods are applied to test various hypotheses based on two different fisheries science data sets. General linear models are introduced in chapter 3. The linear model is subsequently extended to include non-normal and count response variables, namely the generalized linear model, in chapter 3. Hierarchical or mixed-effects models are discussed in chapter 5. The R script used to analyze the various hypotheses have been included in the appendices. After consideration of these models, it is found that a generalized linear mixed model (GLMM) is a more appropriate methodology for these data since it handles both unbalanced and missing data as discussed in chapter 3.

1.1 Research Background

Howell et al. (2013) applied a data transformation to the *Gambusia affinis* data to meet normality and homoscedasticity assumptions. The response or dependent variable was discrete with missing observations. Howell et al. (2013) fitted a general linear model to the data which does not work well with missing data. The data was square root transformed to achieve normality and homoscedasticity. Even though normality was achieved, the issue of missing data was not addressed. These authors did not use an efficient way of modeling the data with missing data. The Albany Angling Association data has repetitions with missing data hence the mixed-effects was applied to the data.

1.2 Research Problem

Missing data is a major problem in fisheries research which may be due to missing a catch, that is a zero catch. As a result a classical ANOVA methods may not be the most appropriate statistical methodology to apply when testing hypotheses in this context.

1.3 Goal of the Research

The aim of this study is to clearly demonstrate and discuss the utility of generalized linear models in the context of two fisheries science studies. This thesis aims to clearly show that well known classical ANOVA techniques are often not the most appropriate approach when assessing the hypotheses of these studies. Generalized linear and mixed models are introduced, model fitting or estimation algorithms discussed, model fit assessment criteria are clearly documented and demonstrated and, appropriate to the context of these studies, hypothesis tests are conducted.

This study investigated which statistical method worked the best with these fisheries studies. Mixed-effects models were applied to the data because the observations were grouped according to one or more levels of experimental units and they also incorporate both fixed effects terms and random effects terms. The observations in the same cluster cannot be considered independent and mixed-effects models constitute a convenient tool for modeling dependence within clustered data (Pinheiro & Bates, 2000, page 153 and 154). These models give intuitive interpretation for the source and the structure of the dependence and can easily handle unbalanced and missing data that are frequently encountered in many areas of scientific investigation, for example fisheries science.

Chapter 2

Classical ANOVA Methods

This chapter utilizes the *Gambusia affinis (G.affinis)* data set to demonstrate various, classical, analysis of variance (ANOVA) models. Analysis of variance refers to a collection of statistical models and methods used to analyze variation in a response variable, such as continuous random variable, measured under conditions defined by discrete factors, that is classification variables, often with nominal levels (Larson, 2008). ANOVA methods can be used to test equality among several means by comparing variance among groups relative to variance within groups. In this case, the primary objective of ANOVA is to test whether the true response means are identical across factor levels. ANOVA is often the most effective method available for analyzing experimental data (Larson, 2008). The one-way ANOVA methodology, discussed in section 2.2.1, is used to determine whether there are any statistically significant differences between the true, or population, means of three or more independent or unrelated groups, or similarly determine the significance of the effects of treatments on the dependent variable. The two-way ANOVA methodology, discussed in section 2.2.2, is used to compare the effects of several levels of two factors. Nested ANOVA, an extension of the ANOVA model allowing variables to be nested within other variables, is discussed in section 2.2.3. Repeated measures ANOVA models, applicable in scenarios where the experimental design is such that multiple observations of the same variable(s) are made on the same subject, item or individual are discussed in section 5.2.

2.1 The *Gambusia Affinis* Data Set

Gambusia affinis is an invasive fish species found in the Sundays River Valley of the Eastern Cape, South Africa. Howell et al. (2013) found that the interaction between sampling events or seasons (late summer to early winter), mean temperature and dams sampled, or location, had a significant effect on the relative abundance of *G. affinis* in five interconnected irrigation impoundments within the Sundays River Valley. Dr Woodford from the University of the Witwatersrand, South Africa, kindly supplied the data set used in this study. This data set

consists of three hundred and eighty (380) observations of twenty six (26) variables.

These data were collected in five different interconnected irrigation impoundments, namely the Avoca (AVO), Disco Chicks (DC), Dunbrody (DB), Sur le Sun (SLS) and Olifantsklip (OLI) impoundments located in the Sundays River Valley. The seventy six (76) sampling events conducted in December of 2011 were removed from the data set as this study was concerned with the late summer to early winter periods only. The sampling was conducted in the late summer and early winter seasons by conducting four monthly sampling trips in February, March, April and June 2012 respectively. February and March represented the mid-to-late summer season and April and June represented the early winter season.

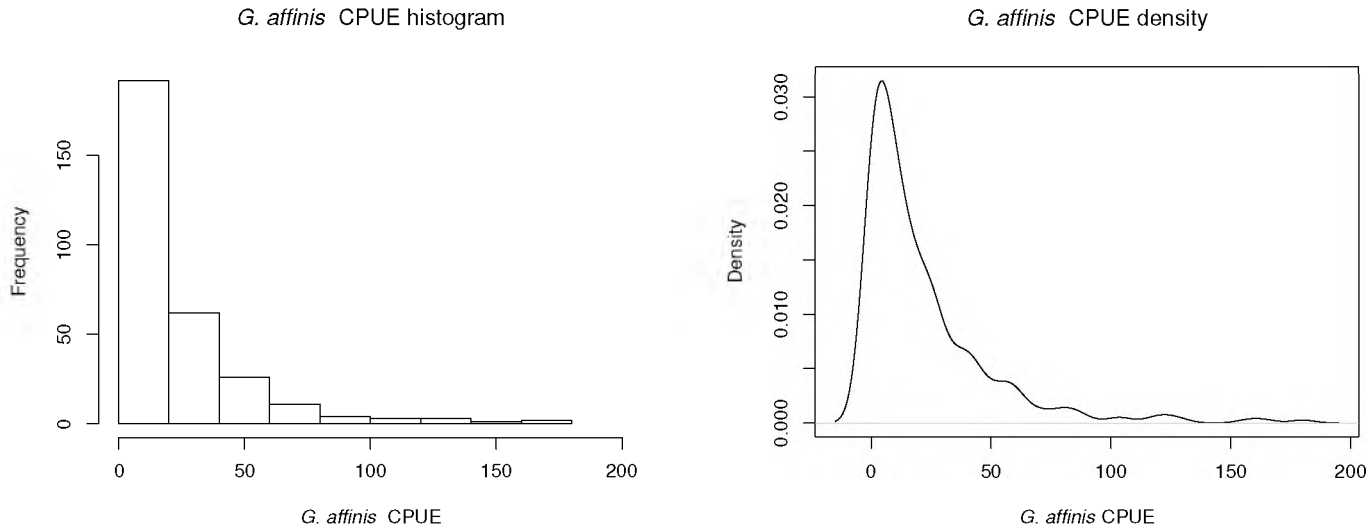
These data were extracted from the excel sheet supplied by Dr Woodford and saved in comma separated format (csv). The csv data were imported into R (R Core Team, 2017) and cleaned. One obvious data capturing error was corrected: observation 241 had a vegetation cover of 25 where all other vegetation cover data were recorded as fractions, that is the value 25 was changed to 0.25. The variables not needed in the subsequent analysis, repeated variables (for example dam and dam 1) and variables that are functions of other variables (for example the $\ln(x)$) were removed from the data set. These variables included date, year, month, CYC, ORM, InDay, vegetation, X164.5, In.gaa, the square root of GLC and square root of ORM. The resulting data set consisted of three hundred and four (304) observations of ten (10) variables. The levels of the factors were set to be similar to those used by Howell et al. (2013). These R instructions can be found in appendix A.

These data were collected by making four scoops or sweeps of a net with a mesh size of 2 mm at four randomly selected sites at each dam. Each site was resampled in subsequent sampling trips. A new variable "site" was added to the data set to facilitate analyses where site is nested within the dam. The reason for doing this is to investigate if there is significant variation among sites within dam habitats. The vegetation cover was estimated at the selected sampling sites along the water edge of the dams. The percentage of aquatic vegetation cover present in each sweep of the net was categorized into four categories; namely 0%, 25%, 50% and 75%. 0% represents no vegetation cover, 25% low vegetation cover, 50% medium vegetation cover and 75% denoted dense vegetation cover. These labels are as per the data in the excel file and not as per Howell et al. (2013). The variable named GAA in the data set represents the relative abundance of *G. affinis* species measured using catch per unit effort (CPUE), namely the number of *G. affinis* fish caught per scoop of the net (Howell et al., 2013). Maunder & Punt (2004) defines CPUE as the catch of fish or animals in numbers or weights taken in a defined period of effort.

The response variable is the number of *G. affinis* caught with explanatory variables mean temperature, dam, sampling event, site and percentage vegetation cover. When the response or dependent variables have a continuous distribution and the conditions of independent or explanatory variables are discrete with more than two classes, whether inherently or by design, then it is appropriate to analyze the data using Analysis of Variance (ANOVA) (Larson, 2008)

because Student's t-test is limited to two categories or populations (McDonald, 2009, page 120).

Descriptive Statistics: The *Gambusia Affinis* Data Set



(a) Histogram. **(b)** Density.

Figure 2.1: Histogram and density for *G. affinis* CPUE.

This section provides descriptive statistics for the variables considered in this analysis. The R code can be found in appendix A.1 on page 143. The density of the *G. affinis* CPUE, figure 2.1 (b), clearly shows that the distribution is non-symmetric. The histogram, figure 2.1 (a), suggests the non-normality of the CPUE since we know that the histogram of a normal distribution shows the highest frequency in the middle at the mean which is approximately equal to the median of the distribution. The distribution is not bell-shaped but positively skewed. The normal Q-Q plot, figure 2.2, shows that these data do not meet the normality assumption since the points are not approximately linear.

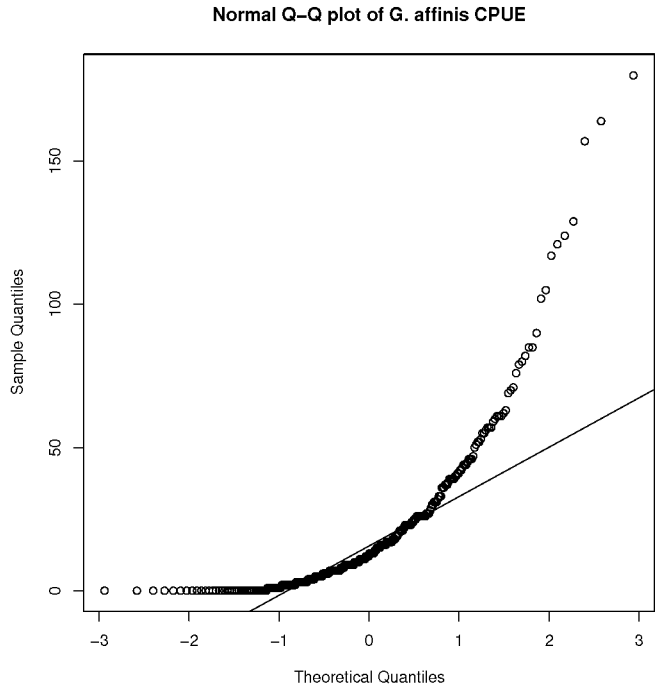


Figure 2.2: Normal Q-Q plot of *G. affinis* CPUE.

Sampling Event or Period

Sampling trips were conducted in February (sampling event 1), March (sampling event 2), April (sampling event 3) and June (sampling event 4) of 2012. Table 2.1 shows the summary statistics for the *G. affinis* CPUE namely; the mean, minimum, maximum, standard error and the 95% confidence interval for true average *G. affinis* CPUE for each sampling event. The lowest variability of *G. affinis* CPUE was found at sampling event 1 since it has the smallest standard error. As can be seen in figure 2.3, the median CPUE are not the same. CPUE in February, April and June are positively skewed while March is symmetric. There are outliers for each sampling event.

Sampling event	Sample size	Mean	Min	Max	Standard error	95% CI for μ	
1: Feb	76	13.9737	0	79	1.6524	10.6819	17.2655
2: Mar	76	27.4211	0	124	2.9231	21.5962	33.2459
3: Apr	76	37.2368	0	180	4.2299	28.8105	45.6632
4: Jun	76	9.1184	0	157	2.3111	4.5145	13.7223

Table 2.1: Summary statistics of *G. affinis* CPUE by sampling event.

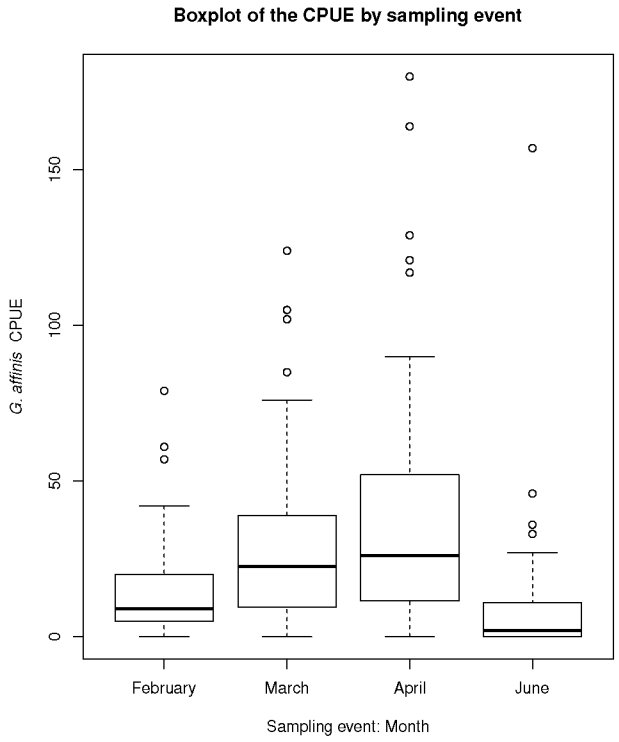


Figure 2.3: Boxplots of the *G. affinis* CPUE by sampling event.

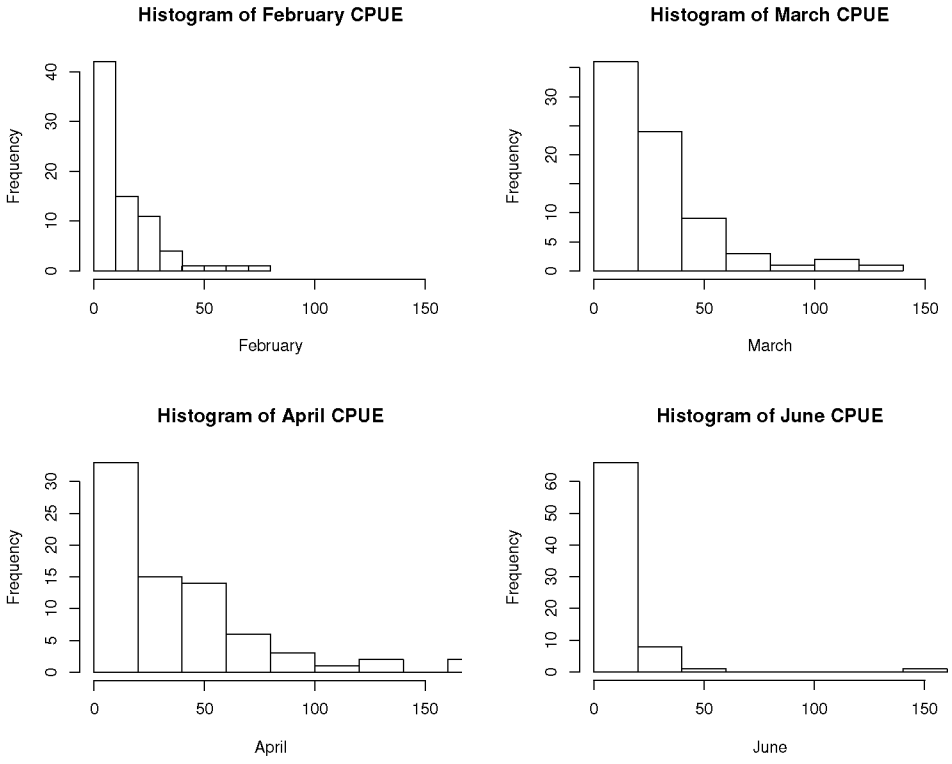


Figure 2.4: Histogram of the *G. affinis* CPUE by sampling event.

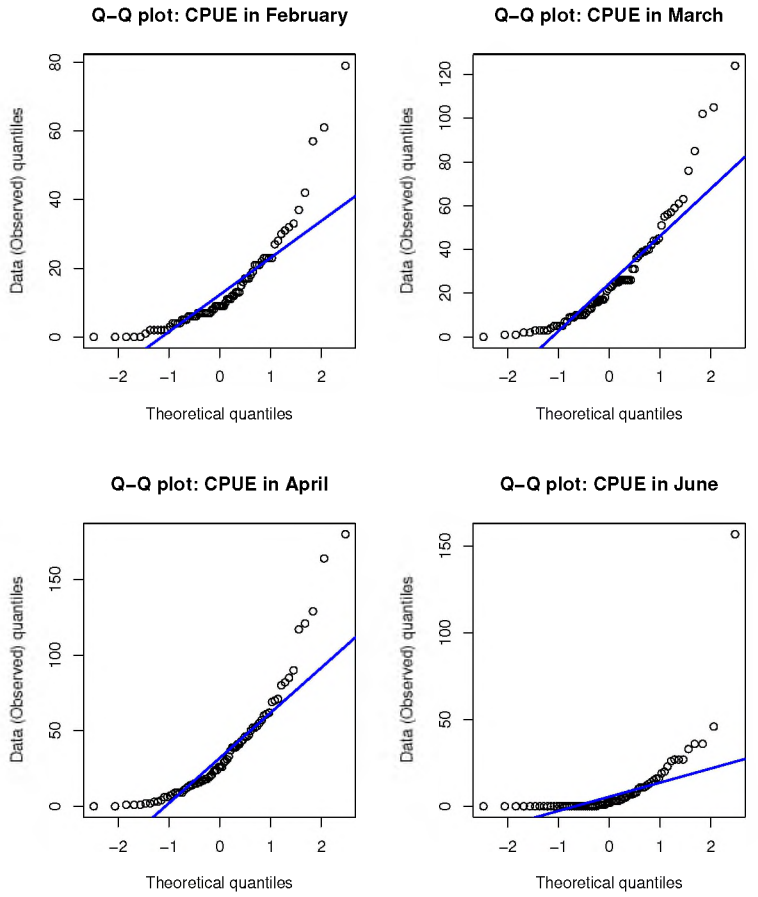


Figure 2.5: Normal Q-Q plots of the *G. affinis* CPUE by sampling event.

The one-way ANOVA requires homogeneity of variances across the levels of the factor and normality of the residual term (Larson, 2008), as discussed in section 2.2.1. These data are not symmetric, not bell-shaped and they are skewed to the right. There are clear departures from normality evident on the quantile-quantile (Q-Q) normal probability plots, figure 2.5. There is evidence on the histograms and Q-Q plots that these data do not meet the normality assumption of a one-way ANOVA. Checking normality of *G. affinis* for all sampling events, using the multivariate Shapiro-Wilk test, revealed that the *G. affinis* CPUE data are not normally distributed (MVW = 0.77083, p-value < 0.001). Hence we conclude that these data are not random samples from normally distributed populations. The data revealed that the true variances are unequal or heterogeneous (Bartlett's K-squared = 67.875, df = 3, p-value < 0.001). Bartlett's test is used when we have one quantitative or measurement variable and one nominal variable, and we want to test the null hypothesis that the variances of the dependent variable are the same for different groups (McDonald, 2009, page 156). Bartlett's test performs poorly with non-normal data set and should not be used unless this has been validated (Larson, 2008).

Dam or Impoundment

Howell et al. (2013) considered five different interconnected irrigation impoundments, or dams, namely Avoca (AVO), Disco Chicks (DC), Dunbrody (DB), Sur le Sun (SLS) and Olifantsklip (OLI) located in the Sundays River Valley. Summary statistics for *G.affinis* CPUE by dam or impoundment are shown in table 2.2. The boxplots of the CPUE by dam, figure 2.6, shows outliers at all dams with extreme outliers at dams DC and DB. The least variability in CPUE is found at dam OLI and SLS since these dams have the smallest standard errors. *G. affinis* CPUE increases from sampling event 1 (February) to sampling event 3 (April), figure 2.7. However the CPUE decreases from sampling event 2 (March) to sampling event 3 (April) at the Olifantsklip dam. The CPUE decreased at all dams from sampling event 3 (April) to sampling event 4 (June).

Dam	Sample size	Dam age	Mean	Min	Max	Standard error	95% CI for μ	
AVO	64	19	25.4844	0	129	3.2053	19.0791	31.8896
DC	64	1	19.0000	0	180	4.6891	9.6278	28.3722
DB	64	33	28.5625	0	157	3.6961	21.1747	35.9503
SLS	48	2	17.6250	0	105	2.9390	11.7124	23.5376
OLI	64	10	17.9375	0	79	2.4048	13.1318	22.7432

Table 2.2: Summary statistics of *G. affinis* CPUE by dam.

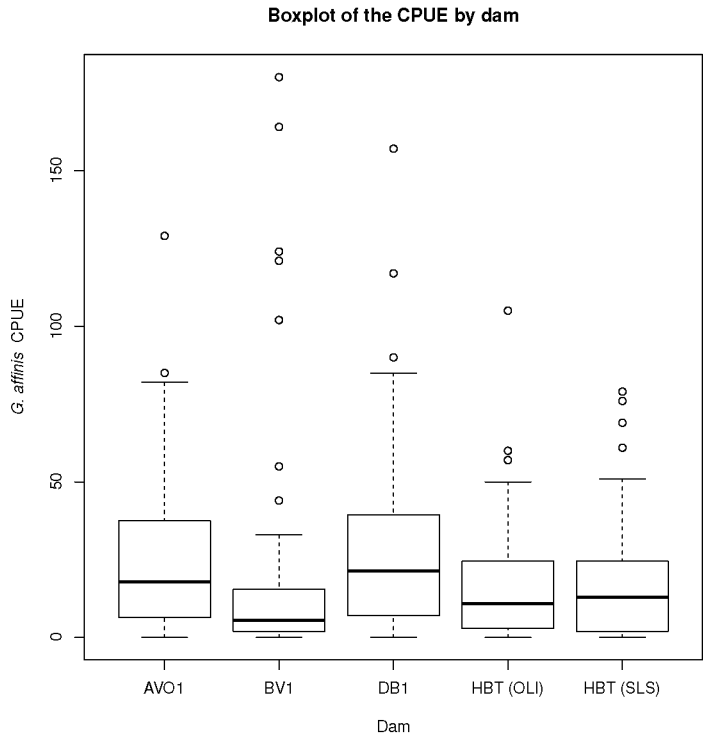


Figure 2.6: Boxplots of the *G. affinis* CPUE by dam.

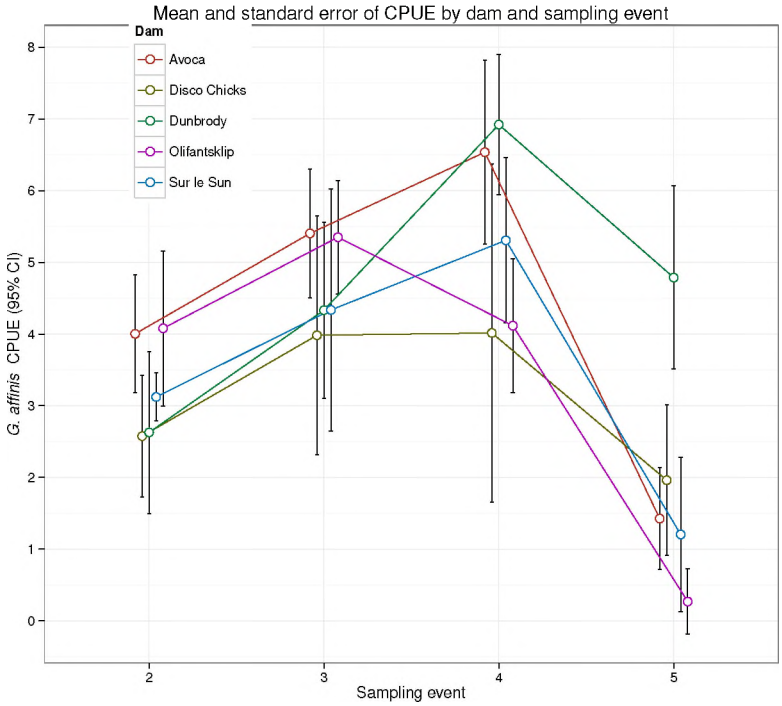


Figure 2.7: Mean and standard error of *G. affinis* CPUE by dam and sampling events.

Percentage Vegetation Cover

53% of the CPUE observations were taken at sites with dense (75%) vegetation cover while 29% of the sites had no vegetation cover, see table 2.3 and figure 2.8. Summary statistics of the CPUE by vegetation cover are shown in table 2.3. There are numerous outliers and it is evident that the CPUE at the different vegetation covers are not symmetrically distributed, see figure 2.9.

(%) Vegetation cover	Sample size	Min	Max	Mean	Standard error	95% CI for μ	
0.00	88	0	180	21.2273	3.6583	13.9559	28.4986
0.25	21	0	42	12.5714	2.6535	7.0364	18.1064
0.50	35	0	129	17.7714	4.1910	9.2542	26.2887
0.75	160	0	157	24.4688	2.0245	20.4704	28.4671

Table 2.3: Summary statistics of *G. affinis* CPUE by percentage vegetation cover.

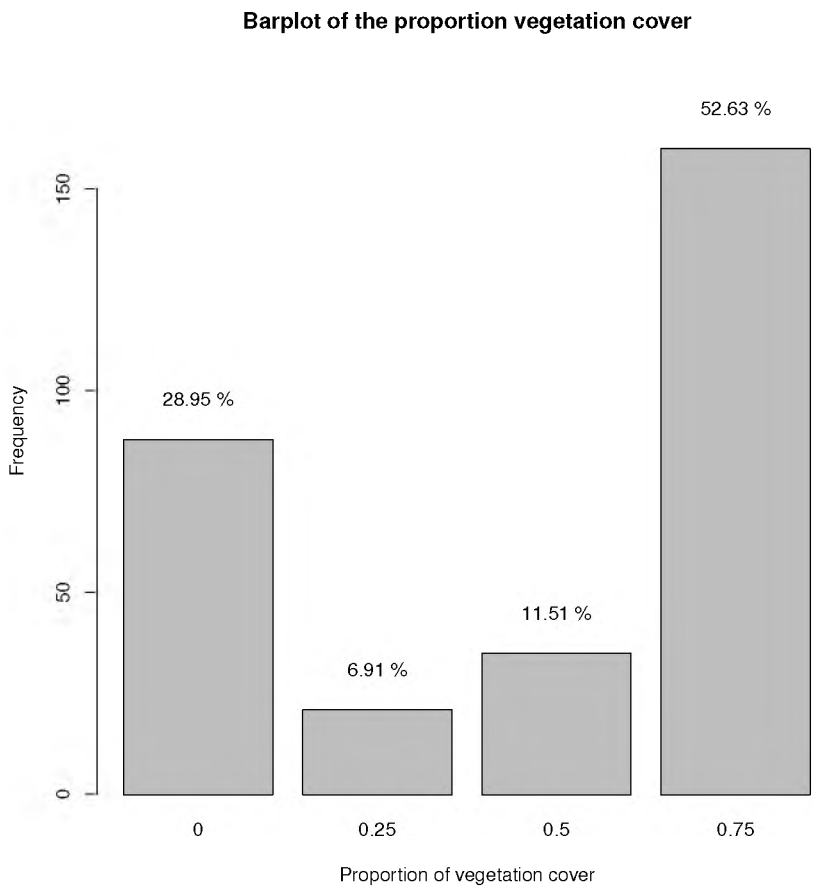


Figure 2.8: Barplot of the proportion of vegetation cover.

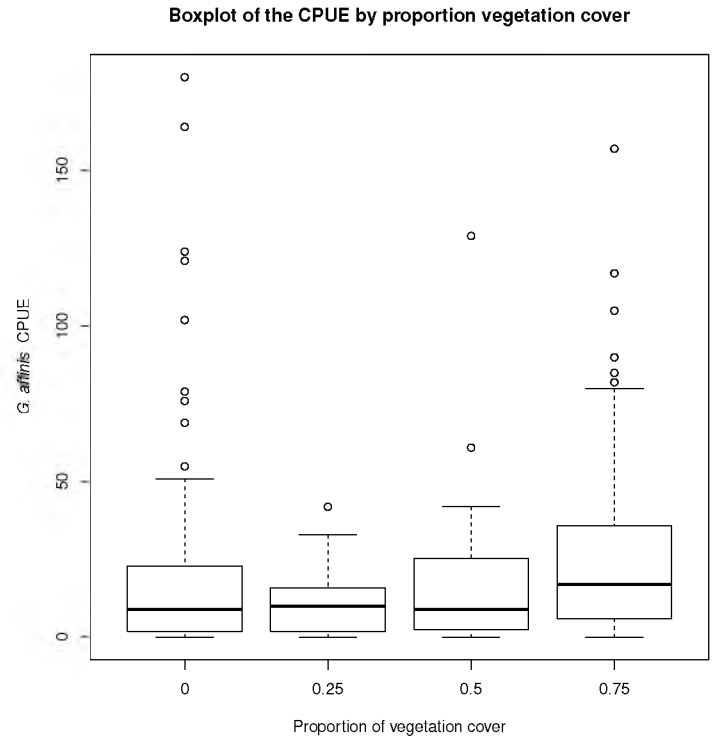


Figure 2.9: Boxplots of the *G. affinis* CPUE by percentage vegetation cover.

2.2 Analysis of Variance (ANOVA)

The experimental units, that is the objects upon which measurements are taken (Wackerly et al., 2008, page 652), are often called subjects although subjects could refer to an animal, a laboratory sample or a piece of industrial equipment. Wackerly et al. (2008, page 652 and 661) defines factors as independent variables which are controlled by the experimenter where the intensity level or distinct subcategory of the factor is called its level. In this context a treatment is a specific combination of factor levels (Wackerly et al., 2008, page 652). The one-way ANOVA model has one measurement, or dependent, variable and one nominal, or independent, variable which has more than two levels of the factor (McDonald, 2009, page 123). A one-way layout is used to compare k populations in which independent random samples are obtained from each of the populations of interest (Wackerly et al., 2008, page 653). A one-way completely randomized design is an experimental design where n relatively homogeneous experimental units are randomly divided into k subgroups of sizes $n_1, \ldots, n_k$ and each experimental unit in the subgroup receives the same treatment, each treatment is applied to only one subgroup and the k treatments are compared (Wackerly et al., 2008, page 652). The one-way ANOVA methodology is used to determine whether there are any statistically significant differences between the true, or population, means of three or more independent or unrelated groups, or similarly determine the significance of the effects of these treatments, or groups, on the dependent variable (Wackerly et al., 2008, page 652). The one-way ANOVA model is discussed in section 2.2.1.

The two-way ANOVA methodology is used to compare the effects of several levels of two factors (Zhang, 2012). A two-way ANOVA is used when there is one measurement variable and two nominal variables. The nominal variables are often called factors or main effects (McDonald, 2009, page 182) and are found in all possible combinations. For example, consider testing the hypothesis that stressed and unstressed rats have the same glycogen content in their gastrocnemius muscle where it is hypothesized that there are sex related differences in glycogen content. The two factors are stress level and sex. In this design, the stressed group contains both male and female rats and the unstressed group contains both male and female rats (McDonald, 2009, page 182). The factors may be fixed, random or a combination of fixed and random, that is the factors are mixed. The experiment may be conducted with replication or without replication, see section 2.2.2. Two-way ANOVA assumes that the observations within each cell are normally distributed and have equal variances. The subset of data occurring at the intersection of one level of every factor being considered is said to be in a cell of the data (McCulloch et al., 2008, page 135).

Fixed factor and random factors are two different types of factors in experimental design and ANOVA models (Larson, 2008). Fixed factor are factors whose levels represent specific populations of interest (Logan, 2011, page 254). For example a factor that comprises 'high', 'medium' and 'low' temperature treatments is a fixed factor and we are only interested in

comparing these three levels of the population. Fixed factors are restricted to specific non-randomly chosen treatment levels and ensures other experiments conducted with these levels are comparable, if the same specified treatments of the factors are used (Logan, 2011, page 254). In contrast to a fixed factor, random factors are factors whose levels are randomly chosen from all the possible levels of the population and are used as random representatives of the populations (Logan, 2011, page 254). Random factors are predictors where the distribution of individual coefficients are explicitly modeled by hyperparameters (Schielzeth & Nakagawa, 2013). For example five random temperature treatments could be used to represent a full spectrum of temperature treatments. In cases like these the conclusions are extrapolated to all the possible treatment levels and for succeeding experiments, a new random set of treatments are typically selected (Logan, 2011, page 254). Whilst fixed factors contrast the effects of the different levels of the factor (Logan, 2011, page 255), with random factors, the ANOVA objective is to make inference about random variation within a population (Larson, 2008). A model with both fixed and random effects is called a mixed-effects model (Pinheiro & Bates, 2000, page 1).

Nested ANOVA is an extension of the ANOVA methodology allowing variables to be nested within other variables. Nested ANOVA is used when we have one dependent variable and two or more categorical variables, at least one of which is nested within another (McDonald, 2009, page 173). Suppose we wish to test the null hypothesis that stressed and unstressed rats have the same glycogen content in their gastrocnemius muscle. In this example suppose that there are several cages of stressed rats and several cages of unstressed rats, with several rats in each cage. How much variation in the glycogen content is there among cages and how much variation in glycogen content is there between the stressed and unstressed groups of rats. In this context we consider the groups as the stressed or unstressed rats and each cage, containing several rats, as a subgroup nested within the stress or unstressed group. Each glycogen content level of a rat would be one observation within a subgroup (McDonald, 2009, page 173). The Department of Statistics at Rhodes University can be considered to have two populations of students; undergraduate and postgraduate students. Each of these groups has a level of study subgroup, namely first year, second year and so on, that is year nested in grade, where grade is defined as undergraduate or postgraduate. Typically a nested ANOVA has one null hypothesis for each level (McDonald, 2009, page 174). In a two-level nested ANOVA, one null hypothesis is that the groups have the same mean and the second null hypothesis is that the subgroups within each group have the same means (McDonald, 2009, page 174). In this context we might test if the true average level of satisfaction with the administration within the Department of Statistics is the same for undergraduate and postgraduate students, and in addition if the true average of the level of satisfaction with the administration within the Department of Statistics is the same amongst first years, second years etc. Nested ANOVA is discussed further in section 2.2.3 on page 42.

Consider a scenario where computer-assisted teaching and conventional teaching, that is

teaching which does not utilize computers, are being used by two different groups of teachers. In this scenario each teacher has his or her own class or group of students. The teaching method is a fixed factor because it has specific non-random levels of interest, namely computer-assisted or non computer-assisted teaching and no generalization to other teaching methods is intended. The teacher is a random factor and the teachers are nested within methods, because each teacher occurs at only one level of method within the design (Jackson & Brashers, 1994). Nested factors are characteristically random factors (Logan, 2011, page 284), of which the levels are randomly selected to represent all possible levels. When the main treatment effect, that is factor A, is a fixed factor, such designs are referred to as a mixed model nested ANOVA, whereas when factor A is a random factor, the design is referred to as a model II nested ANOVA. When all factors are fixed, the design is referred to as a model I mixed model (Logan, 2011, page 284).

When the sample sizes in a nested ANOVA are unequal, the p-values corresponding to the F-statistics may not be very good estimates of the actual probability (McDonald, 2009, page 176). Nested ANOVA should not calculate the F-statistics when the designs are unbalanced (Logan, 2011, page 290). ANOVA models can be balanced or unbalanced depending on the experimental designs as discussed in section 2.2.2. Nested ANOVA assumes that the observations within each subgroups are normally distributed and have equal variances (McDonald, 2009, page 174). Nested factors are typically fitted as random effects (Schielzeth & Nakagawa, 2013), as discussed in section 3.4.

Repeated measures ANOVA models are applicable in scenarios where the experimental design is such that observations are made on the same subject, item or individual more than once (Sullivan, 2008). For example we can measure an athlete's running speed three weeks prior, two weeks prior, one week prior to and on the day of a race. Repeated measures experiments are often done without replication, although sometimes they could be done with replication (McDonald, 2009, page 184). The number of observations obtained under similar experimental conditions, that is the number of observations in a cell or observations at level i of factor A and level j of factor B, are called replications (McCulloch et al., 2008, page 150). When the dependent variable under the repeated measures experiment is continuous the analysis can be performed using the repeated measures ANOVA methodology, but if the outcome is categorical then the test can be performed with a Chi-square test (Sullivan, 2008). The goal of the analysis is to compare responses among the treatments over the given time. Schielzeth & Nakagawa (2013) defines a treatment in this context as the experimental manipulation that is of primary interest in a study and indicate that these variables are typically fitted as fixed factors. This treatment is called the between-subjects treatment since the levels of a treatment can change only between the subjects and all measurements on the same subject will represent the same treatment. Time is called a within-subjects factor because different measurements on the same subject are taken at different times (Littell et al., 2007, page 160). Typically repeated measures experiments are factorial experiments (Sullivan, 2008).

Repeated measures ANOVA methodologies are discussed further in section 5.2.

2.2.1 The One-Way ANOVA

A one-way ANOVA model has one measurement, or dependent, variable and one nominal, or independent, variable which has more than two levels of the factor (McDonald, 2009, page 123). This model is used to test the null hypothesis that all treatments have the same population mean against the alternative hypothesis that at least one population mean differs from the other population means (McDonald, 2009, page 123). The logic behind this method is to calculate the means of the observations within each group and the variance among or between these experimental groups (McDonald, 2009, page 123). The shape of the distribution of the test statistic depends on two degrees of freedom, namely the numerator degrees of freedom that are associated with the among-group, or treatment, variances and the denominator degrees of freedom that are associated the within-group, or residual, variance (McDonald, 2009, page 123). The treatment degrees of freedom are given by the number of groups minus one, whereas the degrees of freedom of the residuals is the total number of observations minus the number of groups. Thus if there are n observations in k groups, the numerator degrees of freedom are $k-1$ and the degrees of freedom of the denominator are $n-k$ (McDonald, 2009, page 124). There is no interaction among variables in a one-way ANOVA since there is only one independent variable or factor.

Consider a scenario where independent random samples of the same variable are taken from three populations, or similarly a scenario where a factor has three levels. In this context we may wish to test if there is a significant difference in the true, or population, means or similarly if there is a significant effect due to the levels of the factor. The data in table 2.4 could be the result of sampling in this scenario (Wackerly et al., 2008, page 26). We may wish to test, at 5% level of significance, the hypothesis that at least one mean is significantly different to the others, that is we may wish to test the hypotheses $H_0 : \mu_A = \mu_B = \mu_C$ against H_1 : The 3 population means are not all equal or similarly there is no effect due to the level and the factor, $H_0 : \alpha_A = \alpha_B = \alpha_C = 0$ against the hypothesis that there is a significant effect due to the level of the factor, $H_1 : \alpha_i$ not all zero (at least one $\neq 0$).

Sample A	Sample B	Sample C
170	224	155
146	196	153
120	163	104
112	231	143
132	195	198

Table 2.4: Example 1: One-way ANOVA.

One-Way Fixed Effects ANOVA Models

Suppose that the observations of the dependent, or response variable, are denoted as Y_{ij} for the j^{th} trial, or observation, at the i^{th} level, $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. In this context we either have k populations or k fixed, that is non-random, levels of a factor. Denote the unknown mean of each population, or level, as μ_i. $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ denote the random sample of size n_i from population i, or level i, for $j = 1, \ldots, n_i$ within each of the k populations, or levels. The one-way ANOVA fixed effects model is defined as follows: for each $i = 1, \ldots, k$ or group

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where ε_{ij} represents the difference between each observation, in each of the k groups, and the corresponding population mean or the natural variation of the sample from each population (Wackerly et al., 2008, page 677). It is assumed that the error terms, ε, are normally distributed random variables with mean zero and constant variance σ^2, that is $\varepsilon_{ij} \sim N(0, \sigma^2)$ for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. As a result $Y_{ij} \sim N(\mu_i, \sigma^2)$. Let μ denote the overall mean, that is $\mu = \frac{1}{k}\sum \mu_i$. Let α_i denote the fixed, that is non-random, effect of the i^{th} population or level, that is $\alpha_i = \mu_i - \mu$, where $\sum \alpha_i = 0$. The classical one-way ANOVA model is then written as

$$\begin{aligned} Y_{ij} &= \mu_i + \varepsilon_{ij} \\ &= \mu + \alpha_i + \varepsilon_{ij}. \end{aligned}$$

In this context $H_0 : \mu_1 = \mu_2, \ldots, = \mu_k$ is equivalent to $H_0 : \alpha_1 = \alpha_2, \ldots, = \alpha_k = 0$ (Wackerly et al., 2008, page 678). In this model each $\mu_i = \mu + \alpha_i$ is considered as a fixed unknown constant, the magnitudes of which we wish to estimate (McCulloch & Searle, 2001, page 6). We are considering just these treatments and there is no thought or consideration of any other treatments and hence these effects are called fixed effects (McCulloch & Searle, 2001, page 6). The primary feature of fixed effects are that they are deemed to be constants representing the effects on the response variable of the various levels of the factor under consideration (McCulloch & Searle, 2001, page 7).

Estimation of the One-Way Fixed Effects ANOVA Model

To conduct a one-way ANOVA, the calculations proceed as follows: For each observation we compute the deviation from the overall mean, that is the individual value minus the overall mean. Squaring each deviation and summing over all observations yields the total sum of squares denoted by SST. SST represents the total variability of the observations from their mean. SST is partitioned into two components, namely:

- The sum of squares between treatments, denoted by SSA, which is obtained by summing the treatment means minus overall mean squared; and

- The sum of squares within treatments, denoted as SSE, which is obtained by summing the individual value minus the treatment mean squared.

The variability among group means is represented by SSA while SSE represents within-group or residual variability. The effective number of independent observations used in forming the sum of squares correspond to the degrees of freedom, df, of each sum of squares. With n observations, the total sum of squares SST has $n-1$ degrees of freedom. The between treatment sum of squares SSA, with $k \geq 2$ treatment groups, has $k-1$ degrees of freedom. The within treatments residual sum of squares SSE has $n-1-k+1 = n-k$ degrees of freedom (Wackerly et al., 2008, page 668). Dividing each sum of squares by its corresponding degrees of freedom yields a quantity called the mean square. If the null hypothesis is true, that is all the treatments have the same population mean, then the between treatments mean square is determined by $MSA = \frac{SSA}{k-1}$ and the residual mean square by $MSE = \frac{SSE}{n-k}$. MSE estimates the error variance, σ^2 (Wackerly et al., 2008, page 669). The mean square variance estimates are used to calculate the observed F value, $F = \frac{MSA}{MSE}$ which has $k-1$ and $n-k$ degrees of freedom (Wackerly et al., 2008, page 669). Large values of the observed F value provide evidence against the null hypothesis of equal treatment population means. The probability that a random variable selected from an F critical value will exceed the observed F value is called the p-value (McDonald, 2009, page 174). The ANOVA table for a one-way classification is given in table 2.5 (Wackerly et al., 2008, page 672).

Source of Variation	*df*	Sum of Squares	Mean Square	F Statistics	p-value
Treatments: Between treatments	$k-1$	SSA	$MSA = \frac{SSA}{k-1}$	$\frac{MSA}{MSE}$	p
Error: Within treatments	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	SST			

Table 2.5: One-way ANOVA table.

Denote the i^{th} group or treatment mean as $\bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{k} Y_{ij}$ and the overall or grand mean as $\bar{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i} Y_{ij}$. The sum of squares between treatments is expressed as

$$SSA = SS_{Treat} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2$$
$$= \sum_{i=1}^{k} n_i\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2.$$

The sum of squares within treatments can be expressed as

$$SSE = SS_{Error} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2.$$

The total sum of squares, $SST = SSA + SSE$, is given by

$$SST = SS_{Total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{..}\right)^2 .$$

Consider the data in table 2.4. The mean of each treatment, $\bar{Y}_{i.}$, and the grand mean of all 15 data points $\bar{Y}_{..}$ are 136, 201.8 and 150.6 respectively. The various sum of squares are computed as:

$$SS_{Treat} = \sum_{i=1}^{k} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 = 5(136 - 162.8)^2 + \cdots + 5(150.6 - 162.8)^2 = 11\,940.$$

$$SS_{Error} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 = (170 - 136)^2 + \cdots + (198 - 150.6)^2 = 9\,536.$$

$$SS_{Total} = SS_{Treat} + SS_{Error} = 11\,940 + 9\,536 = 21\,476.$$

These data are used to complete the one-way ANOVA table as shown in table 2.6. For these data we reject H_0 if $F_{\text{obs}} > F_{0.05,2,12} = 3.89$. The observed F value of 7.5128 is greater than the F critical of 3.89 and the p-value of 0.0077 is less than 5%, the level of significance, and hence we reject the null hypothesis and conclude that at least one population mean is significantly different to the other populations means.

Source of Variation	df	Sum of Squares	Mean Square	F Statistics	p-value
Treatments	$k - 1 = 2$	11 940	5 970.2	7.5128	0.0077
Error	$n - k = 12$	9 536	794.7		
Total	$n - 1 = 14$	21 476			

Table 2.6: One-way ANOVA table for the data in table 2.4.

Assumptions of the One-Way Fixed Effects ANOVA Model

To use the one-way ANOVA model to make inferences about the existence of effects certain assumptions must be met (Sahai & Ageel, 2012, page 11 and 12), namely

1. The errors ε_{ij} are assumed to be normally distributed with mean zero and constant variance σ^2;

2. The errors associated with any pair of observations are assumed to be uncorrelated, that is $E(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$ for $i \neq i'$ and $j \neq j'$;

3. Under the fixed effects model, the effects, $\alpha_i's$ are assumed to be fixed constants subject to the constraint that $\sum_{i=1}^{k} \alpha_i = 0$. This implies that the observations, Y_{ij}, are distributed with mean $\mu + \alpha_i$ and constant variance σ^2; and hence $E(\alpha_i \varepsilon_{ij}) = 0$, for all i and j;

4. The k populations are thus assumed to be normally distributed with means $\mu_1, \mu_2, \ldots, \mu_k$ and constant variance σ^2.

These assumptions provide the theoretical justification for applying ANOVA methods to compare several means. It is important to consider the consequences of applying one-way ANOVA when the assumptions are in question. The normality assumption does not have to be exactly satisfied as long as we are dealing with relatively large samples, that is twenty (20) or more observations from each population, although the consequences of large deviation from normality are somewhat more severe for random factors than for fixed factors (Kleinbaum et al., 2013, page 260). When one or more of these assumptions are in serious question one option is to transform the data, for example by means of log, square root or other transformations so that the transformed dependent variable more closely satisfies the assumptions. Another alternative is to select a more appropriate method of analysis such as non-parametric ANOVA methods (Kleinbaum et al., 2013, page 261).

Violation of these assumptions may invalidate the ANOVA results and hence it is important to examine each assumption (Larson, 2008). The Kruskal-Wallis test is a non-parametric method which is "equivalent" to the one-way ANOVA method but does not make assumptions about normality (van der Laan & Verdooren, 1987). The Kruskal-Wallis test assumes that the samples are from identical populations (Hecke, 2012) and tests the medians of the population.

Assessing the One-Way ANOVA Assumptions

Hypothesis testing procedures for a one-way ANOVA model assumes that the residuals are independent, normally distributed random variables with zero mean and constant variance. Thus the response variable for each of the treatment levels are

1. Normally distributed; and
2. Equally varied; and
3. Independent of one another.

Consider the scenario where we wish to test if the diversity of diatom species in the Rocky Mountain (USA) are affected by zinc and other heavy metals contamination levels, namely low, medium and high (Logan, 2011, page 266 and 268). The results of fitting a one-way ANOVA model to these data are shown in table 2.7. These data provide sufficient evidence that the zinc concentration have a significant effect on the diversity of the diatoms ($F_{obs} = 3.9387$, $df = 3, 30$, p-value $= 0.0176$). Boxplots of the diversity by concentration, figure 2.10, shows no violations of either the normality or homogeneity of variance assumption since the boxplots are not asymmetric and do not vary greatly in size or range. The quantile-quantile plot of the residuals of this model, figure 2.11 (b), shows no evidence of

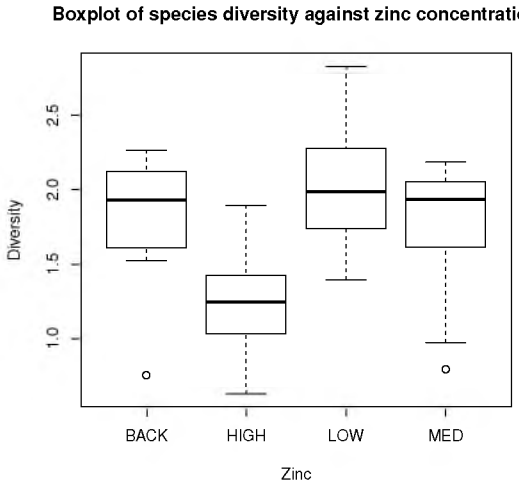
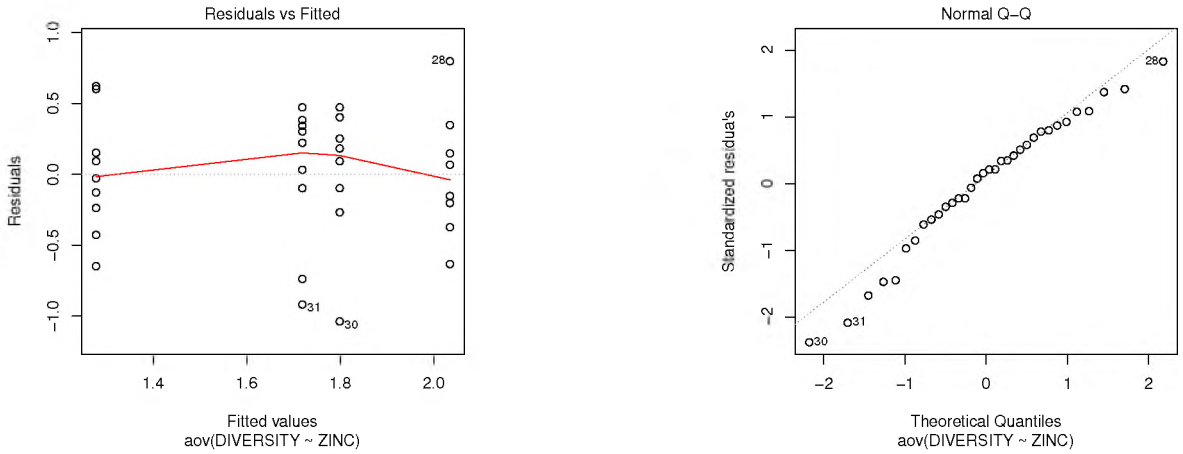


Figure 2.10: Boxplots of species diversity against zinc concentration.



(a) Residuals versus fitted values. **(b)** Normal Q-Q plot.

Figure 2.11: Diagnostics plots for the biodiversity study.

violation of the normality assumption as the observed residuals approximately follows the theoretical normal quantiles (the line). The fitted residuals, figure 2.11 (a), shows that the residuals of this model do not meet the homogeneity of variance assumption since the points are randomly scattered with a particular pattern. The residuals are normally distributed (Shapiro test, $W = 0.9688$, p-value $= 0.43$). There is no evidence that these data violate the model assumptions and as a result inference based on this model is valid.

Source of Variation	*df*	Sum of Squares	Mean Square	F Statistics	p-value
Zinc	3	2.5666	0.8555	3.9387	0.0176
Residuals	30	6.5164	0.2172		

Table 2.7: One-way fixed effects ANOVA applied to the biodiversity data set.

