Zero-Inflated Regression Models with Application to Water Quality Data from Umgeni Water

By

Zibusiso Sandile Hlongwane

A thesis submitted in fulfillment of the requirement for the Masters Degree in Statistics

School of Mathematics, Statistics and Computer Science

University of KwaZulu-Natal

Pietermaritzburg

South Africa

Declaration

I, Zibusiso Sandile Hlongwane, declare that this thesis titled, 'Zero-Inflated Regression Models with Application to Water Quality Data from Umgeni Water' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where I have consulted the published work for others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Mr. Z. Hlongwane Signed____________________ Date____________________

Prof. H. Mwambi Signed____________________ Date____________________

Dr. S. Melesse Signed____________________ Date____________________

Acknowledgements

I would first like to thank my supervisor Prof. Henry Mwambi, for all the encouragement and guidance. This work could not have been accomplished without his continued support. My appreciation to my co-supervisor, Dr. S. Melesse, for expert advice, invaluable input and most especially his time.

I am also grateful for the facilities made available to me by the School of Mathematics, Statistics and Computer Sciences, UKZN. I would like to thank the staff members of the School of Mathematics, Statistics and Computer Sciences, Pietermaritzburg for their conducive environment and encouragement. I also gratefully acknowledge financial support that I received from Umgeni Water. Many thanks extend to Debbie Trollip, Kim Hodgson, Steve Terry, Nirosha Reddy and Lakesh Maharaj for supporting us with real datasets from Umgeni Water.

Special thanks to my late father (Ephraim Hlongwane) for his unconditional love, though he did not live long enough to see this project reach completion. His wisdom and life lessons have brought me this far, I thank him for being my pillar of strength. To my mother (Thembesile Hlongwane), my wonderful sisters (Bongiwe and Thinamangwane) and my brothers (Nhlakanipho, Vumelani and Siyabonga), I thank you for believing in me. Last but not least I would like to express my heartfelt gratitude to the Almighty God for sustaining and providing me with the strength I needed to conquer all challenges thus far.

Contents

List of Figures

List of Tables

Abstract

A common feature of microbiological datasets is their tendency to contain many zero values. Statistical inferences based on such data are likely to be inefficient or wrong unless careful thought is given to how these zeros arose and how best to handle them. Analyzing these data using the classical linear model is mostly inappropriate, even after transformation of outcome variables. Zero-adjusted mixture count models such as zero-inflated and hurdle count models are applied to count data when overdispersion and excess zeros exist. This study considers data collected from four large water treatment plant sites at Umgeni Water in South Africa the province of KwaZulu-Natal. A unique characteristic of the daily incidence data collected from these sites is the occurrence of many zeros. The most common microbiological organisms that are detected during routine water quality checks are *E. coli*, total coliforms, and heterotrophic plate counts (HPC at 37°C). Count data models including traditional (Poisson and negative binomial) models, zero modified models (zero-inflated Poisson and zero-inflated negative binomial) and hurdle models (Poisson logit hurdle and negative binomial logit hurdle) were fitted and compared. Using Akaike information criteria (AIC), the negative binomial logit hurdle (NBLH) and zero-inflated negative binomial (ZINB) models showed the best performance in both datasets. The results show that total chlorine and free chlorine reduces the occurrence of *E. coli*, total coliform counts, and heterotrophic plate counts as expected and are positively correlated with excess zeros. The model further shows that high temperature significantly increases bacterial growth and at low temperatures, the organism would not achieve significant growth. A further important finding is that a declining trend of log

mean positive counts over time was detected as a result that can be interpreted to mean a sustained improvement of water quality.

Chapter 1

Introduction

1.1 Background

It is common to encounter the problem of having a large proportion of zero values in many physical processes, including those in microbiological and environmental studies. Data with a large proportion of zero counts, also called zero-inflated data, appear quite frequently in various fields of research, including health research, agricultural research, ecology and manufacturing (Ridout et al., 1998). In microbiology, zero-inflated data are often found when examining counts of microbiological organisms from water samples, where water samples produce no microbiological organisms.

In this context, the distribution of zero and positive counts in the data set which is being investigated is very important. The Poisson distribution is the basic model that is commonly applied to study count data. However, the equivalence of mean and variance assumption of the Poisson process is invalid for many processes such as in microbiological data because in most cases there are no occurrences of micro-organisms (total coliform, *E. coli* and HPC at 37°C counts) in drinking water which causes the microbiological data set to generally have more zeros than expected. For this reason, overdispersion which is the situation that the variance is greater than the mean seems to be the main feature for microbiological datasets. The negative binomial model is an alternative model for handling over-dispersed count data. An alternative form of expressing the negative binomial model for over-dispersion is to express the variance in terms of the mean. Overdispersion can occur in two ways, namely apparent and real overdispersion. Missing values and outliers requiring interaction terms or misspecified link function can cause apparent overdispersion while the violation of distributional assumptions can cause real overdispersion. It is well known that analyzing these data using classical linear models is mostly inappropriate, even after the transformation of the outcome variables. Zero-inflated models have been developed to handle inflated zero values for the dependent variable which otherwise would lead to the violation of distributional assumptions .