Consider the data shown in table 2.4, the one-way ANOVA example above. The box-

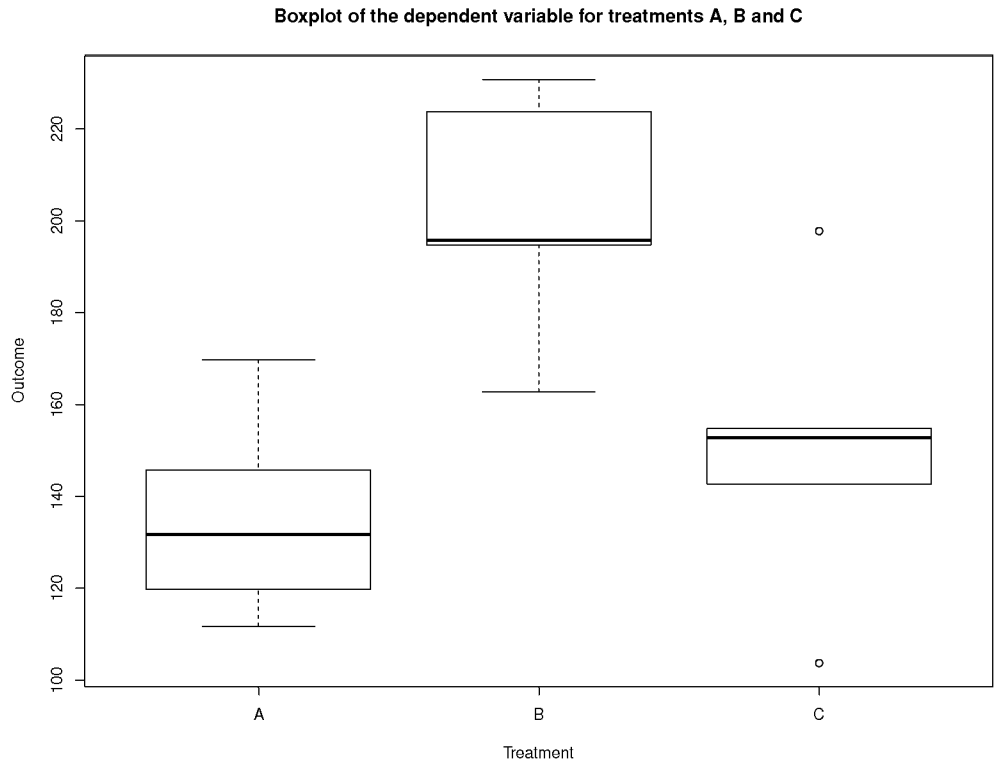


Figure 2.12: Boxplots of the data in table 2.4.

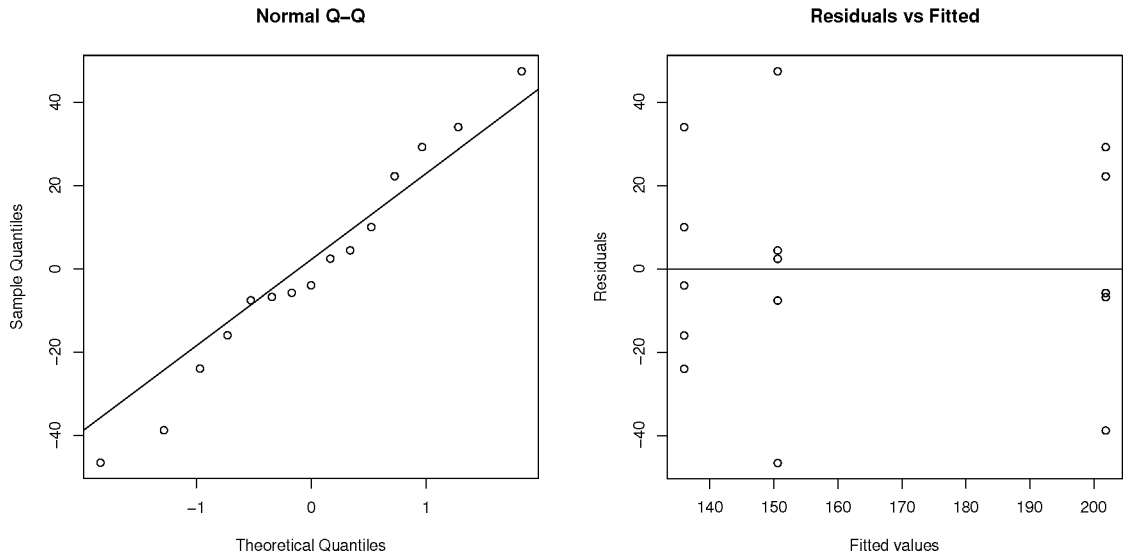
plots of these data, figure 2.12, suggest that the populations have unequal variances since the boxes, that is the data between the lower and upper quartiles, are not of equal size and show that the distribution of these data are not symmetric since the median is not located near the center of the boxes and the whiskers range are not similar. However these provide insufficient evidence that the populations have significantly different variances (Bartlett's K-squared = 0.52533, df = 2, p-value = 0.769). The normal Q-Q plot, figure 2.13 (a), suggest that the residuals of this model do not meet the normality assumption since the residuals deviate from the theoretical quantiles, particularly in the tails. These data provide sufficient evidence that the residuals are normally distributed (Shapiro test, W = 0.9589, p-value = 0.673) and as shown above, the populations have similar variances. Kruskal-Wallis's test is a non-parametric equivalent, or alternative, to the one-way ANOVA, except the Kruskal-Wallis test tests the median and not the mean (Logan, 2011, page 259). Kruskal-Wallis's test assumes that the shape of the distribution of the different groups, populations or treatments are the same. This test is not a good solution to the problem of heteroscedasticity (McDonald, 2009, page 156), that is to say this test should not be considered as an alternative if the one-way ANOVA model violates the homoscedasticity assumption. These data provide sufficient evidence that the medians are significantly different (Kruskal-Wallis chi-squared = 7.02, df = 2, p-value = 0.0299).

One-Way Fixed Effects ANOVA Example: *G. affinis* CPUE by Dam (or Location)

For the purposes of demonstration consider the relative abundance of *G. affinis*, measured as the catch per unit effort (CPUE), that is the number of fish caught per scoop of the net. Here we consider these data as independent observations at each impoundment or dam and ignore the other variables under consideration in the study. The summary statistics for these data are shown in table 2.2 on page 9 and the boxplots of the CPUE by dam is shown in figure 2.6 on page 9. There are large differences in the median CPUE at dams AVO, BV and OLI. The boxplot for BV dam is lower than that for all other dams and the boxplots shows evidence of unequal variances. AVO and DB have long upper whiskers which suggests there are higher CPUE of *G. affinis* at these dams and larger variability than at the other three dams. The median CPUE is different for all the dams and CPUE at AVO, DB, OLI and SLS are all positively skewed. Figure 2.6 shows that there are extreme CPUE outliers at all the dams. We may wish to test the hypothesis that there is significant variability among the dams.

Source of Variation	*df*	Sum of Squares	Mean Square	F Statistics	p-value
Dam	4	6 083	1 520.77	1.9941	0.09538
Error	299	228 025	762.62		
Total	303	234 108			

Table 2.8: One-way fixed effects ANOVA model: *G. affinis* CPUE by dam.

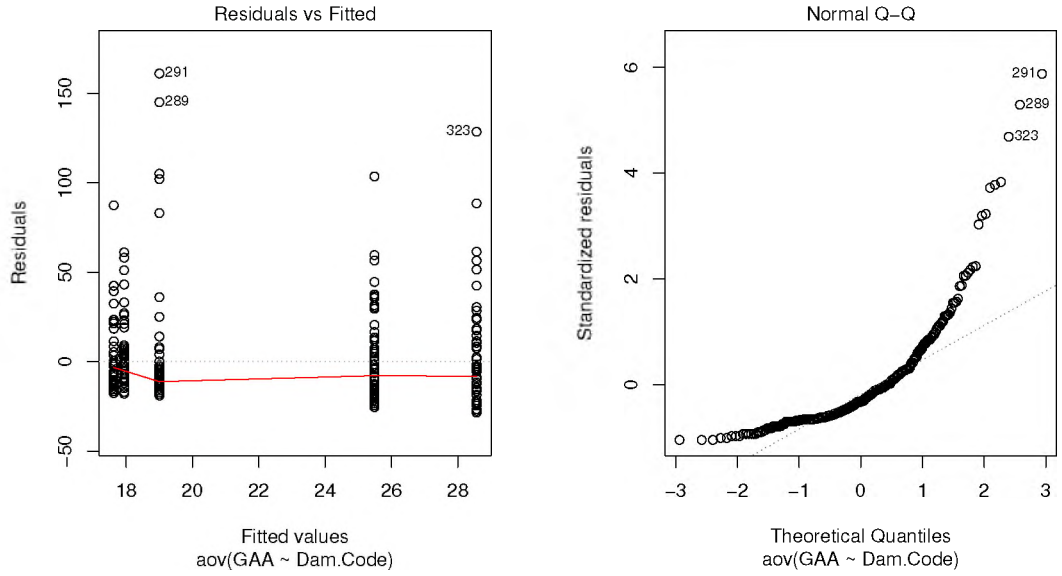


(a) Normal Q-Q plot. **(b)** Residuals versus fitted values.

Figure 2.13: Diagnostics plots for the one-way ANOVA model applied to the data in table 2.4.

Table 2.8 show the results for a one-way fixed effects ANOVA applied to the CPUE across dams. In this context, these data do not provide sufficient evidence that the dam has a

significant effect (F = 1.9941, df = 4, 299, p-value = 0.09538) on the true average *G. affinis* CPUE. This implies that the dams do not have significant different average relative abundances of *G. affinis*. The normal Q-Q plot, figure 2.14 (b), suggest that the residual of this model do not meet the normality assumption since the residuals points deviate from the theoretical quantiles. The fitted residuals, figure 2.14 (a), suggest that the assumption of homoscedasticity has not been met.



(a) Residuals versus fitted values. **(b)** Normal Q-Q plot.

Figure 2.14: Diagnostics plots of *G. affinis* CPUE by dam.

One-Way Random Effects ANOVA Models

Are the k populations, or treatment levels, chosen in the fixed effects model a random sample from some distribution? Suppose for example that a clinical trial is conducted at twenty different clinics in the Eastern Cape of South Africa where all the subjects receive the same dosage of a particular drug. We could represent each observation, in a fixed effects model, as

$$y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

for $i = 1, \ldots, 20$ denoting each specifically chosen clinic and $j = 1, \ldots, n_i$ denoting the j^{th} patient at each clinic. We may wish to test if the mean response to the drug is significantly different across these clinics. However it is not unreasonable to consider a scenario where these clinics are a random sample of all clinics in the Eastern Cape of South Africa. In this

context the model, a one-way random effects ANOVA could be denoted as

$$\begin{aligned} Y_{ij} &= \mu_i + \varepsilon_{ij} \\ &= \mu + \tau_i + \varepsilon_{ij} \end{aligned}$$

where, as before, μ_i denotes the mean of each population or level of the factor and μ the overall, or grand, mean and ε_{ij} represents the natural variation of observation j in population i. τ_i denotes the random effect on the dependent variable by treatment level i, where i is just one of many randomly chosen levels. The levels are chosen randomly so that they can be treated as a representation of the population of all possible treatment levels (McCulloch & Searle, 2001, page 9). In this context it is assumed that $\varepsilon_{ij} \sim N(0, \sigma^2)$ as per the fixed effects model. However in this model τ_i are being treated as random variables and it is customary to assume that these random variables are independently and identically distributed, with zero mean and constant variance σ_τ^2 (McCulloch & Searle, 2001, page 9), that is $\tau_i \sim i.i.d(0, \sigma_\tau^2)$ for all i. It is assumed that τ_i and ε_{ij} are independent. The terms σ_τ^2 and σ^2 are typically termed variance components (McCulloch & Searle, 2001, page 12), since

$$\begin{aligned} Var(Y_{ij}) &= Var(\mu + \tau_i + \varepsilon_{ij}) \\ &= Var(\tau_i) + Var(\varepsilon_{ij}) + 0 \\ &= \sigma_\tau^2 + \sigma^2. \end{aligned}$$

In this context the n_i observations are no longer independent since

$$\begin{aligned} Cov(Y_{ij}, Y_{ik}) &= Cov(\mu + \tau_i + \varepsilon_{ij}, \mu + \tau_i + \varepsilon_{ik}) \\ &= Cov(\tau_i + \varepsilon_{ij}, \tau_i + \varepsilon_{ik}) \\ &= Cov(\tau_i, \tau_i) \\ &= Var(\tau_i) \\ &= \sigma_\tau^2 \neq 0. \end{aligned}$$

Thus observations within the same group, or population i or level i, are correlated with intraclass correlation coefficient

$$\rho_{Y_{ij}, Y_{ik}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2}.$$

In a one-way random effects model the sum of squares are broken down as

$$SS_{Total} = SS_A + SS_{Error}$$

where SS_A denotes the blocking or grouping factor A, SS_{Total} the total and SS_{Error} the error sum of squares. In this model the focus is not on the group means, but the mean-to-mean variability. Thus in this model we test $H_0 : \sigma_\tau^2 = 0$ against $H_1 : \sigma_\tau^2 > 0$ (Logan, 2011, page

255).

Estimation of the One-Way Random Effect ANOVA Model

In this model (McCulloch & Searle, 2001, page 35)

$$\begin{aligned} E\left(Y_{ij}|\mu_i\right) = \mu_i &= \mu + \boldsymbol{\tau}_i \\ Y_i|\tau_i &\sim idd\,\mathrm{N}\left(\mu + \tau_i, \sigma^2\right) \\ \tau_i &\sim iid\,\mathrm{N}\left(0, \sigma_\tau^2\right) \end{aligned}$$

For $\mathbf{Y}_i = [Y_{i1}, \ldots, Y_{in_i}]'$ the model has $\mathbf{Y}_i \sim \mathrm{N}_{n_i}\left(\boldsymbol{\mu}\mathbf{1}_{n_i}, \mathbf{V}_i\right)$ where $\mathbf{V}_i = \sigma^2\mathbf{I}_{n_i} + \sigma_\tau^2\mathbf{J}_{n_i}$, $\mathbf{I}_n$ is an $n_i \times n_i$ identity matrix and $\mathbf{J}_n$ is a $n_i \times 1$ vector of ones. In this context

$$\mathbf{V}_i^{-1} = \frac{1}{\sigma^2}\mathbf{I}_{n_i} - \frac{\sigma_\tau^2}{\sigma^2\left(\sigma^2 + n_i\sigma_\tau^2\right)}\mathbf{J}_{n_i}$$

and $|\mathbf{V}_i| = \left(\sigma^2 + n_i\sigma_\tau^2\right)\sigma^{2(n_i-1)}$. In this context (McCulloch & Searle, 2001, page 36)

$$L\left(\theta\right) = \prod_{i=1}^{k}\left(2\pi\right)^{-n_i/2}\left|\mathbf{V}_i\right|\exp\left\{-\frac{1}{2}\left(\mathbf{y}_i - \boldsymbol{\mu}\mathbf{I}_{n_i}\right)'\mathbf{V}_i^{-1}\left(\mathbf{y}_i - \boldsymbol{\mu}\mathbf{I}_{n_i}\right)\right\}$$

and hence the log-likelihood is given by

$$\begin{aligned} l\left(\theta\right) = &-\frac{1}{2}N\log\left(2\pi\right) - \frac{1}{2}\sum_{i=1}^{k}\log\left(\sigma^2 + n_i\sigma_\tau^2\right) - \frac{1}{2}\left(N-k\right)\log\sigma^2 - \frac{1}{2\sigma^2}\sum_i\sum_j\left(y_{ij}-\mu\right)^2 \\ &+ \frac{\sigma_\tau^2}{2\sigma^2}\sum_i\frac{\left(y_{i\cdot} - n_i\mu\right)}{\sigma^2 + n_i\sigma_\tau^2}. \end{aligned}$$

where $N = n_1 + \cdots + n_k$, the total number of observations. For balanced designs, that is if $n_i = n$ for all i, the log-likelihood is

$$\begin{aligned} l\left(\theta\right) = &-\frac{1}{2}N\log\left(2\pi\right) - \frac{1}{2}k\left(n-1\right)\log\left(\sigma^2\right) - \frac{1}{2}k\log\left(\sigma^2 + n\sigma_\tau^2\right) - \frac{1}{2\sigma^2}\sum_i\sum_j\left(y_{ij}-\mu\right)^2 \\ &+ \frac{n^2\sigma_\tau^2}{2\sigma^2\left(\sigma^2 + n\sigma_\tau^2\right)}\sum_i\left(\bar{y}_{i\cdot} - \mu\right)^2. \end{aligned}$$

Letting $SSA = \sum_i n\left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2$, $SSE = \sum_i\sum_j\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2$ denotes the sums of squares (as per fixed effects model) and $\lambda = \sigma^2 + n\sigma_\tau^2$ yields

$$l\left(\theta\right) - \frac{1}{2}N\log\left(2\pi\right) - \frac{1}{2}k\left(n-1\right)\log\left(\sigma^2\right) - \frac{1}{2}k\log\lambda - \frac{SSE}{2\sigma^2} - \frac{SSA}{2\lambda} - \frac{kn\left(\bar{y}_{\cdot\cdot} - \mu\right)}{2\lambda} = 0.$$

The maximum likelihood estimates are derived in the usual way and yield

$$\frac{\partial l(\theta)}{\partial \mu} = \frac{kn(\bar{y}_{..} - \mu)}{\lambda} = 0$$

that is $\hat{\mu} = \bar{y}_{..}$

$$\frac{\partial l(\theta)}{\sigma^2} = \frac{-k(n-1)}{2\sigma^2} + \frac{SSE}{2\sigma^4} = 0$$

that is $\hat{\sigma}^2 = \frac{SSE}{k(n-1)} = MSE$.

$$\frac{\partial l(\theta)}{\partial \sigma_\tau^2} = \frac{-k}{2\lambda} + \frac{SSA}{2\lambda^2} + \frac{kn(\bar{y}_{..} - \mu)^2}{2\lambda^2} = 0$$

that is $\hat{\lambda} = \frac{SSA}{k} = \left(1 - \frac{1}{k}\right) MSA$

$$\hat{\sigma}_\tau^2 = \frac{\hat{\lambda} - \hat{\sigma}^2}{n} = \frac{\left(1 - \frac{1}{k}\right) MSA - MSE}{n}$$

where $MSA = \frac{SSA}{k-1}$ and $MSE = \frac{SSE}{k(n-1)}$. The unbalanced estimators are derived in a similar fashion, see for example (McCulloch & Searle, 2001, page 42).

Assumptions of the One-Way Random Effects ANOVA Model

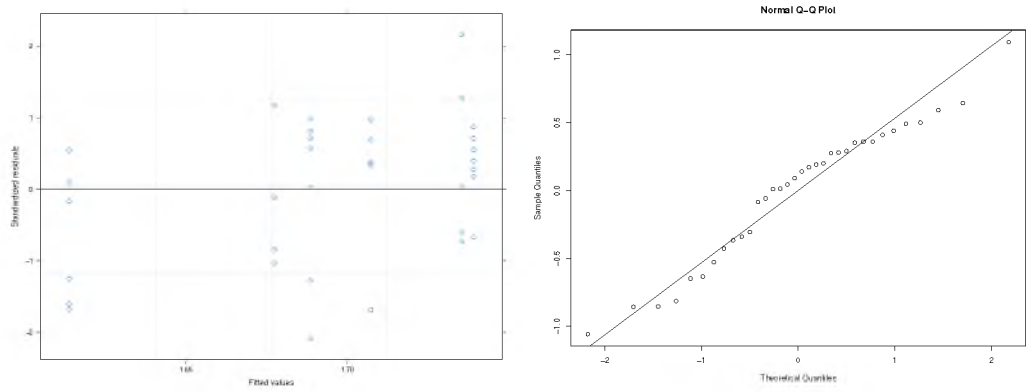
The assumptions of the one-way random effects model are:

1. Under random effects model, the effects $\tau_i's$ are assumed to be randomly distributed with mean zero and variance σ_τ^2. Furthermore, $\tau_i's$ are uncorrelated with each other and each of the $\tau_i's$ and $\varepsilon_{ij}'s$ are also uncorrelated, that is, $E(\tau_i\tau_{i'}) = 0$ for $i \neq i'$ and $E(\tau_i\varepsilon_{ij}) = 0$ for all i and j;

2. Under the random effects model, $\mu_1, \mu_2, \ldots, \mu_k$ are the means of the k randomly selected subpopulations from a population with mean μ and variance σ_τ^2; and

3. All the factor levels must have the same effect in the population of random effects $\tau_i's$, then $\sigma_\tau^2 = 0$ (Sahai & Ageel, 2012, page 12).

Consider the scenario where the interest is if diatom diversity differed across Rocky Mountain streams. Here each stream could be treated as a random factor and we wish to test the hypothesis that there is no added variation in diatom diversity due to stream. Thus we wish to perform one-way random effects analysis of variance of species diversity versus stream (Logan, 2011, page 273 and 274). The fitted one-way random effect ANOVA model, table 2.9, shows that there is no added variance in diatom diversity to streams ($F_{obs} = 1.4108$, $df = 5, 28$, p-value $= 0.2508$). Computing the variance estimates of the random factor require the nlme package in R. The 'lme' function in the nlme package will estimate the random factors



(a) Residuals versus fitted values. **(b)** Normal Q-Q plot.

Figure 2.15: Diagnostics plots for one-way random effects ANOVA model .

for the model and the parameters are estimated by either the maximum likelihood or the restricted maximum likelihood methodologies.

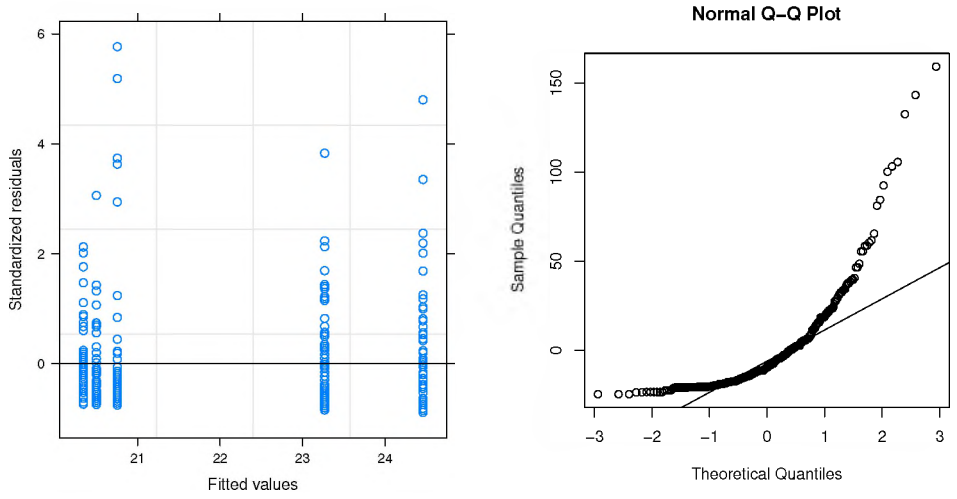
Source of Variation	df	Sum of Squares	Mean Squares	F Statistics	p-value
Stream	5	1.8278	0.3656	1.4108	0.2508
Residuals	28	287.2552	0.2591		

Table 2.9: One-way random effects ANOVA model example.

The results above indicate that most of the variance in diatom diversity is due to differences between sampling stations within streams (ML: 0.2572, REML: 0.2576) and that very little variance is added due to differences between streams (ML: 0.0099, REML: 0.0205) (Logan, 2011, page 274), see appendix A.1. In this context we verify the model quality by running diagnostic checks. The points in the residual plot, figure 2.15 (a), are randomly scattered with no particular pattern, this suggest that the assumption of homoscedasticty has been met. The normal Q-Q plot, figure 2.15 (b), suggest that the residuals of this model do not meet the normality assumption since the points deviate from the theoretical quantiles. The residuals are not normally distributed (Shapiro test, $W = 0.9565$, p-value $= 0.0184$).

One-Way Random Effects ANOVA Example: *G. affinis* CPUE by Dam (or Location)

Consider *G. affinis* CPUE by dam or location, where the dam is considered a random factor and ignore the other variables under consideration in this study. The summary statistics of *G. affinis* CPUE by dam or location are shown in table 2.2 on page 9. The one-way random effects ANOVA results are same the as the one under the fixed effects model, given in table 2.8 on page 21. We wish to test that there is significant variability of CPUE across dam levels. These data provide insufficient evidence that the variance components of CPUE are significantly larger than zero across the dams ($F_{obs} = 1.9941$, $df = 4, 299$, p-value $= 0.0954$).



(a) Residuals versus fitted values. **(b)** Normal Q-Q plot.

Figure 2.16: Diagnostics plots for one-way random effects ANOVA model by dam.

The estimated within standard deviations (ML: 27.6138, REML: 27.61438) showed high variability compared to the between standard deviations, σ_τ^2, (ML: 2.7464, REML: 3.5379), see appendix A.2. The normal Q-Q plot, figure 2.16 (b), suggests that the standardized residuals of this model do not meet the normality assumption since the residuals deviate from the theoretical quantiles. The residuals of this model are not normally distributed (Shapiro test, $W = 0.7308$, p-value < 0.001). The fitted residuals, figure 2.16 (a), show that the assumption of homoscedasticity has not been met since points are randomly scattered with a particular pattern. There is heterogeneity of variances across dams (Bartlett's K-squared $= 36.026$, $df = 4$, p-value < 0.001).

2.2.2 Multi-Factor ANOVA

When there are several factors, the effects of the combination of two or more factors is called the interaction effect while the effect of a single factor is called a main effect. If the change in the mean of the dependent or response variable between two levels of factor A is the same for different levels of factor B, we say that there is no interaction; but if that change is different for different levels of factor B, we say there is an interaction (McCulloch et al., 2008, page 140). When the interaction term is significant, it is advisable not to test the effects of the individual factors (McDonald, 2009, page 183). If the number of observations in each subgroups are the same then the design is said to be balanced, however if they are unequal the design is said to be unbalanced (McDonald, 2009, page 183). If an unadjusted ANOVA methodology is applied to unbalanced data it's effectiveness is strictly limited (McCulloch et al., 2008, page 143).

With balanced design, one factor can be held constant whereas the other is varied independently. When fixed factorial designs are balanced, the total variance in the response variable

could be sequentially partitioned into what is explained by each of the model terms (factors and their interactions) and what is left unexplained (Logan, 2011, page 321). For balanced designs, the total sum of squares is equal to the additive sum of squares of each of the components, including the residuals (Logan, 2011, page 322). For example, when an A by B factorial experiment is conducted with an equal number of observations per treatment combination, the corrected total sum of squares is partitioned as $SS_{Total} = SSA + SSB + SSAB + SSE$, where AB represents the interaction between factor A and factor B. When the design has different combinations with different number of replicates but no empty cells, we referred to it as unbalanced design (Hector et al., 2010). If the proportional number of replicates is not the same across treatments, the design is non-orthogonal and the two explanatory variables are not independent of each other (Logan, 2011, page 322). When explanatory variables are correlated with each other due to imbalance in the number of replicates for different treatment combinations, the values of the sum of squares depends on the position of the factors in the ANOVA model formula (Hector et al., 2010).

Two-way ANOVA aims to compare the effects of several levels of two factors in a factorial experiment with a two-way layout. It is a widely used method in experimental sciences, ranging from biology to psychology (Zhang, 2012). In a factorial experimental design, each factor is crossed with the other factors. Consider two fixed factors, A and B, with a levels for factor A, b levels for factor B and ab levels formed by combinations of A and B. Individual factors are associated with the main effects, whereas crossed factors create the interaction effects (Larson, 2008; Zhang, 2012).

The two-way ANOVA model requires statistical notation to identify specific levels of A and B and their combination, as well as to denote each replicate within each combination. Suppose at the $(i,j)^{\text{th}}$ cell, Y_{ijk} denote the $k = 1, \ldots, n_{ij}$ random observations. The two-way randomized block ANOVA model, is of the form

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

where ε_{ij} are assumed to be independent normally distributed random variables with zero mean and constant variance σ^2_{ij} and μ_{ij} and σ^2_{ij} denote the $(i,j)^{\text{th}}$ cell mean and variance respectively (Zhang, 2012).

Source of Variation	df	Sum of Squares	Mean Square	F Statistics
Factor A: Treatment	$a-1$	SSA	$MSA = SSA/(a-1)$	MSA/MSE
Factor B: Times	$b-1$	SSB	$MSB = SSB/(b-1)$	MSB/MSE
Interaction (AB)	$(a-1)(b-1)$	$SSAB$	$MSAB = SSAB/(a-1)(b-1)$	$MSAB/MSE$
Error	$ab(n-1)$	SSE	$MSE = SSE/ab(n-1)$	
Total	abn	SST		

Table 2.10: Two-way ANOVA table: Sources of variation.

Estimation of the Two-Way Randomized Block ANOVA Model

Partitioning the total sum of squares

$$S^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \bar{Y}_{...}\right)^2}{n-1}$$

where the total sum of squares is the deviation of the individual score from the overall mean $(Y_{ijk} - \bar{Y}_{...})$. $SS_{Total} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \bar{Y}_{...}\right)^2$ can be broken out as

$$SS_{Error} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \bar{Y}_{ij.}\right)^2$$

with $ab(n-1)$ degrees of freedom. This is similar to the within sum of squares in the one-way ANOVA. The degrees of freedom are derived from the fact that there are n cases where $a \times b$ cell means are to be estimated. Consider the sum of squares corresponding to factor A, namely

$$SSA = nb\sum_{i=1}^{a}\left(\bar{Y}_{i..} - \bar{Y}_{...}\right)^2$$

with $(a-1)$ degrees of freedom. Similarly, the sum of squares corresponding to factor B is

$$SSB = na\sum_{j=1}^{b}\left(\bar{Y}_{.j.} - \bar{Y}_{...}\right)^2$$

with $(b-1)$ degrees of freedom. The interaction effect is given by

$$SSAB = n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}\right)^2$$

with $(a-1) \times (b-1)$ degrees of freedom (Zhang, 2012).

Assumptions of the Two-Way ANOVA Model

To use the two-way ANOVA model to make inferences about the existence of effects certain assumptions must be met (Sahai & Ageel, 2012, page 19), namely

1. The random variation around the sample means has the same magnitude at all levels of the factor. The residuals contributing to this variation are free to vary independently of each other and that the residual variation approximates to a normal distribution; and

2. The observations are assumed to be independent and that they have a similar distribution; and

3. We assume that $\varepsilon'_{ijk}\,s$ are independently distributed normal random variables with mean zero and constant variance σ^2.

Two-Way Fixed Effects ANOVA Model with Interaction

Consider a two-way model where $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$, that is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

where

- $i = 1, \ldots, a$ denotes the levels of factor A; and
- $j = 1, \ldots, b$ denotes the levels of factor B; and
- $k = 1, \ldots, n_{ij}$ indicates the observation number within cell i, j.

In this model μ denotes the grand or overall mean, α_i denotes the effect of the fixed effect due to the i^{th} level of factor A, β_j denotes the effect of the fixed effect due to the j^{th} level of factor B and $(\alpha\beta)_{ij}$ denotes the fixed effect due to the interaction of level i of factor A and level j of factor B (Johnson & Wichern, 2014, page 307). Separate hypotheses are associated with each of the main effects and the interaction term (Logan, 2011, page 314). The null hypothesis for factor A is that the true group means across the levels, i, of factor A are all equal, or equivalently there is no effect due to factor A, typically denoted as $H_0 : \alpha_i = 0$, $i = 1, \ldots, a$. The null hypothesis for factor B is that the true group means across the levels, j, of factor B are all the same, or equivalently there is no effect due to factor B, typically denoted as $H_0 : \beta_j = 0$, $j = 1, \ldots, b$. The hypothesis that there is no interaction is typically denoted as $H_0 : (\alpha\beta)_{ij} = 0$. The expected response at the i^{th} level of factor A and the j^{th} level of factor B is given by

$$E\left(Y_{ijk}\right) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}\,.$$

The presence of the interactive term implies that the factor effects are not additive (Johnson & Wichern, 2014; Logan, 2011, pages 308 and 316). As per the one-way ANOVA each observation can be decomposed as

$$Y_{ijk} = Y + \left(\bar{Y}_{i\cdot} - \bar{Y}\right) + \left(\bar{Y}_{\cdot j} - \bar{Y}\right) + \left(\bar{Y}_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}\right) + \left(Y_{ijk} - \bar{Y}_{ij}\right)$$

where

- $\bar{Y}$ denotes the overall average; and
- $\bar{Y}_{i\cdot}$ denotes the overall average for the i^{th} level of factor A; and

- $\bar{Y}_{.j}$ denotes the overall average for the j^{th} level of factor B; and
- $\bar{Y}_{ij}$ denotes the overall average for the i^{th} level of factor A and the j^{th} level of factor B.

Squaring and summing over $k = 1, \ldots, n_{ij}$, that is the observations yields the sum of squares for factors A, B, the interaction term and the error, as shown in table 2.10.

Example: Two-Way Fixed Effects ANOVA

Consider a two-way ANOVA where there are four replicates (Faraway, 2004, page 181 and 182). The factorial design is used by investigators to select a fixed number of levels of each of a number of factors and then run experiments with all possible combinations. The measured responses are the survival times of groups of four animals randomly allocated to each of the twelve combinations of three poisons and four treatments. The experiment was part of an investigation to combat the effects of certain toxic agents. Here we will consider the effects of both poisons and treatments and their influences might not be additive, that is, the difference in survival times between specific treatments may be different for different poisons. The data for the scenario is provided in table 2.11.

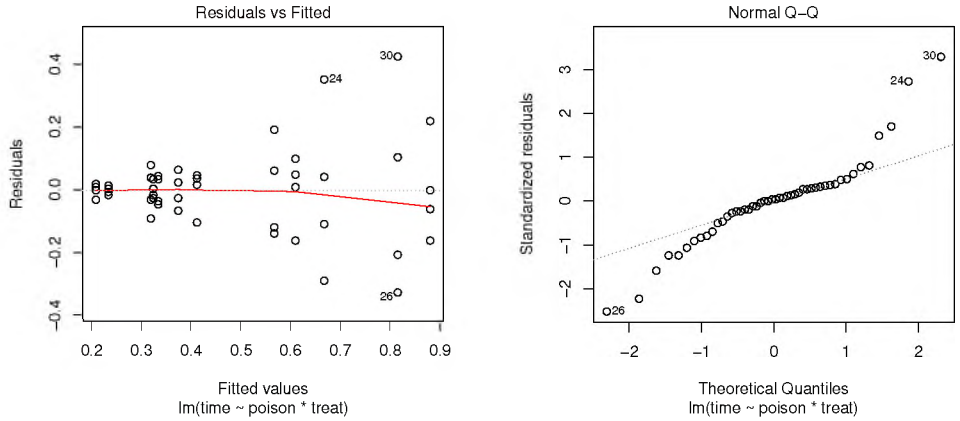
	Treatment A	Treatment B	Treatment C	Treatment D
Poison I	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
Poison II	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
Poison III	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

Table 2.11: Toxic agent data set.

Source of Variation	df	Sum of Squares	Mean Square	F Statistics	p-value
Treatments	3	0.9212	0.3071	13.8056	$3.777e-06$
Poison	2	1.0330	0.5165	23.2217	$3.331e-07$
Interaction: Treatments: Poison	6	0.2501	0.0417	1.8743	0.1123
Residuals	36	0.8007	0.0222		

Table 2.12: Two-way ANOVA table for the data in table 2.11.

The two-way ANOVA results are shown in table 2.12. These results indicate that there is no interaction between the treatments and poisons ($F_{obs} = 1.8743$, df $= 6, 36$, p-value $= 0.1123$). The main effects, treatments ($F_{obs} = 13.8056$, $df = 3, 36$, p-value < 0.001) and poison ($F_{obs} = 23.2217$, $df = 2, 36$, p-value < 0.001) are significant. This means that the mean survival time in the cell which receives the t^{th} treatment are significantly different. The normal Q-Q plot, figure 2.17 (b), shows that the residuals of this model do not meet the normality assumption since the points deviates from the theoretical quantile line. The residuals are not normally distributed (Shapiro test, $W = 0.9123$, p-value $= 0.0016$). The fitted residuals, figure 2.17 (a), shows that the residuals of this model do not meet the homogeneity of variance assumption. There is heterogeneity of variances in the poison (Bartlett'sK-squared $=$ 25.88, df $= 2$, p-value < 0.001) and treatment groups (Bartlett's K-squared $= 13.211$, df $=$ 3, p-value $= 0.0042$).



(a) Residuals versus fitted values. **(b)** Normal Q-Q plot.

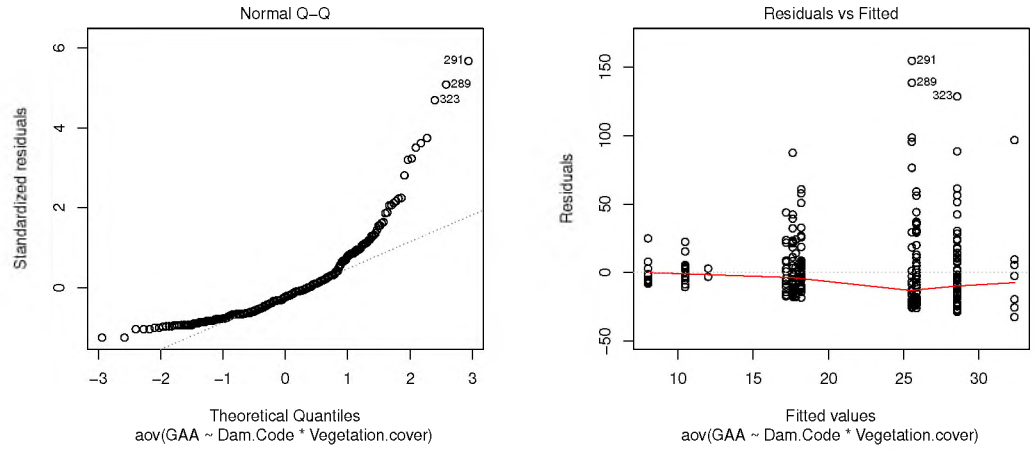
Figure 2.17: Diagnostics plots for a two-way ANOVA model applied to the data in table 2.11.

Two-Way Fixed Effects ANOVA Model : *G. affinis* CPUE by Dam and Percentage Vegetation Cover

Consider a two-way fixed effects ANOVA model of *G. affinis* CPUE by dam and percentage vegetation cover, table 2.13. With these p-values we would conclude that the differences between the levels of both factor A ($F_{obs} = 1.138$, $df = 3, 293$, p-value $= 0.3338$) and factor B ($F_{obs} = 1.999$, $df = 4, 293$, p-value $= 0.0947$) as well as the interactions ($F_{obs} = 1.126$, $df = 3, 293$, p-value $= 0.3386$) are not statistically significant. The results provide evidence that dam and percentage vegetation cover have no significant effect on the true average *G. affinis* CPUE. The normal Q-Q plot, figure 2.18 (a), suggest that the standardized residuals of this model does not meet the normality assumption since the residuals deviates from the theoretical quantiles. The fitted residuals, figure 2.18 (b), shows that the assumption of homoscedasticity has not been met.

Source of Variation	df	Sum of Squares	Mean Square	F Statistics	p-value
Dam	4	6 083	1 520.8	1.999	0.0947
Vegetation cover	3	2 598	865.9	1.138	0.3338
Interaction : Dam : Vegetation cover	3	2 570	856.8	1.126	0.3386
Residuals	293	222 857	760.6		

Table 2.13: Two-way fixed effects ANOVA: *G. affinis* CPUE by dam and percentage vegetation cover.



(a) Normal Q-Q plot. **(b)** Residuals versus fitted values.

Figure 2.18: Two-way fixed effects ANOVA model diagnostics plots: *G. affinis* CPUE by dam and vegetation cover.

Dam	Percentage vegetation cover			
	0	0.25	0.5	0.75
AVO1	2	6	8	48
BV1	38	15	11	0
DB1	0	0	0	64
DB1	0	0	0	64
HBT (OLI)	0	0	0	48
HBT (SLS)	48	0	16	0

Table 2.14: Cross tabulation of dam and percentage vegetation cover.

The interaction degrees of freedom in table 2.13 are incorrect because of the missing cells, this is an incomplete design as shown in table 2.14. R does not generate an error regarding the degrees of freedom, that is the unbalanced design. A block design with at least one zero entry in its incidence matrix is called an incomplete block design. There are nine missing cells, as shown in table 2.14. Since the design is unbalanced and has missing cells, we may use the 'lme' function in the R package nlme for these data. Linear mixed effects models accommodate balanced and unbalanced designs, correlated and hierarchical data which makes

these models the preferred approach to analyzing unbalanced, un-replicated factorial designs (Logan, 2011, page 370).

The Two-Way Random Effects ANOVA Model

Suppose we have two factors A at levels $i = 1, \ldots, a$ and B at levels $j = 1, \ldots, b$ where n_{ij} denotes the number of observations at level i of A and level j of B and denote these observations as $Y_{ij1}, Y_{ij2}, \ldots, Y_{ijn}$ where $n_{ij} > 1$ for all i, j. Consider the model with factor A fixed and B random, namely

$$\begin{aligned} Y_{ijk} &= \mu_{jk} + \varepsilon_{ijk} \\ &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \text{ for i} = 1, 2, \ldots, a, \text{ j} = 1, 2, \ldots, b \text{ and k} = 1, 2, \ldots, n_{ij} \end{aligned}$$

where Y_{ijk} are the experimental responses, μ is the grand or overall mean, α_i is the treatment effect for the i^{th} factor A, β_j is the treatment effect for the j^{th} factor B, $(\alpha\beta)_{ij}$ is the interaction effect for the combination of the i^{th} factor A and the j^{th} factor B and ε_{ijk} denotes the experimental error with zero mean, variance σ^2 (Faraway, 2004, page 179 and 180). The effects of factor A, $\alpha_i \sim N(0, \sigma^2_\alpha)$, the treatment effects for the j^{th} factor B, $\beta_j \sim N\left(0, \sigma^2_\beta\right)$, the interaction effect for the combination of the i^{th} factor A and the j^{th} factor B, $(\alpha\beta)_{ij} \sim N\left(0, \sigma^2_{\alpha\beta}\right)$ and the covariance $Cov\left(Y_{ijk}, Y_{i'j'k'}\right) = \delta_{ii'}\sigma^2_\alpha + \delta_{jj'}\sigma^2_\beta + \delta_{ii'}\delta_{jj'}\sigma^2_{\alpha\beta} + \delta_{ii'}\delta_{jj'}\delta_{kk'}\sigma^2$. Consider estimating σ^2_β in the two-way random effects ANOVA model. A natural estimate is

$$\hat{\sigma}^2_\beta = na(MSB - MSAB)$$

where we test if there is no added variance due to all possible levels of factor A and B, that is, $H_0 : \sigma^2_\alpha = 0$ and also test if there is no added variance due to all possible interactions between all possible levels of A and B, that is, $H_0 : (AB) : \sigma^2_{\alpha\beta} = 0$ (Logan, 2011, page 316 and 317).

Two-Way Random Effects ANOVA Model Example: *G. affinis* CPUE by Dam (or Location) and Percentage Vegetation Cover

Consider *G. affinis* CPUE by dam, where dam is considered as a random factor, utilizing the two-way random effects ANOVA model.

```
# Fit a two-way random effects ANOVA model by dam as a
   random factor and percentange vegetation cover.
# ~1|Dam.Code specifying the model for the random
   effects with Dam.Code as the grouping structure and 1
    indicating that the random effect is constant within
    each group.
```

```
> random.model<-lme(GAA~Dam.Code*Vegetation.cover,
   random=~1|Dam.Code,method="ML",data=gaadata)
> summary(random.model)
# Error message from R
Error in MEEM(object, conLin, control$niterEM) :
Singularity in backsolve at level 0, block 1
```

We cannot estimate the variance components for dam and vegetation cover because of the missing cells or incomplete design. The error message from R emphasizes that the two-way random effects method will not work for these data. Generalized linear models can be a solution to these kind of problems, since it handles both balanced and unbalanced designs as discussed in section 3.5 on page 47.

Two-Way Mixed-Effects ANOVA Model

Consider the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \text{ for } i = 1, 2, \ldots, a, j = 1, 2, \ldots, b, \text{ and } k = 1, 2, \ldots, n_{ij}$$

where $\alpha_i \sim N(0, \sigma_\alpha^2)$, that is factor A is random β_j are constants, that is factor B is fixed and $(\alpha\beta)_{ij} \sim N\left(0, \frac{(b-1)\sigma_{\alpha\beta}^2}{b}\right)$. The constraints under the two-way mixed ANOVA model $\sum_{j-1}^{b} \beta_j = 0$ and $\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij} = 0$. We may wish to test the hypothesis that the population group means are all equal for the levels of factor A, namely $H_0(\text{A}) : \mu_1 = \mu_2 = \cdots = \mu_a = \mu$. If the effects of the i^{th} group are the difference between the i^{th} group mean and the overall mean $(\alpha_i = \mu_i - \mu)$ then the null hypothesis can be expressed as there no effects of any level of this factor pooled over all possible levels of the random factor, that is $H_0(\text{A}) : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$. The hypothesis that for the random factor B there is no added variance due to all possible levels of factor B, that is $H_0(\text{B}) : \sigma_\beta^2 = 0$. The interaction between a fixed factor and a random factor is always considered a random factor. The hypothesis, $H_0(\text{AB}) : \sigma_{\alpha\beta}^2 = 0$, denotes that there is no added variances due to all possible interactions between all possible levels of factor A and B.

Two-Way Mixed Effects ANOVA : *G. affinis* CPUE by Dam (or Location) and Percentage Vegetation Cover

We want to test the hypotheses that the means grouped by one factor are the same and also test if there is an interaction between the two factors dam and percentage vegetation cover. The two-way mixed ANOVA model cannot be fitted because of the missing data, the design

is highly unbalanced. The generalized linear model would be the appropriate method to use for these kind of data and is discussed in section 3.5.

```
# Fit a two-way mixed-effects ANOVA model by dam and
   percentage vegetation cover.
# ~1|Dam.Code specifying the model for the random
   effects with Dam.Code as the grouping structure and 1
    indicating that the random effect is constant within
    each group.
> mixed.model<-lme(GAA ~Dam.Code*Vegetation.cover,
   random = ~1|Dam.Code,method = "ML",data = gaadata)
# Error message from R
Error in MEEM(object, conLin, control$niterEM) :
Singularity in backsolve at level 0, block 1
```

Fitting a Linear Model: *G. affinis* CPUE

Which biotic and abiotic factors have an effect on the *G. affinis* CPUE? For these data we model the linear relationship between a response or dependent variable and one or more explanatory or independent variables. The results of the linear model used to assess the effects of dam age, mean temperature, percentage vegetation cover, *Glossogobius callidus* abundance and *Oreochromis mossambicus* abundance on *G. affinis* CPUE are displayed in table 2.15. The dam age (in years) has a significant effect on *G. affinis* CPUE ($F = 6.9355$, $df = 1, 298$, p-value $= 0.0089$). The vegetation cover ($F_{obs} = 0.0807$, $df = 1, 298$, p-value $= 0.7766$), mean temperature ($F_{obs} = 0.2947$, $df = 1, 298$, p-value $= 0.2947$), *Glossogobius callidus* abundance ($F_{obs} = 0.9752$, $df = 1, 298$,p-value $= 0.3242$) and *Oreochromis mossambicus* abundance ($F_{obs} = 0.2200$, $df = 1, 298$, p-value $= 0.6394$) have no significant effect on the *G. affinis* CPUE in this linear model. The normal Q-Q plot, figure 2.19, shows that the residuals of this model do not meet the normality assumption since the points deviates from the theoretical quantile. The residuals are not normally distributed (Shapiro test, $W = 0.7281$, p-value < 0.001). The fitted residuals, figure 2.19, shows that the residuals of this model do not meet the homogeneity assumption since the points are randomly scattered with a particular pattern. There is heterogeneity of variances across dams (Bartlett's K-squared $= 36.026$, df $= 4$, p-value < 0.001). Based on the graphical illustration in figure 2.19, this model is not appropriate hence other statistical models need to be considered. Further statistical models that would be appropriate for these data are discussed in chapter 3.

Source of Variation	*df*	Sum of Squares	Mean Square	F Statistics	p-value
Dam age	1	5 283	5 283.4	6.9355	0.0089
Mean temperature	1	839	839.4	1.1019	0.2947
Glossogobius callidus	1	743	742.9	0.9752	0.3242
Oreochromis mossambicus	1	168	167.6	0.2200	0.6394
Vegetation cover	1	61	61.4	0.0807	0.7766
Residuals	298	227 013	761.8		

Table 2.15: General linear ANOVA model: *G. affinis* CPUE.

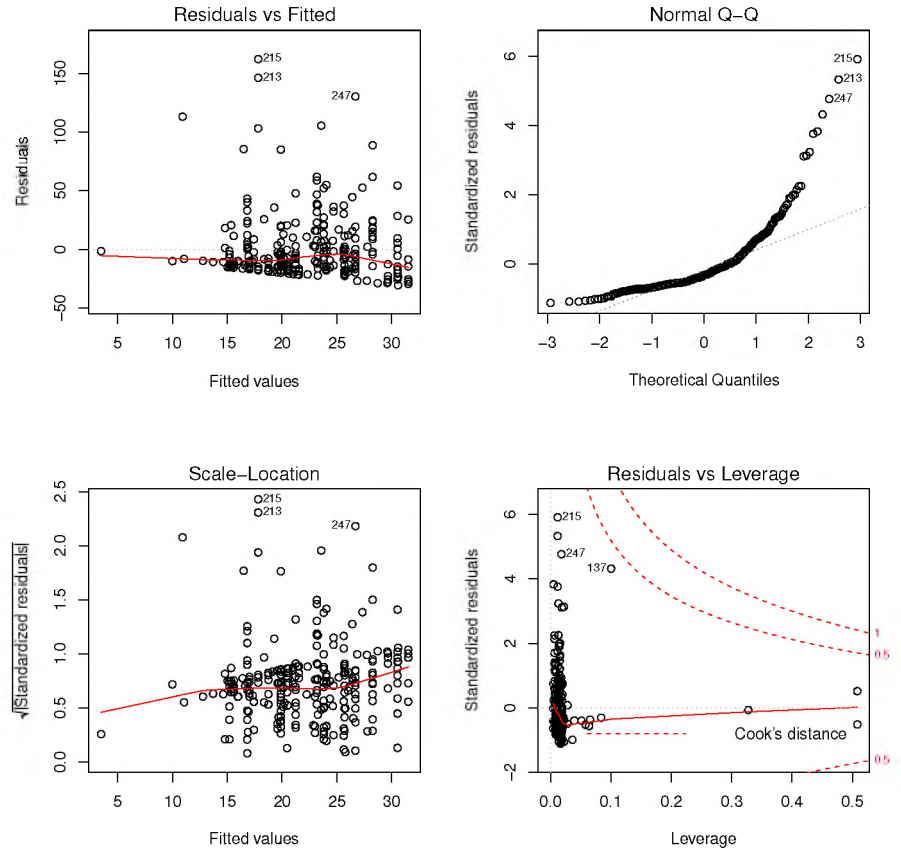


Figure 2.19: Diagnostics plots of the fitted linear model.

2.2.3 Nested ANOVA

A nested factor refers to a factor whose levels are unique within each level of the factor it is nested within and each level is only presented once (Logan, 2011, page 283). For example a fuel reduction burn study design could consists of three burnt sites and three un-burnt sites each containing four quadrats. Each site represents a unique level of a random factor that is nested within the fire treatment, burned or un-burned (Logan, 2011, page 283). Nested ANOVA is widely used in many types of life sciences research especially in the fields of psychology, genetics and ecology (Stephen, 1993, page 473). By nested we mean that

each level of the subgroups occurs in only one level of the groups (Logan, 2011, page 283). Nested factors are typically random factors; the levels are randomly selected to represent all possible levels, for example the sites in the example above (Logan, 2011, page 284). When the main treatment effect, factor A, is a fixed factor, such designs are referred to as a mixed nested ANOVA models whereas when factor A is random, the design is referred to as random nested ANOVA models. When all factors are fixed, the design is referred to as a fixed mixed model (Logan, 2011, page 284). If the higher level nominal variable is a fixed factor and the lower level nominal variable is a random variable, then we are dealing with a mixed effects nested ANOVA (Logan, 2011, page 284). The assumptions in nested ANOVA is that at the uppermost level the sampling units are independent. The two-way nested design is given by

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

where Y_{ijk} is the k^{th} outcome for subject i within level j of the random factor, which is nested within level k of the fixed factor where $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$ and $k = 1, 2, \ldots, n$. μ denotes the grand or overall mean; α_i is the effect associated with the i^{th} level of the fixed factor; $\beta_{j(i)}$ is the effect associated with the j^{th} level of the random factor within i^{th} level of the fixed factor A and ε_{ijk} is the random error (Stephen, 1993, page 474). The error terms are assumed to be normally distributed with a mean of zero, $E(\varepsilon_{ij}) = 0$ and a constant variance of σ^2. $\beta_{j(i)}$ is assumed to be normally distributed with a mean of zero and a variance of σ^2_β and the constraint $\sum_{i=1}^{a} \alpha_i$ is introduced (Radloff, 2008, page 79). For a nested ANOVA we typically use variance components methods to perform the analysis. We can sweep out the common value, the factor A effects, the factor B within A effects and the residuals using the value-splitting technique. Sum of squares can be calculated and summarized in an ANOVA as per table 2.16 (Radloff, 2008, page 81).

Source of Variation	df	Sum of Squares	Mean Square	Expected Mean Square
Factor A	$a-1$	SSA	MSA	$\sigma^2 + n\sigma^2_\beta + bn\sum \alpha_i^2/(a-1)$
Factor B	$a(b-1)$	$SSB(A)$	$MSB(A)$	$\sigma^2 + n\sigma^2_\beta$
Residuals	$ab(n-1)$	SSE	MSE	σ^2
Total	$abn-1$	SST		

Table 2.16: ANOVA table for nested designs.

Estimation of the nested ANOVA model

Partitioning the total sum of squares of the deviation of the individual score from the overall mean $(Y_{ijk} - \bar{Y}_{...})^2$. $SS_{Total} = bn\sum(\bar{Y}_i - \bar{Y})^2 + n\sum\sum\left(\bar{Y}_{j(i)} - \bar{Y}_i\right) + \sum\sum\sum\left(Y_{ijk} - \bar{Y}_{j(i)}\right)^2$ can be broken out as

$$SS_{Error} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \bar{Y}_{j(i)}\right)^2$$

with $ab(n-1)$ degrees of freedom. Consider the sum of squares corresponding to factor A, namely

$$SSA = bn\sum_{i=1}^{a}(\bar{Y}_i - \bar{Y})^2$$

with $(a-1)$ degrees of freedom. Similarly, the sum of squares corresponding to factor B, for example nested in A is

$$SSB(A) = n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{Y}_{j(i)} - Y_i\right)^2$$

with $a(b-1)$ degrees of freedom (Quinn & Keough, 2002, page 214).

We may wish to test the hypothesis that the population group means are all equal, that is $H_0\,(\text{A}) : \mu_1 = \mu_2 = \cdots = \mu_i = \mu$. If the effect of the i^{th} group is the difference between the i^{th}group mean and the overall mean $(\alpha_i = \mu_i - \mu)$ then the null hypothesis can alternatively be written as $H_0\,(\text{A}) : \alpha_1 = \alpha_2 = \cdots = \alpha_i = 0$ the effect of each group equals to zero. If one or more of the α_i are different from zero, the null hypothesis is not true indicating that the treatment does affect the response variable (Logan, 2011, page 285). Similarly we can test the hypothesis that the population group means of B within A are all equal $H_0\,(\text{B}) : \mu_{1(1)} = \mu_{2(1)} = \cdots = \mu_{j(i)} = \mu$ and the effects of each chosen B group equal to zero $H_0\,(\text{C}) : \beta_{1(1)} = \beta_{2(1)} = \cdots = \beta_{j(i)} = 0$.

If A is a random factor, then the model tests if there is no added variance due to differences between all the possible levels of A, that is $H_0\,(\text{A}) : \sigma^2_\alpha = 0$. In a random factor model the hypothesis is typically that the population variance that equals to zero, $H_0(\text{B}) : \sigma^2_\beta = 0$. Testing if there is no added variance due to all possible levels of B within the levels of A where factor B is the nested factor (Logan, 2011, page 285).

Nested ANOVA Model: *G. affinis* CPUE by Dam and Site

Consider the G. affinis data set. Consider site to be nested within dam. We may wish to test the null hypothesis that the levels of site have the same effect on the response within every given level of dam, that is,

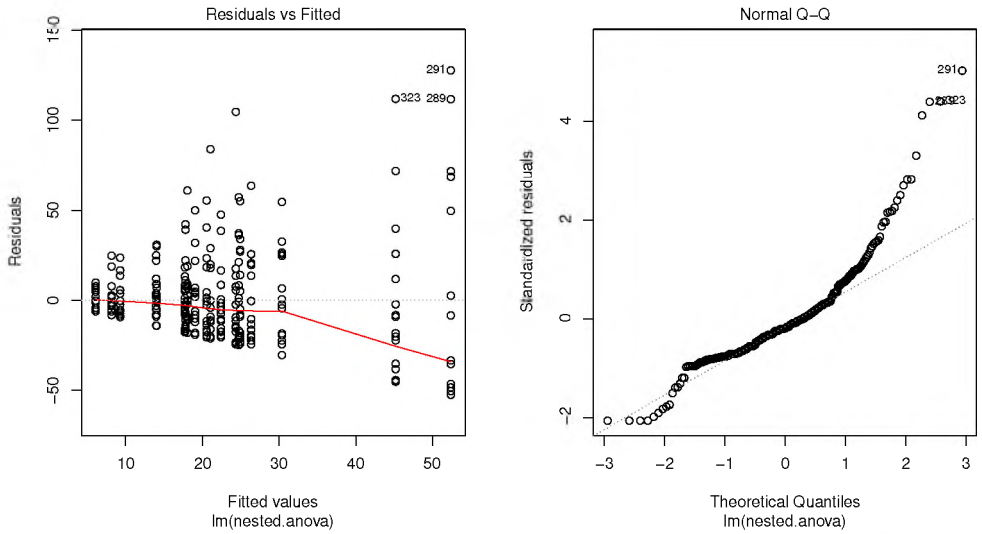
$$H_0^{\beta(\alpha)} : \left\{\beta_{1(i)} = \beta_{2(i)} = \cdots = \beta_{b(i)} = 0\right\} \text{ for every i} = 1, \ldots, a$$

against there is no effect of any of the specifically chosen levels of factor B within any level of factor A. The dam is not statistically significant ($F_{obs} = 0.67$, $df = 4, 285$, p-value $= 0.623$).

Assessing the Two-Way Nested Model Assumptions

In this context we run diagnostic tests to validate the model assumptions. The normal Q-Q plot, figure 2.20 (b), suggest that the residuals of this model do not meet the normality

assumption since the residuals deviates from the theoretical quantiles. The residuals are not normally distributed (Shapiro test, $W = 0.8763$, p-value < 0.001). The fitted residuals, figure 2.20 (a), shows that the assumption of homoscedasticity has not been met since the points are randomly scattered with a particular pattern.



(a) Residuals versus fitted values. **(b)** Normal Q-Q plot.

Figure 2.20: Diagnostics plots of the two-way nested ANOVA model.

The *G. affinis* data set was fitted with different ANOVA models that are discussed in detail in this study. After performing the ANOVA models, we have seen that these models do not fit the data hence we considered other statistical models in chapter 3.

Chapter 3

An Introduction to Generalized Linear Models

3.1 Classical Linear Models

The regression equation of the general linear model is given by $Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. This models the linear relationship between a response, or dependent, variable Y and one or more explanatory or independent or predictor variables denoted by X (Faraway, 2014, page 12). In this context Y denotes the $n \times 1$ vector of the response variables being modeled or explained by linear combinations of the independent variables. X denotes the n observations of the $p-1$ independent variables, assuming the first column of X consists of $1's$ such that the linear model includes an intercept term. $\boldsymbol{\beta}$ denotes the $p \times 1$ vector of parameters, that is $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_{p-1}]'$. $\boldsymbol{\varepsilon}$ denotes the $n \times 1$ vector of random error, or natural variation, terms.

The Classical Linear Regression Model

The classical linear regression model is based on the following assumptions (Johnson & Wichern, 2014, page 362):

1. $E(\boldsymbol{\varepsilon}) = 0$;
2. $\text{Cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}$;
3. $\mathbf{X}$ is of full column rank; and
4. $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2\mathbf{I})$.

Assumptions 1, 2 and 4 are not required to fit the model but are required for inference, namely making confidence statements and testing hypothesis about the model (Johnson &

Wichern, 2014, page 362). Under these assumptions it can be shown (Johnson & Wichern, 2014, page 370) that

1. $\boldsymbol{\mu} = E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$;
2. $\sum_Y = Var(\mathbf{Y}) = \sigma^2\mathbf{I}$;
3. $Y \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$;
4. $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \sim N_p\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$; and
5. $n\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \sim \sigma^2\chi^2_{n-k-2}$, where $\hat{\sigma}^2$ is the maximum likelihood estimator for σ^2.

Estimation and Inference

Several methods can be used for estimating $\boldsymbol{\beta}$. The least squares estimate is the best possible estimate of $\boldsymbol{\beta}$ when the errors, $\boldsymbol{\varepsilon}$, are uncorrelated and have equal variance, that is $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ (Radloff 2008, page 2; Wackerly et al. 2008, page 557).

Ordinary Least Squares

Ordinary least squares (OLS) estimation is a method of estimating $\boldsymbol{\beta}$ which provides an equation used to predict the response for given values of the predictors. Consider the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{X}$ is of full column rank. Assuming that $E(\boldsymbol{\varepsilon}) = 0$, $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$. Denote the estimator of $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ and the predicted values of $\mathbf{y}$ by $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. The vector of observed residuals, or errors, is defined as $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. OLS estimation minimizes the sum of squared errors, that is the minimization of $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$, where

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}}'\boldsymbol{\varepsilon} &= \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \\ &= \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}'\hat{\boldsymbol{\beta}}.\end{aligned}$$

Differentiating with respect to $\boldsymbol{\beta}$ yields

$$\frac{\partial\hat{\boldsymbol{\varepsilon}}'\boldsymbol{\varepsilon}}{\partial\hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}.$$

Setting this equal to zero and solving for $\hat{\boldsymbol{\beta}}$ yields

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

if $\mathbf{X}'\mathbf{X}$ is non-singular. In this context $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is called the ordinary least squares estimate of $\boldsymbol{\beta}$. This estimator is conditionally unbiased in that $E\left(\hat{\boldsymbol{\beta}}|\mathbf{X}\right) = \boldsymbol{\beta}$ (Radloff, 2008, page 3). When $\mathbf{X}$ is not of full rank, $(\mathbf{X}'\mathbf{X})^{-1}$does not exist and the generalized inverse of $\mathbf{X}'\mathbf{X}$ is used (McCulloch & Searle, 2001, page 118).

Maximum Likelihood Estimation

The maximum likelihood estimation method (MLE) finds the set of parameters that makes the observed data most likely to have occurred (Bolker, 2007, page 170). If the likelihood is monotonic, it is often easier to maximize the likelihood (Bolker, 2007, page 170). The derivation of the ordinary least square estimator, $\hat{\boldsymbol{\beta}}$, made no assumption with regards the distribution of the errors, $\boldsymbol{\varepsilon}$, and only require that the expected value of the errors was zero, that is $E(\boldsymbol{\varepsilon}) = 0$ and that $\mathbf{X}$ was of full rank. Assuming that the errors are multivariate normal with mean vector zero and variance covariance matrix $\sigma^2\mathbf{V}$, then $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V})$ and hence

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) &= \prod_{i=1}^{n} f_{Y_i}(y_i|\boldsymbol{\beta}, \sigma^2) \quad \text{(by independence)} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \mid \mathbf{V}\mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}. \end{aligned}$$

The log-likelihood, denoted by $l(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$, is given by

$$\begin{aligned} l(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) &= \ln\left\{L(\boldsymbol{\beta}, \sigma^2|\mathbf{y})\right\} \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\ln\mid\mathbf{V}\mid - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

In order to maximize $l(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$ with respect to $\boldsymbol{\beta}$ we need only minimize

$$\begin{aligned} S &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - 2\boldsymbol{\beta}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

where

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} + 2\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}.$$

Setting this equal to zero and solving for $\hat{\boldsymbol{\beta}}$ yields

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

and if $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ is of full column rank $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ is non-singular and hence

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

$\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator for $\boldsymbol{\beta}$ (Radloff, 2008, page 2). If $\sigma^2\mathbf{V} = \sigma^2\mathbf{I}$ this estimator is the same as the OLS estimator. Similarly maximizing $l(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$ with respect to

σ^2 yields

$$\frac{\partial l\left(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}\right)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)' \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)}{2\sigma^4}$$
$$\hat{\sigma}^2 = \frac{\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)' \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)}{n}.$$

$\hat{\sigma}^2$ denotes the maximum likelihood estimate of σ^2. This estimator is biased since $E\left(\hat{\sigma}^2\right) = \frac{n-k}{n}\sigma^2 \neq \sigma^2$ (McCulloch & Searle, 2001, page 30).

Estimation Bias of the Variance

The bias of an estimator refers to the difference between the expected value of the estimator and the true value, that is Bias $(\theta) = E(\theta) - \theta$ (Wackerly et al., 2008, page 393). The bias of the MLE variance estimator is given by Bias $\left(\hat{\sigma}^2\right) = E\left(\hat{\sigma}^2\right) - \hat{\sigma}^2$. Define an orthogonal projection $\mathbf{A}$, where $\mathbf{A} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$, which satisfies $\mathbf{AA} = \mathbf{A}$, that is the matrix $\mathbf{A}$ is idempotent and $\mathbf{Ay} = \mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}}$. Zhang (2015) indicates that the estimation bias in the variance components originates from the degrees of freedom lost in estimating mean components.

$$E\left(\hat{\sigma}^2\right) = \frac{1}{n} E\left[\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)' \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)\right]$$

Substituting $\mathbf{Ay} = \mathbf{X}\hat{\boldsymbol{\beta}}$ yields

$$\begin{aligned} E\left(\hat{\sigma}^2\right) &= \frac{1}{n} E\left[\left(\mathbf{y} - \mathbf{Ay}\right)' \left(\mathbf{y} - \mathbf{Ay}\right)\right] \\ &= \frac{1}{n} E\left[\mathbf{y}'(I - \mathbf{A})'(I - \mathbf{A})\mathbf{y}\right] \\ &= \frac{1}{n} E\left[\mathbf{y}'(I - \mathbf{A})\mathbf{y}\right] \\ &= \frac{1}{n}\{E(\mathbf{y}'\mathbf{y}) - E(\mathbf{y}'\mathbf{Ay})\}. \end{aligned} \tag{3.1}$$

If $\mathbf{y} \sim N_n\left(0, \mathbf{I}_{n\times n}\right)$ and $\mathbf{A}$ is orthogonal projection then $\mathbf{y}'\mathbf{Ay} \sim \chi^2_{(k)}$ with $k = \text{rank}\ (\mathbf{A})$. If $\mathbf{A}$ is of full rank, then $\mathbf{A} = \mathbf{I}_{n\times n}$ and $\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{Ay}$. Then (Zhang, 2015)

$$E(\mathbf{y}'\mathbf{y}) = n\sigma^2 + (\mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\boldsymbol{\beta})$$

and

$$E(\mathbf{y}'\mathbf{Ay}) = k\sigma^2 + (\mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\boldsymbol{\beta}).$$

Substituting into equation 3.1 above yields

$$\begin{aligned} E(\hat{\sigma}^2) &= \frac{1}{n}\left[\left(n\sigma^2 + (\mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\boldsymbol{\beta})\right) - \left(k\sigma^2 + (\mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\boldsymbol{\beta})\right)\right] \\ &= \frac{1}{n}(n\sigma^2 - k\sigma^2) \\ &= \frac{n-k}{n}\sigma^2 < \sigma^2. \end{aligned}$$

Restricted Maximum Likelihood

Maximum likelihood (ML) and restricted maximum likelihood (REML) estimators are found by maximizing a function of the parameters within the bounds of the parameter space (McCulloch & Searle, 2001, page 21 and 22). REML uses the ML approach applied to linear functions of $\mathbf{y}$, say $\mathbf{A}'\mathbf{y}$, for which $\mathbf{A}'$ is specifically designed such that $\mathbf{A}'\mathbf{y}$ does not contain the fixed effects components of the model (McCulloch & Searle, 2001, page 21). As a result REML allows the variance components to be estimated without being affected by the fixed effects and hence the variance estimates are invariant, or not affected by, the values of the fixed effects (McCulloch & Searle, 2001, page 21). When estimating the variance components using REML the degrees of freedom for the fixed effects are implicitly accounted for, unlike ML. Consider estimating σ^2 for univariate normally distributed data: $y_i \sim N_1(\mu, \sigma^2)$ where $i = 1, \ldots, n$. The REML estimate of σ^2 is $\frac{SS_{yy}}{n-1}$ whereas the ML estimate of σ^2 is $\frac{SS_{yy}}{n}$ (McCulloch & Searle, 2001, page 21). REML estimates are not affected by the process used in estimating the fixed effects since REML methods are designed to be free of the fixed effects portion of a model.

The marginal distribution for $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ equals to the block-design matrix with blocks $\mathbf{V}_i$ on the main diagonal and zeros elsewhere (Zuur et al., 2009, page 119). The REML estimator for the variance components, σ^2, is obtained by maximizing the likelihood function of a set of error contrast $\mathbf{Z} = \mathbf{A}'\mathbf{y}$ where $\mathbf{A}$ is any $(n \times (n-p))$ matrix of full rank with columns orthogonal to the columns of the $\mathbf{X}$ matrix. The vector $\mathbf{Z}$ follows a normal distribution with mean vector zero and covariance matrix $\mathbf{A}'\mathbf{V}\mathbf{A}$, which is not independent of $\boldsymbol{\beta}$.

$$\begin{aligned} f_{\mathbf{z}}(\mathbf{Z}) &= \frac{1}{(2\pi)^{(n-1)/2}|\mathbf{A}'\sigma^2\mathbf{A}|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{A}'\mathbf{y} - \mathbf{A}'\boldsymbol{\mu}\right)'\left(\sigma^2\mathbf{A}'\mathbf{A}\right)^{-1}\left(\mathbf{A}'\mathbf{y} - \mathbf{A}'\boldsymbol{\mu}\right)\right\} \\ &= \frac{1}{(2\pi)^{(n-1)/2}\left((\sigma^2)^{n-1}\right)^{1/2}|\mathbf{A}'\mathbf{A}|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{A}''\left(\mathbf{y} - \boldsymbol{\mu}\right)'\left(\mathbf{A}'\mathbf{A}\right)^{-1}\left(\mathbf{y} - \boldsymbol{\mu}\right)\right\} \\ &= \frac{\exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{y} - \boldsymbol{\mu}\right)'\mathbf{A}\left(\mathbf{A}'\mathbf{A}\right)^{-1}\mathbf{A}'\left(\mathbf{y} - \boldsymbol{\mu}\right)\right\}}{\left(2\pi\sigma^2\right)^{(n-1)/2}|\mathbf{A}'\mathbf{A}|^{1/2}}. \end{aligned}$$

The idea with REML estimators is to find all independent linear combinations of the response, $\mathbf{y}$, such that $\mathbf{A}'\mathbf{X} = \mathbf{0}$. Construct matrix $\mathbf{A}$ with column a, so that $\mathbf{Z} = \mathbf{A}'\mathbf{y} \sim$

$N_{n-1}(0, \mathbf{A}'\mathbf{V}\mathbf{A})$ where

$$f_{\mathbf{z}}(\mathbf{Z}) = \frac{\exp\left\{-\frac{1}{2\sigma^2}\mathbf{y}'\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{y}\right\}}{(2\pi\sigma^2)^{\frac{(n-1)}{2}} \mid \mathbf{A}'\mathbf{A} \mid^{\frac{1}{2}}}.$$

The log likelihood is given by

$$\ell = \ln\{f_{\mathbf{z}}(\mathbf{Z})\} = -\frac{1}{2\sigma^2}\mathbf{y}'\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{y} - \frac{n-1}{2}\ln\sigma^2 - \frac{1}{2}\ln|\mathbf{A}'\mathbf{A}|.$$

We can then proceed to maximize the likelihood based on $\mathbf{A}'\mathbf{y}$ which does not involve any of the fixed effects parameters. But $\mathbf{y}'\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{y} = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)' \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$ for any $\mathbf{A}$ as long it satisfies $\mathbf{A}'\mathbf{X} = \mathbf{0}$. Differentiating the log likelihood

$$\frac{\partial l}{\partial \sigma^2} = \frac{\mathbf{y}'\mathbf{A}(\mathbf{A}'\mathbf{A})\mathbf{A}'\mathbf{y}}{2\sigma^4} - \frac{n-1}{2\sigma^2}$$

and subsequently setting $\frac{\partial l}{\partial \sigma^2} = 0$ yields

$$\begin{aligned}
&\frac{\mathbf{y}'\mathbf{A}(\mathbf{A}'\mathbf{A})\mathbf{A}'\mathbf{y}}{2\sigma^4} - \frac{n-1}{2\sigma^2} = 0 \\
&\mathbf{y}'\mathbf{A}(\mathbf{A}'\mathbf{A})\mathbf{A}'\mathbf{y} = (n-1)\sigma^2 \\
&\sigma^2 = \frac{\mathbf{y}'\mathbf{A}(\mathbf{A}'\mathbf{A})\mathbf{A}'\mathbf{y}}{n-1} \\
&\therefore \hat{\sigma}^2_{\text{unbiased}} = \frac{\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)' \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)}{n-1}
\end{aligned}$$

(Zhang, 2015). REML generally produces less biased estimates of variances in mixed models (McCulloch et al., 2008, page 149). Pinheiro & Bates (2000, page 8) identified ML and REML as the most common estimation methods for estimating parameters in linear mixed models.

3.2 Dummy Variables

A dummy or indicator variable is any variable in a regression model that takes on a finite number of values so that different categories of a nominal variable can be identified (Kleinbaum et al., 2013, page 257). Categorical variables can be dichotomous or polytomous. If the nominal independent variable of interest has k categories, then one must define exactly $k-1$ dummy variables to index these categories, provided that the regression model contains a constant term, that is an intercept β_0 (Kleinbaum et al., 2013, page 257). If the regression model does not contain an intercept then k dummy variables are needed to index the k categories of interest (Kleinbaum et al., 2013, page 257). The number of dummy variables necessary to represent a single attribute is equal to the number of levels or categories in that

variable minus one if the model includes an intercept. For example if there are $k = 4$ categories, the number of dummy variables should be $k - 1 = 4 - 1 = 3$ for a model containing an intercept. For any given attribute variable, none of the dummy variable constructed should be dismissed, that is one dummy variable can not be a constant multiple or a simple linear relation of another (Skrivanek, 2009). Regression models containing dummy variables are easily estimated by the familiar convention of "dropping out" one of the categories. When we include dummy variables in the regression equation, the logic of regression estimation remains the same. The coding of data with categorical variables requires the development of mutually exclusive and exhaustive categories. This rule applies to the creation of dummy variables. Dummy variables assign the binary numbers $'0'$ and $'1'$ to indicate membership in any mutually exclusive and exhaustive category (Skrivanek, 2009). Regression analysis treats all independent variables in the analysis as numerical (Skrivanek, 2009). Using binary $(0, 1)$ coding, dummy variables are always dichotomous variables (Hardy, 1993, page 19).

Suppose that we are predicting a response variable, Y for example *G. affinis* CPUE, as a linear function of a quantitative variable X_1, for example pressure, and water body where water body is a categorical variable with four levels, namely Yarrow, Settlers, White's and Mangazana dams. Define dummy or indicator variables, X_i, as follows:

$$X_2 = \begin{cases} 1 & \text{if the dam is Settlers,} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if the dam is Yarrow} \\ 0 & \text{otherwise} \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{if the dam is White's dam} \\ 0 & \text{otherwise} \end{cases}$$

and

$$X_5 = \begin{cases} 1 & \text{if the dam is Mangazana} \\ 0 & \text{otherwise} \end{cases}$$

The dummy variables X_2, X_3, X_4 and X_5 represents the binary independent variable 'water body'. X_2 takes two values, $'1'$ if the dam is Settlers and $'0'$ if not Settlers. A single dummy variable is needed to represent a variable with two levels. There are three other dummy variables, namely X_3 =Yarrow, X_4 = White's and X_5 =Mangazana dams. However the fourth dummy variable is not needed to represent Mangazana. Setting Yarrow, Settler, White's dam to $'0'$ indicates that the dam is Mangazana. This coding works because there are four levels which are mutually exclusive and exhaustive (Skrivanek, 2009). The linear model can be defined as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon.$$

Since the linear predictor can accommodate quantitative and qualitative predictors with the use of dummy variables and also allows for transformations and combinations of the original predictors (Kleinbaum et al., 2013, page 259).

3.3 The Regression Model Approach to ANOVA

Consider the dummy-regression model

$$Y_i = \alpha + \beta x_i + \gamma d_i + \delta\left(x_i d_i\right) + \varepsilon_i$$

where Y_i denotes income, x_i denotes the years of education, d_i denotes the dummy variable sex, coded as 1 for male and 0 for female for the i^{th} observation and $x_i d_i$ denotes the interaction regressor (Fox, 2015, page 187). This model can be expressed in matrix form as

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{n_1} \\ - - - \\ Y_{n_1+1} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_1} & 0 & 0 \\ - & - & - & - \\ 1 & x_{n_1+1} & 1 & x_{n_1+1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & 1 & x_n \end{bmatrix} \times \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{n_1} \\ - - - \\ \varepsilon_{n_1+1} \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

that is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, for the n_1 female participants, for when d and xd are zero and the remainder of the observation, $n - n_1$, are male. In this context the matrix $\mathbf{X}$ is called the design or model or structure matrix (Fox 2015, pages 189; Kirk 1982, page 177).

In this example $\mathbf{X}$ is not of full column rank, since the first column is equal to the sum of over the other columns. If a column were to be deleted $\mathbf{X}$ would be of full column rank and the associated parameter can be set to zero. For example deleting the last column and setting $\alpha_k = 0$ establishes the last category as the base-line for a dummy-coding scheme (Fox, 2015, page 189). Other solutions to this issue include using different coding schemes (Kirk, 1982, page 187 and 199), using a less than full rank experimental design (Kirk, 1982, page 211) or the full rank experimental design (Kirk, 1982, page 225). Model matrices for dummy regression and ANOVA models are strongly patterned. Consider the one-way fixed effects ANOVA model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

for observations $i = 1, \ldots, n$ and groups $j = 1, \ldots, k$. This model can be written in matrix

form as:

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1,1} \\ --- \\ Y_{12} \\ \vdots \\ Y_{n_2,2} \\ --- \\ \vdots \\ --- \\ Y_{1m} \\ \vdots \\ Y_{n_m,m} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & 1 & 0 & \dots & 0 & 0 \\ - & - & - & - & - & - \\ 1 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & 0 & 1 & & 0 & 0 \\ - & - & - & - & - & - \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ - & - & - & - & - & - \\ 1 & 0 & 0 & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{k-1} \\ \alpha_k \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{n_1,1} \\ --- \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{n_2,2} \\ --- \\ \vdots \\ --- \\ \varepsilon_{1m} \\ \vdots \\ \varepsilon_{n_m m} \end{bmatrix}$$

3.4 The General Linear Model

Consider the linear model given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ denotes the vector of the response variable, $\mathbf{X}$ denotes the design matrix, $\boldsymbol{\beta}$ the vector of parameters and $\boldsymbol{\varepsilon}$ the vector of errors. As per the classical general linear model in section 3.1 it is assumed that $\mathbf{Y}$ are observed values of $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the errors are independent, identically distributed random variables with zero mean and constant variance and to perform inference on the model, these errors are normally distributed. This model can be extended to include cases where components of $\mathbf{X}$ are considered to be random variables, typically assumed to be independent of $\boldsymbol{\varepsilon}$ (McCullagh & Nelder, 1989, page 9). Three types of these linear models are typically considered, namely the fixed effects, random effects and the mixed-effects models (Pinheiro & Bates, 2000, page 3).

Fixed Effects Models

An effect is called fixed if the levels in the study represent all possible levels of the factor, or at least all levels about which inference is to be made (Littell et al., 2007, page 4). This includes regression models where the observed values of the explanatory or independent variable cover the entire region of interest. For example in a blood pressure drug experiment the effects of the drugs are fixed if the five specific drugs are the only candidates for use and if conclusions about the experiment are restricted to those five drugs (Littell et al., 2007, page 4). Factors

can be fitted as fixed effects, but can still be conceptually random in the sense that they represent a random sample of levels rather than distinct treatments (Schielzeth & Nakagawa, 2013). An important feature of fixed effects is that they are deemed to be constants or non-random parameters representing the effects on the response variable, $\mathbf{Y}$, of the various levels of factors under consideration. Consider for a example a two-way fixed effects ANOVA model (Xu et al., 2013) given by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk}.$$

Under the fixed effects model the quantities α_i, β_j, and γ_k are assumed to be non-random parameters. The objective in a fixed effects model is to make inferences about the unknown parameters (Sahai & Ageel, 2012, page 5).

Random Effects Models

A random effect is a predictor variable where we are interested in making inferences about the distribution of values, that is, the variances among the values of the response at different levels rather than in testing the differences of values between particular levels (Schielzeth & Nakagawa, 2013). A factor is considered random if its levels plausibly represent a larger population with a probability distribution (Littell et al., 2007, page 5). Random effects terms have been used in models to represent omitted explanatory variables and random measurement error in the explanatory variables (Agresti et al., 2000). Schielzeth & Nakagawa (2013) defined random effects as the effects estimated at each factor level, but where the distribution of the estimates is explicitly modeled by hyperparameters. The variance of the random effects could be considered the unexplained variance at the level in the sense that the detailed causes of such random effect variance are unknown. Under the random effects model the quantities α_i, β_j and γ_k are assumed to be random variables with zero mean and variances of σ^2_α, $\sigma^2_\beta, \sigma^2_\gamma$ and σ^2_ε (Sahai & Ageel, 2012, page 5).

Random effects are often used in modeling the random variation in the dependent variable at different levels of the data (Schielzeth & Nakagawa, 2013). The objective under the random effects model is to make inferences about the variances and certain functions of them. Random effects are often used for controlling for correlated structure in the data, that is the dependencies between data (Schielzeth & Nakagawa, 2013). Random effects are especially useful when we have lots of levels, for example many species, relatively little data on each species and uneven sampling across species (Schielzeth & Nakagawa, 2013).

Consider the random model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where μ denotes the grand or overall mean in the population and τ denotes the random effects. When the effect is random we typically assume that the distribution of the random effects has

mean zero and variance σ_τ^2. The variance of Y_{ij} is given by $Var\left(Y_{ij}\right) = Var\left(\mu + \tau_i + \varepsilon_{ij}\right) = \sigma_\tau^2 + \sigma^2$ (Littell et al., 2007, page 5).

Mixed-Effects Models

A mixed model contains both fixed and random effects. Consider the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where μ denotes the overall or grand mean, α_i denotes the unknown fixed parameters, β_j denotes the random effects associated with the j^{th} levels and ε_{ij} denotes the random errors, respectively. We typically assume that the random effects, β_j, have zero mean and variance σ_β^2. The random errors, ε_{ij}, are assumed to have zero mean and variance σ^2 (Littell et al., 2007, page 6). Modeling the variance structure is the most powerful and significant feature of mixed models and that is what sets it apart from conventional linear models (Littell et al., 2007, page 6). The role played by the fixed effects parameters is to capture the influence of the explanatory variables on the mean structure of mixed models, exactly as in the standard linear model (Vangeneugden et al., 2004). Mixed model methods primarily use three approaches to variance component estimation, namely (Maunder & Punt, 2004)

1. The procedure is based on expected mean squares from the ANOVA;
2. The maximum likelihood (ML); and
3. The restricted maximum likelihood (REML).

Of these methods the ML is usually discouraged because the variance component estimates are biased and the REML procedure is the most versatile but there are situation for which the ANOVA procedures are preferable (Littell et al., 2007, page 7).

Mixed-effects models are primarily used to describe relationships between a response variable and some covariates in data that are grouped according to one or more classification factors (Pinheiro & Bates, 2000, page 3). Grouped data include longitudinal data, repeated measures data and block designs. In this context repeated measures data means data generated by observing a number of individuals repeatedly under different experimental conditions where the individuals are assumed to constitute a random sample from a population of interest (Pinheiro & Bates, 2000, page 3). A common use of mixed models is in the analysis of longitudinal data, which are defined as data collected on each subject on two or more occasions (Schielzeth & Nakagawa, 2013). Mixed-effects models provide a powerful and flexible tool for analyzing clustered data such as repeated measures data and nested data and is becoming tremendously popular as a framework for the analysis of bio-behavioral data (Pinheiro & Bates, 2000, page 133). It includes both fixed effect parameters associated with one or

more categorical covariates and random effects associated with one or more random factors (Pinheiro & Bates, 2000, page 58). Schielzeth & Nakagawa (2013) stress the importance of consistency that is how the levels of the fixed factors are related to the levels of the random factors as their relationship could be nested or crossed.

The objective in a mixed model is to make inferences about the fixed effect parameters and variances of the random effects (Sahai & Ageel, 2012, page 5). Inference in traditional linear models is based largely on least squares estimation for fixed effects and in analysis of variance sums of squares for estimating variances. When random effects are part of a model we want to estimate variances of the part of the specification of the random effects (Vangeneugden et al., 2004). Consider the linear mixed model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i$$

where $\mathbf{Y}_i$ is the n_i dimensional response vector for subjects i, $1 \leq i \leq n$ where n denotes the number of subjects. $\mathbf{X}_i$ and $\mathbf{Z}_i$ are $(n_i \times p)$ and $(n_i \times q)$ known design matrices corresponding to the fixed and random effects respectively. $\boldsymbol{\beta}$ is the p- dimensional vector containing the fixed effects. $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$ is the q- dimensional vector containing the random effects and $\mathbf{D}$ is the covariance matrix. $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I}_{n_i})$ is an n_i- dimensional vector of measurement error components where $b_1, \ldots, b_N$; $\varepsilon_1, \ldots, \varepsilon_N$ are assumed to be independent (Vangeneugden et al., 2004; Morrell et al., 1997). Since $\mathbf{b}$ and $\boldsymbol{\varepsilon}$ have zero mean vectors, the mean of the data vector is $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and its variance covariance matrix is

$$Var(\mathbf{y}) = E(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'.$$

The variables are modeled as random effects if the primary interest lies in estimating variances, while fixed factors are used for estimating the mean effect of a treatment (Schielzeth & Nakagawa, 2013). Random effects are usually used for controlling correlated structure in the data, that is, dependencies between data and are not estimated independently, whereas fixed effect levels of the same predictor are estimated independently of each other. If the random effect variance is low there is little potential for strong group level fixed effects, although they might still become significant with sufficient data. In the sections that follow estimation methods for linear mixed-effects (LME's) models, based on the likelihood or the restricted likelihood of the parameters are described together with the computational methods used to implement them as per the 'lme' function in R (Zuur et al., 2009, page 107).

Consider two effects, A and B, where A is a fixed, B is random and there is a possible interaction (A $\times$ B) between them. For a given dependent variable, the null hypothesis concerning A is that there is no difference in means among the levels of A in the experiment. For B the null hypothesis is that there is no variability among all possible levels of B, not that there are no differences among levels of that effect included in the experiment. For the interaction term (A $\times$ B) the null hypothesis is that variability among the levels of B is the

same for all levels of A. This differs from the case for fixed effects in that the null hypothesis for an interaction between two fixed effects (A and C) is that the response of the dependent variable is not different among specific levels of A depending upon the particular level of C.

3.5 Generalized Linear Models

A generalized linear model (GLM) consists of three key components, namely (Fox 2015, page 379; McCulloch & Searle 2001, page 136; Faraway 2016, page 113 and 114);

1. A random component which specifies the distribution of the response, Y_i;
2. A linear predictor, that is a linear function of the regressors

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

   and

3. A smooth and invertible link function $g(\cdot)$. The link function describes how the mean of the response is related to the linear predictors (Logan, 2011, page 483 and 484).

Typically the response variable, Y_i, is assumed to consist of n independent measurements from a distribution with a density function from the exponential family, or similar to the exponential family (McCulloch & Searle, 2001, page 137), that is $Y_i \sim$ independent $f_{Y_i}(y_i)$ where $f_{Y_i}(y_i)$, when expressed in canonical form, is of the form (Fox, 2015, page 402)

$$f_{Y_i}(y_i) = \exp\left\{\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i, \phi)\right\}.$$

In this context:

- $f_{Y_i}(y_i, \theta, \phi)$ is the probability function, for a discrete random variable Y_i, or the probability density function for a continuous random variable Y_i;
- $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions that vary from one exponential family to another;
- $\theta = g(\mu_i)$ denotes the canonical parameter for the particular exponential family; and
- ϕ denotes the dispersion parameter and represents the scale.

$a(\cdot)$ and $b(\cdot)$ are specific functions that distinguish one number of the exponential family from the others (Faraway, 2016, page 115). Distributions in the exponential family include the Poisson, binomial and normal (Maunder & Punt, 2004).

Consider $Y_i \sim N_1(\mu, \sigma^2)$, then

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2\right\} \forall\, y.$$

(Fox, 2015, page 402). This distribution can be re-written in the form

$$f_{Y_i}(y_i; \theta, \phi) = \exp\left\{\frac{y_i\theta - \theta^2/2}{\phi} - \frac{1}{2}\left[\frac{y_i^2}{\phi} + \ln(2\pi\phi)\right]\right\}$$

where $\theta = g(\mu)$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$ and $c(y, \phi) = -\frac{1}{2}\left[\frac{y_i^2}{\phi} + \ln(2\pi\phi)\right]$. That is the normal distribution is a member of the exponential family with $a(\cdot)$ and $b(\cdot)$ as specified above.

Family	$a(\phi)$	$b(\theta)$	$c(y, \phi)$
Gaussian	ϕ	$\theta^2/2$	$-\frac{1}{2}\left[y^2/\phi + \log_e(2\pi\phi)\right]$
Binomial	$1/n$	$\log\left(1 + e^{\theta}\right)$	$\log_e\binom{n}{n_y}$
Poisson	1	e^{θ}	$-\log_e y!$
Gamma	ϕ	$-\log_e(-\theta)$	$\phi^{-2}\log_e(y/\phi) - \log_e y - \log_e\Gamma\left(\phi^{-1}\right)$
Inverse-Gaussian	ϕ	$-\sqrt{-2\theta}$	$-\frac{1}{2}\left[\log_e\left(\pi\phi y^3\right) + 1/(\phi y)\right]$

Table 3.1: Characteristics of common univariate distributions in the exponential family.

Maunder & Punt (2004) define GLM as the statistical distribution for the response variable where some linear combination of a set of explanatory variables relate to the expected value of the response variable. GLM's are regarded as the most powerful statistical technique that includes Gaussian linear models including ANOVA, regression, log-linear models for frequency data, logistic regression models and several other models (Maunder & Punt, 2004). When applying GLM's the researcher must choose an appropriate sampling distribution for the response variable from the exponential family, choose an appropriate link function for the distribution and select a set of explanatory variables (Maunder & Punt, 2004).

A key property of the GLM is the linear relationship between some function, denoted as $g(\cdot)$, of the expected value of the response variable, $E(Y_i) = \mu_i$, and the explanatory variables, that is

$$g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$$

where $g(\cdot)$ is a differentiable link function, $\mu_i = E(Y_i)$, $\mathbf{x}_i$ is the vector of size p that specifies the explanatory variables for the i^{th} value of response variable, $\boldsymbol{\beta}$ is a vector of the parameter and Y_i is the i^{th} response (Maunder & Punt, 2004).

GLM Example

Researchers investigated the biogeographical determinants of ant species richness at a regional scale. They used an excerpt of data to contrast inferential and Bayesian approaches (Logan, 2011, page 510). Specifically, ant species richness was modeled against latitude, elevation and habitat type, namely bog or forest, using a Poisson GLM. A Poisson GLM was utilized since the response variable, species richness, was a count. The resulting fitted GLM model is shown in table 3.2. The results revealed that latitude ($\hat{\beta}_2 = -0.2358$, $Z_{obs} = -3.824$, p-value $= 0.0001$) and elevation ($\hat{\beta}_3 = -0.0011$, $Z_{obs} = -3.044$, p-value $= 0.0023$) are declining significantly. While the forest habitat is positively significant ($\hat{\beta}_1 = 0.6354$, $Z_{obs} = 5.315$, p-value < 0.001). These results provide sufficient evidence that the species richness of ants is explained by habitat, latitude and elevation at a regional scale. The normal Q-Q plot, figure 3.1, shows that the residuals of this model meet the normality assumption since the points are approximately linear. The points in the residual plot, figure 3.1, are randomly scattered with no particular pattern, this suggest that the assumption of homoscedasticty has been met. The residuals are normally distributed (Shapiro test, $W = 0.9588$, p-value $= 0.117$), see appendix B. Observation 25 in figure 3.1 is an extreme outlier as it affects both results and assumptions. It is not legitimate to simply drop the outlier as this would affect the whole analyses hence we should try a different model, that is mixed effects model.

Coefficients	Parameter estimate	Standard error	z-value	p-value
Intercept	11.9368	2.6215	4.553	$5.28e-06$
Habitat (forest)	0.6354	0.1196	5.315	$1.07e-07$
Latitude	-0.2358	0.0617	-3.824	0.0001
Elevation	-0.0011	0.0004	-3.044	0.0023

Table 3.2: GLM results: gotelli model.

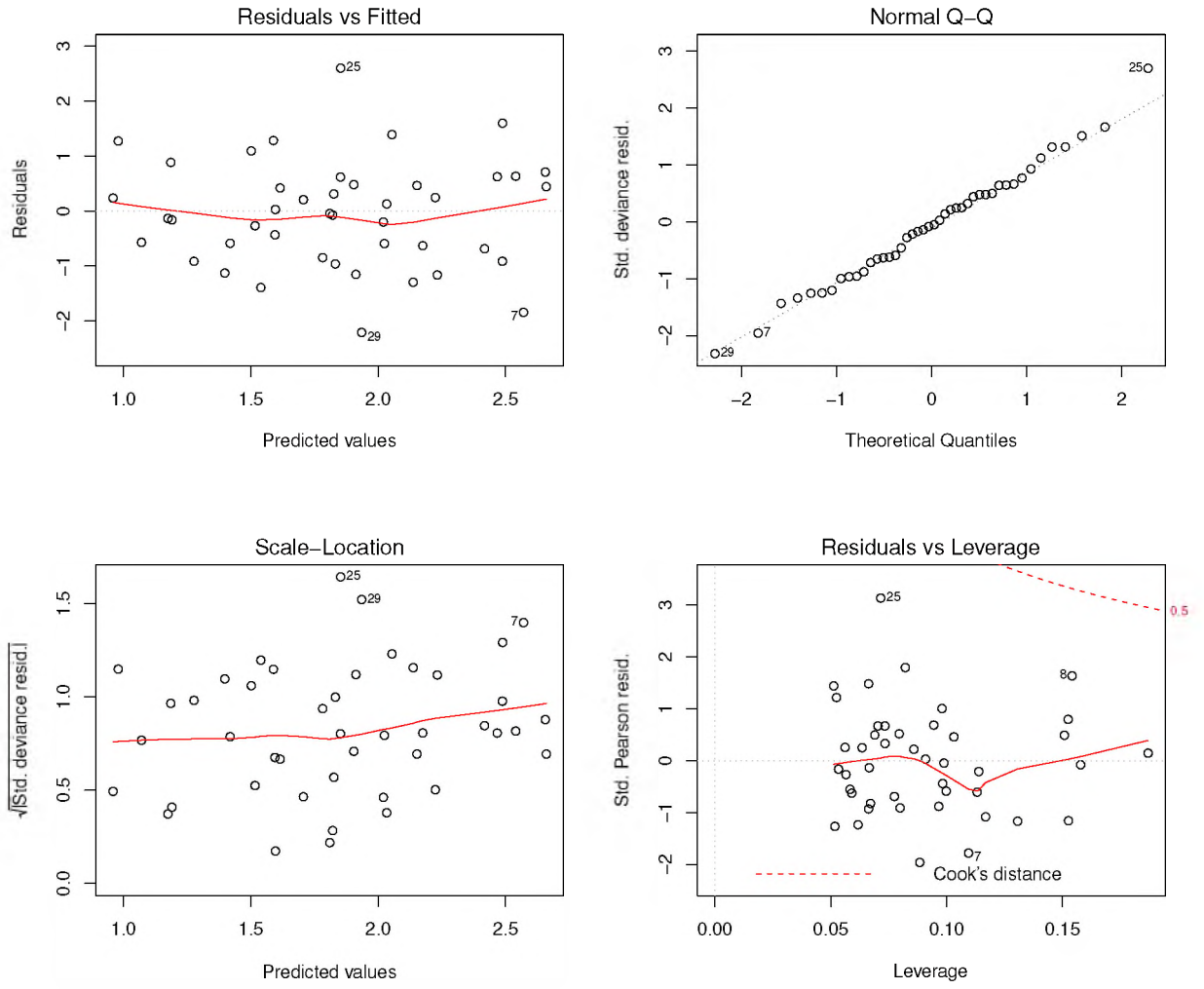


Figure 3.1: Diagnostics plots of the fitted GLM for the gotelli data.

GLM Example: *G. affinis* CPUE

The relative importance of biotic and abiotic factors effects on *G. affinis* CPUE was modeled using a GLM to assess the effects of dam age, mean temperature, *O. mossambicus* abundance, *G. callidus* abundance and percentage vegetation cover on the relative abundance of *G. affinis* CPUE. These factors represents biologically important potential drivers of *G. affinis* population (Howell et al., 2013). A generalized linear model with a Poisson distribution is appropriate for the *G. affinis* CPUE since the response variable is a count. Results from the GLM revealed that dam age ($\hat{\beta}_1 = 0.0165$, $Z_{obs} = 13.912$, p-value < 0.001), mean temperature ($\hat{\beta}_2 = 0.0195$, $Z_{obs} = 7.075$, p-value < 0.001), *O. mossambicus* abundance ($\hat{\beta}_4 = -0.0552$, $Z_{obs} = -6.195$, p-value < 0.001) and *G. callidus* abundance ($\hat{\beta}_5 = -0.4996$, $Z_{obs} = -2.769$, p-value $= 0.0056$) had significant effect on the *G. affinis* CPUE. However percentage vegetation cover ($\hat{\beta}_3 = -0.0794$, $Z_{obs} = -1.764$, $p - value = 0.0777$) had no significant effect on the *G. affinis* CPUE as shown in table 3.3. The normal Q-Q

plot, figure 3.2, shows that the residuals of this model do not meet the normality assumption since the points are not approximately linear. The residuals of this model are not normally distributed (Shapiro test, $W = 0.6437$, p-value < 0.001). The points in the residual plot, figure 3.2, are randomly scattered with a particular pattern, this suggest that the assumption of homoscedasticty has not been met. There is heterogeneity of variances across the dams (Bartlett's K-squared $= 36.026$, $df = 4$, p-value < 0.001), see appendix B.

Coefficients	Parameter estimate	Standard error	z-value	p-value
Intercept	2.5186	0.0612	41.170	$< 2e-16$
Dam age	0.0165	0.0012	13.912	$< 2e-16$
Mean temperature	0.0195	0.0028	7.075	$1.49e-12$
Percentage vegetation cover	−0.0794	0.0450	−1.764	0.0777
O. mossambicus abundance	−0.0552	0.0089	−6.195	$5.82e-10$
G. callidus abundance	−0.4996	0.1804	−2.769	0.0056

Table 3.3: GLM: *G. affinis* CPUE by biotic and abiotic factors.

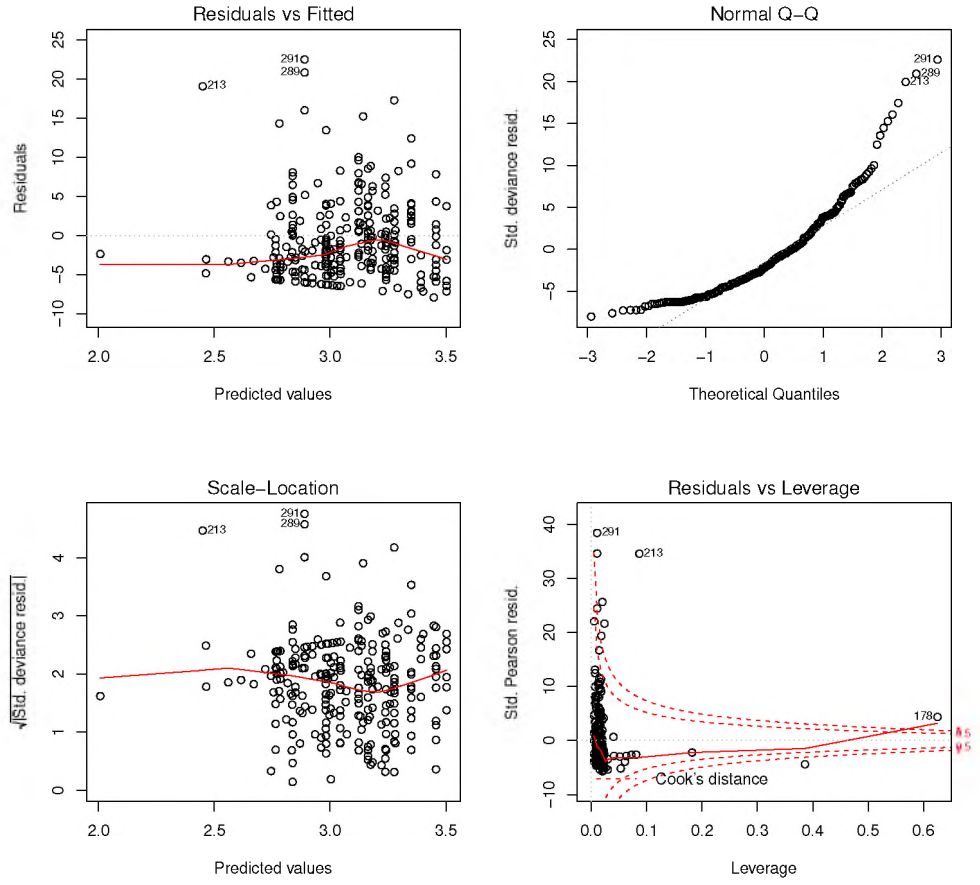


Figure 3.2: Diagnostics plots of the fitted GLM for *G. affinis* CPUE.

GLM Example: Temperature and Pressure Effects on the Tournament CPUE

Temperature and pressure are geographical factors that had an influence on the CPUE measured by bag weight (Hargrove et al., 2015). The bag weight was modeled against minimum temperature, maximum temperature and pressure using a Gaussian GLM with identity link function since the response or dependent variable, bag weight, is continuous. The fitted GLM model is show in table 3.4. The results revealed that the pressure for bag weights are significant ($\hat{\beta}_3 = -0.1412$, $T_{obs} = -2.884$, p-value $= 0.0045$). The minimum ($\hat{\beta}_1 = -0.0813$, $T_{obs} = -1.311$, p-value $= 0.1916$) and maximum ($\hat{\beta}_2 = 0.0073$, $T_{obs} = 0.215$, p-value $= 0.8298$) temperatures are both insignificant as shown in table 3.4. The minimum and maximum temperatures for tournament events are shown in figure 3.5. The Autumn season months, 20 March 2016 and 17 May 2015 had the highest mean pressures between 948 to 952 is shown in figure 3.4. The normal Q-Q plot, figure 3.3, shows that the residuals of this model do not meet the normality assumption since the points are not approximately linear. The residuals are not normally distributed (Shapiro test, $W = 0.9685$, p-value $= 0.0006$). The fitted residuals, figure 3.3, shows that the residuals of this model do not meet the homogeneity of variance assumption since the points are randomly scattered with a particular pattern. There is heterogeneity of variances in the minimum temperature (Bartlett's K-squared $= 21.826$, $df = 10$, p-value $= 0.0160$), maximum temperature (Bartlett's K-squared $= 21.14$, $df = 11$, p-value $= 0.0311$) and pressure (Bartlett's K-squared $= 21.14$, $df = 11$, p-value $= 0.0311$).

Coefficients	Parameter estimate	Standard error	t-value	p-value
(Intercept)	136.1082	46.6142	2.920	0.0031
Minimum temperature	−0.0813	0.0620	−1.311	0.1916
Maximum temperature	0.0073	0.0338	0.215	0.8298
Pressure	−0.1412	0.0481	−2.884	0.0045

Table 3.4: GLM results: Bag weight against minimum, maximum temperature and pressure.

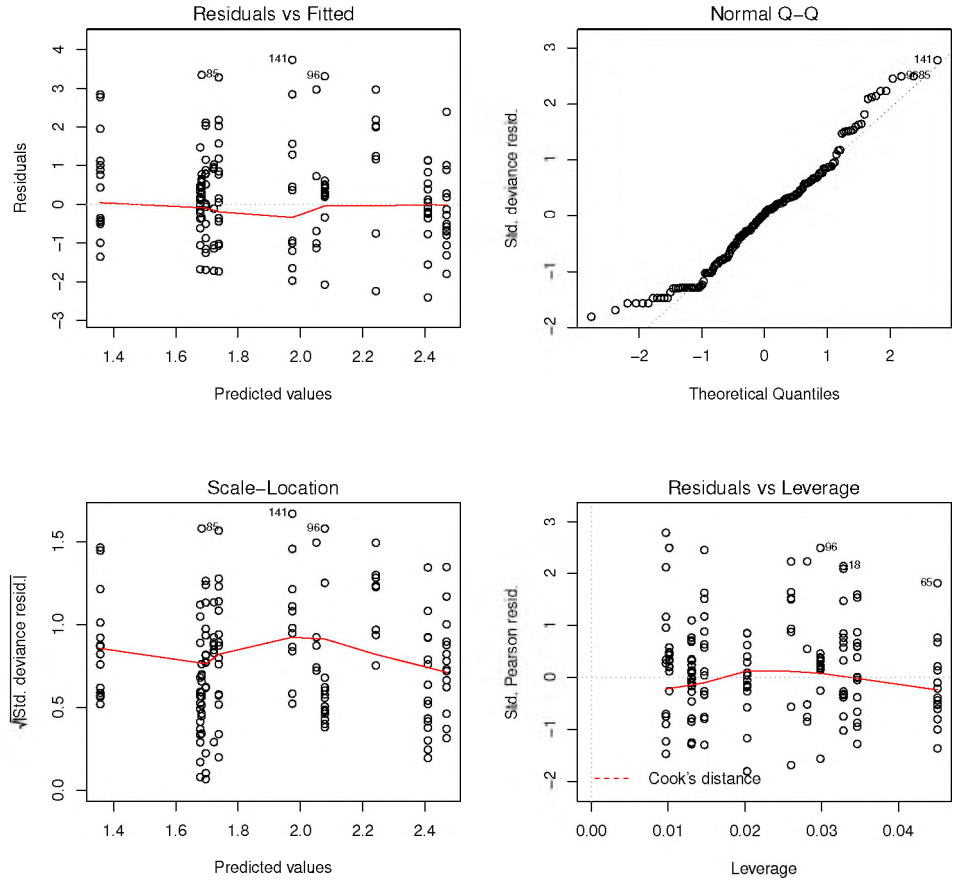


Figure 3.3: Diagnostics plots of the fitted GLM for bag weights.

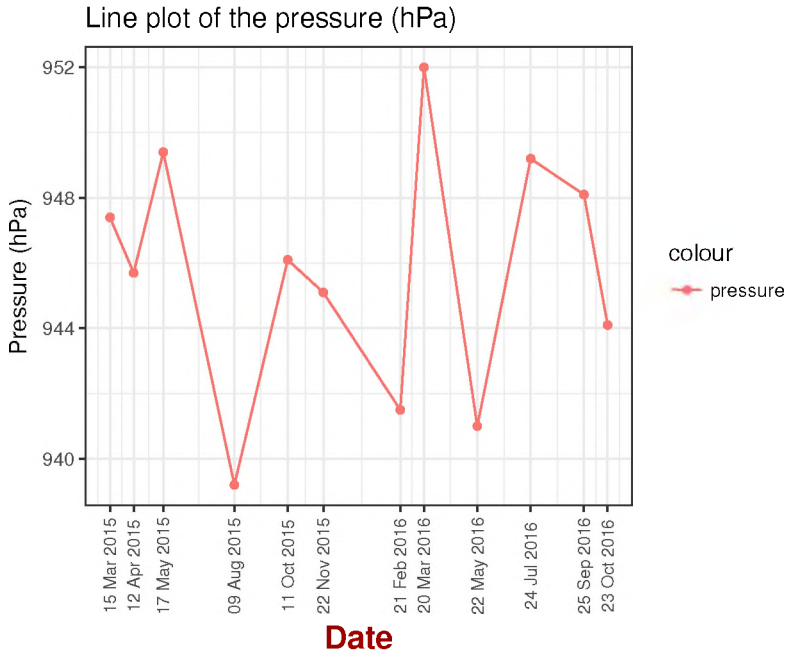


Figure 3.4: Pressure (hPa) on the day of the tournament.

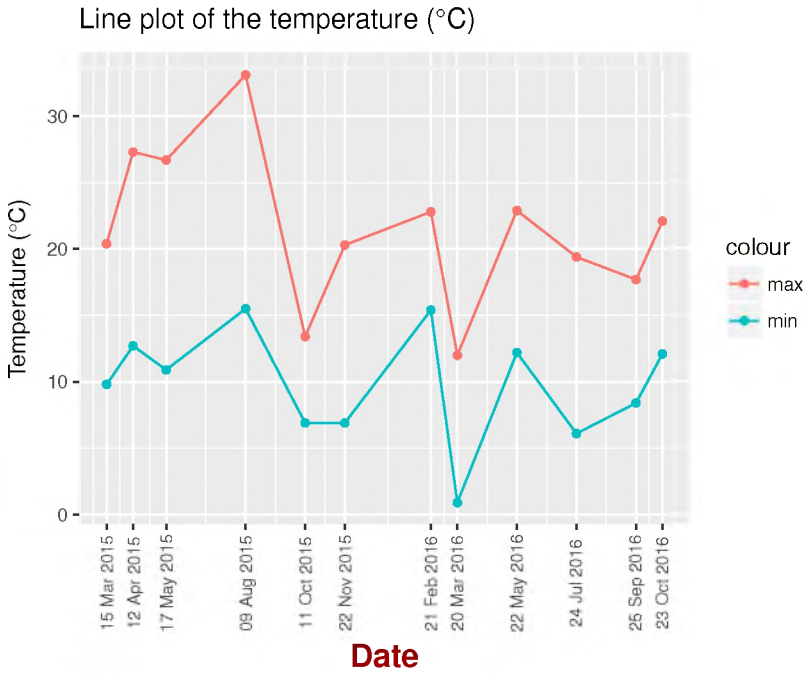


Figure 3.5: Minimum and maximum temperatures on the day of the tournament.

3.6 Generalized Linear Mixed Models

Generalized linear mixed models (GLMM's) are obtained from GLM's by incorporating random effects into the linear predictors (Zuur et al., 2009, page 323). GLMM's include linear mixed models (LMM's) for normal responses as a special case. Schielzeth & Nakagawa (2013) indicate that it is important to consider how the levels of the fixed factors are related to the levels of the random factor in a model with both fixed and random factors. GLMM's extend the GLM approach by allowing some of the parameters in the linear predictor to be treated as random variables (Maunder & Punt, 2004). Random effects have been introduced into models to deal with interactions between continuous and categorical variables (Maunder & Punt, 2004). These models are useful for modeling the dependence among response variables inherent in longitudinal or repeated measures studies, for accommodating overdispersion among binomial or Poisson responses and for producing shrinkage estimators in multi-parameter problems, such as the construction of maps of small area disease rates (Sinha, 2004). GLMM's provide a more flexible approach for analyzing non-normal data when random effects are present (Sinha, 2004). It is usually assumed that the random effects have a multivariate normal distribution whose variance components are to be estimated from the data (Sinha, 2004). Other distributional assumptions on the random effects are made if the random effects are not normally distributed. These assumptions should be validated (Schielzeth & Nakagawa, 2013).

GLMM's model the mean response conditional upon both measured covariates and unobserved random effects. However the inclusion of the unobserved random effects induces corre-

lations among the repeated responses marginally, when averaged over the distribution of the random effects (Sinha, 2004). GLMM can be formulated using the following specification (Fitzmaurice et al., 2008, page 16 and 17):

1. Given a $q \times 1$ vector of random effects $\mathbf{b}_i$, the Y_{ij} are assumed to be conditionally independent and to have exponential family distributions with conditional mean depending upon both fixed and random effects, that is

$$g^{-1}\{E(Y_{ij} \mid \mathbf{b}_i)\} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{b}_i$$

   for some known link function, $g^{-1}(\cdot)$. The conditional variance is assumed to depend on the conditional mean, $Var(Y_{ij} \mid \mathbf{b}_i) = \phi\nu\{E(Y_{ij} \mid \mathbf{b}_i)\}$, where $\nu(\cdot)$ is a known variance function and ϕ is a scale parameter that may be known or may need to be estimated.

2. The random effects, $\mathbf{b}_i$, are assumed to be independent of the covariates, $\mathbf{X}_{ij}$, and to have a multivariate normal distribution with zero mean and $q \times q$ covariance matrix $\mathbf{G}$.

These two components specify a class of GLMM's. The conditional independence assumption in the first component is not necessary, but is commonly made. Any multivariate distribution could be assumed for the $\mathbf{b}_i$, however it is common to assume that the $\mathbf{b}_i$ have a multivariate normal distribution (Fitzmaurice et al., 2008, page 9 and 10). The regression parameters in GLMM's are best understood in terms of the targets of inference. In GLMM's, the target of inference is the individual because the regression coefficients have interpretation in terms of contrasts of the transformed conditional means, $E(Y_{ij} \mid \mathbf{X}_{ij}, \mathbf{b}_i)$. The target inference is the population because the regression parameters have interpretation in terms of the transformed population means, $E(Y_{ij} \mid \mathbf{X}_{ij})$. The special case of linear models, where an identity link function is adopted, the fixed effects in the model for the conditional means is given by

$$E(Y_{ij} \mid \mathbf{X}_{ij}, \mathbf{b}_i) = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{b}_i$$

which also happens to have interpretation in terms of the population means because

$$E(Y_{ij} \mid \mathbf{X}_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta}$$

where averaged over the distribution of the random effects. However for the non-linear link functions usually adopted for discrete data, this relationship no longer holds, since if

$$g^{-1}\{E(Y_{ij} \mid \mathbf{X}_{ij}, \mathbf{b}_i)\} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{b}_i$$

then

$$g^{-1}\{E(Y_{ij} \mid \mathbf{X}_{ij})\} \neq \mathbf{X}_{ij}\boldsymbol{\beta}$$

for any $\boldsymbol{\beta}$ (Fitzmaurice et al., 2008, page 19 and 20). Let $\mathbf{Y}$ be the observed data vector and

conditional on the random effects, $\mathbf{b}$, assume that the elements of $\mathbf{Y}$ are independent and drawn from a distribution in the exponential family,

$$f_{y_i/b}\left(y_i \mid \mathbf{b}, \beta, \phi\right) = \exp\left\{\frac{\left(y_i\theta_i - b\left(\theta_i\right)\right)}{a\left(\phi\right)} + c\left(y_i, \phi\right)\right\}$$

for some known functions a, b, and c. The canonical parameter $\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{b}$, with $\mathbf{x}_i'$ being the i^{th} row of the design matrix $\mathbf{X}$, for the fixed effects, and with $\mathbf{z}_i'$ being the i^{th} row of the design matrix $\mathbf{Z}$, for the random effects. Furthermore, it is assumed that $\mathbf{b} \sim f_b\left(\mathbf{b} \mid \Sigma\right)$ depending on the parameter $\boldsymbol{\Sigma}$. The classical likelihood function can be defined as

$$L\left(\boldsymbol{\beta}, \phi, \Sigma \mid \mathbf{y}\right) = \prod_{i=1}^{n} f_{y_i/b}\left(y_i \mid \mathbf{b}, \beta, \phi\right) f_b\left(\mathbf{b} \mid \Sigma\right) db.$$

The goal is to develop algorithms to calculate the fully parametric maximum likelihood estimates based on the likelihood (Jiang, 2007, page 121). McCulloch (1997) identifies three main algorithms for ML in these models:

- Monte Carlo Expectation Maximization (MCEM);
- Monte Carlo Newton-Raphson (MCNR); and
- Simulation Maximum Likelihood (SML).

The classical maximum likelihood estimating equations for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ can be estimated by Monte Carlo Newton-Raphson (MCNR) and may be expressed in GLM's, GAM's and GLMM's, as

$$E\left[\frac{\partial \ln f_{y_i/t}\left(\mathbf{y} \mid \mathbf{T}, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} \mid \mathbf{y}\right] = 0$$
$$E\left[\frac{\partial \ln f_t\left(\mathbf{T} \mid \boldsymbol{\Sigma}\right)}{\partial \boldsymbol{\Sigma}} \mid \mathbf{y}\right] = 0.$$

GLM's are usually fit using the Newton-Raphson or scoring algorithm (McCulloch, 1997; Sinha, 2004). Jennrich & Schluchter (1986) describes three algorithms for computing maximum likelihood estimates (MLE's) of regression and covariance parameters, namely the Newton-Raphson, Fisher scoring and an algorithm combining scoring with the expected-maximization (EM) algorithm. The Newton-Raphson algorithm is computationally more efficient than the expected maximization algorithm as it converges faster (Kuk & Cheng, 1997).

3.7 The Markov Chain Monte Carlo Algorithm

Markov Chain Monte Carlo (MCMC) is a Bayesian statistical technique that samples parameters according to a stochastic algorithm that converges on the posterior probability

distribution of the parameters, combining information from the likelihood and the posterior distributions (Bolker, 2007, page 310). MCMC algorithms sample sequentially from random values of the fixed effect parameters, the levels of the random effects and random effect parameters, in a way that converges on the distribution of these values (Bolker et al., 2009). Advantages of the MCMC technique include the high flexibility, the arbitary number of random effects and high accuracy. Disadvantages include being very slow, technically challenging and challenges associated with the Bayesian framework (Bolker et al., 2009). The method extends easily to consider multiple random effects, although this requires large data sets. When the data is not normally distributed, not transformable but binary and has more than three random effects Bolker et al. (2009) suggest applying MCMC and MCEM.

The Monte Carlo Newton-Raphson Algorithm

Kuk & Cheng (1997) defined the Monte Carlo Newton-Raphson (MCNR) method as an iterative procedure that can be used to approximate the maximum of a likelihood function in situations where the direct likelihood computation is infeasible because of the existence of unmeasured variables, missing data, or measurement error. A major application of this algorithm is fitting generalized linear models with random effects to clustered binary or count data (Kuk & Cheng, 1997). The MCNR algorithm is a popular method for finding maximum likelihood estimates for incomplete data (Kuk & Cheng, 1997). Let $\mathbf{Y}$ denote the observed incomplete data, $\mathbf{Z}$ the missing data and $\mathbf{X}$ =($\mathbf{Y}, \mathbf{Z}$) the complete data. The log-likelihood function of the observed data, $\mathbf{Y}$, is given by $\ell(\boldsymbol{\theta}; \mathbf{y}) = \ln f(\mathbf{y}; \boldsymbol{\theta})$ whereas the log-likelihood of the complete data, $\mathbf{X}$, is given by $\ell(\boldsymbol{\theta}; \mathbf{x}) = \ln f(\mathbf{y}, \mathbf{z}; \boldsymbol{\theta})$. $l(\boldsymbol{\theta}; \mathbf{y}) = \log f(\mathbf{y}; \boldsymbol{\theta})$ denotes the log-likelihood function based on the observed data $\mathbf{y}$. $l'(\boldsymbol{\theta}; \mathbf{y})$ denotes the $p \times 1$ vector of the first derivatives of $l(\boldsymbol{\theta}; \mathbf{y})$ with respect to the components of $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$. $l''(\boldsymbol{\theta}; \mathbf{y})$ denotes the $p \times p$ matrix of the second derivatives of $l(\boldsymbol{\theta}; \mathbf{y})$ with respect to the components of $\boldsymbol{\theta}$. The Newton-Raphson (NR) iterative update of $\boldsymbol{\theta}$ is given by (Kuk & Cheng, 1997)

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \{l''(\boldsymbol{\theta}^{(k)}; \mathbf{y}\}^{-1} l'(\boldsymbol{\theta}^{(k)}; \mathbf{y}). \tag{3.2}$$

When the data are incomplete it is often not possible to find closed form expressions for $l'(\boldsymbol{\theta}; \mathbf{y})$ and $l''(\boldsymbol{\theta}; \mathbf{y})$. In these cases $l'(\boldsymbol{\theta}; \mathbf{y})$ and $l''(\boldsymbol{\theta}; \mathbf{y})$ are expressed in terms of the conditional expectations of certain functions of the complete data, $\mathbf{X}$, given the observed data $\mathbf{y}$, that is (Kuk & Cheng, 1997)

$$l'(\boldsymbol{\theta}; \mathbf{y}) = E\{l'(\boldsymbol{\theta}; \mathbf{y}, \mathbf{Z}) \mid \mathbf{y}; \boldsymbol{\theta}\}$$

and

$$l''(\boldsymbol{\theta}; \mathbf{y}) = E\{l''(\boldsymbol{\theta}; \mathbf{y}, \mathbf{Z}) \mid y; \boldsymbol{\theta}\} + E\{l'(\boldsymbol{\theta}; \mathbf{y}, \mathbf{Z}) l''(\boldsymbol{\theta}; \mathbf{y}, \mathbf{Z}) \mid y; \boldsymbol{\theta}\} - l'(\boldsymbol{\theta}; \mathbf{y}) l''(\boldsymbol{\theta}; \mathbf{y}).$$

If these conditional expectations cannot be performed analytically, for example if these conditional expectations cannot be expressed in closed form. Kuk & Cheng (1997) suggest Monte Carlo approximations of $l'(\boldsymbol{\theta}^{(k)};\mathbf{y})$ and $l''(\boldsymbol{\theta}^{(k)};\mathbf{y})$ by simulating $\mathbf{Z}_1,\ldots,\mathbf{Z}_M$ from the conditional distribution of $\mathbf{Z}$ given $\mathbf{y}$ where $\boldsymbol{\theta}$ is set to the current estimate, that is the current iterations value, of $\boldsymbol{\theta}$, denoted as $\boldsymbol{\theta}^{(k)}$. The Monte Carlo approximation of $l'(\boldsymbol{\theta}^{(k)};\mathbf{y})$ and $l''(\boldsymbol{\theta}^{(k)};\mathbf{y})$ are given by (Kuk & Cheng, 1997)

$$l'_M(\boldsymbol{\theta}^{(k)};\mathbf{y}) = \frac{1}{M}\sum_{i=1}^{M} l'(\boldsymbol{\theta}^{(k)};\mathbf{y},\mathbf{z}_i) \tag{3.3}$$

$$l''_M(\boldsymbol{\theta}^{(k)};\mathbf{y}) = \frac{1}{M}\sum_{i=1}^{M} l''(\boldsymbol{\theta}^{(k)};\mathbf{y},\mathbf{z}_i) + \{\frac{1}{M}\sum_{i=1}^{M} l'(\boldsymbol{\theta}^{(k)};\mathbf{y},\mathbf{z}_i) l''(\boldsymbol{\theta}^{(k)};\mathbf{y},\mathbf{z}_i) - l'_M(\boldsymbol{\theta}^{(k)};\mathbf{y}) l''_M(\boldsymbol{\theta}^{(k)};\mathbf{y}).$$

Substituting l'_M and l''_M for l' and l'' in equation 3.2 yields the approximate MCNR iterative procedure. If this iterative process converges, at say $\hat{\boldsymbol{\theta}}_M$ for large M, then this should be a good approximation of $\hat{\boldsymbol{\theta}}$, the MLE and hence we use $-\left\{l''_M(\hat{\boldsymbol{\theta}}_M;\mathbf{y})\right\}^{-1}$ to estimate the variance covariance matrix of $\hat{\boldsymbol{\theta}}_M$ (Kuk & Cheng, 1997). The overall statistic under Newton-Raphson (NR) can be constructed as

$$\mathbf{W}(\boldsymbol{\theta}^{(k)}) = \left\{l'_M(\boldsymbol{\theta}^{(k)})\right\}' \hat{\boldsymbol{\sigma}}^{-1} l'_M(\boldsymbol{\theta}^{(k)}) \tag{3.4}$$

where $\hat{\boldsymbol{\sigma}}$ is a suitable estimate of the variance covariance matrix $\boldsymbol{\sigma}$ of $l'_M(\boldsymbol{\theta}^{(k)})$. The distribution of $\mathbf{W}$ under the hypothesis that the gradient $l'(\boldsymbol{\theta}^{(k)})$ is zero follows a χ^2 distribution with v degrees of freedom since $l'_M(\boldsymbol{\theta}^{(k)})$ defined by equation 3.3 is an average over MC replicates and so is asymptotically normal for large M, which suggests that we can use the following level α test $\mathbf{W}(\boldsymbol{\theta}^{(k)}) < \chi^2_v(\alpha)$ (Kuk & Cheng, 1997). If M is large we expect $-l''(\hat{\boldsymbol{\theta}};\mathbf{y})$ to be positive definite. However it is possible that $-l''_M(\boldsymbol{\theta}^{(k)};\mathbf{y})$ is not positive definite if $\boldsymbol{\theta}^{(k)}$ is far away from $\hat{\boldsymbol{\theta}}$ which suggests that non-positive definiteness of $-l''_M(\boldsymbol{\theta}^{(k)};\mathbf{y})$ can result in the non-convergence of equation 3.4 (Kuk & Cheng, 1997).

Algorithm 3.1 The Newton-Raphson Algorithm

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \left\{l''_{M1}(\boldsymbol{\theta}^{(k)};\mathbf{y})\right\}^{-1} l'_M(\boldsymbol{\theta}^{(k)};\mathbf{y}) \tag{3.5}$$

where

$$l''_{M1}(\boldsymbol{\theta}^{(k)};\mathbf{y}) = \frac{1}{M}\sum_{i=1}^{M} l''(\boldsymbol{\theta}^{(k)};\mathbf{y},\boldsymbol{z_i})$$

This can be interpreted as the MC approximation and is often called complete information.

The Monte Carlo Expectation Maximization Algorithm

The expectation maximization (EM) algorithm is used to obtain ML estimates for models that yield analytically formidable likelihood equations, that is for models where it is difficult

to maximize the observed likelihood function directly (Kuk & Cheng, 1997). The EM is a routine requiring two primary iterative calculations, namely the computation of a particular conditional expectation of the log-likelihood, the E-step, and the maximization of this expectation over the relevant parameters, the M-step (Levine & Casella, 2001). The EM algorithm is most useful in situations where it is difficult to maximize the observed log-likelihood function $l(\boldsymbol{\theta}; \mathbf{y}) = \ln f(\mathbf{y}; \boldsymbol{\theta})$ directly whereas the log-likelihood function $l(\boldsymbol{\theta}; \mathbf{x}) = \ln f(\mathbf{y}; \mathbf{z}; \boldsymbol{\theta})$ based on the complete data can be maximized easily (Kuk & Cheng, 1997). EM is a standard technique for LMM's and this procedure requires simulations from the conditional distribution of the missing data given the observed data (Kuk & Cheng, 1997).

The current estimate of the parameter $\boldsymbol{\theta}$, is given by $\boldsymbol{\theta}^{(k)}$. An approximation of the complete data log-likelihood function is obtained by taking the conditional expectation, the E-step:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = E\{l(\boldsymbol{\theta}; \mathbf{X}) \mid \mathbf{y}; \boldsymbol{\theta}^{(k)}\} = \int l(\boldsymbol{\theta}; \mathbf{y}; \mathbf{z}) f(\mathbf{z} \mid \mathbf{y}; \boldsymbol{\theta}^{(k)}) \mathbf{dz}.$$

In the M-step of the algorithm, $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ is maximized as a function of $\boldsymbol{\theta}$ to obtain the updated estimate $\boldsymbol{\theta}^{(k+1)}$. A Monte Carlo approximation of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ is

$$Q_M(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \frac{1}{M} \sum_{i=1}^{M} l(\boldsymbol{\theta}; \mathbf{y}, \mathbf{z}_i).$$

The subsequent M-step used to obtain $\boldsymbol{\theta}^{(k+1)}$ usually requires iterations unless a closed form formula exists for the maximize of $Q_M(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ (Kuk & Cheng, 1997). The EM algorithm converges at a slower linear rate than NR algorithm (Kuk & Cheng, 1997).

MC Approximation in the EM Algorithm

Set up the EM algorithm where the random effects, $\mathbf{u}$, are considered as the missing data. The complete data, $\mathbf{W} = (\mathbf{Y}, \mathbf{u})$, and associated log-likelihood is given by

$$\ln L_W = \sum_i \ln f_{y_i/u}(y_i \mid \mathbf{u}, \boldsymbol{\beta}, \boldsymbol{\phi}) + \ln f_u(\mathbf{u} \mid \mathbf{D}). \tag{3.6}$$

There are two advantages in this process: The $\boldsymbol{u}'s$ and the $Y_i's$ are independent and the M step of the EM algorithm maximizes equation 3.6 with respect to $\boldsymbol{\beta}$, $\boldsymbol{\phi}$ and $\mathbf{D}$. The M step with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ uses only $f_{y/u}$, and it is similar to a standard GLM computation with the values of $\mathbf{u}$ assumed as known. The EM algorithm can be used to conduct MLE's as follows

1. Choose the initial values $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\phi}^{(0)}$, $\mathbf{D}^{(0)}$ and set $m = 0$;

2. Calculating the expectations evaluated under $\boldsymbol{\beta}^{(m)}$, $\boldsymbol{\phi}^{(m)}$, $\mathbf{D}^{(m)}$:

- $\boldsymbol{\beta}^{(m+1)}$ and $\boldsymbol{\phi}^{(m+1)}$ maximizes $E\left[\ln f_{y|u}(\mathbf{y} \mid \mathbf{u}, \boldsymbol{\beta}, \boldsymbol{\phi}) \mid \mathbf{y}\right]$
- $\mathbf{D}^{(m+1)}$ maximizes $E[\ln f_u(\mathbf{u} \mid \mathbf{D}) \mid \mathbf{y}]$
- Set $m = m + 1$.

3. When convergence is achieved, $\boldsymbol{\beta}^{(m+1)}$, $\boldsymbol{\phi}^{(m+1)}$ and $\mathbf{D}^{(m+1)}$ are MLE's of the corresponding parameters (McCulloch, 1997).

Chapter 4

Analysis of GLM Fit

This section discusses statistics and techniques useful in fitting and assessing the goodness-of-fit of a GLM model. Dobson & Barnett (2008, page 19) suggests the following model fitting process:

1. Model specification, where the model is specified in two parts:
   (a) The link function which links the response and predictor variables; and
   (b) The probability distribution of the response variable.
2. Estimation of the parameters in the model;
3. Checking the adequacy of the model, that is assessing how well the model fits or summarizes the data;
4. Inference: computing confidence intervals, testing hypotheses about the parameters in the model and interpreting the results.

Assessing the link function

We have to investigate whether the link function is appropriate for a particular distribution. In a Poisson regression we may want to examine whether the default log-link of multiplicative effects is appropriate compared with an identity link representing additive effects (Hardin et al., 2007, page 50). For a binomial regression we may want to compare the logit link, which is symmetric about one half, with the complementary log-log link, which is asymmetric about one half (Hardin et al., 2007, page 50). Two link functions can be compared by entrenching them in a parametric family of link functions, for example the Box-Cox family of power transforms (Hardin et al., 2007, page 50)

$$g(\mu; \lambda) = \frac{\mu^{\lambda} - 1}{\lambda}$$

and yields the log-link at $\lim_{\lambda \to 0} g(\mu; \lambda)$ and the identity link at $\lambda = 1$. Similarly the family

$$g(\mu; \lambda) = \ln \left\{ \frac{(1-\mu)^{-\lambda} - 1}{\lambda} \right\}$$

gives the logit link at $\lambda = 1$ and the complementary log-log link at $\lim_{\lambda \to 0} g(\mu; \lambda)$ (Hardin et al., 2007, page 50).

Deviance

When using the maximum likelihood (ML) approach a standard method of assessing the fitted model is to compare this model to the fully specified model. The fully specified model is the most general model containing the maximum number of parameters, typically as many parameters as observations (Hardin et al., 2007, page 48). The deviance, D, is defined in terms of the likelihoods of the fitted model, denoted by L_m, and the full model, denoted by L_f, as (Hardin et al., 2007, page 48)

$$D = -2 \ln \left(\frac{L_m}{L_f} \right) = 2 \left\{ \ln (L_f) - \ln (L_m) \right\}.$$

In the general linear model, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the deviance is given by (Dobson & Barnett, 2008, page 48)

$$\frac{1}{\sigma^2} \sum (y_i - \hat{y}_i)^2 = \frac{1}{\sigma^2} \left(y - X\hat{\beta} \right)' \left(y - X\hat{\beta} \right).$$

In this context the scaled deviance $\sigma^2 D = \sum (y_i - \hat{y}_i)^2 = SS_{Error}$. The scaled deviance is generalized in GLM's (Dobson & Barnett, 2008, page 48) as

$$S = \frac{D}{\phi}$$

where D is the deviance and ϕ is the scale parameter. The deviance is calculated for each family using the canonical parameter, the function of $\theta(\mu)$. The main aim in assessing deviance is to determine the utility of the parameters added to the null model, that is in determining the usefulness or benefit of adding parameters to the null model (Dobson & Barnett, 2008, page 48). The deviance of a GLM's is calculated as

$$D(\mu; \hat{\mu}) = 2 \sum_i \left[y \left\{ \theta(y_i) - \theta(\hat{\mu}_i) \right\} - b \left\{ \theta(y_i) \right\} + b \left\{ \theta(\hat{\mu}_i) \right\} \right]$$

where $\theta()$ denotes the canonical parameter and $b()$ the cumulant (Hardin et al., 2007, page 48). In more concise terms the deviance can be expressed as (Hardin et al., 2007, page 48)

$$D = 2\phi \left\{ \ln (L_f) - \ln (L_m) \right\}.$$

The difference in deviance statistics between the saturated, or full, and fitted models captures the distance between the predicted values and the outcomes and hence models can be compared based on this statistics (Hardin et al., 2007, page 49). The smaller the deviance the better the model (Zuur et al., 2009, page 218). Care must be taken, particularly for non-nested models. The deviance as defined above, is the log of the ratio of the likelihoods, or equivalently as the difference of the log of the likelihoods. The deviance for each observation is typically denoted as the deviance residual and hence the deviance is the sum of the squared deviance residuals. For example in logistic regression (Hardin et al., 2007, page 48)

$$\ln(L) = \sum_i \{y_i \ln(\hat{\pi}_i) + (1 - y_i) \ln(1 - \hat{\pi}_i)\}$$

and hence the deviance residual for the i^{th} observation is given by (Hardin et al., 2007, page 48)

$$d_i = S_i \sqrt{-2 \{y_i \ln(\hat{\pi}_i) + (1 - y_i) \ln(1 - \hat{\pi}_i)\}}$$

where

$$S_i \begin{cases} 1 & \text{if } y_i = 1 \\ 0 & \text{if } y_i = 0 \end{cases}$$

and hence $D = \sum_{i=1} \hat{d}_i^2$. Pearson residual, denoted by r_p, that is the letter r , "standardize" each observation by subtracting the mean and dividing by the standard deviation. For example in logistic regression (Hardin et al., 2007, page 54):

$$r_p = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i (1 - \hat{\pi}_i)}}$$

Pearson's statistics is defined as

$$\sum_{i=1}^{n} r_i^p.$$

If $f(\mathbf{y}; \boldsymbol{\theta})$ is the density function or probability distribution for the observation $\mathbf{y}$ given the parameter $\boldsymbol{\theta}$, then the log-likelihood expressed as a function of the mean value parameter, $E(\mathbf{Y}) = \boldsymbol{\mu}$ (McCullagh & Nelder, 1989, page 24) that is

$$l(\mathbf{y}; \boldsymbol{\mu}) = \ln f(\mathbf{y}; \boldsymbol{\mu}).$$

The log-likelihood based on a set of independent observations $y_1, \ldots, y_n$ is just the sum of the individual contributions, that is $l(\mathbf{y}; \boldsymbol{\mu}) = \sum_{i=1}^{n} \ln f_i(\mathbf{y}_i; \boldsymbol{\mu})$ where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. The univariate normal density with known variance σ^2 is given by

$$f(y; \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y - \mu)^2\right]$$

so the log-likelihood is

$$l(y;\mu) = \sum_{i=1}^{n} \left\{ \frac{(y_i^2 - y_i^2)/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right) \right\}.$$

Setting $\mu = y$ gives the maximum log-likelihood, $l(y;y) = \sum_{i=1}^{n} \left\{ -\frac{1}{2}\ln\left(2\pi\sigma^2\right) \right\}$ so the Gaussian deviance is calculated as

$$\begin{aligned} D &= 2\sigma^2 \sum_{i=1}^{n} \left\{ -\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{(y_i^2 - y_i^2)/2}{\sigma^2} + \frac{y_i^2}{2\sigma^2} + \frac{1}{2}\ln\left(2\pi\sigma^2\right) \right\} \\ &= 2\sigma^2 \sum_{i=1}^{n} \left\{ -\frac{y_i^2 - 2y_i\mu_i + \mu_i^2}{2\sigma^2} \right\} \\ &= \sum_{i=1}^{n} (y_i - \mu_i)^2 . \end{aligned}$$

The Gaussian deviance is identical to the sum of squared residuals (Hardin et al., 2007, page 70).

Residual Analysis

The residual deviance is defined as twice the difference between the log-likelihood of a model that provides a perfect fit for the model under study, that is the Poisson GLM deviance (Zuur et al., 2009, page 217)

$$\begin{aligned} D &= 2\log\left[L(\mathbf{y};\mathbf{y})\right] - 2\log\left[L(\mathbf{y};\boldsymbol{\mu})\right] \\ &= 2\sum_{i} \left(y_i \log\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i) \right). \end{aligned}$$

Null deviance is the residual deviance in a model that only contains an intercept and represents the deviance explained by the overall mean of the response (Zuur et al., 2009, page 217 and 218). The residual deviance is referred to as the total deviance explained by the model. When a model is not well fitted, this can be detected by considering the deviance residuals, higher deviance residuals indicate poorer fit. Residual analysis is characterized by the consideration of various different residuals (Hardin et al., 2007, page 53). By default R uses the deviance residuals for model checking as these have distributional properties that are closer to the residuals from a Gaussian linear regression model than other alternatives (Zuur et al., 2009, page 230). The three most common residuals considered in GLM's includes deviance residuals, Pearson residuals and response residuals.

Deviance Residuals

If the deviance is used as a measure of inconsistency of a GLM, then each unit contributes a quantity d_i to that measure, so that $\sum d_i = D$, hence we define

$$r_{D_i} = \text{sign}(y_i - \mu) \times \sqrt{d_i}$$

where the term sign stands for positive and negative. The positive sign is only used if $Y_i > \mu_i$ and negative is applied when $Y_i < \mu_i$ (Zuur et al., 2009, page 229 and 230). We have a quantity that increases with $y_i - \mu_i$ and $\sum r_D^2 = D$. For the Poisson distribution

$$r_{D_i} = \text{sign}\,(y - \mu)\left\{2\left[y \log\left(\frac{y}{\mu}\right) - y + \mu\right]\right\}^{\frac{1}{2}}.$$

If the residual deviance value is close to its degrees of freedom, it indicates a reasonable fit to the data (Hardin et al., 2007, page 55).

Pearson Residuals

Pearson's residual are defined (Hardin et al., 2007, page 54) as:

$$r_p = \frac{y - \hat{\mu}}{\sqrt{Var(\hat{\mu})}}$$

that is the raw residual scaled by the estimated standard deviation of Y. Pearson statistic is used not so much as a goodness-of-fit but as a measure of residual variation (McCullagh & Nelder, 1989, page 37).

Response Residuals

Response residuals are the deviation, $y_i - \hat{y}_i$, considered in general linear model (Hardin et al., 2007, page 53). These residuals are simply the difference between the observed and fitted outcome, that is $r_i^R = y_i - \hat{y}_i$.

4.1 GLM Residuals and Diagnostics

In linear regression the fit of the model is assessed using diagnostics based on the sums of squared residuals (Faraway, 2016, page 135). The deviance $D = \frac{1}{\sigma^2}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$ is typically expressed in terms of the "hat" matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ (Dobson & Barnett,

2008, page 83) as

$$\begin{aligned} D &= \frac{1}{\sigma^2}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right) \\ &= \frac{1}{\sigma^2}\mathbf{y}'\left(\mathbf{I}-\mathbf{H}\right)\mathbf{y} \end{aligned}$$

D has a $\chi^2_{(n-p)}$ distribution with non-centrality parameter $\lambda = \frac{1}{\sigma^2}\left(\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{I}-\mathbf{H}\right)\left(\mathbf{X}\boldsymbol{\beta}\right)$, however since $\left(\mathbf{I}-\mathbf{H}\right)\mathbf{X} = 0$ then $\lambda = 0$, that is D has a central $\chi^2_{(n-p)}$ distribution (Dobson & Barnett, 2008, page 83). In this model the leverages, h_i, are the diagonal components of $\mathbf{H}$ and represent the potential of the point i to influence the fit. These distances are solely a function of $\mathbf{X}$ and their effect depends on $\mathbf{y}$ since $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ (Faraway, 2016, page 124).

As discussed above, in GLM's there are different types of residuals and hence sum of squared residuals. Consider the deviance statistics defined in the previous section, that is the ratio of the likelihood of the saturated and fitted model. This likelihood ratio test assumes that the model assumptions are valid and hence should not be used to assess the validity of the model (Dobson & Barnett, 2008, page 84). In the linear general model, Cook's distances are defined for the i^{th} observation as

$$D_i = \frac{\left(\hat{y}-\hat{y}_{(i)}\right)^{i^{th}}\left(\hat{y}-\hat{y}_{(i)}\right)}{\rho\hat{\sigma}^2}$$

where the subscript (i) denotes the fitted model where the i^{th} observation has been omitted from the data set. These distances can be plotted against the half-normal quantiles to reveal influential observations (Faraway, 2014, page 70).

Model Assessment

Diagnostic checks are typically conducted to assess heteroscedasticity, validate assumptions on the distribution of random effects and outlier detection (Hardin et al., 2007, page 49). In linear models the variance σ^2 is estimated independently of the parameters (Zuur et al., 2009, page 217). This is often not the case when GLM's are fitted using the binomial, Poisson or negative binomial distribution in the analysis of discrete response data (Hardin et al., 2007, page 165). When the variation of a fitted GLM is greater than that predicted by the model, that is the variance of the response is greater than the nominal variance (Hardin et al., 2007, page 165). Overdispersion only affects discrete models as continuous models fit the scale parameter ϕ (Hardin et al., 2007, 165). For example in the Poisson model it is assumed that the mean is equal to the variance. If there is no overdispersion then the residual mean deviance should be approximately one. Overdispersion causes incorrect standard errors of the estimates of β and the selection of overly complex models (Zuur et al., 2009, page 224). Zuur et al. (2009, page 224) suggests two options of assessing if a model is overdispersed. The first is based on the χ^2 approximation of the residual or scaled deviance, namely if the

model is overdispersed $S = \frac{D}{\phi} \sim \chi^2_{(n-p)}$. As a result we estimate ϕ with $\hat{\phi} = \frac{D}{n-p}$. If $\hat{\phi}$ is approximately one then it can safely be assumed that the model is not overdispersed and the model validation process should be initiated (Zuur et al., 2009, page 224). The second option is to utilize a different estimator which is based on Pearson residuals, for example adjust the model for overdispersion using quasi-likelihood (Zuur et al., 2009, page 224).

In GLM's the estimated hat matrix is of the form

$$\hat{\mathbf{H}} = \hat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{W}}^{\frac{1}{2}}$$

where h_i is the i^{th} diagonal of this matrix (Hardin et al., 2007, page 49). In GLM's Cook's distance can be approximated by (Hardin et al., 2007, page 49)

$$C_i = \left(\hat{\boldsymbol{\beta}}^{\star}_{(i)} - \hat{\boldsymbol{\beta}}\right)' \mathbf{I} \left(\hat{\boldsymbol{\beta}}^{\star}_{(i)} - \hat{\boldsymbol{\beta}}\right)$$

where $\mathbf{I}$ denotes the Fisher information matrix, that is the inverse Hessian, (Hardin et al., 2007, page 41) and $\hat{\boldsymbol{\beta}}^{\star}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{\left(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X}\right)^{-1}\mathbf{x}'_i\hat{\mathbf{r}}^R_{(i)}}{1-\hat{h}_{(i)}}$ are the one-step jackknife-estimated coefficient vectors, that is $\hat{\mathbf{V}} = \text{diag}\left\{V\left(\hat{\mu}\right)\right\}$ matrix, $\mathbf{X}$ is the $(n \times p)$ matrix of covariates, $\hat{h}_{(i)}$ is the i^{th} diagonal of the hat matrix, $\hat{\boldsymbol{\beta}}$ is the estimated coefficient vector using all the observations and $\hat{r}^R_i$ is the estimated response residual (Hardin et al., 2007, page 49). These distances are used to detect observations with excessive influence, by for example classifying those observations whose Cook's distances are greater than $\frac{4}{n-p-1}$ as extreme or problematic and investigation the observation whose distances exceed $\frac{4}{n}$.

Model Selection

Model selection criteria should consider both the fit of the model to the data and complexity of the model (Johnson & Omland, 2004). A set of candidate models should be fitted to the observed data. The most heavily parameterized, or full, model should be assessed in terms of the goodness-of-fit of the model to the data using, for example, the χ^2 or G tests or a parametric bootstrap approach (Johnson & Omland, 2004). If this heavily parameterized model provides a reasonable fit to the data then other, less heavily parameterized models in the candidate set are fit to the data. These multiple models are then compared simultaneously in order to determine which explanatory variables are important and hence determine which model is best supported by the data. In linear regression this could be achieved using for example the forward or backward stepwise regression approach based on the appropriate F-test (Johnson & Omland, 2004). The model selection process is designed to avoid the issues associated with these multiple F-tests, for example the selection of sub-optimal models which are selected as a result of the hierarchical order in which these models are constructed (Johnson & Omland, 2004). Model selection criteria are used to rank the various candidate models and to weigh the relative support for each model using the negative log-likelihood

scores as a measure of lack of fit and a penalty term representing the complexity of the model (Johnson & Omland, 2004). Akaike's information criteria (AIC), Schwartz criteria (SC) or the Bayesian information criteria (BIC) are often employed as model selection criteria. AIC is defined as $AIC = -2L + 2K$ where L denotes the maximum log-likelihood of the model and K is the number of parameters in the model (McCarthy, 2007, page 47). AIC scores computed using REML are not comparable to AIC scores computed using ML (Zuur et al., 2009, page 121). Poor fitting models and complex models have higher AIC scores, thus the lower the AIC the better the model (Zuur et al., 2009, page 122). The quality of fit for the different models can be evaluated by comparing the deviance of these models to that of the null model, that is the model fitted without explanatory variables (Zuur et al., 2009, page 218). The hypothesis testing approach drops the least significant term and then refits the model. The updated model is then investigated to ascertain if there are still non-significant terms in the model (Zuur et al., 2009, page 221). In this context

$$R^2 = 1 - \frac{\text{Residual deviance}}{\text{Null deviance}}$$

where R^2, the coefficient of determination, refers to the relative predictive power of the model. The models do not need to be nested and can have the same or different number of parameters. The hypothesis testing procedure can be used to assess the model fit. There are three options for hypothesis testing, namely the t-statistic, F-statistics and the likelihood ratio test. In the likelihood ratio approach, two models with the same random effects structure, fitted using REML or ML estimation, can be compared using the likelihood criteria (Zuur et al., 2009, page 126).

Model Adequacy

Model adequacy indicates how well the model fits the data or whether there were any violation of the model assumptions. The model fit can be assessed using goodness-of-fit tests based on the deviance or Pearson χ^2 statistics. Graphical methods can be utilized to assess model fit, for example quantile-quantile plots and partial residual plots. Partial residual plots are useful in assessing the model assumptions, such as linearity. However if a partial residual plot is non-linear, it indicates that a linear assumption may not be appropriate for the model (Johnson & Omland, 2004). There are multiple concepts and issues in goodness-of-fit testing. A statistical measure is needed to judge which model is the best among many candidate models. After selecting the best model, we need to determine how good the selected model is.

4.2 Fitting and Assessing a GLM

This section demonstrates various models and model selection, hypothesis testing and model validation. The full, or saturated, and reduced models can be compared using ANOVA methodologies. In this example the model fitting the parameter estimates are computed using the Poisson log-likelihood.

The Sinclair Data Set

A researcher is investigating the association between predation, sex and health in Serengeti wildebeest (Logan, 2011, page 515). The Sinclair data set consists of 226 cross classified wildebeest carcasses from the Serengeti by three variables: sex (male or female), cause of death (predation and non-predation) and bone marrow type (solid white fatty, opaque gelatinous, translucent gelatinous, where solid fatty, indicates a healthy animal which is not undernourished). The boxplots, figure 4.1, shows that the carcasses by sex and the carcasses by death have unequal variances and the frequency of carcasses are asymmetric. The marrow bone type boxplots for solid white fatty (SWT) and translucent gelatinous (TG) seem to have equal variance but opaque gelatinous (OG) bone marrow type seem to have unequal variances. These data provide sufficient evidence that there is homogeneity of variances in the wildebeest variables, namely death (Bartlett's K-squared $= 0.86743$, $df = 1$, p-value $= 0.3517$) , sex (Bartlett's K-squared $= 0.2695$, $df = 1$, p-value $= 0.6037$) and marrow type (Bartlett's K-squared $= 0.6361$, $df = 2$, p-value $= 0.7273$). The normal Q-Q plot, figure 4.2, suggests that there is a violation of normality since the normal Q-Q plot is not approximately linear. The ANOVA model in table 4.1 demonstrates that the sex ($F_{obs} = 0.0029$, $df = 1, 7$, p-value $= 0.9589$), death ($F_{obs} = 1.1412$, $df = 1, 7$, p-value $= 0.3208$) and marrow type ($F_{obs} = 2.3031$, $df = 2, 7$, p-value $= 0.1704$) are not statistically significant. The interaction of wildebeest variables has not been considered in this model.

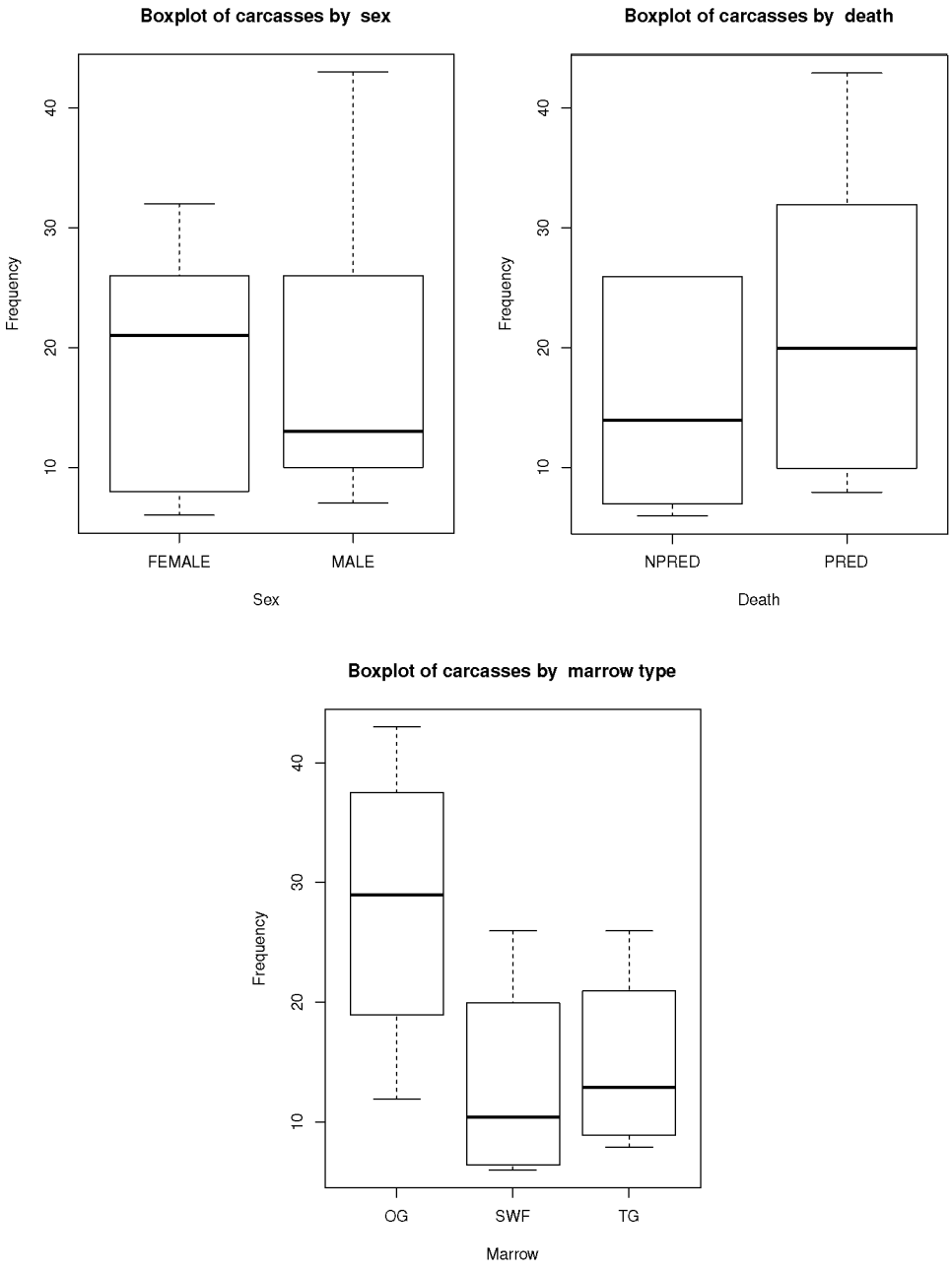


Figure 4.1: Boxplots of the various variables in the Sinclair data set.

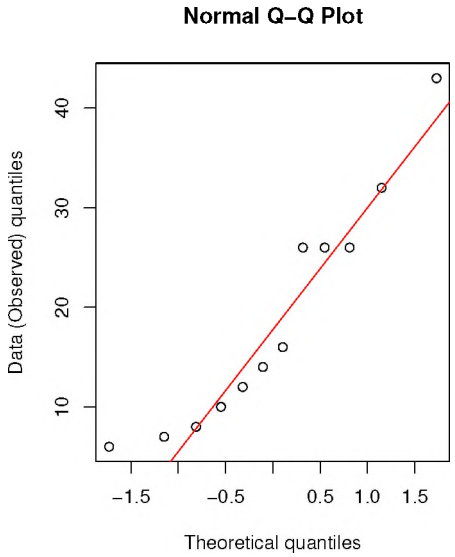
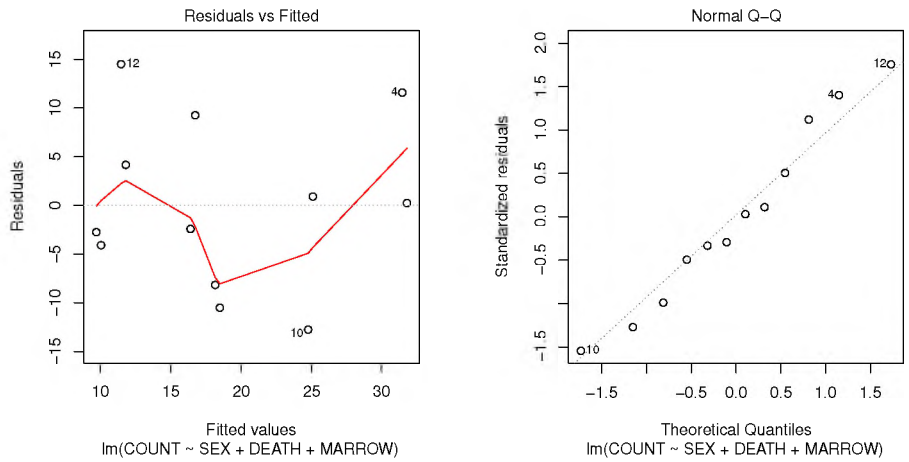


Figure 4.2: Normal Q-Q plot for carcasses: Sinclair data set.

Source of Variation	df	Sum of Squares	Mean Square	F Statistics	p-value
Sex	1	0.33	0.33	0.0029	0.9589
Death	1	133.33	133.333	1.1412	0.3208
Marrow	2	538.17	269.083	2.3031	0.1704
Residuals	7	817.83	116.833		

Table 4.1: ANOVA model: Wildebeest deaths.



(a) Residuals versus fitted values. **(b)** Normal Q-Q plot.

Figure 4.3: Diagnostics plots from fitting a model.

The normal Q-Q plot, figure 4.3 (b), suggests that the residuals of this model meet the normality assumption since the points do not deviate from the theoretical quantiles. The residuals are normally distributed (Shapiro test, $W = 0.9639$, p-value $= 0.8382$). The fitted

residuals, figure 4.3(a), shows that the residuals of this model meet the homogeneity of variance assumption since the points are randomly scattered with no particular pattern. These data do not meet the ANOVA assumptions and hence a log-linear model was utilized to assess the researchers hypothesis. Log-linear models are referred to as GLM's that relate the log of the expected values to a linear combination of the variables and their interactions (Quinn & Keough, 2002, page 400). For hypothesis testing, we are fitting these models hierarchically and fitting various combinations of log-linear models starting with the saturated model. The dredge function in R assesses the fit of all possible model combinations using the AIC to compare the models (Logan, 2011, page 516). The models summaries shown in table 4.2 are based on comparing the fit of each model to that of the saturated model. The model with the smaller AIC will be preferred. We fitted a GLM model having carcasses death as the response or dependent variable that is explained by sex, death and marrow type. The Poisson distribution was selected since the dependent variable is a count. The Poisson distribution is useful for a response variable that is a discrete variable, for example if you count the number of animals on a farm that are infected with a disease (Zuur et al., 2009, page 205). Counts are always non-negative and tend to be heterogeneous and both comply with the Poisson distribution. The Poisson logarithmic link function, also called the log link, ensures that the fitted values are always non-negative (Zuur et al., 2009, page 211). The comparison of the fit of model 64 and the saturated model 128 is a test of the null hypothesis that there is no three way interaction.

Model number	(Int)	DEA	MAR	SEX	DEA:MAR	DEA:SEX	MAR:SEX	DEA:MAR:SEX	df	logLik	AIC	delta	weight
128	3.258	+	+	+	+	+	+	+	12	−27.613	79.2	0.00	0.435
12	2.944	+	+		+				6	−34.243	80.5	1.26	0.231
48	2.971	+	+	+	+		+		9	−31.845	81.7	2.46	0.127
64	3.072	+	+	+	+	+	+		10	−31.207	82.4	3.19	0.088
16	2.953	+	+	+	+				7	−34.234	82.5	3.24	0.086
32	2.976	+	+	+	+	+			8	−34.191	84.4	5.16	0.033
4	3.146	+	+						4	−49.003	106.0	26.78	0.000
40	3.173	+	+	+			+		7	−46.605	107.2	27.98	0.000
8	3.155	+	+	+					5	−48.994	108.0	28.76	0.000
56	3.195	+	+	+		+	+		8	−46.562	109.1	29.90	0.000
24	3.178	+	+	+		+			6	−48.951	109.9	30.68	0.000
3	3.341		+						3	−52.561	111.1	31.90	0.000
39	3.367		+	+			+		6	−50.164	112.3	33.10	0.000
7	3.350		+	+					4	−52.553	113.1	33.88	0.000
2	2.741	+							2	−62.529	129.1	49.83	0.000
6	2.750	+		+					3	−62.520	131.0	51.82	0.000
22	2.773	+		+		+			4	−62.477	133.0	53.73	0.000
1	2.936								1	−66.088	134.2	54.95	0.000
5	2.944			+					2	−66.079	136.2	56.93	0.000

Table 4.2: Model selection table: Wildebeest log-linear model.

The difference in fit between models 128, 12, 48, 16 and 32 is inconsequential. A biologist might argue that model 48, the reduced model which has omitted both a two way interaction between death and sex and the three way interaction and has a lower AIC of 81.7 is the most appropriate for these data. Any difference between this reduced model and the saturated model could be due to either two way or three way interaction or both.

Log-linear models for contingency tables with three variables include three main effects (sex, marrow, death), three two variable interactions (sex : marrow, sex : death, marrow : death) also called conditional independence and one three variable interaction (sex : marrow : death). There is a large number of full and reduced models for testing the different interactions and main effects. Comparing the full, or saturated, and reduced models in a hierarchical manner is the most common method of analyzing and presenting the results of log-linear modeling (Quinn & Keough, 2002, page 399 and 400). The goodness-of-fit of a range of possible models is tested using the deviance and AIC. The hierarchical step function in R has been used to generate the set of possible models, as shown below, where we wish to test the hypothesis that there is no association between cause of death, sex and marrow in Serengeti wildebeest. For example marrow : death is significant ($df = 2$, Deviance $= 42.676$, $AIC = 109.901$, p-value < 0.001) demonstrates that there is an association between cause of death and marrow for any sex, that is for either male or female, whether a wildebeest is taken by a predator or not is independent of which marrow type they have. The fit of a model can be judged through it's deviance and the model with the smallest deviance is the better fit. Based on the deviance, the saturated model would be selected since it has the lowest deviance (df$= 2$, Deviance $= 7.1883$, $AIC = 82.414$, p-value$= 0.02748$). The normal Q-Q plot, figure 4.4, shows that the residuals of this model meet the normality assumption since the points are approximately linear. The fitted residuals, figure 4.4, shows that the residuals of this model meet the homogeneity of variance assumption. The residuals of the model are normally distributed (Shapiro test, $W = 0.9768$, p-value $= 0.9675$). The null hypothesis is rejected (p-value $= 0.02748$), there is an association between cause of death, sex and marrow type in Serengeti wildebeest, see appendix C.1.

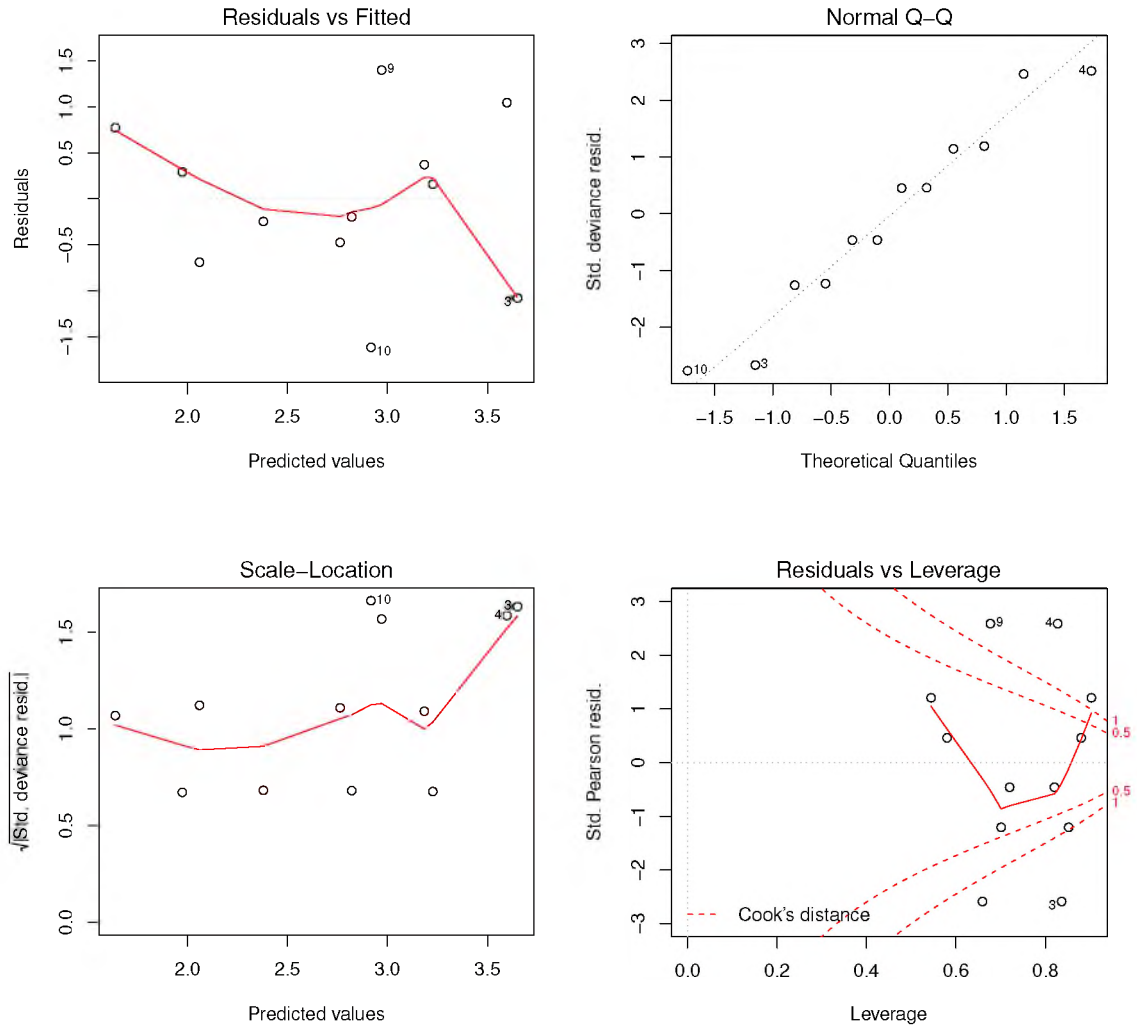


Figure 4.4: Diagnostics plots of Sinclair model.

Chapter 5

Hierarchical or Mixed-Effects Models

Hargrove et al. (2015) considered catch per unit effort (CPUE) data that was collected during fishing tag and release events. Maunder & Punt (2004) used three data sets to explore the effectiveness of bag limit restrictions where catch and effort data were obtained from boat surveys. The catch records were used to calculate the combined weight of all fish caught per individual angler per day, that is the average bag weight for each individual. This tournament sampling procedure was used to draw comparisons of populations across different water bodies, seasons and years. Factors affecting fishing quality may depend on the targeted angler group, geographical location, species and changing angler attitudes (Hargrove et al., 2015). A discrete distribution, such as Poisson or negative binomial may be the most appropriate distribution if the catch is recorded as the number of individual fish (Zuur et al., 2009, page 205). A continuous distribution may be more appropriate if the catch is recorded as weights (Zuur et al., 2009, page 205). Logistic regression may be used to predict the winner of a tournament (Maunder & Punt, 2004). Consider

$$p(m) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

where $p(m)$ denotes the probability of the dependent variable, for example bag weight, and η denotes a linear combination of the independent variables, for example $\eta = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$. In this context β_0 denotes the intercept, β_1, β_2 and β_3 are the regression coefficients associated with the independent variables X_{1i}, X_{2i} and X_{3i} which represent bag size, average fish weight per angler and the total weight number of fish per angler for the i^{th} tournament.

5.1 The Albany Angling Association Tournament Data Set

This data set was provided by the Albany Angling Association (AAA) of the Eastern Cape, South Africa. Climate data associated with the dates on which the AAA events took place were obtained from South African Weather Service. Hargrove et al. (2015) found that temperature had an influence on the catch during fishing tournaments. The AAA data set includes information about the participants, that is anglers at each fishing event, as well as the individuals catch reported as the number and weight of the anglers three best, defined as heaviest, fish. This study is based on these angling tournament data comprising twenty nine (29) anglers at tournaments held in the Grahamstown area of the Eastern Cape province. The tournaments were held at four water bodies or dams and consisted of a total of one hundred and seventy one (171) observations of nine (9) variables. The angling events were held between March 2015 and October 2016 in different months at different locations (table 5.1). In 2015, two events were held at Mangazana dam where a total of twenty six (26) anglers participated, one event was held at Yarrow dam where a total of fourteen (14) anglers participated and three events were held at Settlers dam where a total of thirty one (31) anglers participated. In 2016, one event was held at Mangazana dam with fifteen (15) anglers, one event was held at Settlers dam with nineteen (19) anglers, one event at White's dam with sixteen (16) anglers and three events were held at Yarrow dam with a total of fifty (50) anglers. Anglers at these events were all members of the Albany Angling Association. During this period the club had twenty one (21) adult male members, two (2) adult women and six (6) juniors, that is anglers under the age of eighteen. The date and venue of each event was chosen by the AAA chair. Water bodies of an adequate size to accommodate the large number of anglers were used. The geographical locations of the angling tournament venues were determined using Google maps and are shown in figure 5.1.

Water body	Date	Min temp	Max temp	Males	Females	Juniors	Total anglers
Yarrow	17 May 2015	6.1°C	19.4°C	11	1	2	14
	20 Mar 2016	12.2°C	22.9°C	11	1	3	15
	22 May 2016	10.9°C	26.7°C	15	1	2	18
	24 Jul 2016	0.9°C	12.0°C	12	1	4	17
Settlers	9 Aug 2015	6.9°C	13.4°C	4	1	0	5
	11 Oct 2015	9.8°C	20.4°C	9	2	0	11
	15 Mar 2015	15.5°C	33.1°C	12	2	1	15
	21 Feb 2016	15.4°C	22.8°C	15	1	3	19
Mangazana	12 Apr 2015	12.7°C	27.3°C	13	2	2	17
	22 Nov 2015	6.9°C	20.3°C	8	1	0	9
	23 Oct 2016	12.1°C	22.1°C	11	1	3	15
White's dam	25 Sep 2016	8.4°C	17.7°C	13	0	3	16

Table 5.1: Summary of AAA fishing events.

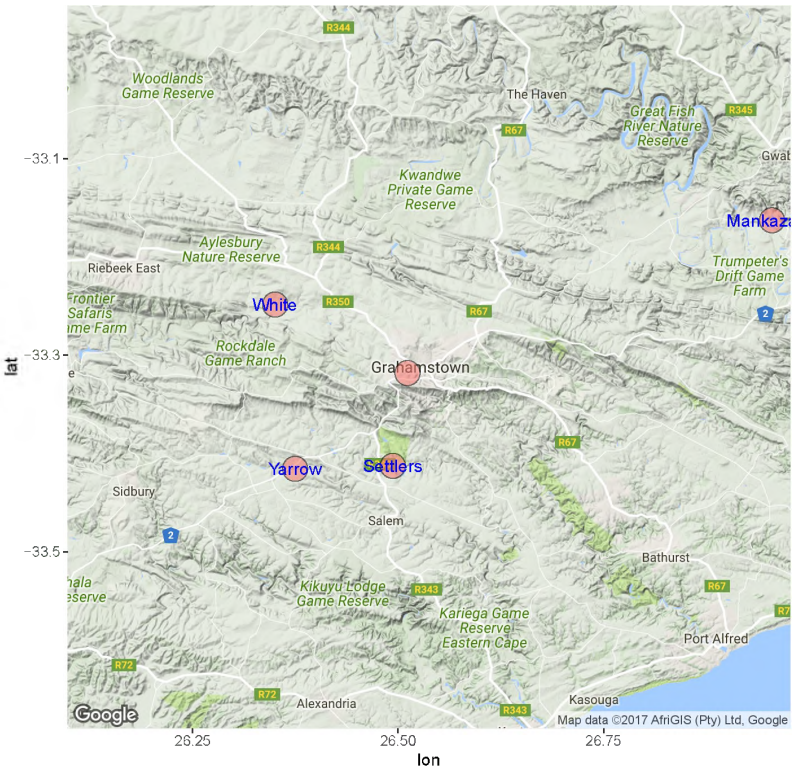
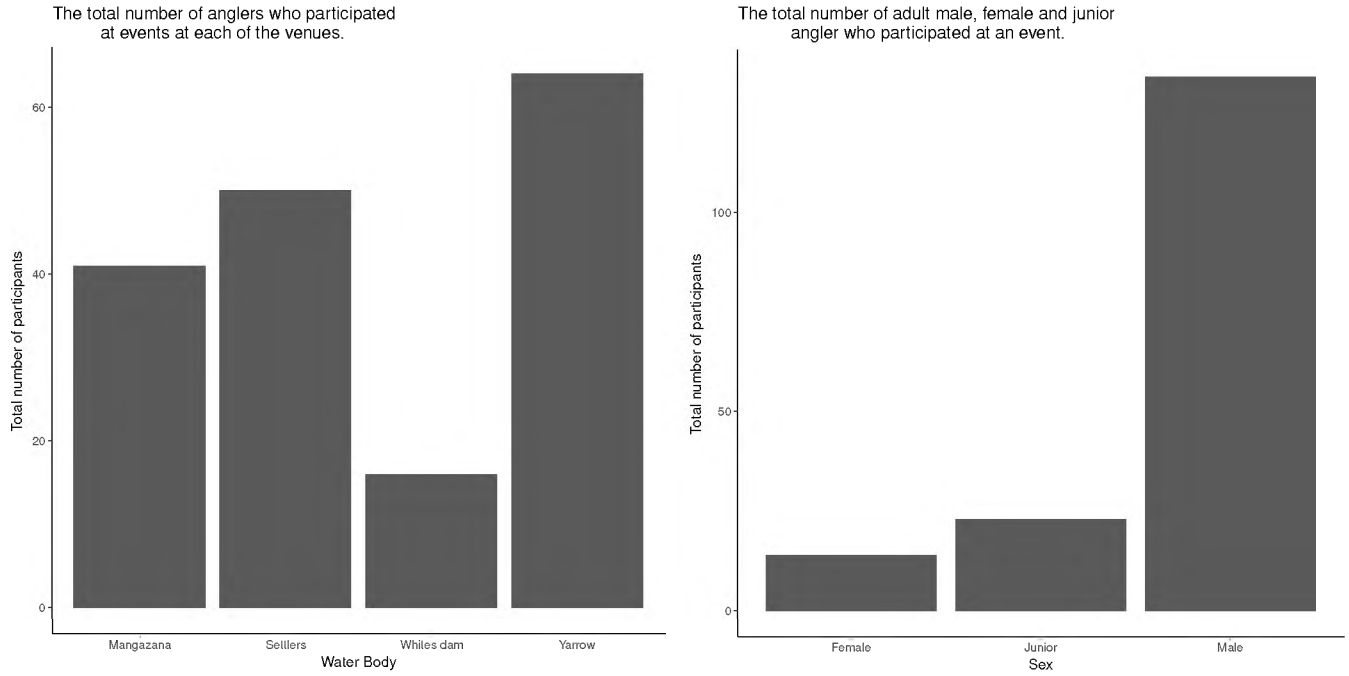


Figure 5.1: Map of the AAA tournament venues, March 2015 to October 2016.

Descriptive Statistics: The AAA Data Set

A total of one hundred (100) observations were made for events held in 2016 and seventy one (71) for the events held in 2015. One hundred and thirty four (134) of these observations are of adult males, fourteen (14) of adult females and twenty three (23) of juniors (figure 5.2 (b)). Six events were held in 2015 and 2016. Sixty four (64) observations were made at Yarrow dam, fifty (50) at Settlers dam, forty one (41) at Mangazana and sixteen (16) at White's dams (figure 5.2 (a)). Table 5.1 summarizes the climatic conditions and composition of the anglers at these events.



(a) Venue. **(b)** Group.

Figure 5.2: Bar graph of the total number of participants by venue and sex.

Testing the normality of the residuals of one-way ANOVA model for fish weight, using the Multivariate Shapiro-Wilk test revealed that the fish CPUE data are not normally distributed ($MVW = 0.915$, p-value < 0.001). The normal Q-Q plots, figure 5.3, have been used to assess the normality of the various fish weight variables in the data set. The normal Q-Q plots suggest that these variables are not normally distributed populations. The boxplots, figure 5.4, show outliers and fish two weight and fish three weight are positively skewed. The boxplots by water bodies are shown in figure 5.5, they are asymmetric with unequal medians. Figure 5.6 shows histogram for each fish number weight. The data shows non-symmetric distributions of all the fish weights and bag weight. The histogram of fish two weight showed an outlier that falls between 3.0 and 3.5 while the histogram for fish three weight had the highest frequency of 70 for no fish.

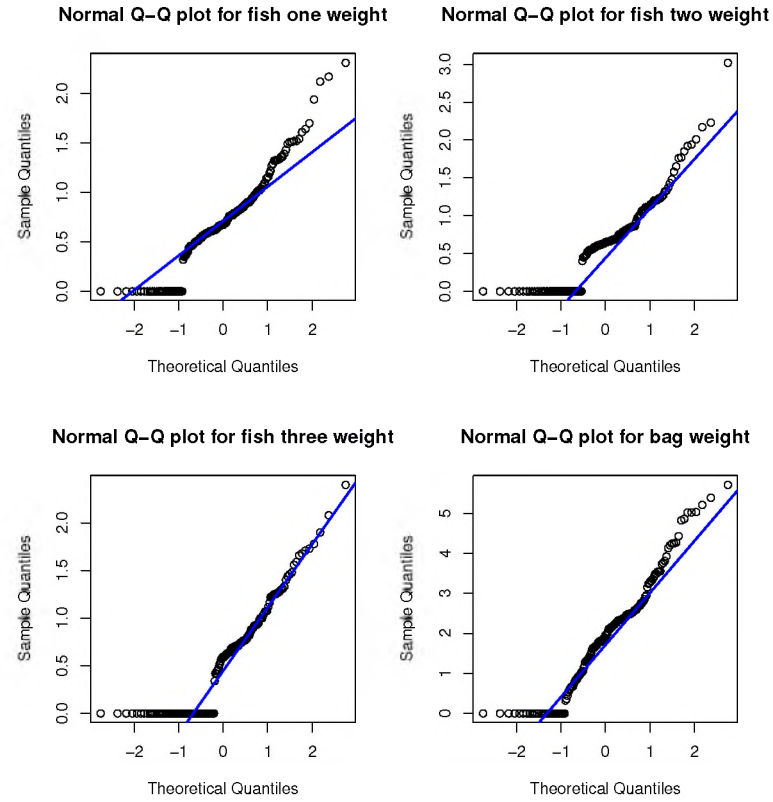


Figure 5.3: Normal Q-Q plots of each fish weighed and total bag weight.

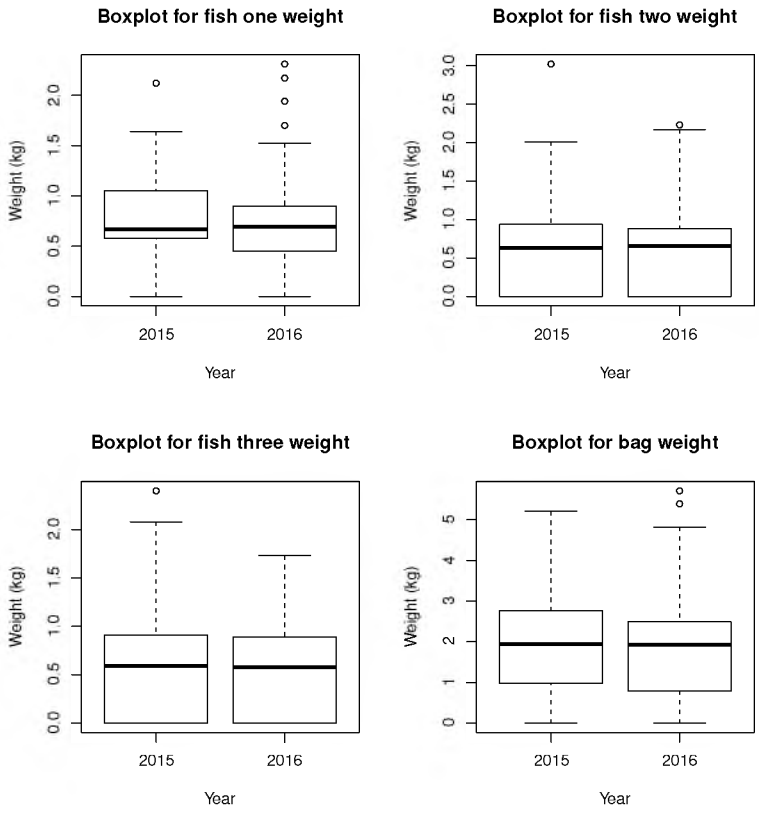


Figure 5.4: Boxplots of the weight of each fish weighed and the total bag weight for the tournaments in 2015 and 2016.

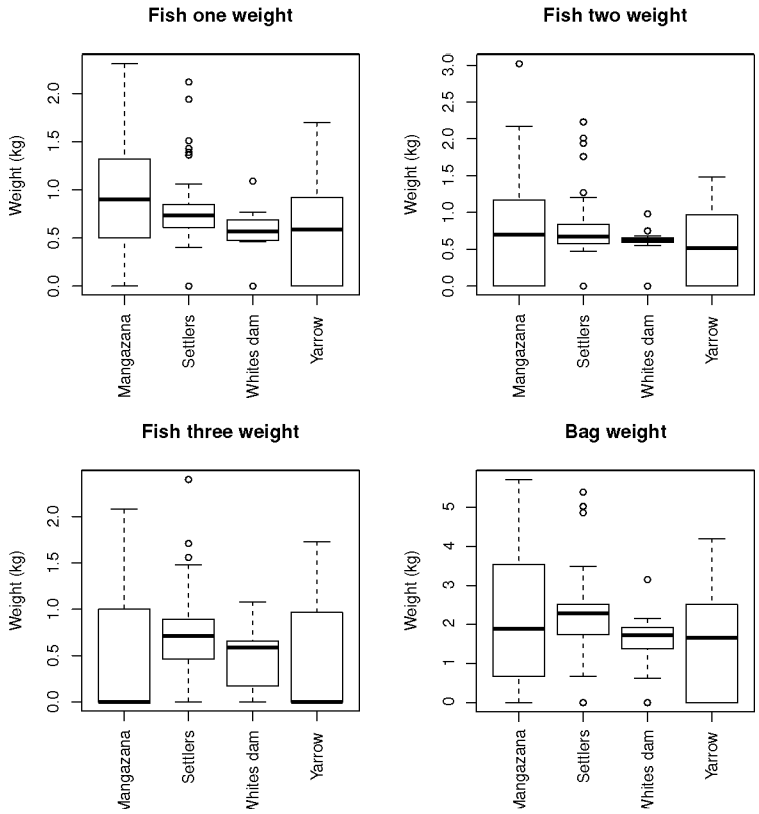


Figure 5.5: Boxplots of the weight of each fish weighed and the total bag weight in the different water bodies.

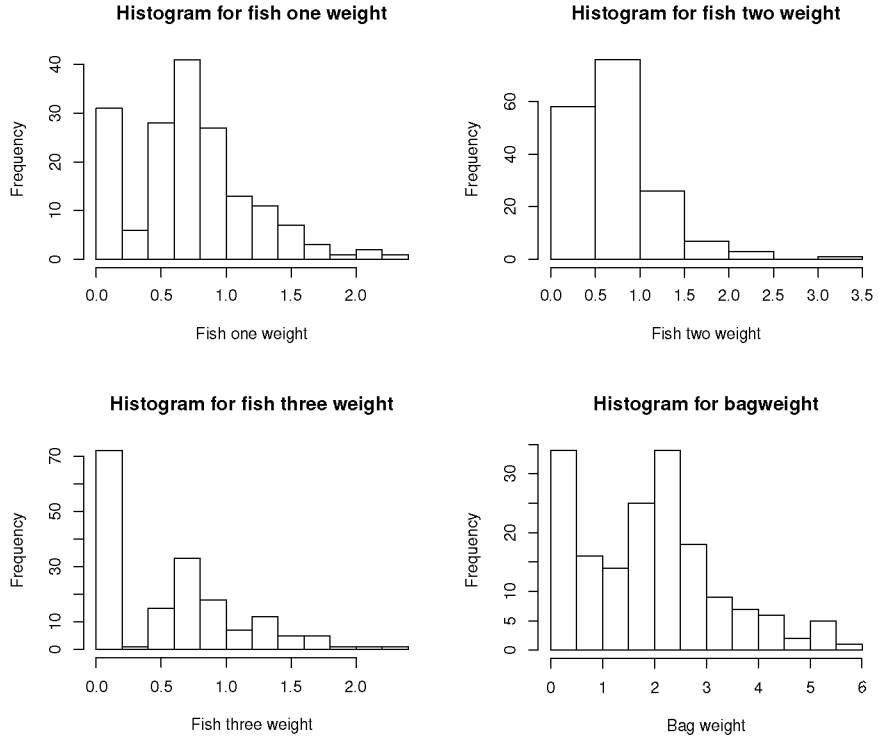


Figure 5.6: Histogram of the weight of each fish weighed and the total bag weight for the fish weight.

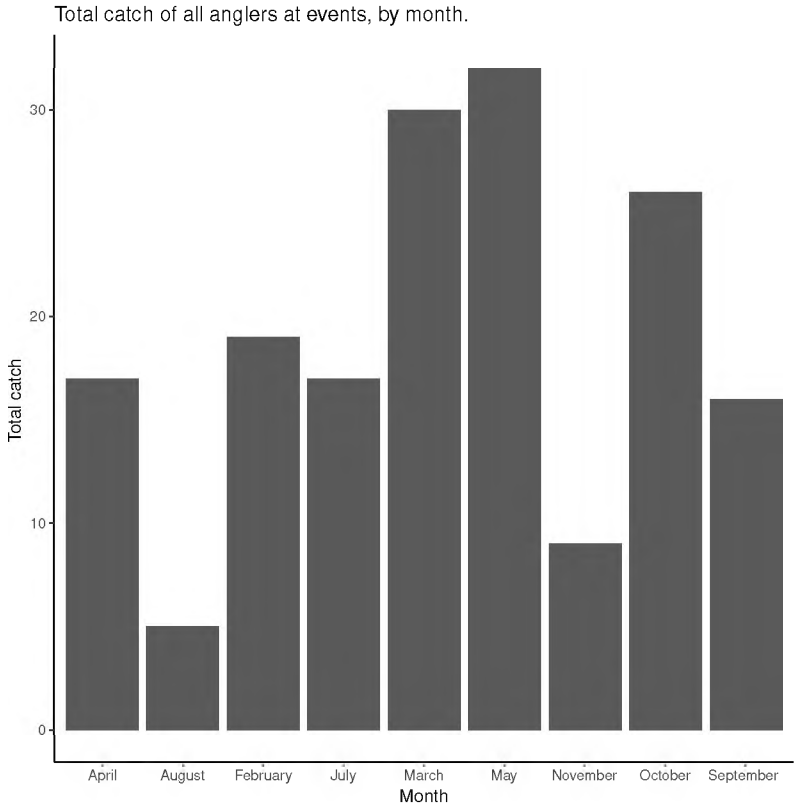


Figure 5.7: Bar graph of the total catch at events, by month.

Four fishing events were held at Yarrow dam, one in 2015 and three in 2016. Four events were held at Settlers dam, three in 2015 and one in 2016. Three events were held at Mangazana dam, two in 2015 and one in 2016. One event was held in White's dam in 2016. The number of anglers, standard error of the bag weight and a 95% confidence interval for the true average bag weight for each event are shown in table 5.3. The number of fish caught per angler at each event are shown in table 5.4. The lowest number of anglers participated at events held at White's dam. The most fished dam was Yarrow (37.43%) followed by Settlers (29.24%) and Mangazana (23.98%).

Water body	Total number of anglers	Relative frequency (%)
Yarrow	64	37.43
Settlers	50	29.24
Mangazana	41	23.98
White's dam	16	9.36
Total	171	

Table 5.2: Total number of participants, by venue.

Water body	Date	Sample size	Mean bag weight	Standard error	95% CI for μ	
Yarrow	17 May 2015	14	0.7800	0.3017	0.1283	1.4317
	20 Mar 2016	15	2.3000	0.2456	1.7733	2.8268
	22 May 2016	18	1.4833	0.3339	0.7788	2.1879
	24 Jul 2016	17	1.6571	0.3018	1.0173	2.2968
Settlers	9 Aug 2015	5	2.2240	0.7738	0.0756	4.3724
	11 Oct 2015	11	2.3418	0.2845	1.7078	2.9758
	15 Mar 2015	15	2.2627	0.2706	1.6822	2.8431
	21 Feb 2016	19	2.0453	0.2974	1.4205	2.6691
Mangazana	12 Apr 2015	17	1.9453	0.3631	1.1755	2.7151
	22 Nov 2015	9	3.4211	0.5494	2.1541	4.6881
	23 Oct 2016	15	1.6793	0.4917	0.6247	2.7339
White's	25 Sep 2016	16	1.6250	0.1739	1.2544	1.9956

Table 5.3: Summary statistics of the bag weight at the various events.

Yarrow Dam			
17 May 2015			
0 fish	1 fish	2 fishes	3 fishes
8	3	1	2
20 Mar 2016			
0 fish	1 fish	2 fishes	3 fishes
1	0	1	13
22 May 2016			
0 fish	1 fish	2 fishes	3 fishes
5	3	4	6
24 Jul 2016			
0 fish	1 fish	2 fishes	3 fishes
3	4	2	8

(a) Yarrow dam.

Settlers Dam			
9 Aug 2015			
0 fish	1 fish	2 fishes	3 fishes
0	1	2	2
11 Oct 2015			
0 fish	1 fish	2 fishes	3 fishes
0	0	1	10
15 Mar 2015			
0 fish	1 fish	2 fishes	3 fishes
0	1	2	12
21 Feb 2016			
0 fish	1 fish	2 fishes	3 fishes
4	0	0	15

(b) Settlers dam.

Mangazana			
12 Apr 2015			
0 fish	1 fish	2 fishes	3 fishes
3	2	3	9
22 Nov 2015			
0 fish	1 fish	2 fishes	3 fishes
1	1	1	6
23 Oct 2016			
0 fish	1 fish	2 fishes	3 fishes
5	4	2	4

(c) Mangazana dam.

White's dam			
25 Sep 2015			
0 fish	1 fish	2 fishes	3 fishes
1	1	2	12

(d) White's dam.

Table 5.4: The number of anglers and their bag size at the various events.

5.2 Repeated Measures

An experimental unit is the smallest unit experimental material to which a factor or combination of factors may be applied (Larson, 2008). For example twenty rats are randomly assigned to each of four doses of a potential carcinogen: none, low, medium and high. The rats are kept in individual cages under the same environmental conditions in the same room. Each rat has its assigned dose stirred into its daily meal for four weeks. The number of tumors found in each rat is recorded at the end of the four week period (McDonald, 2009, page 128). In repeated measures designs, the experimental unit could be a person or a species of animal where the repeated measurements are taken sequentially in time or repeated surveys conducted under different experimental conditions. For example consider a situation where a researcher tests subjects before fitness starting an experiment and after two weeks, four weeks and six weeks of endurance training.

The term repeated measures refers to experimental designs where there are several individuals and several measurements taken on each individual, with more than one observatory on the same individual or sampling unit (Littell et al., 2000). Repeated measures is the term used when the same entities or sampling units participate in all conditions of an experiment or provide data at multiple times (Field et al., 2012, page 550). Classical ANOVA methods discussed in chapter 2 are typically applied when different sampling units, for example people, take part under different experimental conditions (Field, 2013, page 550).

If repeated measurements of the same unit are taken it is not unreasonable to assume that the observations of the same unit are correlated (Littell et al., 2000). Repeated measures ANOVA is primarily concerned with the within-subjects effects and are often referred to as within-subjects designs (Verma, 2015, page 73). The measurements might be affected by within-subject characteristics such as age or genetic factors. In a repeated measures ANOVA, we wish to test the hypotheses that there are significant differences in means over time. For example when testing the effect of alcohol on a persons ability to drive, it is not unreasonable to assume that the participatory drivers have different tolerances to alcohol and that this affects each drivers driving ability, where each participants ability to drive is assessed after consuming two units of alcohol, after three units of alcohol etc. Statistical analysis of repeated measures data thus need to address the covariance between measures of the same sampling unit as ignoring or avoiding the covariance may result in ineffective influence (Littell et al., 2000). The main objective of repeated measures analysis is to model within subjects variance which describes changes in the average response over the repeated measures, for example time, and assesses how these changes are related to covariates of interest, for example the number of units of alcohol consumed (Fitzmaurice et al., 2012, page 611). Quite clearly the repeated measurements of the drivers ability are not independent, since the observations are made on the same subject, and hence violate the assumption of the classical ANOVA methods. As a result the conventional or classical ANOVA F tests will lack accuracy (Field, 2013, page 551).

The relationship between the repeated measurements of the response under the different treatment levels or conditions is typically assumed to be similar between pairs of experimental conditions, that is the level of dependence between experimental conditions is approximately equal. Homogeneity, or equality, of variances over time is the most important assumption underlying repeated measures ANOVA (McDonald, 2009, page 156). This assumption is termed sphericity (Field, 2013, page 551). Sphericity refers to the equality of variances of the differences between treatment levels and is a less restrictive form of compound symmetry which refers to the scenario where the variance across conditions are equal and the covariances between pairs of conditions are equal (Field, 2013, page 551). Repeated measures analysis is not robust to this assumption so when there is violation power decreases and a corresponding increase in probability of a type II error occurs (Verma, 2015, page 20). When the sphericity assumption is not satisfied, inferences using procedures that make the sphericity assumption will be incorrect. Mauchly's test assesses the hypothesis that the variances are significantly different. The sphericity assumption is only relevant when there are more than two levels of the within-subjects factor (Quinn & Keough, 2002, page 282). When repeated measures are nested within the experimental units the experimental design is referred to as a hierarchical design or multilevel model or mixed-effects models (Wagner et al., 2006). Multilevel data structures have a hierarchical structure in which the response variables are measured at the lowest level of the hierarchy and these responses are modeled as a function of predictor variables measured at this level and higher levels of the hierarchy, for example measurements taken on individual fish (the lowest level of the hierarchy) that are nested within lakes (or dams) or streams, the higher level of the hierarchy (Wagner et al., 2006). Multilevel or mixed-effects models overcome the estimation issues associated with ordinary least squares applied to multilevel models in that they estimate standard errors correctly and hence result in improved estimation of the fixed effects components in the multilevel data structures (Wagner et al., 2006).

In the AAA data, the number of fish an angler weighs or the total bag weight of an angler are repeated measurements since these anglers fished multiple competitions. However these response variables are nested within location, that is the dam at which the competition took place. It is likely that the density of fish and the nature of the population of these fish, that is if these fish are typically large or heavy fish or small or lighter fish, are correlated with the location or water body and hence this variable can be considered as a random effect. The catch, namely the number of fish or the total weight of these fish, is nested within the location, modeled as a random effect. In addition these locations or dams represent a random sample from the, reasonable large, population of dams in the Eastern Cape and hence the results of the analysis can be generalized to other dams in the Eastern Cape.

Mixed-effects models make use of all the available data and account for correlation between repeated measurements on the same subjects (Lindstrom & Bates, 1990). Mixed-effects models handle missing data more appropriately than generalized linear models (Pinheiro &

Bates, 2000, page 133). In a mixed-effects model, each individual's vector of responses is modeled as a parametric function, hence the effects are random variables with a multivariate normal distribution (Lindstrom & Bates, 1988). Lindstrom & Bates (1990) proposed a method of estimating the parameters which are difficult to compute manually. This method implements the Expectation Maximization (EM) and Newton-Raphson (NR) algorithms for matrix decomposition and also for estimating parameters in mixed-effects models for repeated measures data (Lindstrom & Bates, 1988).

The repeated measures on the same individual are expected to be positively correlated with each other. The n repeated measures of sampling unit i are collected into the vector $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})'$. Define the variance-covariance matrix to be the 2-dimensional array of variances and covariances, namely;

$$Cov\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} = \begin{pmatrix} Var(Y_{i1}) & Cov(Y_{i1}, Y_{i2}) & \ldots & Cov(Y_{i1}, Y_{in}) \\ Cov(Y_{i2}, Y_{i1}) & Var(Y_{i2}) & \ldots & Cov(Y_{i2}, Y_{in}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Y_{in}, Y_{i1}) & Cov(Y_{in}, Y_{i2}) & \ldots & Var(Y_{in}) \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{in} \\ \sigma_{21} & 22 & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_{nn} \end{pmatrix}$$

where $Cov(Y_{ij}, Y_{ik}) = \sigma_{jk}$. The correlation, unlike the variance, is a measure of dependence that is free of the scales of measurements of Y_{ij} and Y_{ik} (Fitzmaurice et al., 2012, page 29).

5.3 Classical Repeated Measures ANOVA

Repeated measures ANOVA designs require that the outcome variable is quantitative and, as a result of the assumptions made of the residual term, normally distributed (Verma, 2015, page 22). The covariates are required to be discrete, or qualitative, variables (Littell et al., 2000). Repeated measures ANOVA designs typically require the sphericity assumption. The assumption of constant correlation of repeated measures is often unrealistic as repeated measures often become less correlated with increasing time from treatment (Sullivan, 2008). Repeated measures ANOVA models can only handle longitudinal studies in which all subjects have the same number of repeated measurements (Verma, 2015, page 22). The purpose of repeated measures ANOVA is to investigate the behavioral trend of subjects in relation to the criterion or covariable variables over a period of time (Verma, 2015, page 73). Repeated measures ANOVA relates the dependent variable to a set of covariates, the treatment groups, over time and compares the mean outcome at multiple time points or between groups (Littell et al., 2007, page 160). Dependency, or correlation, among responses measured in the same experimental unit is the defining feature of a repeated measures design (Sullivan, 2008). Let Y_{ij} denote one-way mixed-effects repeated measures ANOVA model given by (Davis, 2002, page 104)

$$Y_{ij} = \mu_{ij} + \alpha_{ij} + \varepsilon_{ij}$$

where μ_{ij}, a fixed effect, denotes the mean at time j for the i^{th} subject randomly selected from the population and α_{ij}, a random effect, denotes the consistent departure of Y_{ij} from μ_{ij} for the i^{th} subject. Under hypothetical repetitions from the same individual, Y_{ij} has the mean $\mu_{ij} + \alpha_{ij}$. The last component, denoted by ε_{ij}, represents the departure of Y_{ij} from $\mu_{ij} + \alpha_{ij}$ for individual i at time j. At time j the means and variances of the random effects, α_{ij}, are assumed to be $E(\alpha_{ij}) = 0$ and $Var(\alpha_{ij}) = \sigma^2_{\alpha_j}$. As a result any non-zero mean is absorbed in μ_{ij} and the variance at time j is constant over individuals. In addition at given time j, it is assumed that $E(\varepsilon_{ij}) = 0$ and $Var(\varepsilon_{ij}) = \sigma^2_{\varepsilon_j}$.

Consider a situation where subjects are randomly assigned to one of g treatments and measurements are made at t equally spaced times on each subject. Let Y_{ijk} denote the measurement at time k on the j^{th} subject assigned to treatment i. A statistical model for these repeated measures might be given by

$$Y_{ijk} = \mu + \alpha_i + \gamma_k + (\alpha\gamma)_{ik} + \varepsilon_{ijk} \text{ where } i = 1, \ldots, g,\ j = 1, \ldots, n_i \text{ and } k = 1, \ldots, t.$$

In this model $\mu+\alpha_i+\gamma_k+(\alpha\gamma)_{ik}$ denotes the mean for treatment i at time k, containing effects for treatment, time and the treatment by time interaction. ε_{ijk} denotes the random error associated with the measurement at time k on the j^{th} subject that is assigned to treatment i (Littell et al., 2007, page 161).

One-Way Repeated Measures ANOVA

Consider a repeated measures design where repeated measures are taken over one factor, for example testing a persons reaction time after the consumption of one, two,...,g units of alcohol. Similarly a researcher might be interested in assessing the effect of g treatments by testing the effect of these treatments on the experimental units. Students might be concerned about the consistency of assessment between different lectures or assessors. Suppose that ten randomly selected essays from a class are each assessed by four lectures and the marks recorded. In this design each subject, or essay in this example, is exposed to all levels of a qualitative variable, namely the lectures who mark each essay. The response variable is a qualitative variable, that is the mark assigned to each essay by each lecturer. In this example the mark for essay i assigned by lecturer j could be modeled as

$$Y_{ij} = \mu_{ij} + \alpha_{ij} + \varepsilon_{ij}$$

where Y_{ij} denotes the mark for essay i, where $i = 1, \ldots, 10$, awarded by lecturer j, where $j = 1, \ldots, 4$. In this context Y_{ij} is a continuous, normally distributed response variable measured at t time points, or occasions, that is there are 4 repeated observations of essay i, for each of 10 essays or experimental units. A one-way repeated measures ANOVA aims is to compare the treatments with respect to differences in the outcome. The treatment

factor is a between-subjects factor and has no repeated measures. Here repeated measures or assessments are taken on each subject within each treatment over time. A repeated measures ANOVA can be used to test for significant differences in the means over time that is in the average mark.

The following assumptions must be assessed before using this design (Verma, 2015, page 24):

1. The independent variable should be categorical and the dependent variable should be measured on an interval or ratio scale;

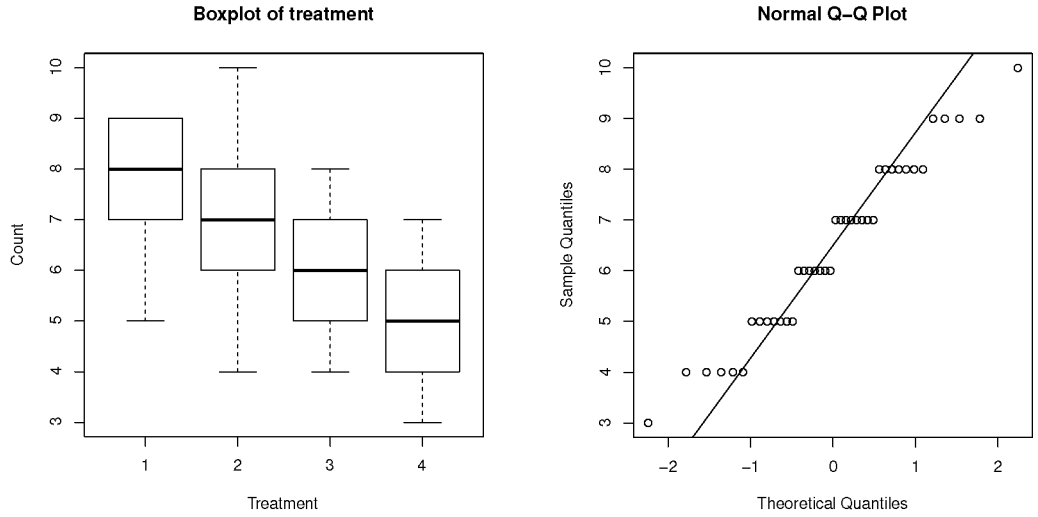
2. Observations obtained on the dependent variable must be independent from each other;

3. The data for the dependent variable obtained on the subjects in each treatment condition must follow a normal distribution; and

4. Sphericity should exist among the data. The sphericity assumption is satisfied if correlations among the repeated measurements of dependent variables are all equal.

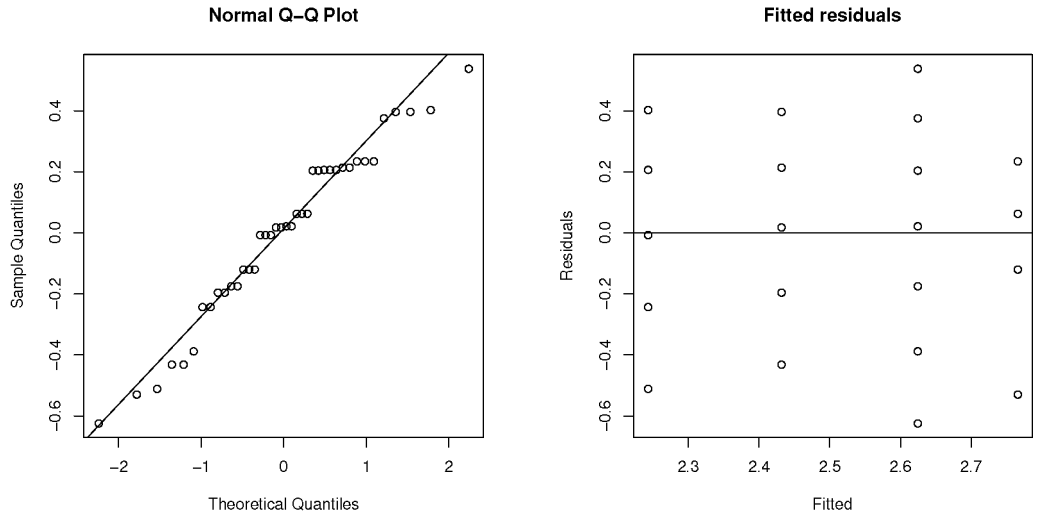
One-Way Repeated Measures ANOVA Example

Consider a scenario where we have four levels of treatment groups from the same subject. In this scenario, a treatment group might be a medical treatment where same subjects are assigned to each treatment group but the outcome is measured repeatedly over time (Paul, 2016). The goal is to compare treatments with respect to the differences in the outcome. We may wish to test for treatment effects. The subject can be included as a factor and the design is a one-way repeated measures design. The interaction effect is not included since there is only one observation of each subject in each condition and hence there are not enough degrees of freedom. The boxplots, figure 5.8 (a), show evidence that homogeneity of variance assumption is violated since the boxplots are of unequal size. However the hypothesis appears to meet the homogeneity of variances assumption (Bartlett's K-squared $= 1.9759$, $df = 3$, p-value $= 0.5774$). The normal Q-Q plot, figure 5.8 (b), show that these data do not meet the normality and linearity assumptions since the points deviates away from the theoretical quantiles. The square root transformation was applied to the dependent variable to improve normality, however there is clear evidence that homogeneity of variance is likely violated, figure 5.9 (a) and (b). The Shapiro-Wilk test revealed that the residuals of the square root transformed data do not violate the normality assumption ($W = 0.96582$, p-value=0.2634).



(a) Boxplots of the treatment data. **(b)** Normal Q-Q plot.

Figure 5.8: Graphical assessment of the normality, linearity and homogeneity of variance assumptions.



(a) Normal Q-Q plot. **(b)** Residuals versus fitted values.

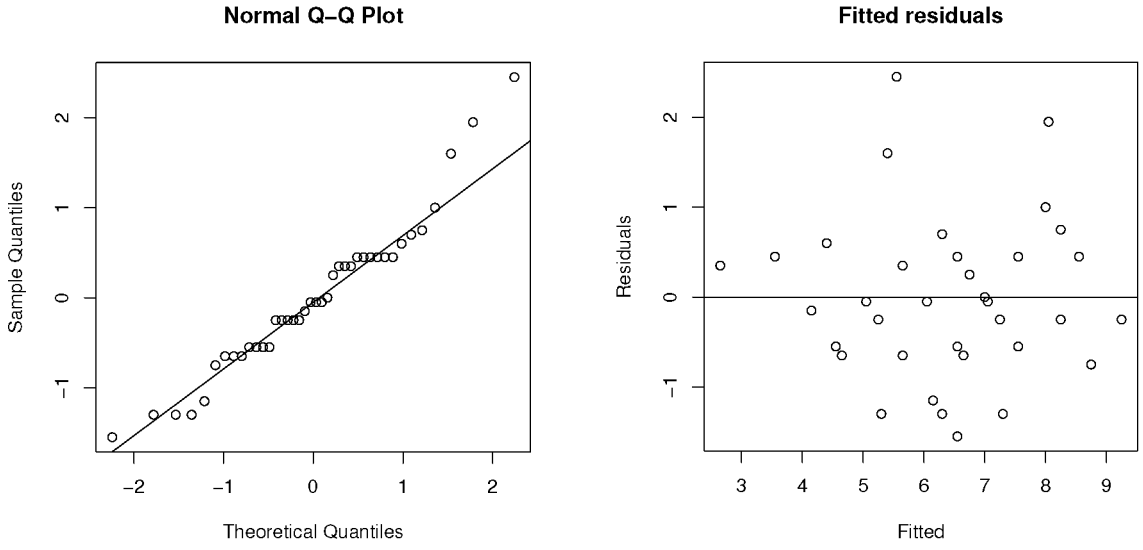
Figure 5.9: Model diagnostics plots for one-way repeated measures design, square root transformed dependent variable.

The univariate (subject/treatment) notation in R indicates that the treatment factor is the repeated measures factor over the subjects. The model has two main effects and no interaction effect as shown in table 5.5. Both main effects, treatment ($F_{obs} = 12.241$, $df = 3, 27$, p-value $= 3.06e-05$) and within-subjects ($F_{obs} = 5.077$, $df = 9, 27$, p-value $= 0.0005$) are highly significant. The multivariate approach method is used to perform ANOVA when there is repeated measures in the data (Paine, 1996). One benefit of using multivariate approach is that it provides ways of testing the sphericity assumption as well as providing corrected

versions of the F test assuming the sphericity assumption has been violated. Mauchly's test for sphericity is insignificant ($T_{obs} = 0.30915$, p-value $= 0.10905$) as shown in table 5.5. Both Greenhouse-Geisser and Huynh-Feldt Corrections estimates suggest that sphericity was not met since they are both less than 0.75 (Logan, 2011, page 368). The normal Q-Q plot, figure 5.10 (a), suggest that the residuals for this model meet the normality assumption since the points do not deviate from the theoretical quantiles. The fitted residuals, figure 5.10 (b), shows that the residuals of this model meet the homogeneity of variance assumption since the points are randomly scattered with no particular pattern. The sphericity assumption was not met (as discussed above).

Source of Variation	df	Sum of Squares	Mean Square	F Statistics	p-value	
Treatment	3	38.9	12.967	12.241	$3.06e-05$	
Within subject	9	48.4	5.378	5.077	0.0005	
Residuals	27	28.6	1.059			
Mauchly test for Sphericity						
Source	t-test	p-value	GG eps	GG-p	HF eps	HF-p
rfactor	0.30915	0.10905	0.71657	0.0002157	0.9487151	$3.066742e-05$

Table 5.5: One-way repeated measures ANOVA example.



(a) Normal Q-Q plot.

(b) Residuals versus fitted values.

Figure 5.10: One-way repeated measures ANOVA model diagnostics plots.

Two-Way Repeated Measures ANOVA

Consider a repeated measures design where repeated measures are taken over two factors, the two factors simply means that two independent variables are manipulated in the experiment. For example do men and women want different things in relationships? The participants

are heterosexual women who came to speed dating night and over the course of the evening they speed dated all nine men. Each woman rated nine different people who varied in their attractiveness and personality (Field, 2013, page 8). There are two repeated measures variables: looks with three levels namely attractive, average or ugly and personality with three levels namely lots of charisma, have some charisma or be dull. In a repeated measures design, two or more treatments are applied to n replicates several times. The ANOVA table 5.6 represent a simple repeated measures design with k treatments applied to n replicates per treatment and measurements are made on each replicate at t times. In most repeated measures designs, the treatment by time interaction, not the overall treatment effect, is the primary effect of interest since the interaction is generally the best test of influences (Field, 2013, page 8). In a two-way repeated measures ANOVA, we combine each independent variable with its time interval resulting in columns for each pairing. The two-way repeated measures ANOVA model is given by

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \gamma_{i(j)} + \varepsilon_{ijk}$$

where Y_{ijk} represents the observations taken at the k^{th} time point from j^{th} group of the i^{th} subject, α_j denotes the main effect of factor A subject to $\sum \alpha_j = 0$, β_k denotes the main effect of factor B subject to $\sum \beta_k = 0$, $\gamma_{i(j)}$ denotes the effect of subject nested in factor A, $(\alpha\beta)_{jk}$ denotes the interaction effect of factor A and factor B and ε_{ijk} denotes the error term which is assumed to follow a normal distribution, $\varepsilon_{ijk} \sim N(0, \sigma^2)$. A two-way ANOVA table with repeated measures is shown in table 5.6. The two-way repeated measures ANOVA computations are given by the following equations:

$$SSA = nb \sum_{j=1}^{b} (\bar{Y}_{j.} - \bar{Y})^2$$

$$SSB = na \sum_{i=1}^{a} (\bar{Y}_{..k} - \bar{Y})^2$$

$$SSWA = b \sum_{j} \sum_{k}^{n} (\bar{Y}_{j.} - \bar{Y})^2$$

$$SSAB = n \sum_{j} \sum_{k}^{n} (\bar{Y}_{.jk} - \bar{Y}_{j.} - \bar{Y}_{..k} + \bar{Y})^2.$$

Source of Variation	df	SS	MS	F Statistics
Factor A	$a-1$	SSA	$MSA = \frac{SSA}{a-1}$	$\frac{MSA}{MSE}$
Factor B	$b-1$	SSB	$MSB = \frac{SSB}{b-1}$	$\frac{MSB}{MSE}$
AB interaction	$(a-1)(b-1)$	$SSAB$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$\frac{MSAB}{MSE}$
Subjects within A	$a(n-1)$	$SSWA$	$MSWA = \frac{SSWA}{a(n-1)}$	$\frac{MSWA}{MSE}$
Error	$a(n-1)(b-1)$	SSE	$MSE = SSE/a(n-1)(b-1)$	
Total	$abn-1$	SST		

Table 5.6: Two-way repeated measures ANOVA table.

The two-way repeated measures ANOVA assumptions are similar to that of one-way repeated measures ANOVA, namely:

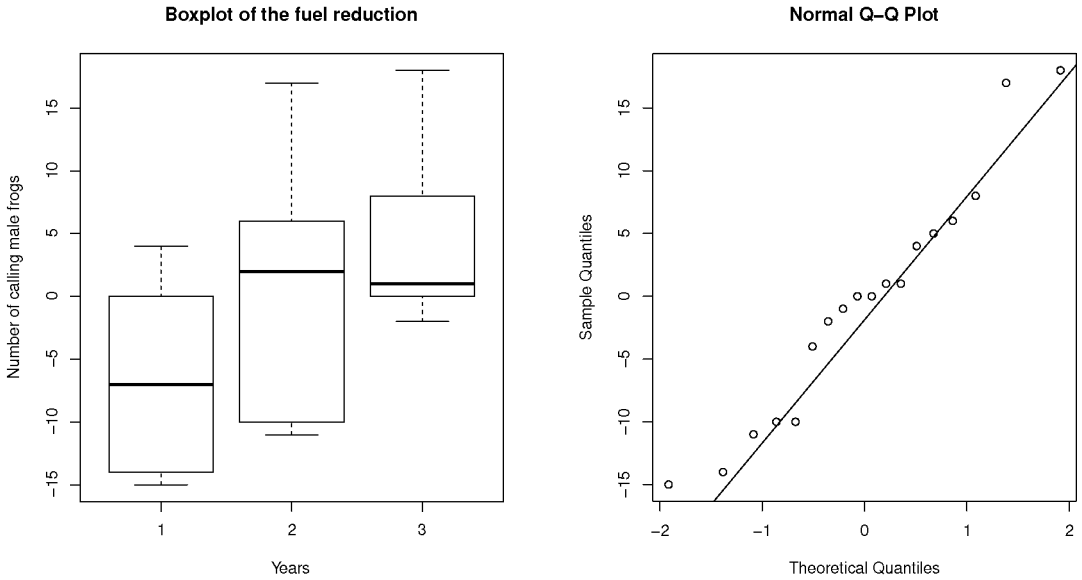
1. The independent variable should be categorical and the dependent variable should be measured on an interval or ratio scale;
2. Observations obtained on the dependent variable must be independent from each other;
3. The data for the dependent variable obtained on the subjects in each treatment condition must follow a normal distribution; and
4. Sphericity should exist among the data. The sphericity assumption is satisfied if correlations among the repeated measurements of dependent variables are all equal.

In this design, one needs to test the normality and sphericity assumptions. The observations of the dependent variable obtained on the subjects in each treatment combination must be normally distributed.

Two-Way Repeated Measures ANOVA Example: One Observation Per Cell

Researchers examined the effects of fuel reduction burning on the abundance of a species of frog in Western Australia. They used six drainages within a catchment which represent the subjects or blocks. In each drainage they had a matched burnt site and unburnt site and the response variable for the experiment was the difference in the number of calling male frogs between the burnt and un-burnt site in each drainage. This variable was recorded three times for the pre-burn experiment in 1992 and two times for the post-burn experiment in 1993 and 1994. The hypothesis of interest in this classical repeated measures design was that there was no difference between years in the mean difference in the number of calling male frogs between burnt and unburnt catchments (Quinn & Keough, 2002, page 266). The boxplots, figure 5.11 (a), shows no evidence of unequal variance hence the homogeneity of variance assumption for this data is met. There is evidence that these data meet the homogeneity of variances (Bartlett's K-squared $= 0.7264$, $df = 2$, p-value $= 0.6954$). The normal Q-Q plot, figure 5.11 (b), shows that these data meet the normality assumptions since the points do not

deviate from the theoretical quantiles line. The resulting ANOVA table, table 5.7, indicates that there was a significant effect in the years ($F_{obs} = 9.66$, $df = 2$, p-value $= 0.0046$) , that is time prior or post fuel reduction burn, on the differences in number of males calling between burnt and unburnt sites. This test indicates significant variation between drainages ($F_{obs} = 9.99$, $df = 5, 10$, p-value $= 0.001$). The normal Q-Q plot, figure 5.12 (a), shows that the residuals of the model do not meet the normality assumption since the points deviates from the theoretical quantiles line. The residuals of this model are normally distributed (Shapiro test $W = 0.93878$, p-value $= 0.4824$). The boxplot, figure 5.12 (b), shows that the residuals of this model do not meet the homogeneity of variance assumption since the boxplot is asymmetric. There is homogeneity of variances in the number calling male frogs (Bartlett's K-squared $= 2.5346$, $df = 5$, p-value $= 0.7713$).

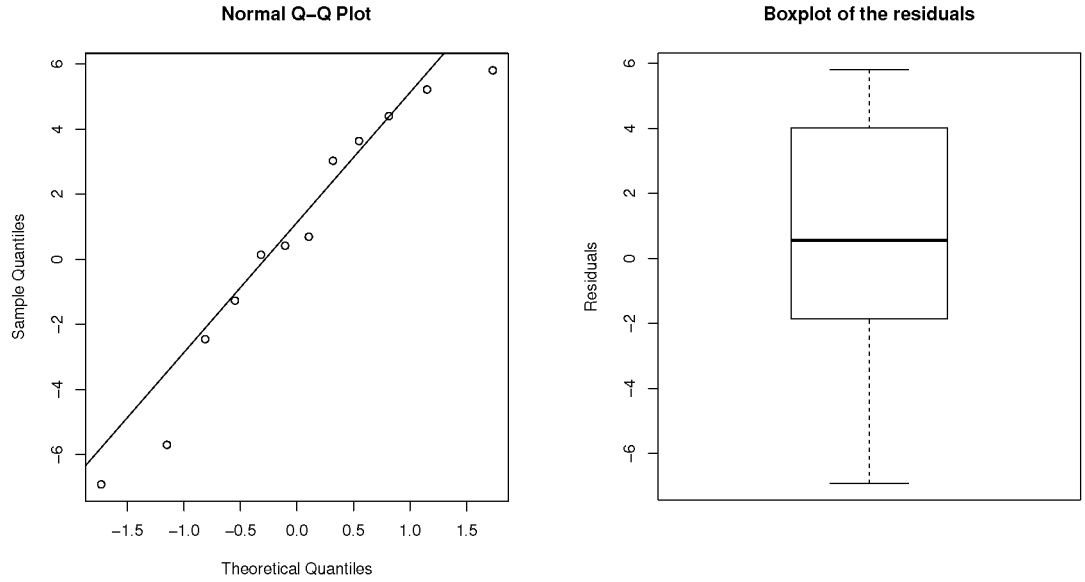


(a) Boxplots of the number of calling male frogs. **(b)** Normal Q-Q plot.

Figure 5.11: Graphical assessment of the number of calling male frogs.

Source of Variation	df	Sum of Squares	Mean Square	F Statistics	p-value
Year	2	369.4	184.72	9.66	0.0046
Block (drainage)	5	955.6	191.1	9.99	0.001
Residuals	10	191.2	19.12		

Table 5.7: Two-way repeated measures ANOVA: Frog example.



(a) Normal Q-Q plot. **(b)** Boxplot of the residuals.

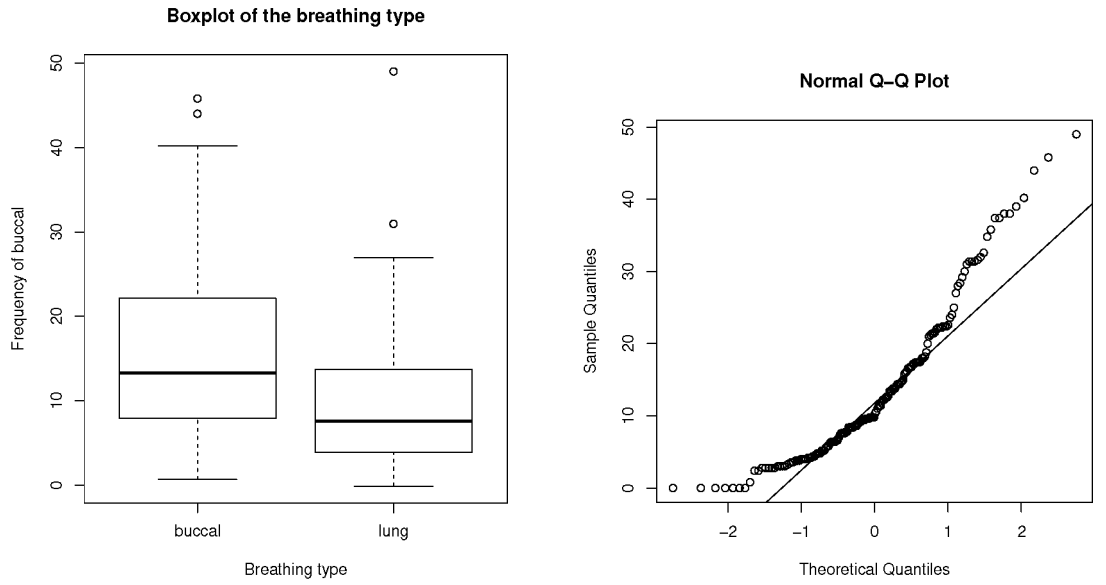
Figure 5.12: Diagnostics plots: Frog example.

Two-Way Repeated Measures ANOVA Example: Including Interaction

Consider a scenario where a researcher is investigating the impact of oxygen stress on the ventilation patterns of cane toads. In anticipation of variability in ventilation patterns between individual toads, each oxygen concentration level was measured for each individual. As a result the individual toads represent the blocks (toads) and the oxygen levels represent a within block treatment. Individual toads were also categorized according to their typical predominant mode of breathing and therefore breathing type represents a between block treatment. Ventilation patterns were measured as the frequency of breathing (Logan, 2011, page 421). The response variable is given by the frequency of buccal breathing in the data set. The boxplot, figure 5.13 (a), suggest no significant difference in average frequency of buccal breathing. The normal Q-Q plot, figure 5.13 (b), shows evidence that the data do not meet the normality assumption since the points deviates far away from the theoretical quantile line. The summary statistics of breathing type, oxygen level, standard error of the frequency and 95% confidence interval for the true average frequency for each ventilation are shown in table 5.8.

Breathing type	Oxygen level	Sample size	Mean	Standard error	95% CI of μ	
Buccal	0	13	4.9401	1.1771	4.2297	5.6523
Buccal	5	13	4.7236	1.1669	4.0185	5.4287
Buccal	10	13	4.3934	0.9779	3.8024	4.9843
Buccal	15	13	4.0723	1.1978	3.3484	4.7961
Buccal	20	13	3.3966	1.1401	2.7071	4.0861
Buccal	30	13	3.4766	1.4045	2.6278	4.3253
Buccal	40	13	2.6953	0.9291	2.1339	3.2568
Buccal	50	13	2.4809	0.9212	1.9243	3.0376
Lung	0	8	1.5496	1.4731	0.3180	2.7811
Lung	5	8	2.3684	1.7281	0.9221	3.8139
Lung	10	8	3.3347	1.5601	2.0304	4.6381
Lung	15	8	3.0538	1.5397	1.7665	4.3410
Lung	20	8	3.3036	1.0480	2.4275	4.1798
Lung	30	8	3.0619	0.9709	2.2502	3.8736
Lung	40	8	2.9207	1.7136	1.4882	4.3533
Lung	50	8	2.5565	0.9036	1.8010	3.3119

Table 5.8: Summary statistics of breathing type and oxygen level.

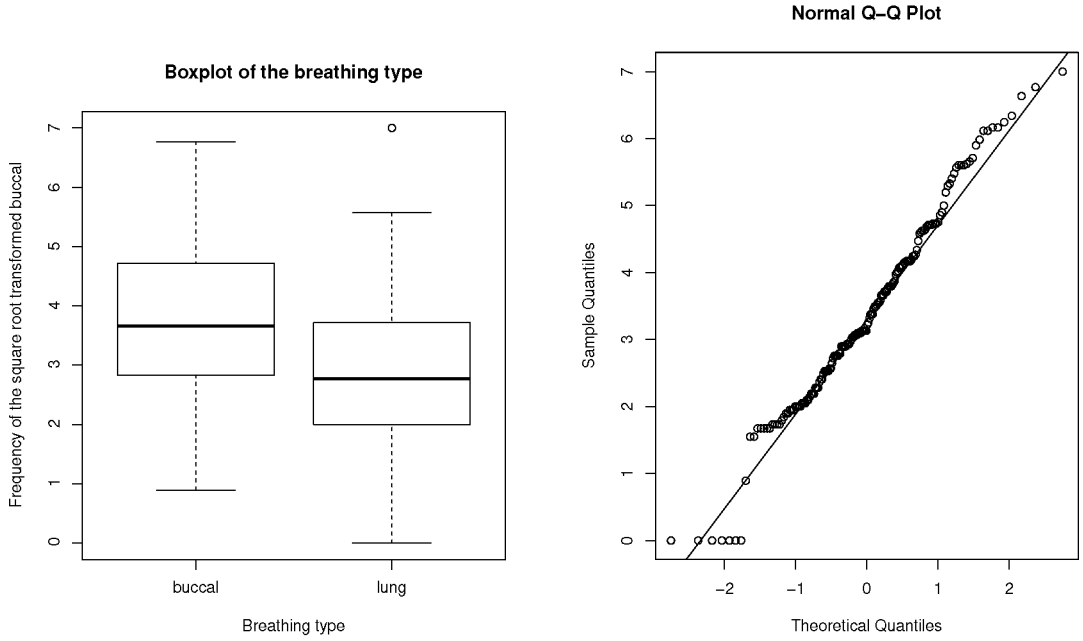


(a) Boxplots of breathing type. **(b)** Normal Q-Q plot for buccal breathing.

Figure 5.13: Boxplots and normal Q-Q plot for breathing type.

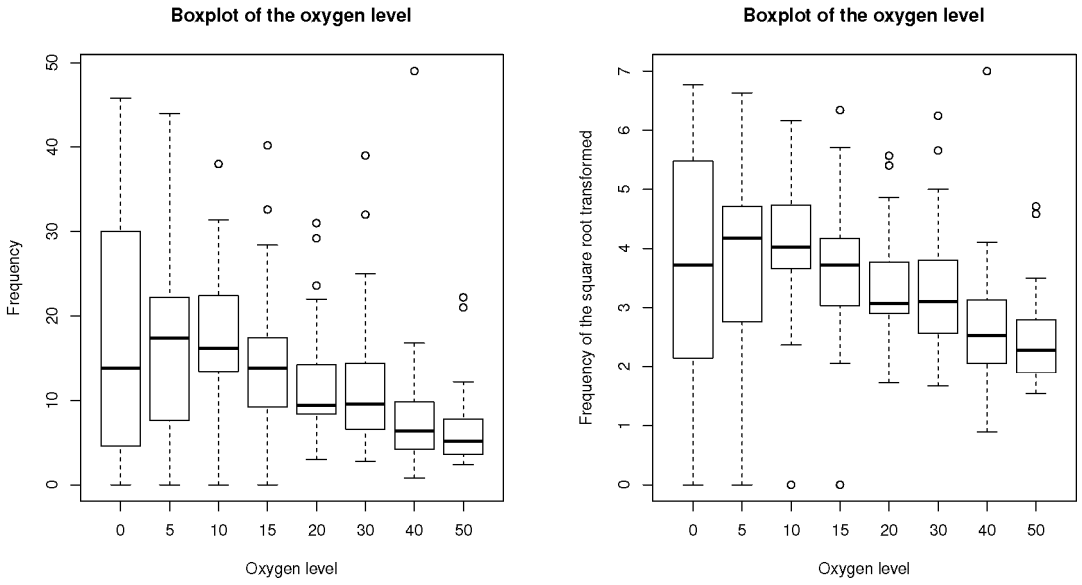
The frequency of buccal breathing was square root transformed in in an attempt to meet the normality and homogeneity of variance assumptions. The boxplots, figure 5.14 (a), shows that the transformed data probably meet the homogeneity of variance assumption and the normal Q-Q plot, figure 5.14 (b), shows evidence that the normality assumption is not violated by the transformed data. The square root transformed data appears to meet assumptions better than the raw data. The boxplots, figure 5.15 (a), shows evidence that homogeneity of variance assumption is violated since the boxplots are of unequal size. The boxplots of the

square root transformed data, figure 5.15 (b), shows that these data may meet the parametric assumptions better than the raw data. There is homogeneity of variances in the breathing type (Bartlett's K-squared = 0.11861, $df = 1$, p-value = 0.7305).



(a) Boxplots of the breathing type. **(b)** Normal Q-Q plot of the data.

Figure 5.14: Diagnostics plots of the square root transformed frequency of breathing.



(a) Boxplots of raw data. **(b)** Boxplots of the square-root transformed data.

Figure 5.15: Boxplots of the frequency of buccal breathing by oxygen level.

A linear model was fit to produce an appropriate ANOVA table to test the hypotheses that

there are no effects of breathing type, oxygen concentration or interaction of the pattern of ventilation, that is the frequency of buccal breathing. There is a significant breathing type by oxygen level interaction ($F_{obs} = 10.693$, $df = 7, 133$, p-value < 0.001). Testing the residuals of the fitted model, using Shapiro-Wilk test revealed that the residuals of this model are not normally distributed ($W = 0.98141$, p-value=0.0438). The normal Q-Q plot, figure 5.16, shows that the residuals of this model meet the normality assumptions.

Source of Variation	df	Sum of Squares	Mean Square	F Statistics	p-value
Breathing type	1	39.92	39.92	5.762	0.0268
Oxygen level	7	38.97	5.567	7.392	$1.71e-07$
Interaction: Breathing type:Oxygen level	7	56.37	8.053	10.693	$1.23e-10$
Residuals	133	100.17	0.753		

Table 5.9: ANOVA model for mullens data.

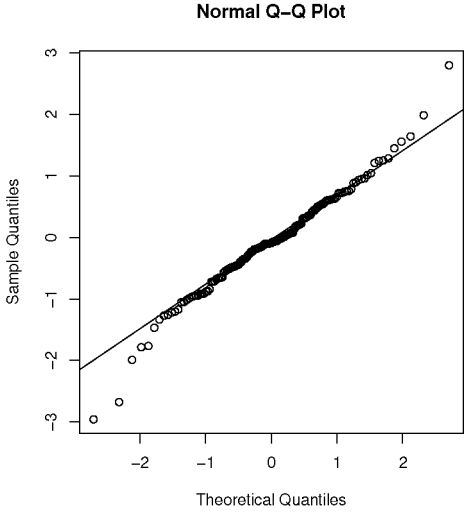


Figure 5.16: Normal Q-Q plot of the residuals.

5.4 Linear Mixed Model for Repeated Measures

This section develops the GLMM, as discussed in section 3.6, to a minimally sufficient level that will allow us to efficiently use the lme4 and lmerTest R package, the builtin functions with optional Satterthwaite degrees of freedom, such that we can approximate. We want to model correlations among the residuals and correlation between repeated measurements of fish weights in the AAA data set. Linear mixed-effects models (LMM's) have a great flexibility in model effects and handle missing data (Pinheiro & Bates, 2000, page 133). In a model with fixed and random factors, it is important to consider how the levels of the fixed factor are related to the levels of the random factor (Schielzeth & Nakagawa, 2013). We assume a Completely Randomized Design (CRD) for fish weights in g treatments groups, with n_i subjects assigned to group i.

Let Y_{ijk} denote the value of the response measured at time k on subject j in group i, $i = 1, \ldots, g$, $j = 1, \ldots, n_i$ and $k = 1, \ldots, t$. Assuming that the random effects are normally distributed and the fixed effect of the GLMM specifies that the expected value of Y_{ijk} as $E(Y_{ijk}) = \mu_{ijk}$. The expected value μ_{ijk} is usually modeled as a function of the treatment, for example the water body and other fixed effects or covariates. The random effects of the model specifies the covariance structure of the observations. We assume that observations on different subjects are independent, which is permissible as a result of the CRD. Thus $Cov(Y_{ijk}, Y_{i'j'l}) = 0$ if $i \neq i'$ or $j \neq j'$. We assume that variances and covariances of measures on a single subject are the same within each of the groups. The general covariance structure is denoted as $Cov(Y_{ijk}, Y_{ijl}) = \sigma_{k,l}$, where $\sigma_{k,l}$ is the covariance between measures at times k and l on the same subject, and $\sigma_{k,k} = \sigma_k^2$ denotes the variance at time k for a particular subject. Let $\mathbf{Y}_{ij} = (Y_{ij1}, Y_{ij2}, \ldots, Y_{ijt})'$ denote the vector of data at times $1, 2, \ldots, t$ for subject j in group i. In matrix notation the model can be expressed as $\mathbf{Y}_{ij} = \boldsymbol{\mu}_{ij} + \boldsymbol{\varepsilon}_{ij}$ where $\boldsymbol{\mu}_{ij} = (\mu_{ij1}, \mu_{ij2}, \ldots, \mu_{ijt})'$ denotes the vector of means and $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij1}, \varepsilon_{ij2}, \ldots, \varepsilon_{ijt})'$ the vector of errors, respectively, for subject j in group i. The matrix representation of the expectation and variance of $\mathbf{Y}_{ij}$ are $E(\mathbf{Y}_{ij}) = \boldsymbol{\mu}_{ij}$ and $Var(\mathbf{Y}_{ij}) = \mathbf{V}_{ij}$ where $\mathbf{V}_{ij}$ is a $t \times t$ matrix with $\sigma_{k,l}$ in row k and column l.

In a CRD the treatments are assigned to the experimental units at random. This is appropriate when the units are homogeneous. A univariate LMM for the fish repeated measures data may be given by

$$Y_{ijk} = \boldsymbol{\mu} + \boldsymbol{\lambda}\mathbf{x}_{ij} + \boldsymbol{\alpha}_i + \mathbf{d}_{ij} + \boldsymbol{\tau}_k + (\boldsymbol{\alpha\tau})_{ik} + \boldsymbol{\varepsilon}_{ijk} \tag{5.1}$$

where $\boldsymbol{\mu}$ is a constant common to all the observations, $\boldsymbol{\lambda}$ is a fixed coefficient on the covariate $\mathbf{x}_{ij}(fishes)$ for angler j in factor group i, $\boldsymbol{\alpha}_i$ is a parameter corresponding to treatment i, $\boldsymbol{\tau}_k$ is a parameter corresponding to time k and $(\boldsymbol{\alpha\tau})_{ik}$ is an interaction parameter corresponding to treatment i and time k. $\mathbf{d}_{ij}$ is a normally distributed random variable with mean zero and variance σ_d^2 corresponding to angler j in treatment group i, ε_{ijk} is a normally distributed random variable with mean zero and variance σ_e^2 independent of d_{ij}, corresponding to angler j in treatment group i at time k. The univariate equation 5.1 is simple and efficient method when the sample size is not sufficiently large (Dhakal, 2016, page 6). Then

$$\begin{aligned} E(\mathbf{Y}_{ijk}) &= \boldsymbol{\mu}_{ijk} = \boldsymbol{\mu} + \boldsymbol{\lambda}\mathbf{x}_{ij} + \boldsymbol{\alpha}_i + \boldsymbol{\tau}_k + (\boldsymbol{\alpha\tau})_{ik} \\ Var(\mathbf{Y}_{ijk}) &= \boldsymbol{\sigma}_d^2 + \boldsymbol{\sigma}_e^2 \end{aligned}$$

where

$$Cov(\mathbf{Y}_{ijk}, \mathbf{Y}_{ijl}) = \boldsymbol{\sigma}_d^2 + Cov(\boldsymbol{\varepsilon}_{ijk}, \boldsymbol{\varepsilon}_{ijl}).$$

The model for $\mathbf{Y}_{ijk}$ can be expressed in matrix notation as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon} \tag{5.2}$$

where $\mathbf{X}$ is a design matrix for the fixed effects, $\mathbf{Z}$ is the design matrix for the random effects and $\mathbf{Y}$ is the response or dependent variable. Here we assume that $\mathbf{b}$ and $\boldsymbol{\varepsilon}$ are normally distributed and independent variables with the following parameters $b_i \sim N(0, \sigma_b^2)$ and $\varepsilon_i \sim N(0, \sigma^2)$. The vector of random effects and errors are assumed to be multivariate normal with $E(\mathbf{b}) = 0$, $Var(\mathbf{b}) = \mathbf{G}$, and $E(\boldsymbol{\varepsilon}) = 0$, $Var(\boldsymbol{\varepsilon}) = \mathbf{R}$. As a result

$$Var(\mathbf{Y}) = \mathbf{ZGZ}' + \mathbf{R}$$

where $\mathbf{ZGZ}'$ represents the between variables covariance structure and $\mathbf{R}$ represent the within variation. The GLM specifies that the continuous response variable, $\mathbf{Y}$, has a multivariate normal distribution with mean response vector $E(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$. The best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ is given by (Littell et al., 2000)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{Y}$$

and the covariance matrix of the sampling distribution of $\boldsymbol{\beta}$ is

$$V(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}.$$

However the estimate $\hat{\boldsymbol{\beta}}$ and its covariance matrix $V(\hat{\boldsymbol{\beta}})$ are both functions of $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$ (Littell et al., 2000). In a mixed-effects models we want to describe a relationship between response variable and some of the covariates that have been measured or observed along with the response variable. The covariates must be categorical, representing observational units in the data set.

The Random Intercept Model

Considering data with two levels of hierarchy, where the repeated measurements are collected over time (Level 1) and nested within the subjects (Level 2). The linear regression model equation 5.2 includes a random effect variable. We can write the conditional mean response (Dhakal, 2016, page 7) as

$$\begin{aligned} E(\mathbf{Y}|\mathbf{b}) &= E(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}) \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{b}. \end{aligned}$$

The population averaged or marginal mean of $\mathbf{Y}_i$, taken mean over the random effects, $\mathbf{b}_i$, is

$$\begin{aligned} E(\mathbf{Y}_i) &= E\left\{E(\mathbf{Y}_i|\mathbf{b}_i)\right\} \\ &= E(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i) \\ &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i E(\mathbf{b}_i). \end{aligned}$$

The marginal variance of each response, $\mathbf{Y}_{ij}$, is given by

$$\begin{aligned} Var(\mathbf{Y}_{ij}) &= Var(\mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_{ij}) \\ &= Var(\mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_{ij}) \\ &= Var(\mathbf{b}_i) + Var(\boldsymbol{\varepsilon}_{ij}) \\ &= \sigma_b^2 + \sigma^2 \end{aligned}$$

σ_b^2 denotes the between-group variability, and σ^2 denotes the within-group variability. The marginal covariance between any two responses, $\mathbf{Y}_{ij}$ and $\mathbf{Y}_{ik}$, is given by:

$$\begin{aligned} Cov(\mathbf{Y}_{ij}, \mathbf{Y}_{ik}) &= Cov(\mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_{ij}, \mathbf{X}_{ik}\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_{ik}) \\ &= Cov(\mathbf{b}_i + \boldsymbol{\varepsilon}_{ij}, \mathbf{b}_i + \boldsymbol{\varepsilon}_{ik}) \\ &= Cov(\mathbf{b}_i, \mathbf{b}_i) \\ &= \sigma_b^2. \end{aligned}$$

Thus σ_b^2 denotes the intraclass covariance, that is the covariance between every pair of observations in the same class and $\frac{\sigma_b^2}{(\sigma_b^2+\sigma^2)}$ is the intraclass correlation coefficient. Richter (2006) defined intraclass correlation as the proportion of criterion variance between units that are organized into groups. The correlation between any pair of response is given by

$$\begin{aligned} Corr(\mathbf{Y}_{ij}, \mathbf{Y}_{ik}) &= \frac{Cov(\mathbf{Y}_{ij}, \mathbf{Y}_{ik})}{\sqrt{Var(\mathbf{Y}_{ij})}\sqrt{Var(\mathbf{Y}_{ik})}} \\ \rho &= \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}. \end{aligned}$$

If $\sigma_b^2 = 1$, the correlation becomes 0. The marginal covariance of $\mathbf{Y}_i$, averaged over the distribution of the random effects $\mathbf{b}_i$, is given by

$$\begin{aligned} Cov\left(\mathbf{Y}_i\right) &= Cov\left(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i\right) \\ &= \mathbf{Z}_i Cov\left(\mathbf{b}_i\right)\mathbf{Z}_i' + Cov\left(\boldsymbol{\varepsilon}_i\right) \\ &= \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}. \end{aligned}$$

Parameter Estimation

The conditional model can be written as multivariate normal distribution (MVN), that is

$$\mathbf{Y}_i|\mathbf{b}_i \sim MVN\left(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \mathbf{R}_i\right).$$

The corresponding marginal model is

$$\mathbf{Y}_i \sim MVN\left(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i\right) \tag{5.3}$$

where $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$. If $\mathbf{V}_i$ is known the MLE of the fixed effect $\boldsymbol{\beta}$ can be obtained by finding the $\hat{\boldsymbol{\beta}}$ that maximizes the log-likelihood function of multivariate marginal function equation 5.3 which is given by (Dhakal, 2016, page 11) as

$$l = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\log|\mathbf{V}_i| - \frac{1}{2}\left[\sum_{i=1}^{N}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})\right]$$

where $n = \sum_{i=1}^{N} n_i$ and N denotes the total number of subjects. Under the multivariate model assumptions, maximizing l yields (Dhakal, 2016, page 11)

$$\hat{\boldsymbol{\beta}} = \left(\sum\nolimits_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum\nolimits_{i=1}^{N}\left(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{Y}_i\right).$$

It can be shown that $\hat{\boldsymbol{\beta}}$ is unbiased, that is $E\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$ (Dhakal, 2016, page 11). The sampling distribution of $\hat{\boldsymbol{\beta}}$ is multivariate normal with mean $\boldsymbol{\beta}$ and covariance (Dhakal, 2016, page 11)

$$Cov(\boldsymbol{\beta}) = \hat{\left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}}.$$

Since the random effects, $\mathbf{b}_i$, are random variables with a multivariate normal distribution, we predict the random effects rather than estimating them (Dhakal, 2016, page 11). The prediction is done by predicting the conditional mean given the data, that is

$$\hat{\mathbf{b}_i} = E(\mathbf{b}_i|\mathbf{Y}_i) = \mathbf{G}\mathbf{Z}_i'\mathbf{V}_i^{-1}\left(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right).$$

This is known as the Best Linear Unbiased Predictor (BLUP) (Dhakal, 2016, page 11).

5.5 Hierarchical Linear Models for Repeated Measures

Gavin & Hofmann (2002) defined hierarchical linear model (HLM) as a statistical technique that examine relationships involving predictors at two or more levels and an outcome at a single level, generally at the lowest level represented by the predictors. This method handles data where observations are dependent and correctly models correlated errors. Hierarchical linear models (HLM's) are a complex form of ordinary least squares (OLS) regression that is used to analyzed variance in the outcome variables when the predictor variables are at varying hierarchical levels (Gavin & Hofmann, 2002). The model is known by several names, including multilevel, mixed level, random effects, random coefficient and complex covariance components modeling (Dhakal, 2016, page 12). The influence of predictors at both the individual and group levels on an individual level outcome can be assessed by the use of HLM. The model can also assess the moderating effects of group level variables on relationships between individual level variables. There are several computational algorithms that exist for

estimating HLM, namely EM, Fisher scoring, Iterative generalized least squares (IGLS) and restricted generalized least squares (RGLS) (Woltman et al., 2012). Denote the groups as $j = 1, \ldots, J$ and assume that there are n_j observations in group j. The hierarchical linear model can be written in matrix form as (Woltman et al., 2012)

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\varepsilon}_j$$

where the matrices $\mathbf{X}_j$ and $\mathbf{Z}_j$ are the known fixed effects and random effects regressor matrices. $\mathbf{X}_j$ is an $n_j \times p$ vector matrix of fixed effects variables where the columns of $\mathbf{X}_j$ are the values of the explanatory variables for group j. $\mathbf{Z}_j$ is an $n_j \times q$ vector matrix of random variables where the columns of $\mathbf{Z}_j$ represent a subset of the columns of $\mathbf{X}_j$.

Mixed-Effects Example

These data arise from a clinical trial of an interactive, multimedia program known as Beat the Blues (BtheB) designed to deliver cognitive behavioral therapy to depression patients via a computer terminal (Hothorn & Everitt, 2014, page 159). The computer based intervention consists of nine sessions followed by eight therapy sessions each lasting about fifty minutes. In a randomized controlled trial, patients with depression recruited in primary care were randomized to either the beating the blues or to treatment as usual (TAU) therapies. A number of outcome measures were used in the trial but here we concentrate on the beck depression inventory (BDI). Measurements on the BDI variable were made on five occasions. The measurements were taken prior to treatment, two months after treatment began and at one, three and six month follow-ups, that is at three, five and eight months after treatment (Hothorn & Everitt, 2014, page 159). The effect of taking antidepressant drug and length of the current episode of depression, that is less or more than six months, were assessed. The data set is rearranged from the wide form in which they appear in BtheB data frame into long form which will separate repeated measurements and associated covariate values will appear as a separate row in a data frame, appendix F. An additional of random term 'subject' in the data frame is used to identify the source of repeated measurements. The results shows that time ($\hat{\beta}_2 = -0.7128$, $T_{obs} = -4.882$, p-value < 0.001) and BDI ($\hat{\beta}_1 = 0.6141$, $T_{obs} = 7.792$, p-value < 0.001) are statistically significant, see appendix D.1.

Both the random intercept and random intercept slope models were fitted, see appendix D.2. BDI is the response or dependent variable that is explained by explanatory or independent variables and subject as a random factor. The smaller AIC score of 1886.7 indicates that the simple random intercept model is adequate for these data. The normal Q-Q plot, figure 5.17, shows that the residuals of this model meet the normality assumption since the points are approximately linear. The points in the residual plot, figure 5.17, are randomly scattered with no particular pattern, this suggest that the assumption of homoscedasticty has been met. The residuals are not normally distributed (Shapiro test, $W = 0.9596$, p-value < 0.001).

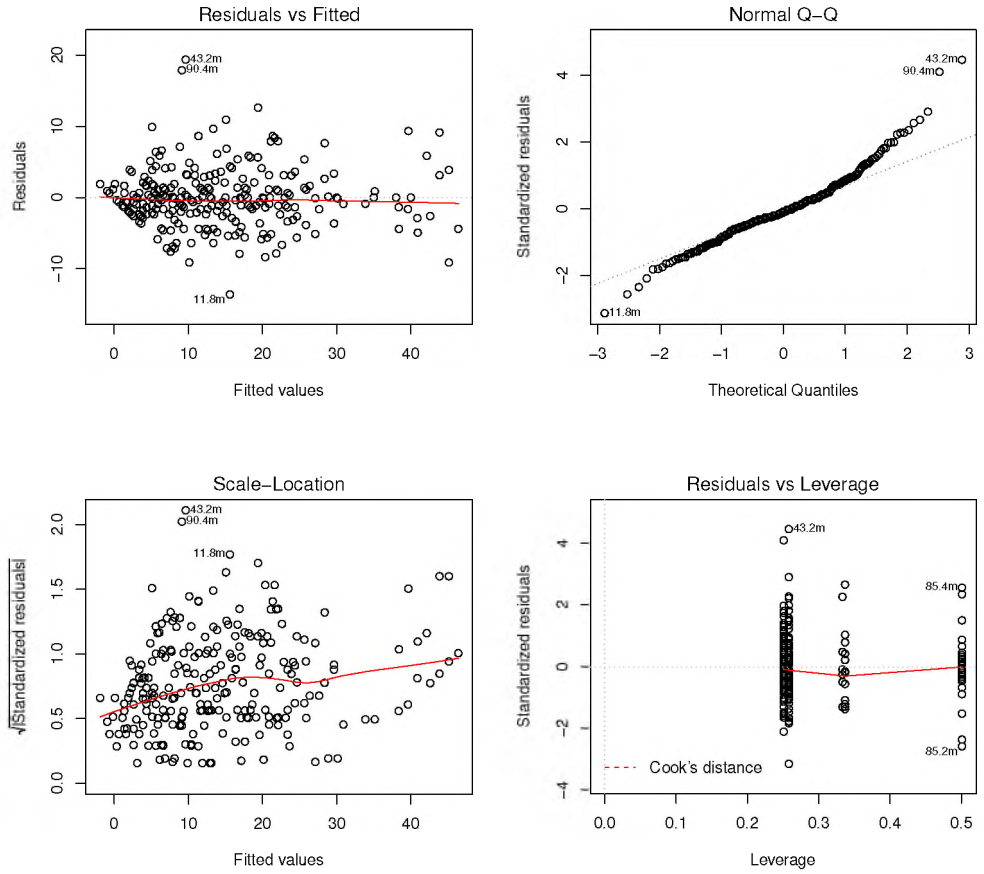


Figure 5.17: Diagnostics plots for the BtheB model.

General Assumptions Underlying the Application of HLM's

The assumptions underlying the HLM refers to the distributions of the error components and their relationships to each other and to the predictor variables. A normal distribution and a constant variance within each unit are required. The joint distribution of the unit error components are assumed to be multivariate normal with a constant covariance matrix across units. All the error components should be independent of the predictor variable and have population means of zero (Richter, 2006).

When developing a hypothesis for Gaussian repeated measures with missing data, it is often assumed that the missing data are "missing at random" (Catellier & Muller, 2000). Missing at random occurs when the probability of missing an observation depends on the outcome measures that have been observed in the past (Hothorn & Everitt, 2014, page 170). Hypothesis testing under the HLM provides sophisticated options and the relevant individual parameters could be tested by comparing the ratio of the parameter and it's standard error to a standard normal distribution or Student's t-distribution. Under the HLM the tests for random effects commonly use a statistic that follows a Chi-square distribution (Richter, 2006). The model allows hypotheses with multiple parameters to be tested using a χ^2 distributed statistics by

conducting individual and multiple parameter tests by comparing nested models. In nested models, one model is the full model whereas the other model is nested under the full hypothesis by fitting one or more parameters to zero. Nested models have a significant function in model building because the more complex models can be easily tested against the fit of simpler models (Richter, 2006).

Let $\mathbf{y}$ denote the response variable and $\mathbf{u}$ the unobserved random effects. The hglm package in R fits a HLM

$$\mathbf{y}|\mathbf{u} \sim f_m(\mu, \phi)$$

and

$$\mathbf{u} \sim f_d(\varphi, \lambda)$$

where f_m and f_d are specified distributions for the mean and dispersion components of the model (Rönnegård et al., 2010). The conditional log-likelihood for $\mathbf{y}$ given $\mathbf{u}$ has the form of a GLM

$$l(\theta, \phi; \mathbf{y}|\mathbf{u}) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

where θ is the canonical parameter, ϕ is the dispersion term, μ is a conditional mean of $\mathbf{y}|\mathbf{u}$ where $\eta = g(\mu)$ is the link function for the GLM. The linear predictor is given by $\eta' = \eta + \nu$ where $\eta = X\beta$ and $\nu = \nu(\mu)$. The link function $\nu(\mu)$ is specified so that random effects occur linearly in the linear predictor. The hierarchical likelihood or h-likelihood is defined as

$$h = l(\theta, \phi; \mathbf{y}|\mathbf{u}) + l(\alpha, v)$$

where $l(\alpha; v)$ is the log density for v with parameter α (Rönnegård et al., 2010).

5.6 Hierarchical Generalized Linear Models

Hierarchical generalized linear models (HGLM's) is an extension of GLM to hierarchical data that allow HLM's to include models that have non-normal error terms and a non-linear structure (Kamata, 2001). They are typically presented as a 2-level formulation of a multilevel item response model. The multilevel item response is referred to as a hierarchical data structure since item responses are nested within respondents (Kamata, 2001). The 1-level HGLM can be extended to a 2-level latent regression model that permits an investigation of the variation of fish weights across water bodies and the interactive effect of the temperature and pressure variables. Latent regression models are utilized for studying the relationship between a latent or unobserved outcome and observed covariates (Kamata, 2001). Three types of parameters that can be estimated in a 2-level hierarchical analysis, namely the fixed effects, random level-1 coefficients and variance-covariance components.

The General Two-level Model

The 2-level model consists of two sub-models at level-1 and level-2. For example, if the study consists of data on fish weights nested within water bodies, the level-1 model would represents fish weights variables and level-2 model would capture the water body level factors. In this context there are $i = 1, \ldots, n_i$ level-1 fish units nested within $j = 1, \ldots, J$ level-2 water bodies.

Level-1 Model

Level-1 is represented as the outcome for case i within j as

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}$$

where Y_{ij} is the dependent variable measured for i^{th} level-1 unit nested within the j^{th} level-2 unit, β_{ij} denote the level-1 coefficients; X_{ij} is the level-1 predictor for observation i in unit j, β_{0j} is the intercept for the j^{th} level-2 unit, β_{1j} represents the regression coefficient associated with X_{ij} for the j^{th} level-2 unit, r_{ij} is the level-1 random effect and σ^2 is the variance of r_{ij}, that is the level-1 variance (Raudenbush & Bryk, 2002, page 35). Here we assume that the random term $r_{ij} \sim N(0, \sigma^2)$.

Level-2 Models

Each of the level-1 coefficients, β_{ij}, defined in level-1 model is considered as an outcome variable in the level-2 model. Consider:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + b_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + b_{1j}$$

model combination form:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} W_j + \gamma_{11} X_{ij} W_j + b_{0j} + b_{1j} X_{ij} + r_{ij}$$

where $\gamma_{00}, \ldots, \gamma_{11}$ are the level-2 coefficients and are referred fixed effects; W_j is a level-2 predictor, b_{0j}, and b_{1j} are level-2 random effects. We assume that $E(r_{ij}) = 0$, $Var(r_{ij}) = \sigma^2$ and for each unit j, the vector $(b_{0j}, b_{1j}, \ldots, b_{qj})'$ is distributed as a multivariate normal where the elements of b_{qj} have a mean of zero and variance of $Var(b_{qj}) = T$. The level-2 variance covariance components are obtained as:

$$Var \begin{bmatrix} b_{0j} \\ b_{1j} \end{bmatrix} = \begin{bmatrix} \gamma_{00} & \gamma_{01} \\ \gamma_{10} & \gamma_{11} \end{bmatrix} = T$$

and $Cov(b_{0j}, r_{ij}) = Cov(b_{1j}, r_{ij}) = 0$ (Raudenbush & Bryk, 2002, page 35).

The REML method is generalized for the estimation of dispersion to the wider class and shows how the joint fitting of models for mean and dispersion can be expressed by two interconnected GLM's (Richter, 2006). The method persists models with any combination of a GLM distribution for the response with any conjugate distribution for the random effects. It also allows the use of quasi-likelihood in place of likelihoods for either or both of the means and dispersion models. The algorithms for fitting these models can be reduced to the fitting of a 2-dimensional set of GLM's, 1-dimension being the mean and dispersion and algorithm for MCNR with the other fixed and random effects (Richter, 2006).

5.7 Missing Data

Missing data is a major issue in many applied studies. Missing data may arise due to many circumstances, including the unavailability of covariance measurements, survey non-response, respondents refusing to answer certain items on a questionnaire and loss of data. Ibrahim et al. (2005) suggested methods for overcoming issues arising from missing data. These include statistical methods such as LMM that handle missing data more appropriately (Ibrahim et al., 2005). Advanced computational technologies exist for handling missing data problems, including Monte Carlo EM algorithm. EM has very good properties when the data are missing at random (MAR). In a situations where there are missing data, the EM algorithm can be applied as a computational technique for obtaining MLE's (Regoeczi & Riedel, 2003). EM is particularly well suited for assigning missing values where there are few continuous variables, as is the case with fisheries (Ibrahim et al., 2005). As a general approach to missing data, the EM algorithm will produce estimates of the standard error that are too low and consequently overestimate the correlations because it treats the imputed data as if they were real numbers (Regoeczi & Riedel, 2003). ML methods can be used assuming the multivariate normality of the data and the missing values are treated as nuisance parameters.

Data are said to be missing at random if, conditional on the observed data, the failure to observe a value does not depend on the data that are unobserved. For example suppose that y_i is completely observed while some components of $\mathbf{x}_i$ may be missing. The missing values of $\mathbf{x}_i$ are missing at random (MAR) if, conditional on the observed data, the probability of observing $\mathbf{x}_i$ is independent of the values of $\mathbf{x}_i$ that would have been observed, but this probability is not necessarily independent of y_i and the observed values of $\mathbf{x}$ (Ibrahim et al., 2005). MAR missingness depends only on the $\mathbf{x}_i's$ and not on the y_i and as a result an analysis will lead to unbiased estimates. However if missingness depends on y_i, then an analysis will result in biased estimates (Ibrahim et al., 2005). For MAR, the missingness of a covariable cannot depend on unobserved covariable values. For example whether a predictor is observed cannot depend on another predictor when the latter is missing but it can depend on the latter when it is observed (Ibrahim et al., 2005).

If the failure to observe a value does not depend on any data, either observed or missing, the data are said to be missing completely at random (MCAR). In this case the observed data are a random sample of all the data. For example, in a logistic regression suppose that y_i is completely observed, whereas some components of $\mathbf{x}_i$ are missing for subject i. The missing values of $\mathbf{x}_i$ are MCAR if the probability of observing $\mathbf{x}_i$ is independent of y_i and is independent of the values of $\mathbf{x}_i$ that are observed or would have been observed (Ibrahim et al., 2005).

Chapter 6

Results and Discussion

The objective of this chapter is to utilize AAA tournament catch data set to determine which statistical method yields to the best results. The R code for the analysis is included in appendix E. According to (McCullagh & Nelder, 1989), GLM's are the most common method for standardizing catch and effort data. In the context of the AAA data set, a GLM model with a Gaussian family distribution and identity link function was used since the total bag weight response variable is continuous. The environmental variables were included in the model in order to determine their influence on total bag weight. The resulting GLM, table 6.1, revealed that the pressures ($\hat{\beta}_3 = -0.2067$, $T_{obs} = -2.406$, p-value = 0.0173) and minimum temperatures ($\hat{\beta}_1 = -0.1771$, $T_{obs} = -1.986$, p-value = 0.0487) of the tournament angling events had a negative significant effect on the bag weight. The change in temperature prior to the event, change in pressure prior to the event, maximum temperature, Yarrow, Settlers and White's dam had insignificant effects on the total bag weight. Increasing change in temperature and pressure of 0.2197 and 0.0177 respectively, this indicates that there are higher chances of heavier bags.

Coefficients	Parameter estimate	Standard error	t-value	p-value
Intercept	201.2375	83.2118	2.418	0.0167
Minimum temperature	−0.1771	0.0892	−1.986	0.0487
Maximum temperature	−0.0816	0.0916	−0.891	0.3741
Pressure	−0.2067	0.0859	−2.406	0.0173
Settlers dam	0.1336	0.3477	0.384	0.7013
White's dam	−0.6212	0.4781	−1.297	0.1965
Yarrow dam	−0.4230	0.3405	−1.242	0.2159
Change temperature	0.2197	0.1843	1.192	0.2351
Change pressure	0.0177	0.0791	0.223	0.8235

Table 6.1: GLM model: Total bag weight.

The GLM model diagnostics plots are shown in figure 6.1. The normal Q- Q shows that the residuals of this model do not meet the normality and linearity assumptions since the

points deviates from the theoretical quantile line. The residuals are not normally distributed (Shapiro test, $W = 0.977$, p-value $= 0.0061$). The fitted residuals shows that the residuals of this model do not meet the homogeneity of variance assumption. There is heterogeneity of variances (Bartlett's K-squared $= 18.131$, $df = 3$, p-value $= 0.0004$). The diagnostics plots shows that the GLM is not appropriate for these data.

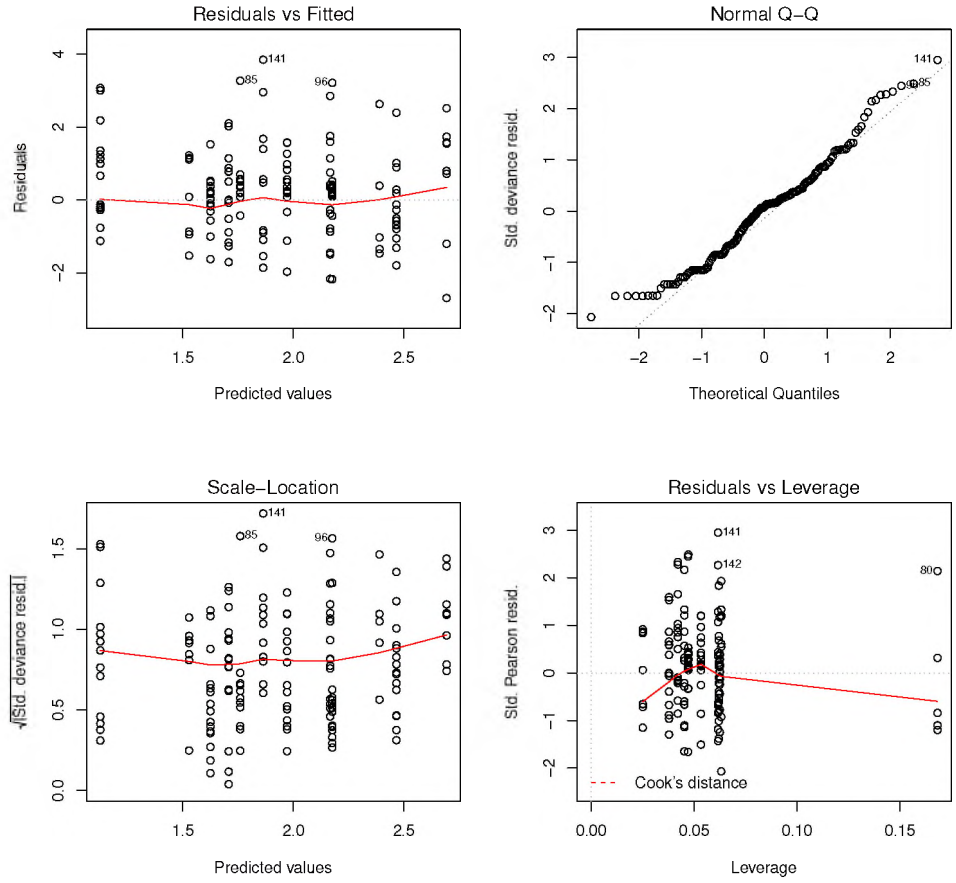


Figure 6.1: Diagnostics plots of the GLM model.

The stepwise backward feature was used to reduce the number of variables by dropping insignificant variables from the full model. The fit GLM was redone using the Gaussian distribution with identity link function. The reduced model in table 6.2 shows that the pressure variable is the only statistically significant variable for this model ($\hat{\beta}_2 = -0.1391$, $T_{obs} = -2.907$, p-value $= 0.0042$). The model assessments, figure 6.2, shows that the reduced model do not fit the data adequately as the normal Q-Q plot is not approximately linear and the fitted residuals do not meet the homoscedasticity assumption. The residuals of the model are not normally distributed (Shapiro test, $W = 0.9667$, p-value $= 0.0004$). These data provide evidence that there heterogeneity of variances (Bartlett's K-squared $= 21.14$, $df = 11$, p-value $= 0.0311$). The GLM does not perform well with these kind of data. In this study AIC was used to decide between alternative models, choosing the model with the smaller AIC. The AIC value was used in defining the measure of fit and the number of parameters measures

the complexity of the model.

Coefficients	Parameter estimate	Standard error	t-value	p-value
Intercept	134.2049	45.6384	2.941	0.0037
Minimum temperature	−0.0718	0.0437	−1.645	0.1018
Pressure	−0.1391	0.0479	−2.907	0.0042

Table 6.2: Reduced model: Total bag weight.

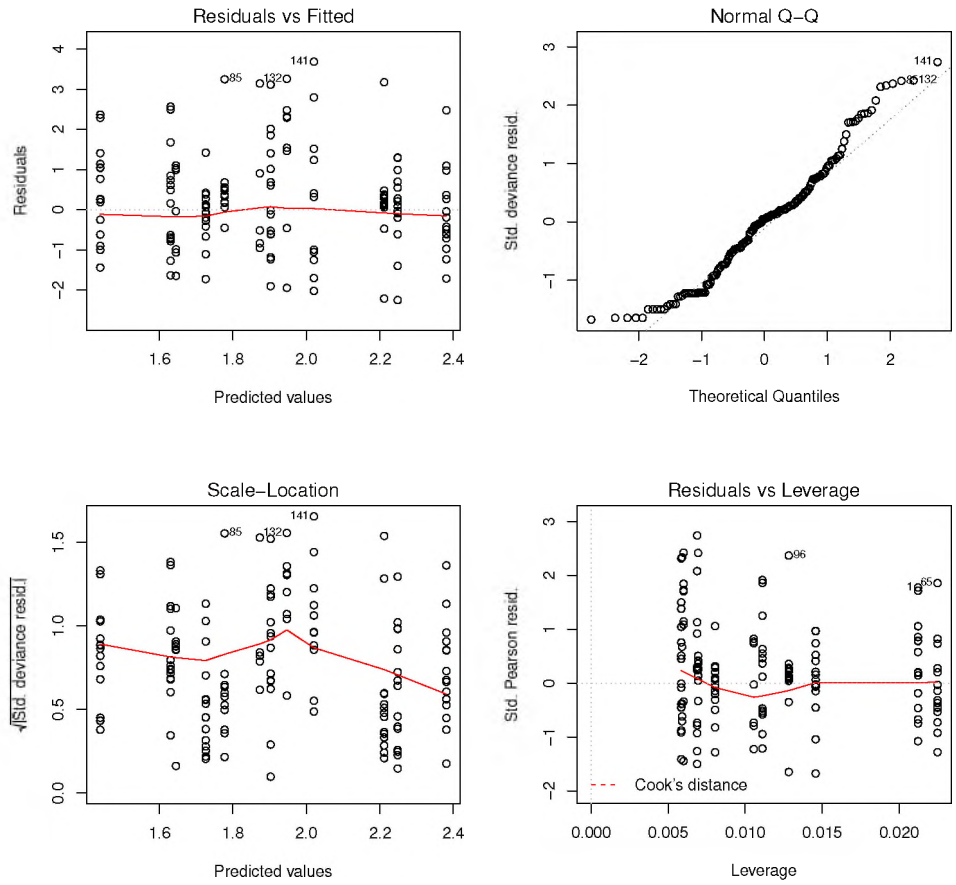


Figure 6.2: Diagnostics plots of the reduced model: Total bag weight.

The GLM model failed to meet the assumptions and hence a mixed-effects model was considered. The 'lme' function in the R package nlme was used to fit a linear mixed-effects model. The first argument indicates that the response variable is the bag weight. The second argument indicates that there is a single random effect for each group and the grouping factor is given by the variable person (ps). The third argument indicates that the data will be found in the object named repeated measures (rm). The estimation methods were specified as ML and REML respectively. The resulting linear mixed-effects model are found in appendix E.1. The REML estimates for the parameters have been calculated as $\hat{\sigma}_\varepsilon = 1.102$ and the random intercept $\hat{\sigma} = 0.755$ with a corresponding log-restricted-likelihood of -284.9038. To examine the ML estimates we used the same argument as for mod.lme except for the method. A

convenient way of fitting alternative models is the 'update' function. The resulting linear mixed-effects models are shown in appendix E.2.

The ML estimates for the parameters have been calculated as $\hat{\sigma}_\varepsilon = 1.0721$ and the random intercept $\hat{\sigma} = 0.7387$ with a corresponding log-likelihood of -271.9898. The ML parameter estimates of the random effects standard deviations are smaller than the corresponding REML estimates. This occurs because REML estimation generally produces larger estimates for the random effect variances (Pinheiro & Bates, 2000, page 153 and 154). The fixed effects estimates obtained by ML and REML are similar, thought they are not identical (Pinheiro & Bates, 2000, page 155). Pressure has a significant effect on the total bag weight ($\hat{\beta}_6 = -0.1568$, $T_{obs} = -2.1732$, p-value $= 0.0315$). The diagnostics plots, figure 6.3, shows that the model violated the normality and homogeneity of variance assumptions. These data provide evidence that there is heterogeneity of variances (Bartlett's K-squared $=$ 18.131, $df = 3$, p-value $= 0.0004$). The normal Q-Q plot indicates that these data are not approximately linear and the fitted residuals shows persisting pattern suggesting that the fitted model is not adequate for modeling these data. The residuals are not normally distributed (Shapiro test, $W = 0.9472$, p-value < 0.001).

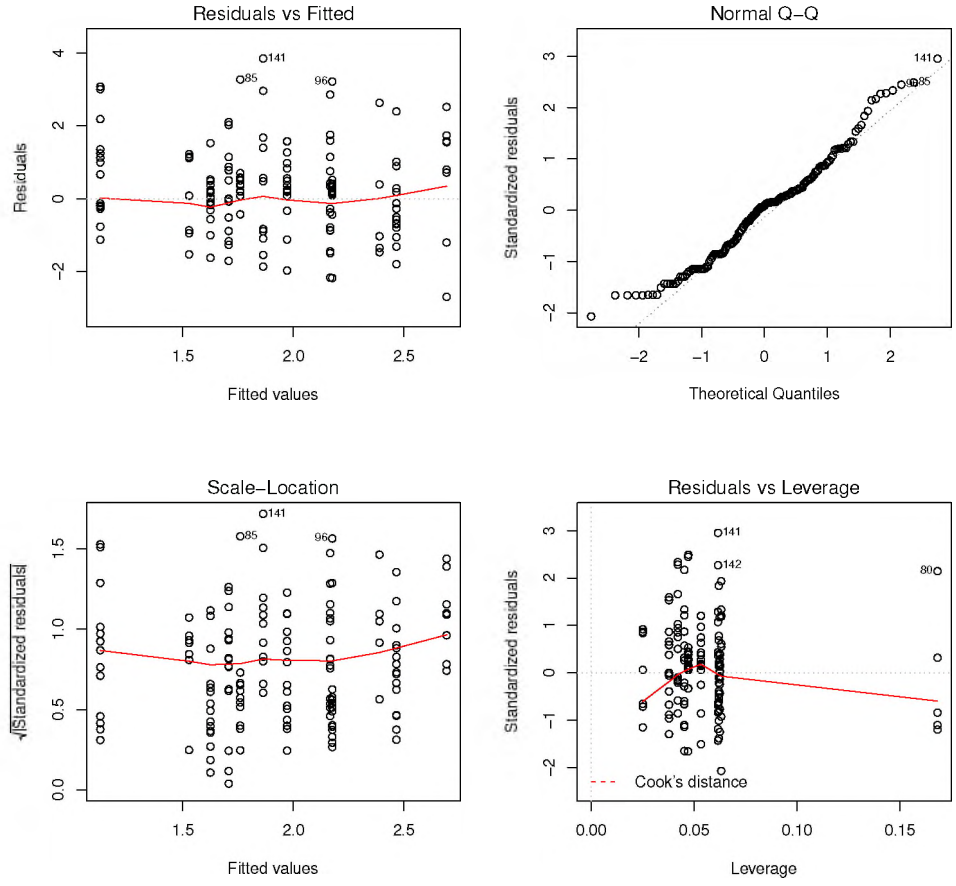


Figure 6.3: Diagnostics plots for model fitted with lme function.

A GLMM was fitted using the 'lmer' function to incorporate random effects. When conside-

ring person (ps) as the only random term in the model, the variance component for person is 0.5457 and residual is 1.1493. The variance components for the model fitted with the 'lme' function are the same as the model fitted with 'lmer' function under both ML and REML estimation methods, see appendix E.3. The slope for minimum temperatures is -0.1083, indicating that bag weight decreases as temperature decreases. The random intercept, person, is normally distributed with mean zero and variance 0.5457. The residuals are normally distributed with mean zero and variance 1.1493. These two variance can be used to estimate the correlation between observations from the same subject: $\frac{0.5457}{(0.5457+1.1493)} = 0.3219$. The fitted random intercept and slope models are shown in appendix E.3. Again pressure is the only significant variable ($\hat{\beta}_3 = -0.1503$, $T_{obs} = -2.563$, p-value $= 0.0114$). There is a high correlation of 78% between pressure and minimum temperature in the random intercept and slope model. The 'anova' function is used is used to compare fitted models based on their AIC, see appendix E.4. The AIC score of the random intercept model is considerable smaller and the p-value is essentially zero, so we prefer the random intercept model since it has a smaller AIC score of 565.98, see appendix . The diagnostics plots, figure 6.4, shows that the residuals for this model meet the normality and linearity assumptions since the normal Q-Q plot do not deviates from the theoretical quantiles. The fitted residual plot, figure 6.4, shows that the residuals of this model meets the homoscedasticity assumption since the residuals are randomly scattered with no particular pattern. Cook's distance suggest that none of the observations are influential in the fitted model. These results revealed that GLMM is the best model for these data.

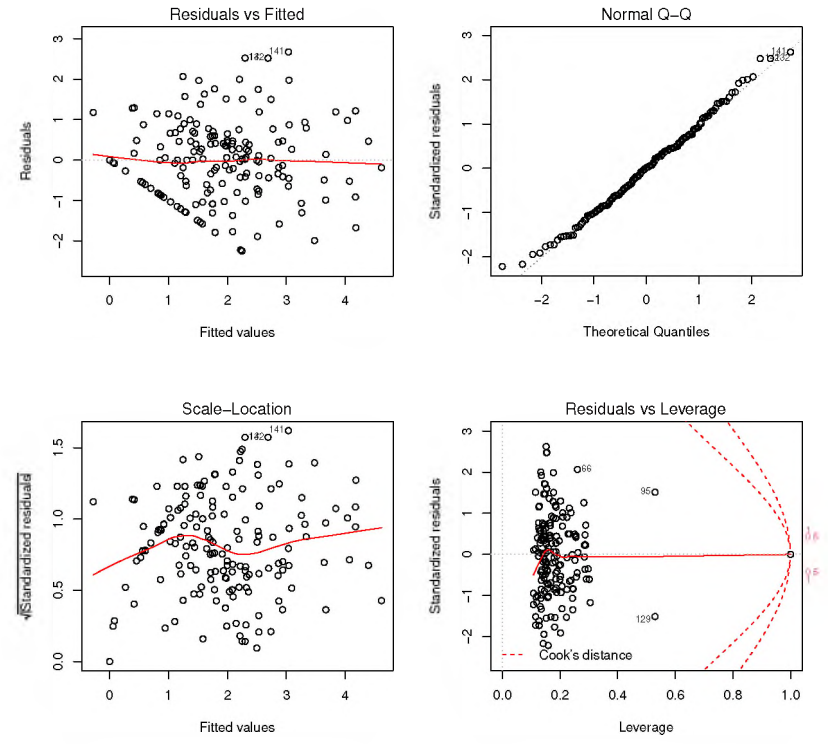


Figure 6.4: Model validation graphs for the mixed-effects model.

Chapter 7

Conclusion

ANOVA methods are commonly utilized in the analysis of structured experimental designs, they work well for balanced designs having discrete independent or explanatory variables but are not widely applicable when the data are unbalanced and some predictors are continuous (Raudenbush & Bryk, 2002). For the AAA study the best model was determined.

Determining the best modeling method to use in a particular situation should lead to the development of better predictions. The reason that these modeling methods were selected was that it is not entirely clear what environmental influences are most significant for predicting total bag weight. Fisheries data is typically better modeled with the generalized linear mixed-effects model than a generalized linear model as there is missing data. The generalized linear mixed model was utilized for these data since it incorporates random effects into the linear predictors. This study had demonstrated how repeated measures data was modeled using mixed-effects model. The flexibility of mixed-effects models makes it appropriate for a tremendous variety of designs, including nested designs and repeated measures designs (Raudenbush & Bryk, 2002). Future studies of fisheries research should consider utilizing mixed-effects models particularly in cases where missing data are observed.

Bibliography

Agresti, A., Booth, G. J., Hobet, P. J., & Caffo, B. (2000). Random-Effects Modeling of Categorical Response Data. *Sociological Methodology*, 30(1), 27–80.

Bolker, B. M. (2007). *Ecological Models and Data in R.* Princeton University Press.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized Linear Mixed Models: A Practical Guide for Ecology and Evolution. *Trends in Ecology & Evolution*, 24, 127–135.

Catellier, D. J. & Muller, K. E. (2000). Tests for Gaussian Repeated-Measures with Missing Data in Small Samples. *Statistics in Medicine*, 19(8), 1 101–1 114.

Davis, C. S. (2002). *Normal Theory Methods: Repeated-Measures ANOVA*. Springer.

Dhakal, P. (2016). *Hierarchical Modeling of Biological Rhythm Data: Classical and Bayesian Approaches.* PhD thesis, New Mexico Institute of Mining and Technology.

Dobson, A. J. & Barnett, A. (2008). *An Introduction to Generalized Linear Models.* CRC Press, third edition.

Faraway, J. J. (2004). *Practical Regression and ANOVA Using R.* Chapman & Hall/CRC.

Faraway, J. J. (2014). *Linear Models with R.* CRC Press, second edition.

Faraway, J. J. (2016). *Extending the Linear Model with R: Generalized Linear, Mixed-Effects and Non-parametric Regression Models.* CRC Press, second edition.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics.* Sage, fourth edition.

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R.* Sage, third edition.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longitudinal Data Analysis.* CRC Press.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied Longitudinal Analysis.* John Wiley & Sons, second edition.

Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models*. Sage Publications, third edition.

Gavin, M. B. & Hofmann, D. A. (2002). Using Hierarchical Linear Modeling to Investigate the Moderating Influence of Leadership Climate. *The Leadership Quarterly*, 13(1), 15–33.

Hardin, J. W., Hilbe, J. M., & Hilbe, J. (2007). *Generalized Linear Models and Extensions*. Stata Press, second edition.

Hardy, M. A. (1993). *Regression with Dummy Variables*, volume 93. Sage.

Hargrove, J. S., Weyl, O. L., Allen, M. S., & Deacon, N. R. (2015). Using Tournament Angler Data to Rapidly Assess the Invasion Status of Alien Sport Fishes *Micropterus spp.* in Southern Africa. *PloS One*, 10(6), 1–14.

Hecke, T. V. (2012). Power Study of ANOVA versus Kruskal-Wallis Test. *Journal of Statistics and Management Systems*, 15(2-3), 241–247.

Hector, A., Von Felten, S., & Schmid, B. (2010). Analysis of Variance with Unbalanced Data: An Update for Ecology & Evolution. *Journal of Animal Ecology*, 79(2), 308–316.

Hothorn, T. & Everitt, B. S. (2014). *A Handbook of Statistical Analyses Using R*. CRC press.

Howell, D. H., Woodford, D. J., Weyl, O. L., & Froneman, W. (2013). Population Dynamics of the Invasive Fish, *Gambusia affinis*, in Irrigation Impoundments in the Sundays River Valley, Eastern Cape, South Africa. *Water SA*, 39(4), 485–490.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., & Herring, A. H. (2005). Missing Data Methods for Generalized Linear Models: A Comparative Review. *Journal of the American Statistical Association*, 100(469), 332–346.

Jackson, S. A. & Brashers, D. E. (1994). *Random Factors in ANOVA*, volume 98. Sage.

Jennrich, R. I. & Schluchter, M. D. (1986). Unbalanced Repeated-Measures Models with Structured Covariance Matrices. *Biometrics*, 42(4), 805–820.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Science & Business Media.

Johnson, J. B. & Omland, K. S. (2004). Model Selection in Ecology and Evolution. *Trends in Ecology & Evolution*, 19(2), 101–108.

Johnson, R. A. & Wichern, D. W. (2014). *Applied Multivariate Statistical Analysis*. Prentice-Hall New Jersey, sixth edition.

Kamata, A. (2001). Item Analysis by the Hierarchical Generalized Linear Model. *Journal of Educational Measurement*, 38(1), 79–93.

Kirk, R. E. (1982). *Experimental Design.* Wiley Online Library, second edition.

Kleinbaum, D., Kupper, L., Nizam, A., & Rosenberg, E. (2013). *Applied Regression Analysis and Other Multivariable Methods.* Nelson Education, fifth edition.

Kuk, A. Y. & Cheng, Y. W. (1997). The Monte Carlo Newton-Raphson Algorithm. *Journal of Statistical Computation and Simulation*, 59(3), 233–250.

Larson, M. G. (2008). Analysis of Variance. *Circulation*, 117(1), 115–121.

Levine, R. A. & Casella, G. (2001). Implementations of the Monte Carlo EM Algorithm. *Journal of Computational and Graphical Statistics*, 10(3), 422–439.

Lindstrom, M. J. & Bates, D. M. (1988). Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. *Journal of the American Statistical Association*, 83(404), 1 014–1 022.

Lindstrom, M. J. & Bates, D. M. (1990). Nonlinear Mixed-Effects Models for Repeated-Measures Data. *Biometrics*, 46(3), 673–687.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Tutorial in Biostatistics: Modelling Covariance Structure in the Analysis of Repeated-Measures Data. *Statistics in Medicine*, 19(13), 1 793–1 819.

Littell, R. C., Stroup, W. W., Milliken, G. A., Wolfinger, R. D., & Schabenberger, O. (2007). *SAS for Mixed Models.* SAS Institute, second edition.

Logan, M. (2011). *Biostatistical Design and Analysis Using R: A Practical Guide.* John Wiley & Sons.

Maunder, M. N. & Punt, A. E. (2004). Standardizing Catch and Effort Data: A Review of Recent Approaches. *Fisheries Research*, 70(2), 141–159.

McCarthy, M. A. (2007). *Bayesian Methods for Ecology.* Cambridge University Press.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models.* Chapman & Hall, second edition.

McCulloch, C. E. (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 92(437), 162–170.

McCulloch, C. E. & Searle, S. R. (2001). *Generalized Linear and Mixed Models.* Wiley Online Library.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models.* Wiley, second edition.

McDonald, J. H. (2009). *Handbook of Biological Statistics*. Sparky House Publishing Baltimore, MD, second edition.

Morrell, C. H., Pearson, J. D., & Brant, L. J. (1997). Linear Transformations of Linear Mixed-Effects Models. *The American Statistician*, 51(4), 338–343.

Paine, M. D. (1996). Repeated-Measures Designs. *Environmental Toxicology and Chemistry*, 15(9), 1439–1441.

Paul, G. (2016). Repeated Measures ANOVA. `http://www.gribblelab.org/stats/notes/RepeatedMeasuresANOVA.pdf`, accessed February 2017.

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus*. Springer, third edition.

Quinn, G. P. & Keough, M. J. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radloff, S. E. (2008). *Mathematical Statistics III, Paper 2, An Introduction to the General Linear Model, Course Notes*. Rhodes University.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, volume 1. Sage, second edition.

Regoeczi, W. C. & Riedel, M. (2003). The Application of Missing Data Estimation Models to the Problem of Unknown Victim/Offender Relationships in Homicide Cases. *Journal of Quantitative Criminology*, 19(2), 155–183.

Richter, T. (2006). What is Wrong with ANOVA and Multiple Regression? Analyzing Sentence Reading Times with Hierarchical Linear Models. *Discourse Processes*, 41(3), 221–250.

Rönnegård, L., Shen, X., & Alam, M. (2010). hglm: A Package for Fitting Hierarchical Generalized Linear Models. *The R Journal*, 2(2), 20–28.

Sahai, H. & Ageel, M. I. (2012). *The Analysis of Variance: Fixed, Random and Mixed Models*. Springer Science & Business Media.

Schielzeth, H. & Nakagawa, S. (2013). Nested by Design: Model Fitting and Interpretation in a Mixed Model Era. *Methods in Ecology and Evolution*, 4(1), 14–24.

Sinha, S. K. (2004). Robust Analysis of Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 99(466), 451–460.

Skrivanek, S. (2009). The Use of Dummy Variables in Regression Analysis. `https://www.moresteam.com/WhitePapers/download/dummy-variables.pdf`, accessed February 2010.

Stephen, W. (1993). *Hierarchical Linear Models and Experimental Design*, volume 137. CRC Press, revised edition.

Sullivan, L. M. (2008). Repeated-Measures. *Circulation*, 117(9), 1 238–1 243.

van der Laan, P. & Verdooren, L. R. (1987). Classical Analysis of Variance Methods and Non-Parametric Counterparts. *Biometrical Journal*, 29(6), 635–665.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2004). Applying Linear Mixed Models to Estimate Reliability in Clinical Trial Data with Repeated-Measurements. *Controlled Clinical Trials*, 25(1), 13–30.

Verma, J. (2015). *Repeated-Measures Design for Empirical Researchers*. John Wiley & Sons.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical Statistics with Applications*. Brooks/Cole, seventh edition.

Wagner, T., Hayes, D. B., & Bremigan, M. T. (2006). Accounting for Multilevel Data Structures in Fisheries Data Using Mixed Models. *Fisheries*, 31(4), 180–187.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An Introduction to Hierarchical Linear Modeling. *Tutorials in Quantitative Methods for Psychology*, 8(1), 52–69.

Xu, L.-W., Yang, F.-Q., & Qin, S. (2013). A Parametric Bootstrap Approach for Two-Way ANOVA in Presence of Possible Interactions with Unequal Variances. *Journal of Multivariate Analysis*, 115, 172–180.

Zhang, J. T. (2012). An Approximate Degrees of Freedom Test for Heteroscedastic Two-Way ANOVA. *Journal of Statistical Planning and Inference*, 142(1), 336–346.

Zhang, X. (2015). A Tutorial on Restricted Maximum Likelihood Estimation in Linear Regression and Linear Mixed-Effects Model. `http://web.mit.edu/xiuming/www/docs/tutorials/ReML.pdf`, accessed December 2016.

Zuur, A., Ieno, E., Walker, N., Saveliev, A., & Smith, G. (2009). *Mixed-Effects Models and Extensions in Ecology with R*. Springer, New York.

Appendix A

Appendix A: R Code for *G. Affinis*

```
#####################
# Author: B. Sidumo
# Last update:10/11/2017
# Workable dataset is on "raw data 1", starting at row 80.
# These data saved as Gambusian raw data.xlsx.
# Gambusian raw data.xlsx exported to Gambusian raw data.csv
# 22/02/2017
# These data are from G.Affins by Weyl,Froneman,Howell and Woodford
# Original data : Gambusian raw data.xlsx kindly supplied by
# Dr Woodford (University of the Witwatersrand, South Africa)
#####################
#--------------------
# Data set setup
#--------------------
# Typo's corrected:
# Row 241 : Veg cover of 25 which is 0.25
# Clear R work history
rm(list=ls()) # You want to clear out old variables before building
    new ones
# Set working directory
# setwd("L:/profile/Desktop/Project/")
# The data file is named: "Gambusia exported raw data.csv"
gaadata <-read.csv(file=file.choose(),header=TRUE)
# Data structure: check
str(gaadata)
# Data frame : 380 observations of 26 variables
names(gaadata)
levels(gaadata$Dam.Code)
```

```
levels(gaadata$Dam.Code.1)
# Dam.Code.1 == Dam.Code : same variables
# Dam names in the paper : DC, OLI, SLS, AVO, DB
# dam names in the data : "AVO1" "BV1" "DB1" "HBT (OLI)" "HBT (SLS)
  "
levels(gaadata$Dam.Code) <- cbind("AVO","DC","DB","SLS","OLI")
# Four sampling events
# Remove sampling events in 2011 data
gaadata <- gaadata[gaadata$YY==12,]
gaadata$Sampling.Event <- factor(gaadata$Sampling.Event)
# Check if the number of observations are balanced by sampling
  event
table(gaadata$Sampling.Event)
# Create a new variable "Site"
site <- rep(gl(19,4),4)
gaadata <- data.frame(gaadata,site = site)
# Removed all the unneeded variables in the data set
gaadata <- gaadata[,c("Dam.Code", "Sampling.Event", "Dam.Age..years
  .","Mean.Temp", "GAA", "GLC..GOB.", "GIA..GIL.", "X", "sqrtGAA",
   "site")]
# Renaming the column number 8 in the data set
names(gaadata)[8]
names(gaadata)[8]<-"Vegetation.cover"
# Mean temperature :  temperature recorded in each selected dam
# Variables : "Mean.Temp" , "lnTemp"
# Gambusian affins (Mosquito fish)
# Variables : "GAA" , "sqrtGAA" , "ln.gaa"
#-------------------------------
# Descriptive Statistics
#-------------------------------
# Basic line and point graph with error bars representing either
  the standard error of the mean, or 95% confidence interval.
## Install a package "ggplot2" and "gcookbook"
# Install a packages:
library(ggplot2)
library(gcookbook)
library(lattice)
library(plyr)
# http://www.cookbook-r.com/Manipulating_data/Summarizing_data/
# summarySE(): Summarizes data.
```

```
# Gives count, mean, standard deviation, standard error of the mean
    , and confidence interval (default 95%).
#   data: a data frame.
#   measurevar: the name of a column that contains the variable to
    be summariezed
#   groupvars: a vector containing names of columns that contain
    grouping variables
#   na.rm: a boolean that indicates whether to ignore NA's
#   conf.interval: the percent range of the confidence interval (
    default is 95%)
summarySE <- function(data=gaadata, measurevar= gaadata$GAA,
    groupvars= Dam.Code, na.rm=FALSE,
                          conf.interval=.95, .drop=TRUE) {
  library(plyr)
  # New version of length which can handle NA's: if na.rm==T, don't
      count them
  length2 <- function (x, na.rm=FALSE) {
    if (na.rm) sum(!is.na(x))
    else       length(x)
  }
  # This does the summary. For each group's data frame, return a
      vector with
  # N, mean, and sd
  datac <- ddply(data, groupvars, .drop=.drop,
                 .fun = function(xx, col) {
                   c(N    = length2(xx[[col]], na.rm=na.rm),
                     mean = mean   (xx[[col]], na.rm=na.rm),
                     sd   = sd     (xx[[col]], na.rm=na.rm)
                   )
                 },
                 measurevar
  )
  datac$se <- datac$sd / sqrt(datac$N)  # Calculate standard error
      of the mean
  # Confidence interval multiplier for standard error
  # Calculate t-statistic for confidence interval:
  # e.g., if conf.interval is .95, use .975 (above/below), and use
      df=N-1
  ciMult <- qt(conf.interval/2 + .5, datac$N-1)
  # t*se
```

```
  datac$ci <- datac$se * ciMult
  # lower bound: bar(x) - t*(se)
  print(datac$mean)
  datac$Lower <- datac$mean - datac$ci
  # print(lower.bound)
  # Upper bound: bar(x) + t*(se)
  datac$Upper <- datac$mean + datac$ci
  # Rename the "mean" column
  datac <- rename(datac, c("mean" = measurevar))
  return(datac)
}
#-----------------------------------------
# G. affinis CPUE by Sampling Events
#-----------------------------------------
## Generating summary statistics of G.affinis CPUE by sampling
   event
summarySE(gaadata, measurevar = "GAA", groupvars = c("Sampling.
   Event"))
# Mininum and maximum G.affinis CPUE by sampling event
tapply(gaadata$GAA, gaadata$Sampling.Event, min) #tapply group the
   values by a unique combination.
tapply(gaadata$GAA, gaadata$Sampling.Event, max)
# Subset argument works on the rows
str(gaadata$Sampling.Event) # is currently an integer, we Want a
   factor.
gaadata$Sampling.Event <- factor(gaadata$Sampling.Event, labels = c
   ("Feb","Mar","Apr","Jun"),levels = c(2,3,4,5))
February <- subset(gaadata$GAA, gaadata$Sampling.Event=="Feb")
March <- subset(gaadata$GAA, gaadata$Sampling.Event=="Mar")
April <- subset(gaadata$GAA, gaadata$Sampling.Event=="Apr")
June <- subset(gaadata$GAA, gaadata$Sampling.Event=="Jun")
se<-cbind(February,March,April,June)
# Generating sampling event factors
gl(1,4,labels = c("February","March","April","June"))
# Boxplot of G. affinis CPUE by Sampling events
boxplot(se, main="Boxplot of the CPUE by sampling event", xlab="
   Sampling event: Month",ylab=expression(paste(italic('G. affinis'
   ),"  CPUE")))
dev.print(file="boxp_se.eps")
```

```
## Histogram of G. affinis CPUE by Sampling events
par(mfrow=c(2,2)) # Combine the pictures
hist(February, main = "Histogram of February CPUE", xlab = "
   February",xlim = c(0,160))
hist(March, main = "Histogram of March CPUE", xlab = "March",xlim =
    c(0,160))
hist(April, main = "Histogram of April CPUE", xlab = "April",xlim =
    c(0,160))
hist(June, main = "Histogram of June CPUE", xlab = "June",xlim = c
   (0,160))
dev.print(file="hist_se.eps")
par(mfrow=c(1,1)) # Save the figure as one picture

# Normality Q-Q plots for G. affinis CPUE by Sampling event
   variable
par(mfrow=c(2,2))
qqnorm(February, main= "Q-Q plot: CPUE in February", xlab="
   Theoretical quantiles",ylab ="Data (Observed) quantiles")
qqline(February, lwd=2 ,col="blue")
qqnorm(March, main= "Q-Q plot: CPUE in March",xlab="Theoretical
   quantiles",ylab ="Data (Observed) quantiles")
qqline(March, lwd=2 ,col="blue")
qqnorm(April, main= "Q-Q plot: CPUE in April",xlab="Theoretical
   quantiles",ylab ="Data (Observed) quantiles")
qqline(April, lwd=2 ,col="blue")
qqnorm(June, main= "Q-Q plot: CPUE in June",xlab="Theoretical
   quantiles",ylab ="Data (Observed) quantiles")
qqline(June, lwd=2 ,col="blue")
par(mfrow=c(1,1))
#-----------------------------------------------------------
# Testing Normality and Homogeneity (Univariate)
#-----------------------------------------------------------
## Multivariate Shapiro Test for normality by sampling events
library(mvShapiroTest)
mvShapiro.Test(se)
# http://www.cookbook-r.com/Statistical_analysis/Homogeneity_of_
   variance/
library(car)
## Testing Homogeneity of sampling events using non-parametric test
   , Bartlett test
```

```
bartlett.test(gaadata$GAA ~ gaadata$Sampling.Event, data = gaadata)
#-----------------------------
# Dam or (Location)
#-----------------------------
# Generating summary statistics of G.affinis CPUE by dam
summarySE(gaadata, measurevar = "GAA", groupvars = c("Dam.Code"))
# Minimum and maximum
tapply(gaadata$GAA, gaadata$Dam.Code, min)
tapply(gaadata$GAA, gaadata$Dam.Code, max)
# Boxplot for G.affinis CPUE by dam
boxplot(gaadata$GAA ~ gaadata$Dam.Code,main ="Boxplot of the CPUE
   by dam",xlab = "Dam",ylab=expression(paste(italic('G. affinis'),
    " CPUE")))
hist(gaadata$GAA,xlab = expression(paste(italic('G. affinis'),"
   CPUE")),main=expression(paste(italic('G. affinis')," CPUE
   histogram"))
      ,xlim = c(0,200))
plot(density(gaadata$GAA),main =expression(paste(italic('G. affinis
   '),"  CPUE density")),
     xlab =expression(paste(italic( 'G. affinis'), " CPUE")))
dev.off()
qqnorm(gaadata$GAA,main ="Normal Q-Q plot of G. affinis CPUE")
qqline(gaadata$GAA)
#-----------------------------------
# Percentage Vegetation Cover
#-----------------------------------
## Generating summary statistics of G.affinis CPUE by percentage
   vegetation cover
summarySE(gaadata, measurevar = "GAA",groupvars = c("Vegetation.
   cover"))
# Boxplot for G. affinis CPUE by percantage vegetation cover
boxplot(gaadata$GAA ~ gaadata$Vegetation.cover,main="Boxplot of the
    CPUE by proportion vegetation cover",
        xlab="Proportion of vegetation cover",ylab=expression(paste
           (italic('G. affinis')," CPUE")))

frequencies <- table(gaadata$Vegetation.cover)
frequencies
percentages <- round(table(gaadata$Vegetation.cover)/sum(table(
   gaadata$Vegetation.cover))*100,2)
```

```
percentages
x<-barplot(table(gaadata$Vegetation.cover), main="Barplot of the
    proportion vegetation cover", xlab="Proportion of vegetation
    cover", ylab="Frequency",ylim=c(0,180))
text(x,frequencies+10, paste(percentages,"%"))
dev.print(file="Barplot_vgc.eps")
#-----------------------------------------------
# One-Way ANOVA model: GAA over dam.code
#-----------------------------------------------
oneway.test(gaadata$GAA ~ gaadata$Dam.Code)
# Assuming equal variances
oneway.test(gaadata$GAA ~ gaadata$Dam.Code, var.equal = TRUE)
anova2<-anova(lm (gaadata$GAA ~ gaadata$Dam.Code))
## Non-parametric test: Kruskal-Wallis
kruskal.test(gaadata$GAA~gaadata$Dam.Code)
#-----------------------------
# Fit the linear model
#-----------------------------
# G.affins(response variable) against: "dam age(years), temperature
    , GIA..GIL, GLC..GOB, & Vegetation cover"(explanatory variables)
fit1 <- lm(gaadata$GAA~ gaadata$Dam.Age..years. + gaadata$Mean.Temp
    + gaadata$GLC..GOB. + gaadata$GIA..GIL. + gaadata$Vegetation.
    cover)
summary(fit1)
# Assessing the model fit
par(mfrow=c(2,2))
plot(fit1)
par(mfrow=c(1,1))
# Testing the residuals of the fitted model
shapiro.test(fit1$residuals)
bartlett.test(GAA~Dam.Age..years.,gaadata)
# Install Package
library(ggplot2)
# From reference : "Winston Chang" >>>
# http://www.cookbook-r.com/Graphs/Plotting_means_and_error_bars_(
    ggplot2)/
GAASummaryMean<-summarySE(gaadata, measurevar = "GAA",groupvars = c
    ("Dam.Code","Sampling.Event"))
# Plot is of ln(Gaa+1)
sqrtGAASummaryMean<-summarySE(gaadata, measurevar = "sqrtGAA",
```

```
    groupvars = c("Dam.Code","Sampling.Event"))
# 95% confidence interval of the mean
# They are one per figure so
pd<-position_dodge(0.5)
ylabs=expression(paste(italic('G. affinis'), "  CPUE (95% CI)"))
## Avoca (AVO)
ggplot(sqrtGAASummaryMean|sqrtGAASummaryMean$Dam.Code=="AVO",|,
        aes(x =Sampling.Event, y =sqrtGAA, colour =Dam.Code)) +
  geom_errorbar(aes(ymin =sqrtGAA-ci, ymax =sqrtGAA+ci), width=.1,
     colour="black",position = pd) +
  geom_line(position = pd) +
  geom_point(position = pd, size=3) +
  xlab("Sampling event") +
  ylab(ylabs) +
  ggtitle("Mean and standard error of CPUE by dam and sampling
     event")

# Disco Chicks (DC)
ggplot(sqrtGAASummaryMean|sqrtGAASummaryMean$Dam.Code=="DC",|,
        aes(x=Sampling.Event, y=sqrtGAA, colour=Dam.Code)) +
  geom_errorbar(aes(ymin=sqrtGAA-ci, ymax=sqrtGAA+ci), width=.1) +
  geom_line() +
  geom_point()+
  xlab("Sampling event") +
  ylab(ylabs)+
  ggtitle("Mean and standard error of CPUE by dam and sampling
     event")

#  Sur le Sun (SLS)
ggplot(sqrtGAASummaryMean|sqrtGAASummaryMean$Dam.Code=="SLS",|,
        aes(x=Sampling.Event, y=sqrtGAA, colour=Dam.Code)) +
  geom_errorbar(aes(ymin=sqrtGAA-ci, ymax=sqrtGAA+ci), width=.1) +
  geom_line() +
  geom_point()+
  xlab("Sampling event") +
  ylab(ylabs)+
  ggtitle("Mean and standard error of CPUE by dam and sampling
     event")

# Olifantsklip (OLI)
```

```
ggplot(sqrtGAASummaryMean|sqrtGAASummaryMean$Dam.Code=="OLI",|,
       aes(x=Sampling.Event, y=sqrtGAA, colour=Dam.Code)) +
  geom_errorbar(aes(ymin=sqrtGAA-ci, ymax=sqrtGAA+ci), width=.1) +
  geom_line() +
  geom_point()+
  xlab("Sampling event") +
  ylab(ylabs)+
  ggtitle("Mean and standard error of CPUE by dam and sampling
     event")

# Dunbrody
ggplot(sqrtGAASummaryMean|sqrtGAASummaryMean$Dam.Code=="DB",|,
       aes(x=Sampling.Event, y=sqrtGAA, colour=Dam.Code)) +
  geom_errorbar(aes(ymin=sqrtGAA-ci, ymax=sqrtGAA+ci), width=.1) +
  geom_line() +
  geom_point()+
  xlab("Sampling event") +
  ylab(ylabs)+
  scale_colour_hue(name="Dam",     # Legend label, use darker colors
                   breaks=c("AVO"),
                   labels=c("Avoca"),
                   l=40) +                    # Use darker colors,
                      lightness=40
  ggtitle("CPUE")

# All dams
pd <- position_dodge(0.2) # move them .05 to the left and right
ggplot(sqrtGAASummaryMean, aes(x=Sampling.Event, y=sqrtGAA, colour=
   Dam.Code, group=Dam.Code)) +
  geom_errorbar(aes(ymin=sqrtGAA-ci, ymax=sqrtGAA+ci), width=.1,
     position=pd, colour="black") +
  geom_line(position=pd) +
  geom_point(position=pd, size=3, shape=21, fill="white") + # 21 is
      filled circle
  xlab("Sampling event") +
  ylab(ylabs) +
  scale_colour_hue(name="Dam",     # Legend label, use darker colors
                   breaks=c("AVO", "DC", "DB", "OLI", "SLS"),
                   labels=c("Avoca", "Disco Chicks", "Dunbrody", "
                      Olifantsklip", "Sur le Sun"),
```

```
                    l=40) +                    # Use darker colors,
                       lightness=40
  ggtitle("Mean and standard error of CPUE by dam and sampling
     event") +
  expand_limits(y=0) +                         # Expand y range
  scale_y_continuous(breaks=0:20*1) +          # Set tick every 1
     unit CPUE
  theme_bw() +
  theme(legend.justification=c(1,0),
        #       legend.position=c(2,6))              # Position
           legend in bottom right
        #       legend.position="top")             # Position
           legend in bottom right
        legend.position=c(0.3,0.7))             # Position legend
            in almost top almost left
#-----------------------------
# One-way ANOVA Example
#-----------------------------
# One-way ANOVA example from Wackerly Dennis
A<-c(170, 146, 120, 112, 132)
B<-c(224, 196, 163, 231, 195)
C<-c(155, 153, 104, 143,198)
ybarA<-mean(A)
ybarB<-mean(B)
ybarC<-mean(C)
grandmean<-mean(allpoints)
allpoints<-c(170,146, 120, 112, 132,224, 196, 163, 231, 195,155,
   153, 104, 143,198)
Sample<-c(rep("A",5),rep("B",5),rep("C",5))
treats<-data.frame(allpoints,Sample)
mmod<-(lm(allpoints~Sample,data=treats))
summary(mmod)
anova(mmod)
## Diagnostic plots for the one-way random effects model:
qqnorm(resid(mmod),main = "Normal Q-Q plot")
qqline(resid(mmod))
plot(fitted(mmod),resid(mmod),xlab = "Fitted values",ylab = "
   Residuals",main = "Residuals vs Fitted")
abline(0,0)
# Boxplot for generated sample data:
```

```
boxplot(allpoints~Sample,main="Boxplot of the dependent variable
    for treatments A, B and C",xlab="Treatment",ylab="Outcome")
# Residuals plots (Diagnostics)
plot(mmod)
## Non-parametric test for normality
kruskal.test(allpoints~Sample,data = treats)
shapiro.test(allpoints)
## Testing equality of variances
bartlett.test(allpoints~Sample,data = treats)
#-------------------
# One-way ANOVA
#-------------------
## Test the null hypothesis that no variation among dams
dams<-aov(GAA ~ Dam.Code,data = gaadata)
anova(dams)
## Model Diagnostics plots
plot(dams)
#-------------------------------------------------------------
# One-way random effect ANOVA example from Murray Logan
#-------------------------------------------------------------
library(nlme)
library(lmerTest)
medley<-read.csv(file = file.choose(),header = TRUE)

# Fit one-way random effect model
medley.model1<-lme(DIVERSITY~1,random = ~1|STREAM,method = "ML",
    data =medley)
summary(medley.model1)
VarCorr(medley.model1)
medley.model2<-lme(DIVERSITY~1,random = ~1|STREAM,method = "REML",
    data =medley)
summary(medley.model2)
VarCorr(medley.model2)
## Model diagnostics plots
plot(medley.model1)
qqnorm(resid(medley.model1))
qqline(resid(medley.model1))

# Testing the normality of residuals
shapiro.test(medley.model1$residuals)
```

```
#-------------------------
# Fit the ANOVA model
#-------------------------
medly2<-read.csv(file = file.choose(),header = TRUE)
medly.aov<-aov(DIVERSITY ~ ZINC, medly2)
anova(medly.aov)
# Boxplots for diversity against zinc
boxplot(DIVERSITY ~ ZINC,medly2,xlab="Zinc",ylab="Diversity",main="
   Boxplot of species diversity against zinc concentration")
# residual plot
plot(medly.aov)
## Model assumptions
qqnorm(resid(medly.aov))
qqline(resid(medly.aov))
# Testing normality
shapiro.test(resid(medly.aov))
#-------------------------------------------------------------
## One-way random effect ANOVA example by dam or location
#-------------------------------------------------------------
## Estimate random factors and residuals
## Two methods for estimating parameters: ML and REML
model.random1<-lme(GAA ~ 1, random = ~1|Dam.Code,method = "ML",
   data = gaadata)
model.random2<-lme(GAA ~ 1, random = ~1|Dam.Code,method = "REML",
   data = gaadata)

# Model assessment
plot(model.random1)
plot(model.random2)
qqnorm(resid(model.random1))
qqline(resid(model.random1))
qqnorm(resid(model.random2))
qqline(resid(model.random2))
# Testing the residuals of the fitted model
shapiro.test(model.random1$residuals)
bartlett.test(GAA~Dam.Code,gaadata)
#-------------------------
# Two-way ANOVA Models
#-------------------------
```

```
Dam.Code <- as.factor(gaadata$Dam.Code)
## Two-way ANOVA model by dam and sampling event
anova(lm(gaadata$GAA ~ gaadata$Dam.Code * gaadata$Sampling.Event))
# Two_way ANOVA example from Faraway
library(faraway)
data(rats)
mod<-lm(time~poison*treat, data = rats)
anova(mod)

## Model diagnostics
plot(mod)
# Construct boxplots
boxplot(rats$time~rats$treat,xlab="Treatment",ylab="Time")
boxplot(rats$time~rats$poison,xlab="Poison",ylab="Time")
plot(fitted(mod),residuals(mod),xlab = "Fitted",ylab = "Residuals")
qqnorm(residuals(mod), main = "")
# Testing the residuals
shapiro.test(mod$residuals)
bartlett.test(time~poison,rats)
bartlett.test(time~treat,rats)
#-----------------------------------------------------------------------
# Two-way random effects ANOVA model: G. affinis CPUE by dam
#-----------------------------------------------------------------------
gaadata$Vegetation.cover<-as.factor(gaadata$Vegetation.cover)
gaadata$Dam.Code<-as.factor(gaadata$Dam.Code)
# Fit two-way random effects model
random.model<-lme(GAA ~ Dam.Code*Vegetation.cover, random = ~1|Dam.
    Code,method="ML",data = gaadata)
# Check the design
table(gaadata$Dam.Code,gaadata$Vegetation.cover)
# Fixed effects: fit the factorial linear model
fixed.model<-aov(GAA ~Dam.Code*Vegetation.cover,data = gaadata)
summary(fixed.model)

## Model diagnostics plots
qqnorm(resid(fixed.model))
qqline(resid(fixed.model))
plot(fitted(fixed.model),resid(fixed.model),xlab = "Fitted",ylab =
    "Residuals")
abline(0,0)
```

```
#-----------------
## Mixed Anova
#-----------------
mixed.model<-lme(GAA ~Dam.Code*Vegetation.cover,random = ~1|Dam.
    Code,method = "ML",data = gaadata)
#-------------------
# NESTED ANOVA
#-------------------
# One approach to fit a nested model ANOVA is to use a mixed model.
# Dam.Code is a random effect & site is a  fixed effect.
## Nested model
library(nlme)
library(lme4)
# Fit the nested random effect model
nested.lme<-lme(GAA~Dam.Code,random=~1|site,gaadata,method="REML")
summary(nested.lme)
plot(nested.lme)
nested.lmer<-lmer(GAA~Dam.Code+(1|site),gaadata,REML = T)
summary(nested.lmer)
plot(nested.lmer)
# No issues from residuals in both methods
# Using the aov fuction for a nested anova.
# aov is designed for balanced designs.
# Balanced designs can be checked with the replications function.
replications(y ~ Dam.Code+site, data = gaadata)
# Fit nested ANOVA with aov function:
nested.anova<-aov(GAA ~ Dam.Code+Error(site), data = gaadata)
nested.anova
summary(nested.anova)

# Model diagnostics plots
par(mfrow=c(2,2))
plot(lm(nested.anova))
par(mfrwo=c(1,1))
# Testing the residuals of the fitted model
shapiro.test(nested.anova$Within$residuals)
bartlett.test(GAA~Dam.Code,gaadata)
```

Listing A.1: Appendix: R code for the G. affinis data

A.1

```
# Install the nlme package
> library(nlme)
# Fit a one-way random effects ANOVA model by stream as
    random factor.
# The first argument indicates that the response is
   diversity and that there is a single fixed effect,
   the intercept.
# The argument on the right hand side of the '|' sign
   is a nominal variable.
# ~1|STREAM specifies the model for the random effects
   and that STREAM is the grouping structure. The 1
   indicates that the random effect is constant within
   each group.
# The method = "ML" or method = "REML" specifies which
   estimation method to be used.
> medley.model1<-lme(DIVERSITY~1,random = ~1|STREAM,
   method = "ML",data =medley)
Linear mixed-effects model fit by maximum likelihood
 Data: medley
 AIC       BIC       logLik
57.5001  62.0792  -25.7501
Random effects:
Formula: ~1 | STREAM
       (Intercept)  Residual
StdDev:  0.0996     0.5071
Fixed effects: DIVERSITY ~ 1
             Value  Std.Error  DF   t-value   p-value
(Intercept)  1.6936   0.0977   28   17.3287         0
Number of Observations: 34
Number of Groups: 6
# Extract variance and correlation components for ML
   method
> VarCorr(medley.model1)
STREAM = pdLogChol(1)
             Variance    StdDev
(Intercept) 0.0099       0.0996
Residual    0.2572       0.5071
```

```
# REML method
>medley.model2<-lme(DIVERSITY~1,random = ~1|STREAM,
   method = "REML",data =medley)
Linear mixed-effects model fit by REML
  Data: medley
  AIC       BIC      logLik
60.2562  64.7457  -27.1281
Random effects:
Formula: ~1 | STREAM
         (Intercept)  Residual
StdDev:    0.1433      0.5075
Fixed effects: DIVERSITY ~ 1
               Value  Std.Error  DF  t-value   p-value
(Intercept)   1.6933  0.1053    28  16.0795         0
Number of Observations: 34
Number of Groups: 6
# Extract variance and correlation components for REML
   method
> VarCorr(medley.model2)
STREAM = pdLogChol(1)
             Variance   StdDev
(Intercept) 0.0205      0.1433
Residual    0.2576      0.5075
```

A.2

```
# Fit a one-way random effects ANOVA model
> dams<-aov(GAA ~ Dam.Code,data = gaadata)
> libray(nlme)
# ML method
> model.random1<-lme(GAA ~ 1, random = ~1|Dam.Code,
   method = "ML", data = gaadata)
# ~1|Dam.Code specifying the model for the random ef
   There is no evidence that these data violate the
   model assumptions and as a result inference based on
   this model is valid.fects with Dam.Code as the
   grouping structure and 1 indicating that the random
   effect is constant within each group.
```

```
Linear mixed-effects model fit by maximum likelihood
   Data: gaadata
    AIC      BIC     logLik
2888.602  2899.753  -1441.301
Random effects:
 Formula: ~1 | Dam.Code
        (Intercept) Residual
StdDev:  2.7464     27.6138
Fixed effects: GAA ~ 1
             Value  Std.Error   DF   t-value  p-value
(Intercept) 21.8656   2.0103    299  10.8768        0
Number of Observations: 304
Number of Groups: 5

# REML method
> model.random2<-lme(GAA ~ 1, random = ~1|Dam.Code,
   method = "REML", data = gaadata)
Linear mixed-effects model fit by REML
   Data: gaadata
    AIC      BIC     logLik
2885.264   2896.405  -1439.632
Random effects:
Formula: ~1 | Dam.Code
       (Intercept)  Residual
StdDev:  3.537864   27.61438
Fixed effects: GAA ~ 1
             Value  Std.Error   DF   t-value   p-value
(Intercept)  21.8398  2.2421    299  9.7409        0
Number of Observations: 304
Number of Groups: 5
```

A.3

```
# Fit a two-way nested ANOVA model
aov(formula = GAA ~ Dam.Code + Error(site), data =
   gaadata)
Grand Mean: 21.9375
Stratum 1: site
```

```
Terms:
                   Dam.Code  Residuals
Sum of Squares    6 083.08   31 756.67
Deg. of Freedom          4          14
Residual standard error: 47.6270
Estimated effects may be unbalanced

Stratum 2: Within
Terms:
                     Residuals
Sum of Squares       196 268.1
Deg. of Freedom            285

Residual standard error: 26.2423
> summary(nested.anova)
Error: site
           Df Sum Sq  Mean Sq  F value Pr(>F)
Dam.Code    4   6 083   1 521     0.67   0.623
Residuals  14  31 757   2 268

Error: Within
               Df  Sum Sq   Mean Sq  F value Pr(>F)
Residuals     285  196 268    688.7
```

Appendix B

Appendix B: R Code for GLM Fit

```
# 06/11/2017
rm(list = ls())
gaadata <-read.csv(file=file.choose(),header=TRUE)
#-----------------------------
## Fitting the glm model
#-----------------------------
# G.affins(response variable) against: "dam age(years), temperature
    , GIA..GIL, GLC..GOB, & Vegetation cover"(explanatory variables)
# Response is discrete , family is poisson
glm.model2<-glm(gaadata$GAA ~ gaadata$Dam.Age..years.+gaadata$Mean.
    Temp+gaadata$Vegetation.cover+gaadata$GLC..GOB.+gaadata$GIA..GIL
    .,
                family = "poisson",data = gaadata)
summary(glm.model2)
# Model diagnostics plots
par(mfrow=c(2,2))
plot(glm.model2)
par(mfrow=c(1,1))
# Testing the residuals of the fitted model
shapiro.test(glm.model2$residuals)
bartlett.test(GAA~Dam.Age..years.,gaadata)
#
    ------------------------------------------------------------------


## The GLM fit to bag weight: CPUE against Temperature and pressure
#
    ------------------------------------------------------------------
```

```
tdata<-read.csv(file=file.choose(),header = TRUE)
glmfit<-glm(Bweight~Minimum+Maximum+Pressure,data=tdata,family=
    gaussian)
summary(glmfit)
# Model diagnostics plots
par(mfrow=c(2,2))
plot(glmfit)
par(mfrow=c(1,1))
# Testing the residuals of the fitted model
shapiro.test(glmfit$residuals)
bartlett.test(Bweight~Minimum,tdata)
bartlett.test(Bweight~Maximum,tdata)
bartlett.test(Bweight~Pressure,tdata)
#-----------------------------------------------
## Construct line graphs for temperatures
#-----------------------------------------------
## The simple way!
library(ggplot2)
xLab<-c("17/05/2015","20/03/2016","22/05/2016","24/07/2016","9/08/
    2015","11/10/2015","15/03/2015","21/02/2016","12/04/2015","22/11
    /2015","23/10/2016","25/09/2016")
x<- seq(from=1, to=12, by =1)
min<-Temperature$Minimum
max<-Temperature$Maximum
press<-Temperature$Pressure
mytemp<-data.frame(min,max,press)
lab = expression(paste("Temperature (",degree,"C)"))
mainlab<-expression(paste("Line plot of the temperature (",degree,"
    C)"))
mytemp2 <-data.frame(mytemp,xLab =as.Date(xLab, format = "%d/%m/%Y"
    ))
str(mytemp2)
## Using ggplot2 package
library(ggplot2)
tp<-ggplot(mytemp,aes(Temperature$Date)) +
  #  scale_x_continuous(name="Date", breaks=c(1:12), labels=xLabels
      ) +
  #  scale_x_continuous(name="Date", labels=xLabels) +
   geom_line(aes(y=min,colour="min")) +
   geom_line(aes(y=max,colour="max")) +
```

```
  geom_point(aes(y=min,colour="min")) +
  geom_point(aes(y=max,colour="max")) +
  ggtitle(mainlab) +
  labs(x="Date",y=lab)+
  theme(axis.title.x = element_text(face="bold", colour="#990000",
      size=15),
        axis.text.x  = element_text(angle=90, vjust=0.5, size=8))

mytemp2 <-data.frame(mytemp,xLab =as.Date(xLab, format = "%d/%m/%Y"
  ))
str(mytemp2)
tp <-
  ggplot(mytemp2,aes(xLab, min)) +
  scale_x_date(date_labels = "%d %b %Y", breaks = sort(mytemp2$xLab
    ))+
  geom_point(aes(y=mytemp2$min,colour="min")) +
  geom_point(aes(y=mytemp2$max,colour="max")) +
  geom_line(aes(y=mytemp2$min, colour="min")) +
  geom_line(aes(y=mytemp2$max,colour="max")) +
  ggtitle(mainlab) +
  labs(x="Date",y=lab)
## From: http://www.cookbook-r.com/Graphs/Axes_(ggplot2)/
tp + theme(axis.title.x = element_text(face="bold", colour="#990000
  ", size=15),
           axis.text.x  = element_text(angle=90, vjust=0.5, size=8)
             )
#-----------------------------------------------------------
## Construct the Line Pressure for tournament days
#-----------------------------------------------------------
xLab<-c("17/05/2015","20/03/2016","22/05/2016","24/07/2016","9/08/
  2015","11/10/2015","15/03/2015","21/02/2016","12/04/2015","22/11
  /2015","23/10/2016","25/09/2016")
x<- seq(from=1, to=12, by =1)
min<-Temperature$Minimum
max<-Temperature$Maximum
press<-Temperature$Pressure
mytemp<-data.frame(min,max,press)
mytemp2 <-data.frame(mytemp,xLab =as.Date(xLab, format = "%d/%m/%Y"
  ))
```

```
ggplot(mytemp, aes(Temperature$Date))+
  geom_line(aes(y=press, colour="press"))+
  geom_point(aes(y=press, colour="press"))+
  ggtitle("Line plot of the pressure (hPa)")+xlab("Date")+
  ylab("Pressure (hPa)")
mytemp2 <-data.frame(mytemp,xLab =as.Date(xLab, format = "%d/%m/%Y"
  ))
tp <-
  ggplot(mytemp2,aes(xLab, min)) +
  scale_x_date(date_labels = "%d %b %Y", breaks = sort(mytemp2$xLab
    ))+
  geom_point(aes(y=mytemp2$press, colour="pressure")) +
  geom_line(aes(y=mytemp2$press, colour="pressure")) +
  theme_bw()+
  ggtitle("Line plot of the pressure (hPa)") +
  xlab("Date") +
  ylab("Pressure (hPa)")
## From: http://www.cookbook-r.com/Graphs/Axes_(ggplot2)/
tp + theme(axis.title.x = element_text(face="bold", colour="#990000
  ", size=15),
           axis.text.x  = element_text(angle=90, vjust=0.5, size=8)
             )
#
```

```
gotelli<-read.csv(file = file.choose(),header = T)
str(gotelli)
gotelli.glm<-glm(Srich~Habitat+Latitude+Elevation,family = poisson,
  data = gotelli)
summary(gotelli.glm)
# Model diagnostics plots
par(mfrow=c(2,2))
plot(gotelli.glm)
par(mfrow=c(1,1))
# Testing the residuals of the fitted model
shapiro.test(gotelli.glm$residuals)
bartlett.test(Srich~Elevation,gotelli)
bartlett.test(Srich~Habitat,gotelli)
bartlett.test(Srich~Latitude,gotelli)
```

Appendix C

Appendix C: R Code for Analysis of GLM Fit Methodology

```
# 03/10/2017
## Analysis of GLM fit example from Murray Logan
rm(list=ls())
sinclair<-read.csv(file = file.choose(),header = TRUE)
str(sinclair)
#-----------------------------
# Descriptive statistics
#-----------------------------
qqnorm(sinclair$COUNT,xlab = "Theoretical quantiles",ylab ="Data (
    Observed) quantiles")
qqline(sinclair$COUNT,col="red")
boxplot(sinclair$COUNT~sinclair$SEX,xlab="Sex",ylab="Frequency",
    main="Boxplot of carcasses by  sex")
boxplot(sinclair$COUNT~sinclair$MARROW,xlab="Marrow",ylab="
    Frequency",main="Boxplot of carcasses by  marrow type")
boxplot(sinclair$COUNT~sinclair$DEATH,xlab="Death",main="Boxplot of
     carcasses by  death",ylab="Frequency")
# Testing homogeneity of variances
bartlett.test(COUNT~DEATH,sinclair)
bartlett.test(COUNT~SEX,sinclair)
bartlett.test(COUNT~MARROW,sinclair)
## Fit ANOVA model
model.anova<-lm(COUNT~SEX+DEATH+MARROW,data = sinclair)
anova(model.anova)

## Model diagnostics plots
```

```
par(mfrow=c(2,2))
plot(model.anova)
par(mfrow=c(1,1))
# Testing the residuals of the fitted model
shapiro.test(model.anova$residuals)
#-----------------------
#Fitting GLMs
#-----------------------
## Full or satured model
sinclair.glm <- glm(COUNT~SEX*MARROW*DEATH, family = poisson, data =
    sinclair)
summary(sinclair.glm)
#Generate a set of models with combinations (subsets) of terms in
    the global model, with optional rules for model inclusion.
library(MuMIn)
options(na.action = "na.fail")
dredge(sinclair.glm, rank="AIC")
# Compare and contrast to a hierarchical approach
## Complete and  Conditional Independence
# Dropping ABC (sex:marrow:death)
sinclair.glm1<-update(sinclair.glm, ~.-SEX:MARROW:DEATH, family =
    poisson,data = sinclair)
# The first argument is the result of a fit, and the second an
    updating formula.
# The place holder ~ separates the response from the predictors.
# The dot . refers to the right hand side of the original formula,
    so here we simply remove SEX:MARROW:DEATH
summary(sinclair.glm1)
# Model comparison
anova(sinclair.glm,sinclair.glm1,test = "Chisq")
drop1(sinclair.glm,test = "Chisq")

## Dropping AB (sex:marrow)
sinclair.glm2<-update(sinclair.glm1, ~.-SEX:MARROW, family = poisson
    ,data = sinclair)
# Model comparison
anova(sinclair.glm1,sinclair.glm2,test = "Chisq")
## Drop criteria
drop1(sinclair.glm2,test = "Chisq")
```

```
## Dropping AC (sex:death)
sinclair.glm3<-update(sinclair.glm1, ~.-SEX:DEATH, family = poisson,
   data = sinclair)
# Model comparison
anova(sinclair.glm1,sinclair.glm3,test = "Chisq")
## Drop criteria
drop1(sinclair.glm3,test = "Chisq")

## Dropping BC (marrow:death)
sinclair.glm4<-update(sinclair.glm1, ~.-MARROW:DEATH, family =
   poisson,data = sinclair)
# Model comparison
anova(sinclair.glm1,sinclair.glm4,test = "Chisq")
## Drop criteria
drop1(sinclair.glm4,test = "Chisq")

## Model diagnostics plots
par(mfrow=c(2,2))
plot(sinclair.glm3)
# Testing the residuals of the fitted model
shapiro.test(sinclair.glm3$residuals)
```

90G__thesis_GLM_Analysis_of_GLM_Fit.R

C.1

```
# Model 1
> drop1(sinclair.glm,test = "Chisq")
Single term deletions
Model: COUNT ~ SEX * MARROW * DEATH
                Df Deviance     AIC     LRT Pr(>Chi)
<none>               0.0000 79.226
SEX:MARROW:DEATH  2   7.1883 82.414 7.1883  0.02748 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

 # Model 2 (Dropping SEX:MARROW)
Model:
COUNT ~ SEX + MARROW + DEATH + SEX:DEATH + MARROW:DEATH
            Df Deviance      AIC      LRT  Pr(>Chi)
<none>           13.156   84.382
```

```
SEX:DEATH      1   13.243  82.468  0.0866    0.7685
MARROW:DEATH   2   42.676 109.901 29.5199 3.889e-07 ***
---
# Model 3 (Dropping SEX:DEATH)
Model: COUNT ~ SEX + MARROW + DEATH + SEX:MARROW +
   MARROW:DEATH
             Df Deviance     AIC     LRT  Pr(>Chi)
<none>              8.465  81.690
SEX:MARROW     2   13.243  82.468  4.7779   0.09173 .
MARROW:DEATH   2   37.985 107.210 29.5199 3.889e-07 ***
---
# Model 4 (Dropping DEATH:MARROW)
Model: COUNT ~ SEX + MARROW + DEATH + SEX:MARROW + SEX:
   DEATH
           Df Deviance    AIC    LRT Pr(>Chi)
<none>           37.898 109.12
SEX:MARROW  2   42.676 109.90 4.7779  0.09173 .
SEX:DEATH   1   37.985 107.21 0.0866  0.76853
---
```

Appendix D

Appendix D: R Code for Repeated Measures

```
# 06/11/2017
# Tournament data provided by Albany Angling Association
## One fish species in each tournament day
## Water bodies: Mangazana, Settlers, White dam and Yarrow.
## Fishing months in the year 2016: Jul, Oct, May, Mar, Feb and Sep.
## Fishing months in 2015: Mar, Apr, May, Aug, Oct and Nov.
rm(list=ls()) # Clear objects from the workspace
tdata<-read.csv(file=file.choose(),header = TRUE)
f1<-tdata$F1
f2<-tdata$F2
f3<-tdata$F3
bw<-tdata$Bweight
wb<-tdata$WaterBody ## Water body
ps<-tdata$Person ## Anglers

# Data frame: 171 observations of 24 variables
str(tdata) # Data structure checked
levels(tdata$Category)
# Table for a single variable
table(tdata$WaterBody)
#-----------------------------
# Descriptive Statistics
#-----------------------------
## Normality Q-Q plot of fish CPUE
par(mfrow=c(2,2)) # Combine the pictures
qqnorm(tdata$F1,main="Normal Q-Q plot for fish one weight")
```

```
qqline(tdata$F1,lwd=2,col="blue")
qqnorm(tdata$F2,main="Normal Q-Q plot for fish two weight")
qqline(tdata$F2,lwd=2,col="blue")
qqnorm(tdata$F3,main="Normal Q-Q plot for fish three weight")
qqline(tdata$F3,lwd=2,col="blue")
qqnorm(tdata$Bweight,main="Normal Q-Q plot for bag weight")
qqline(tdata$Bweight,lwd=2,col="blue")
par(mfrow=c(1,1)) # Save the figure as one picture

## Boxplots CPUE by years
par(mfrow=c(2,2))
boxplot(tdata$F1 ~ tdata$Year,xlab="Year",main= "Boxplot for fish
   one weight",ylab="Weight (kg)")
boxplot(tdata$F2 ~ tdata$Year, xlab= "Year",main= "Boxplot for fish
    two weight",ylab="Weight (kg)")
boxplot(tdata$F3 ~ tdata$Year, xlab="Year",main= "Boxplot for fish
   three weight",ylab="Weight (kg)")
boxplot(tdata$Bweight ~ tdata$Year, xlab="Year",main= "Boxplot for
   bag weight",ylab="Weight (kg)")
par(mfrow=c(1,1))

## Boxplot CPUE by waterbody for the years
par(mfrow=c(2,2))
boxplot(tdata$F1 ~ tdata$WaterBody,main="Fish one weight",ylab="
   Weight (kg)",las=3)
boxplot(tdata$F2 ~ tdata$WaterBody,main="Fish two weight",ylab="
   Weight (kg)",las=3)
boxplot(tdata$F3 ~ tdata$WaterBody,main="Fish three weight",ylab="
   Weight (kg)",las=3)
boxplot(tdata$Bweight ~ tdata$WaterBody,main="Bag weight",ylab="
   Weight (kg)",las=3)
par(mfrow=c(1,1))

# Histogram of fish CPUE
par(mfrow=c(2,2))
hist(tdata$F1,xlab = "Fish one weight",main = "Histogram for fish
   one weight")
hist(tdata$F2,xlab = "Fish two weight",main = "Histogram for fish
   two weight")
hist(tdata$F3,xlab = "Fish three weight",main = "Histogram for fish
```

```
    three weight")
hist(tdata$Bweight,xlab = "Bag weight",main = "Histogram for
   bagweight")
par(mfrow=c(1,1))

## Check normality using non-parametric method: Multivariate
   Shapiro-Wilk test
library(mvShapiroTest)
f1<-tdata$F1 # Extracting fish weights from the main data set
f2<-tdata$F2
f3<-tdata$F3
bgw<-tdata$Bweight
fishes<-cbind(f1,f2,f3,bgw) # Combining individual fish weights
mvShapiro.Test(fishes)
# Bar graph for the distribution of recorded data from 4
   waterbodies for the year.
library(ggplot2)
# ggplot() is used to construct the initial plot object, and is
   almost always followed by + to add component to the plot.
# aesthetics: measurement variables (x,y)

# Waterbody
ggplot(tdata,aes(x=WaterBody))+geom_bar()+
   theme_classic()+
  ylab("Total number of participants")+
  xlab("Water body")+
  ggtitle("The total number of anglers who participated
          at events at each of the venues.")

# Category or sex CPUE
ggplot(Data,aes(x=Category))+geom_bar()+
  theme_classic()+
  xlab("Sex")+
  ylab("Total number of participants")+
  ggtitle("The total number of adult male, female and junior
           angler who participated at an event.")

# Fishing Months CPUE
ggplot(tdata,aes(x=Month))+geom_bar()+
  theme_classic()+
```

```
  ylab("Total catch")+
  xlab("Month")+
  ggtitle("Total catch of all anglers at events, by month.")

## Generating summary statistics for tournaments
summarySE <- function(data=gaadata, measurevar= gaadata$GAA,
   groupvars= Dam.Code, na.rm=FALSE,
                      conf.interval=.95, .drop=TRUE) {
  library(plyr)
  # New version of length which can handle NA's: if na.rm==T, don't
     count them
  length2 <- function (x, na.rm=FALSE) {
    if (na.rm) sum(!is.na(x))
    else       length(x)
  }
  # This does the summary. For each group's data frame, return a
     vector with
  # N, mean, and sd
  datac <- ddply(data, groupvars, .drop=.drop,
                 .fun = function(xx, col) {
                   c(N    = length2(xx[[col]], na.rm=na.rm),
                     mean = mean   (xx[[col]], na.rm=na.rm),
                     sd   = sd     (xx[[col]], na.rm=na.rm)
                   )
                 },
                 measurevar
  )
  datac$se <- datac$sd / sqrt(datac$N)  # Calculate standard error
     of the mean
  # Confidence interval multiplier for standard error
  # Calculate t-statistic for confidence interval:
  # e.g., if conf.interval is .95, use .975 (above/below), and use
     df=N-1
  ciMult <- qt(conf.interval/2 + .5, datac$N-1)
  # t*se
  datac$ci <- datac$se * ciMult
  # lower bound: bar(x) - t*(se)
  print(datac$mean)
  datac$Lower <- datac$mean - datac$ci
  # print(lower.bound)
```

```
  # Upper bound: bar(x) + t*(se)
  datac$Upper <- datac$mean + datac$ci
  # Rename the "mean" column
  datac <- rename(datac, c("mean" = measurevar))
  return(datac)
}
summarySE(tdata,measurevar = "Bweight", groupvars = c("WaterBody","
    Month","Year","Date"))
##  Extracting the tournament days, months, year and climate by row
     numbers in the main data set.
temp<-subset(tdata, Date==24,Month=July,Pressure=Pressure,Year=Year
    ,select = c(Minimum,Maximum,Date,Month,Pressure,Year))
temp[4,]
July.temp<-temp[4,]

temp.may<-subset(tdata,Date==22,Month=May,Pressure=Pressure,Year=
    Year,select = c(Minimum,Maximum,Date,Month,Pressure,Year))
temp.may[5,]
May.temp<-temp.may[5,]

temp.mar<-subset(tdata,Date=20,Month=March,Pressure=Pressure,Year=
    Year,select = c(Minimum,Maximum,Date,Month,Pressure,Year))
temp.mar[36,]
Mar.temp<-temp.mar[36,]

Temp.May<-subset(tdata,Date=17,Month=May,Pressure=Pressure,Year=
    Year,select = c(Minimum,Maximum,Date,Month,Pressure,Year))
Temp.May[51,]
May.seventeen<-Temp.May[51,]

Temp.Mar<-subset(tdata,Date=15,Month=March,Pressure=Pressure,Year=
    Year,select = c(Minimum,Maximum,Date,Month,Pressure,Year))
Temp.Mar[65,]
March.temp<-Temp.Mar[65,]

Temp.Aug<-subset(tdata,Date=9,Month=August,Pressure=Pressure,Year=
    Year,select = c(Minimum,Maximum,Date,Month,Pressure,Year))
Temp.Aug[80,]
Aug.temp<-Temp.Aug[80,]
```

```
Temp.Oct<-subset(tdata,Date=11,Month=October,Pressure=Pressure,Year
    =Year,select = c(Minimum,Maximum,Date,Month,Pressure,Year))
Temp.Oct[85,]
Oct.temp<-Temp.Mar[85,]

Temp.Feb<-subset(tdata,Date=21,Month=February,Pressure=Pressure,
    Year=Year,select = c(Minimum,Maximum,Date,Month,Pressure,Year))
Temp.Feb[96,]
Feb.temp<-Temp.Mar[96,]

Temp.Apr<-subset(tdata,Date=12,Month=April,Pressure=Pressure,Year=
    Year,select = c(Minimum,Maximum,Date,Month,Pressure,Year))
Temp.Apr[115,]
Apr.temp<-Temp.Apr[115,]

Temp.Nov<-subset(tdata,Date=22,Month=November,Pressure=Pressure,
    Year=Year,select = c(Minimum,Maximum,Date,Month,Pressure,Year))
Temp.Nov[132,]
Nov.temp<-Temp.Nov[132,]

Temp.October<-subset(tdata,Date=23,Month=October,Pressure=Pressure,
    Year=Year,select = c(Minimum,Maximum,Date,Month,Pressure,Year))
Temp.October[141,]
October.temp<-Temp.October[141,]

Temp.Sep<-subset(tdata,Date=25,Month=September,Pressure=Pressure,
    Year=Year,select = c(Minimum,Maximum,Date,Month,Pressure,Year))
Temp.Sep[156,]
Sept.temp<-Temp.Sep[156,]
Temperature<-rbind(May.temp,July.temp,Mar.temp,May.seventeen,March.
    temp,Aug.temp,
                   Oct.temp,Feb.temp,Apr.temp,Nov.temp,October.temp
                       ,Sept.temp)
#-----------------------------------------
# One-way repeated measures ANOVA
#-----------------------------------------
mydata<-read.csv(file = file.choose(),header = TRUE)
str(mydata)
#-----------------------
# Descriptive statistics
```

```
#-----------------------------
boxplot(dv~treatment,main="Boxplot of treatment",ylab="Count",xlab=
    "Treatment",mydata)
qqnorm(mydata$dv)
qqline(mydata$dv)
# Testing normality and homogeneity
bartlett.test(dv~treatment,mydata)

# Square root transformation
#-------------------------
# Fit the ANOVA model
#-------------------------
mydata$dv2<-sqrt(mydata$dv)
mydata$treatment<-as.factor(mydata$treatment)
mydata$subject<-as.factor(mydata$subject)
model.rm<-aov(dv2 ~ treatment,data = mydata)
summary(model.rm)
# Construct normal Q-Q plot
qqnorm(residuals(model.rm))
qqline(residuals(model.rm))
boxplot(dv2~treatment,main="Boxplot of treatment",ylab="Count",xlab
    ="Treatment",mydata)
# Plot residuals
plot(fitted(model.rm),resid(model.rm),xlab ="Fitted",ylab = "
    Residuals",main="Fitted residuals")
abline(0,0)
shapiro.test(model.rm$residuals)
# Fit the one-way repeated measures ANOVA model after data
    transformation
model.rm2<-aov(dv2~treatment+Error(subject/treatment),data = mydata
    )
summary(model.rm2)

#-----------------------------------------
# One-way repeated measures ANOVA model
#-----------------------------------------
## Multivariate approach
# Reorganize the data into a format in which each row represents a
    single subject and columns represent levels of the treatment
    factor.
```

```
response<-with(mydata,cbind(dv2[treatment==1],dv2[treatment==2],dv2
    [treatment==3],dv2[treatment==4]))
# Multivariate model using the lm() function
rm.model<-lm(response~1) # No between subjects here #Multivariate
    model
rfactor<-factor(c("r1","r2","r3","r4")) # design of the study

library(car)
rm.modelAOV<-Anova(rm.model,idata = data.frame(rfactor),idesign = ~
    rfactor,type = 3)
summary(rm.modelAOV,multivariate = F)
# rm.model is our multivariate model defined above.
# idata=data.frame(rfactor) passes information about the within-
    subjects variable.
# idesign= rfactor, passes information about the within-subjects
    design.
# The variable that rfactor describes is the repeated-measures
    variable.
# type="III", instructs Anova() to calculate the Type-"III" sums of
     squares when forming the ANOVA table.

#-----------------------------------------------------------------
# Two way repeated measures ANOVA: one observation per cell
#-----------------------------------------------------------------
driscoll<-read.csv(file = file.choose(),header = T)
str(driscoll)
driscoll$YEAR<-as.factor(driscoll$YEAR)
# Assessing normality and homogeneity of variance assumptions
boxplot(CALLS~YEAR,driscoll,main="Boxplot of the fuel reduction",
        xlab="Years",ylab="Number of calling male frogs")
# Testing homogeneity of variances
bartlett.test(CALLS~YEAR,driscoll)
# Testing normality
qqnorm(driscoll$CALLS)
qqline(driscoll$CALLS)
# Fit the repeated measures
driscoll.aov<-aov(CALLS~YEAR+Error(BLOCK),driscoll)
summary(driscoll.aov)

# Convert the data to wide format
```

```
dris.rm <- reshape(driscoll, timevar = "YEAR",v.names = "CALLS",
   idvar = "BLOCK",direction = "wide")
# Fit the simple MANOVA
dris.lm <- lm(cbind(CALLS.1, CALLS.2, CALLS.3) ~ 1, dris.rm)
# Create a data frame that defines the intra-block design
idata <- data.frame(YEAR = as.factor(c(1, 2, 3)))
# Install package
library(car)
# Use the Anova (car) function to estimate the MANOVA test
   statistics
driscoll.aov2<- Anova(dris.lm,idata = idata,idesign = ~YEAR)
summary(driscoll.aov2)

# Assessing the residuals of the model fit
qqnorm(driscoll.aov$Within$residuals)
qqline(driscoll.aov$Within$residuals)
shapiro.test(driscoll.aov$Within$residuals)
boxplot(driscoll.aov$Within$residuals,main="Boxplot of the
   residuals"
        ,xlab="",ylab="Residuals")
bartlett.test(CALLS~BLCK, driscoll)

#
   ------------------------------------------------------------------

# Two way reapeted measures Example including interaction from
   Logan
#
   ------------------------------------------------------------------

# raw data
mullens<-read.csv(file = file.choose(),header = T)
str(mullens)

boxplot(FREQBUC~BRTH.TYP, mullens,main="Boxplot of the breathing
   type"
        ,xlab="Breathing type",ylab="Frequency of buccal")
boxplot(FREQBUC~O2LEVEL, mullens,main="Boxplot of the oxygen level"
        ,xlab="Oxygen level",ylab="Frequency")
qqnorm(mullens$FREQBUC)
```

```
qqline(mullens$FREQBUC)

# Transformed response
mullens$O2LEVEL<-as.factor(mullens$O2LEVEL)
mullens$SFREQBUC<-sqrt(mullens$FREQBUC)
boxplot(SFREQBUC~BRTH.TYP, mullens, main="Boxplot of the breathing
    type"
        ,xlab="Breathing type",ylab="Frequency of the square root
            transformed buccal")
# Testing homogeneity of variances
bartlett.test(SFREQBUC~BRTH.TYP, mullens)
bartlett.test(SFREQBUC~O2LEVEL, mullens)
# Testing normality
qqnorm(mullens$SFREQBUC)
qqline(mullens$SFREQBUC)
boxplot(SFREQBUC~O2LEVEL, mullens, main="Boxplot of the oxygen level"
        ,xlab="Oxygen level",ylab="Frequency of the square root
            transformed")

# Install a package
# Summary statistics
library(doBy)
summaryBy(SFREQBUC ~BRTH.TYP*O2LEVEL, data = mullens, FUN =
    function(x) {
  c(n = sum( !is.na(x) ),m = mean(x), s = sd(x),lower = mean(x) -
      qt(0.95/2 + .5,
sum(!is.na(x))-1)*sd(x)/sqrt(sum(!is.na(x))),upper = mean(x) + qt
    (0.95/2 + .5, sum(!is.na(x))-1)*sd(x)/sqrt(sum(!is.na(x)))) } )
# Fit a linear model
mullens.aov<-aov(SFREQBUC~BRTH.TYP*O2LEVEL+Error(TOAD), mullens)
summary(mullens.aov)
## Ploting residuals of the fitted model
qqnorm(mullens.aov$Within$residuals)
qqline(mullens.aov$Within$residuals)
shapiro.test(mullens.aov$Within$residuals)

#-----------------------------------------------------------
# Fitting the mixed-effects model example: Hothorn
#-----------------------------------------------------------
BtheB<-read.csv(file = file.choose(),header = T)
```

```
str(BtheB)
# Install a package
library(HSAUR)
# Add subject variable to the data
BtheB$subject<-factor(rownames(BtheB))
nobs<-nrow(BtheB)
BtheB_long<-reshape(BtheB, idvar = "subject",
                    varying = c("bdi.2m","bdi.4m","bdi.6m","bdi.8m"
                      ),direction = "long")
BtheB_long$time<-rep(c(2,4,6,8),rep(nobs,4))

# Fit the random intercept model using lmer
library(lmerTest)
library(nlme)
BtheB.lmer<-lmer(bdi~bdi.pre+time+treatment+drug+length+(1|subject)
    ,
                    BtheB_long,na.action = na.omit,REML = F)
summary(BtheB.lmer)
# Fit a random intercept and slope model
BtheB.lmer2<-lmer(bdi~bdi.pre+treatment+drug+length+(time|subject),
                    BtheB_long,na.action = na.omit,REML = F)
summary(BtheB.lmer2)
# Model comparison
anova(BtheB.lmer,BtheB.lmer2)
# Model diagnostics plots
par(mfrow=c(2,2))
plot(lm(BtheB.lmer))
# Testing the residuals of the fitted model
shapiro.test(residuals(BtheB.lmer))
```

91G__thesis_TournamentData_Tournament_data.R

D.1

```
# Random intercept model
Linear mixed model fit by maximum likelihood t-tests
   use Satterthwaite approximations to degrees of
   freedom [lmerMod]
Formula: bdi ~ bdi.pre+time+treatment+drug+length+(1 |
   subject)
Data: BtheB_long
```

```
 AIC      BIC    logLik   deviance  df.resid
1886.7   1919.4   -934.4   1868.7       271

Scaled residuals:
   Min      1Q      Median    3Q       Max
-2.7664  -0.4748  -0.0977   0.4002   3.7397

Random effects:
Groups      Name        Variance Std.Dev.
subject  (Intercept) 47.25    6.874
Residual             25.11    5.011
Number of obs: 280, groups:  subject, 97

Fixed effects:
             Estimate Std. Error  df  t value Pr(>|t|)
(Intercept)   3.7675   2.2033  109.50   1.710   0.0901.
bdi.pre       0.6141   0.0788  102.87   7.792   5.50e-12
time         -0.7128   0.1451  203.58  -4.882   2.12e-06
treatmentTAU  2.5641   1.6541   97.22   1.551   0.1242
drugYes      -2.8491   1.7048   98.51  -1.671   0.0978
length<6     12.2583   8.8520  144.21   1.385   0.1683
length>6      0.5724   1.6329  100.29   0.351   0.7267

Correlation of Fixed Effects:
            (Intr) bdi.pr time   trtTAU drugYs lngt<6
bdi.pre     -0.595
time        -0.251  0.017
treatmntTAU -0.349 -0.136 -0.020
drugYes     -0.315 -0.227 -0.026  0.320
length<6     0.064 -0.218  0.028  0.081 -0.020
length>6    -0.228 -0.264 -0.039  0.009  0.153  0.137

# Random intercept and slope model
Formula: bdi ~ bdi.pre+treatment+drug+length+(time |
   subject)
Data: BtheB_long
 AIC      BIC      logLik    deviance    df.resid
1904.9   1941.2   -942.4     1884.9       270

Scaled residuals:
```

```
   Min      1Q      Median     3Q      Max
-2.0264  -0.4900  -0.0603   0.4100  3.6353

Random effects:
Groups     Name        Variance  Std.Dev.  Corr
subject  (Intercept) 46.057    6.7865
         time          0.691    0.8313    -0.13
Residual              23.334    4.8305
Number of obs: 280, groups:  subject, 97

Fixed effects:
             Estimate Std. Error  df  t value Pr(>|t|)
(Intercept)    1.5439  2.1787   96.58  0.709   0.4803
bdi.pre        0.6292  0.0797   99.94  7.890 3.92e-12
treatmentTAU   2.7766  1.6865   97.44  1.646   0.1029
drugYes       -3.2216  1.7364   98.44 -1.855   0.0665
length<6      12.7351  8.6816  110.65  1.467   0.1452
length>6       0.10520 1.65910  99.31 0.063   0.9496

 Correlation of Fixed Effects:
             Intr) bdi.pr trtTAU drugYs lngt<6
bdi.pre      -0.612
treatmntTAU  -0.368 -0.134
drugYes      -0.340 -0.217  0.321
length<6      0.075 -0.226  0.085 -0.023
length>6     -0.253 -0.255  0.013  0.155  0.140
```

D.2

```
> anova(BtheB.lmer,BtheB.lmer2)
Data: BtheB_long Models:
object: bdi ~ bdi.pre + time + treatment + drug +
   length + (1 | subject)
..1: bdi ~ bdi.pre + treatment + drug + length + (time
   | subject)
        Df   AIC    BIC   logLik   deviance Chisq Chi Df Pr
           (>Chisq)
object 9 1886.7 1919.4 -934.36   1868.7
```

```
..1    10 1904.9 1941.2 -942.44   1884.9      0      1
   1
```

Appendix E

Appendix E: R Code for Mixed-Effects Models

```
rm(list=ls())
#06/11/2017
## 12 fishing days
## Repeated measures data set
rm<-read.csv(file = file.choose(),header = TRUE)
minT<-rm$Minimum
maxT<-rm$Maximum
press<-rm$Pressure
f1<-rm$F1
f2<-rm$F2
f3<-rm$F3
bw<-rm$Bweight
wb<-rm$WaterBody ## Water body
ps<-rm$Person ## Anglers
Temp<-cbind(rm$T1,rm$T2,rm$T3,rm$T4,rm$T5) # Prior fishing
    temperatures
Pressures<-cbind(rm$P1,rm$P2,rm$P3,rm$P4,rm$P5) # Prior fishing
    pressures

#-------------
# GLM Fit
#-------------
# Gaussian family
# what is the change in the average pressure of the previous 5 days
changeP <- apply(press-Pressures,1,mean)
changeT <- apply((minT+maxT)/2-Temp,1,mean)
```

```
# or as a factor: up /down or high low?
mod<-glm(bw~minT+maxT+press+wb+changeT+changeP,family = "gaussian",
    data = rm)
summary(mod)
par(mfrow=c(2,2))
plot(mod)
#save the plots
par(mfrow=c(1,1))
# Testing the residuals
shapiro.test(mod$residuals)
# Testing homogeneity of variances
bartlett.test(bw~wb,rm)

# Refit the model
# Drop insignificant variables
mod.glm<-glm(bw~press,family = "gaussian",data = rm)
summary(mod.glm)
par(mfrow=c(2,2))
plot(mod.glm)
# Testing normality
shapiro.test(mod.glm$residuals)
# testing homogeneity of variances
bartlett.test(bw~press,rm)
#-----------------------------------
# Fit the linear mixed-effects model
#-----------------------------------
library(lme4) # intentionally not reporting p-values
library(lmerTest) ## package that will give us p-values
library(nlme)
mod.lme<-lme(bw~maxT+minT+wb+press+changeP+changeT,random = ~1|ps,
    rm,method = "REML")
summary(mod.lme)
par(mfrow=c(2,2))
# The plot method for the lme class is the primary tool for
    obtaining diagnostic plots for Assumption
plot(lm(mod.lme))
# Testing the normality of the residuals
shapiro.test(mod.lme$residuals)
# Testing homogeneity of variances
bartlett.test(bw~wb,rm)
```

```
# ML method
mod.lme2<-update(mod.lme,method = "ML")
summary(mod.lme2)

# GLMM Fit
## Fit the glmm with lmer() function and estimate parameters with
   REML method.
# lmer() residuals use standardized residuals rather than raw
   residuals
mod.lmer<-lmer(bw~minT+maxT+press+wb+changeP+changeT+(1|ps),rm,REML
   =F)
summary(mod.lmer)
## Model diagnostics plots
par(mfrow=c(2,2))
plot(lm(mod.lmer))

# Fit a random intercept and slope model
mod.lmer2<-lmer(bw~minT+maxT+press+changeP+changeT+(wb|ps),rm,REML=
   F)
summary(mod.lmer2)
# Model comparison
anova(mod.lmer,mod.lmer2)
```

92G__thesis_Repeated_Repeated_Measures.R

E.1

```
mod.lme<-lme(bw~maxT+minT+wb+press+changeP+changeT,
   random = ~1|ps,rm,method = "REML")
Linear mixed-effects model fit by REML
 Data: rm
  AIC       BIC       logLik
591.8075  625.7711  -284.9038
Random effects:
 Formula: ~1 | ps
        (Intercept)  Residual
StdDev:   0.7549     1.1022
Fixed effects: bw ~ maxT + minT + wb + press + changeP
   + changeT
               Value  Std.Error  DF  t-value  p-value
(Intercept) 152.4782  69.9092   134  2.1811   0.0309
```

```
maxT          -0.0484    0.0771    134   -0.6285   0.5307
minT          -0.1085    0.0749    134   -1.4479   0.1500
Settlers      -0.0185    0.2907    134   -0.0637   0.9493
Whites dam    -0.5195    0.4044    134   -1.2847   0.2011
Yarrow        -0.3811    0.2882    134   -1.3223   0.1883
press         -0.1561    0.0722    134   -2.1744   0.0314
changeP        0.0326    0.0658    134    0.4957   0.6209
changeT        0.1383    0.1550    134    0.8911   0.3740

Correlation:
              (Intr) maxT   minT   wbSttl wbWhtd wbYrrw
                press changP
maxT          -0.688
minT          -0.704  0.322
wbSettlers     0.325 -0.300 -0.304
wbWhites dam -0.241  0.504  0.169  0.093
wbYarrow       0.395 -0.359 -0.001  0.524  0.203
press         -1.000  0.676  0.702 -0.324  0.231 -0.399
changeP        0.259 -0.111  0.191  0.324  0.034  0.398
   -0.268
changeT        0.813 -0.894 -0.529  0.466 -0.418  0.423
   -0.805  0.333

Standardized Within-Group Residuals:
     Min        Q1      Med      Q3       Max
-2.0477   -0.7242  0.0445  0.5719  2.6887

Number of Observations: 171
Number of Groups: 29
```

E.2

```
mod.lme2<-update(mod.lme,method = "ML")
Linear mixed-effects model fit by maximum likelihood
 Data: rm
  AIC           BIC         logLik
565.9796     600.5379    -271.9898
Random effects:
```

```
 Formula: ~1 | ps
        (Intercept)    Residual
StdDev: 0.7387          1.0721
Fixed effects: bw ~ maxT + minT + wb + press + changeP
   + changeT
             Value     Std.Error  DF    t-value p-value
(Intercept) 152.3083  69.8708    134    2.1799  0.0310
maxT         -0.0483   0.0770    134   -0.6269  0.5318
minT         -0.1083   0.0749    134   -1.4459  0.1505
Settlers     -0.0192   0.2906    134   -0.0659  0.9475
Whites dam   -0.5194   0.4042    134   -1.2841  0.2010
Yarrow       -0.3812   0.2881    134   -1.3234  0.1880
press        -0.1568   0.0721    134   -2.1732  0.0315
changeP       0.0326   0.0658    134    0.4962  0.6206
changeT       0.1379   0.1549    134    0.8903  0.3749

Standardized Within-Group Residuals:
  Min         Q1      Med       Q3      Max
-2.1046   -0.7437  0.0454  0.5875  2.7620

Number of Observations: 171
Number of Groups: 29
```

E.3

```
# Random intercept model
Linear mixed model fit by maximum likelihood t-tests
   use Satterthwaite approximations to degrees of
   freedom [lmerMod]
Formula: bw ~ minT + maxT + press + wb + changeP +
   changeT + (1 | ps)
Data: rm
 AIC       BIC    logLik    deviance    df.resid
566.0     600.5    -272.0     544.0        160

Scaled residuals:
  Min       1Q       Median    3Q      Max
-2.1046  -0.7438  0.0454  0.5875  2.7621
```

```
Random effects:
Groups   Name        Variance Std.Dev.
 ps       (Intercept) 0.5457   0.7387
 Residual             1.1493   1.0721
Number of obs: 171, groups:  ps, 29

Fixed effects:
            Estimate  Std. Error  df     t value Pr(>|t|)
(Intercept) 152.3083   68.0073   152.45   2.240   0.0266
minT         -0.1083    0.0729   152.43  -1.486   0.1395
maxT         -0.0483    0.0741   153.48  -0.644   0.5205
press        -0.1568    0.0702   152.44  -2.233   0.0270
Settlers     -0.0192    0.2828   152.23  -0.068   0.9461
Whites dam   -0.5194    0.3934   153.94  -1.320   0.1887
Yarrow       -0.3812    0.2804   154.29  -1.360   0.1759
changeP       0.0326    0.0640   148.40   0.510   0.6110
changeT       0.1379    0.1508   153.18   0.915   0.3618

Correlation of Fixed Effects:
            (Intr) minT    maxT  press  Sttl   Whtd
               Yrrw changP
minT        -0.704
maxT        -0.688  0.322
press       -1.000  0.702  0.676
Settlers     0.325 -0.305 -0.300 -0.324
Whitesdam   -0.241  0.169  0.504  0.231  0.093
Yarrow       0.395 -0.001 -0.359 -0.399  0.524  0.203
changeP      0.259  0.191 -0.111 -0.268  0.324  0.034
   0.398
changeT      0.813 -0.529 -0.894 -0.805  0.466 -0.418
   0.423 0.333

# Random intercept and slope model
Formula: bw ~ minT + maxT + press + changeP + changeT +
    (wb | ps)
Data: rm
 AIC      BIC   logLik  deviance  df.resid
573.5    626.9  -269.7    539.5       154
Scaled residuals:
```

```
  Min       1Q     Median     3Q      Max
-2.32497 -0.65587  0.02939  0.54744  2.40173
Random effects:
Groups   Name          Variance Std.Dev.  Corr
ps       (Intercept)   1.1103   1.0537
Settlers               0.5852   0.7650  -0.55
Whites dam             0.8035   0.8964  -0.99  0.69
Yarrow                 0.3064   0.5536  -0.72  0.98 0.83
Residual               1.0283   1.0141
Number of obs: 171, groups:  ps, 29

Fixed effects:
            Estimate    Std. Error df  t value Pr(>|t|)
(Intercept) 145.25496 56.62817 143.67   2.565  0.0113
minT         -0.05969  0.06536 145.99  -0.913  0.3626
maxT         -0.04141  0.05225 128.17  -0.792  0.4296
press        -0.15032  0.05865 143.82  -2.563  0.0114
changeP       0.05316  0.05751 147.56   0.924  0.3568
changeT       0.11180  0.10541 121.16   1.061  0.2910

Correlation of Fixed Effects:
        (Intr) minT   maxT   press  changP
minT    -0.781
maxT    -0.562  0.291
press   -1.000  0.779  0.549
changeP  0.132  0.248  0.099 -0.141
changeT  0.763 -0.560 -0.816 -0.756  0.208
```

E.4

```
> anova(mod.lmer,mod.lmer2)
Data: rm Models:
object: bw ~ minT + maxT + press + wb + changeP +
   changeT + (1 | ps)
..1: bw ~ minT + maxT + press + changeP + changeT + (wb
    | ps)
       Df    AIC   BIC   logLik  deviance Chisq Chi Df
          Pr(>Chisq)
```

```
object 11 565.98 600.54 -271.99  543.98
..1    17 573.47 626.88 -269.74  539.47 4.5066     6
       0.6085
```

Appendix F

Appendix F: BtheB Data Set

drug,length,treatment,bdi.pre,bdi.2m,bdi.4m,bdi.6m,bdi.8m
No,>6,TAU,29,2,2,NA,NA
Yes,>6,BtheB,32,16,24,17,20
Yes,<6,TAU,25,20,NA,NA,NA
No,>6,BtheB,21,17,16,10,9
Yes,>6,BtheB,26,23,NA,NA,NA
Yes,<6,BtheB,7,0,0,0,0
Yes,<6,TAU,17,7,7,3,7
No,>6,TAU,20,20,21,19,13
Yes,<6,BtheB,18,13,14,20,11
Yes,>6,BtheB,20,5,5,8,12
No,>6,TAU,30,32,24,12,2
Yes,<6,BtheB,49,35,NA,NA,NA
No,>6,TAU,26,27,23,NA,NA
Yes,>6,TAU,30,26,36,27,22
Yes,>6,BtheB,23,13,13,12,23
No,<6,TAU,16,13,3,2,0
No,>6,BtheB,30,30,29,NA,NA
No,<6,BtheB,13,8,8,7,6
No,>6,TAU,37,30,33,31,22
Yes,<6,BtheB,35,12,10,8,10
No,>6,BtheB,21,6,NA,NA,NA
No,<6,TAU,26,17,17,20,12
No,>6,TAU,29,22,10,NA,NA
No,>6,TAU,20,21,NA,NA,NA
No,>6,TAU,33,23,NA,NA,NA
No,>6,BtheB,19,12,13,NA,NA
Yes,<6,TAU,12,15,NA,NA,NA

Yes,>6,TAU,47,36,49,34,NA
Yes,>6,BtheB,36,6,0,0,2
No,<6,BtheB,10,8,6,3,3
No,<6,TAU,27,7,15,16,0
No,<6,BtheB,18,10,10,6,8
Yes,<6,BtheB,11,8,3,2,15
Yes,<6,BtheB,6,7,NA,NA,NA
Yes,>6,BtheB,44,24,20,29,14
No,<6,TAU,38,38,NA,NA,NA
No,<6,TAU,21,14,20,1,8
Yes,>6,TAU,34,17,8,9,13
Yes,<6,BtheB,9,7,1,NA,NA
Yes,>6,TAU,38,27,19,20,30
Yes,<6 ,BtheB,46,40,NA,NA,NA
No,<6,TAU,20,19,18,19,18
Yes,>6,TAU,17,29,2,0,0
No,>6,BtheB,18,20,NA,NA,NA
Yes,>6,BtheB,42,1,8,10,6
No,<6,BtheB,30,30,NA,NA,NA
Yes,<6,BtheB,33,27,16,30,15
No,<6,BtheB,12,1,0,0,NA
Yes,<6,BtheB,2,5,NA,NA,NA
No,>6,TAU,36,42,49,47,40
No,<6,TAU,35,30,NA,NA,NA
No,<6,BtheB,23,20,NA,NA,NA
No,>6,TAU,31,48,38,38,37
Yes,<6,BtheB,8,5,7,NA,NA
Yes,<6,TAU,23,21,26,NA,NA
Yes,<6,BtheB,7,7,5,4,0
No,<6,TAU,14,13,14,NA,NA
No,<6,TAU,40,36,33,NA,NA
Yes,<6,BtheB,23,30,NA,NA,NA
No,>6,BtheB,14,3,NA,NA,NA
No,>6,TAU,22,20,16,24,16
No,>6,TAU,23,23,15,25,17
No,<6,TAU,15,7,13,13,NA
No,>6,TAU,8,12,11,26,NA
No,>6,BtheB,12,18,NA,NA,NA
No,>6,TAU,7,6,2,1,NA
Yes,<6,TAU,17,9,3,1,0

Yes,<6,BtheB,33,18,16,NA,NA
No,<6,TAU,27,20,NA,NA,NA
No,<6,BtheB,27,30,NA,NA,NA
No,<6,BtheB,9,6,10,1,0
No,>6,BtheB,40,30,12,NA,NA
No,>6,TAU,11,8,7,NA,NA
No,<6,TAU,9,8,NA,NA,NA
No,>6,TAU,14,22,21,24,19
Yes,>6,BtheB,28,9,20,18,13
No,>6,BtheB,15,9,13,14,10
Yes,>6,BtheB,22,10,5,5,12
No,<6,TAU,23,9,NA,NA,NA
No,>6,TAU,21,22,24,23,22
No,>6,TAU,27,31,28,22,14
Yes,>6,BtheB,14,15,NA,NA,NA
No,>6,TAU,10,13,12,8,20
Yes,<6,TAU,21,9,6,7,1
Yes,>6,BtheB,46,36,53,NA,NA
No,>6,BtheB,36,14,7,15,15
Yes,>6,BtheB,23,17,NA,NA,NA
Yes,>6,TAU,35,0,6,0,1
Yes,<6,BtheB,33,13,13,10,8
No,<6,BtheB,19,4,27,1,2
No,<6,TAU,16,NA,NA,NA,NA
Yes,<6,BtheB,30,26,28,NA,NA
Yes,<6,BtheB,17,8,7,12,NA
No,>6,BtheB,19,4,3,3,3
No,>6,BtheB,16,11,4,2,3
Yes,>6,BtheB,16,16,10,10,8
Yes,<6,TAU,28,NA,NA,NA,NA
No,>6,BtheB,11,22,9,11,11
No,<6,TAU,13,5,5,0,6
Yes,<6,TAU,43,NA,NA,NA,NA