In general, hurdle and zero-inflated models are used for modelling count data with a preponderance of zeros. The hurdle and other zero-inflated models are two component models in which one component models the probability of zero counts and the other component uses a truncated Poisson/ negative binomial distribution that modifies an ordinary distribution by conditioning on a positive outcome (Dalrymple et al., 2003). The zero-inflated model has a distribution that is a mixture of a binary distribution that is degenerate at zero and an ordinary count distribution such as Poisson or negative binomial. The hurdle model considers the zeros to be completely separate from the non-zero values. The zero-inflated model is similar to the hurdle model; however, it permits some of the zeros to be analyzed along with the non-zeros (Hua et al., 2014). The choice of the zero-inflated model in this thesis is guided by the researcher's beliefs about the source of the zeros. There are two distinct processes driving the zeros, one is sampling zeros which occur by chance and can be assumed to be as a result of a dichotomous process, and the other one is structural zeros (true zeros) which are inevitable and are part of the counting process. Beyond this substantive consideration, the choice should be based on the model providing the closest fit between the observed and predicted values.

1.2 Literature Review

Zero-inflated Poisson (ZIP) models were developed by Lambert (1992) to handle zero-inflated count data. Zero-inflated models combine two sources of zero outcomes which are called true zeros and excess zeros. Greene (2008) investigated zero-inflated models as modifications of the Poisson and the negative binomial models. He also presented the test procedure to separate the zero inflation and over-dispersion. Fahrmeir and Osuna Echavarría (2006) developed structured additive regression models for over-dispersed and zero-inflated data. Boucher et al. (2009) presented different risk classification models for panel count data based on the zero-inflated Poisson distribution. Mullahy (1986) first discussed in the econometric literature hurdle count data models, which were also called two-part models by Heilbron (1994). Gurmu (1998) introduced a generalized hurdle model for the handling of overdispersion and also under-dispersion which is the situation where the variance is less than the mean. Ridout et al. (1998) reviewed some zero-inflated models and hurdle models and gave an example of biological count data. Saffari et al. (2012) suggested using a hurdle negative binomial regression model to overcome the problem of over-dispersion. They introduced a censored hurdle negative binomial model on count data with many zeros. In this work, they also described several extensions of the models and presented an application to water quality data when comparing the models. They reviewed the development of zero-inflated models and paid attention to the fact that there are very few applications on microbiological frequency data in the literature. Bermúdez and Karlis (2011) extended this work based on Bayesian inference by using multivariate Poisson regression models with their zero inflated versions. Mouatassim and Ezzahid (2012) compared the Poisson model to the zero-inflated model and applied his approach to water quality data set. Mouatassim et al. (2012) analyzed operational risk to the zero-inflated data and assessed the impact of the zero-inflated Poisson distribution on the operational capital charge. They concluded that the zero-inflated and hurdle models had the most consistent performance at any combination of dispersion and zero-inflation in the simulation study.

1.2.1 Zero Inflation Due to True Zeros

True zeros means that a zero data value indicates the absence of the object being measured. When true zeros lead to an excess of zeros, zero-inflated models such as the two-part (also known as conditional or hurdle models) or mixture models are recommended (Lambert, 1992; Barry and Welsh, 2002). The negative binomial has also been advocated for modeling data sets with many zeros because of its ability to account for overdispersion (Warton, 2005). However, Barry and Welsh (2002) and Warton (2005) demonstrated that the excess number of zeros often exceeds those expected under a negative binomial distribution. For count data, there are two parts in the modelling approach, whereby the first part is a binary outcome model (i.e. Bernoulli), and the second part is a truncated count model (e.g. Poisson or negative binomial) (Cameron and Trivedi, 1998). This approach assumes that zeros arise from a single process with a set of covariates. One of its computational benefits are that it is possible to fit these models in two parts, for example, fitting zeros using a logistic regression separately from fitting non-zeros using a truncated Poisson (Barry and Welsh, 2002; Dobbie and Welsh, 2001).

Mixture models are combinations of probability distributions chosen for their ability to represent two or more real ecological processes. The ZIP mixture model used to model count data is a mixture of a point mass at zero and a Poisson distribution. With this approach, zeros may arise from one of two processes and their related covariates, a zero process from which only zero values are observed and a Poisson process in which non-zero and a proportion of the zero values, appropriate to the Poisson distribution are observed (Lambert, 1992). The interpretation of mixture model parameters is less straightforward than the two-part model. For example, to get the true estimate of relative mean abundance from the ZIP one must multiply the estimated relative mean number of individuals at a site by the probability that the relative mean number of individuals at a site is generated through a Poisson distribution. Where there is zero inflation and overdispersion caused by large counts of individuals (e.g. total coliform counts and *E. coli*), the use of a zero-inflated negative binomial (ZINB) mixture model has been shown to be appropriate (Barry and Welsh, 2002).

1.2.2 Zero Inflation Due to False Zeros

If false zeros are present in the data, a zero-inflated mixture modeling approach is required because we are interested in modelling two processes, that is a process leading to true zeros and a process leading to false zeros (Bolker, 2008). Failing to take account of false zero observations in the analysis may have substantial impacts on the ability to accurately infer relationships between site occupancy and habitat attributes or management actions (Martin et al., 2005). The zero-inflated binomial model and its extensions provide an appropriate framework for analyzing data that are collected for these purposes and which are likely to contain false-zero observation error.

1.3 Introduction of Microbiota

Water quality management is a critical component of overall integrated water resources. Human health depends on safe drinking water more than any other thing, and most of the problems in developing countries are mainly due to the lack of safe drinking water (Sharp et al., 2006). Drinking water should meet the accepted quality standards which imply that water should be wholesome and clean. It should be free from any micro-organisms, parasites and from any substances which, in number or concentration, constitute a potential risk to human health. Besides, the water of poor quality can also be harmful from an economic perspective as resources must be directed towards improving and purifying the water supply system. For these reasons, there is growing pressure to improve water treatment and water quality management at catchment scale to ensure safe drinking water at reasonable costs (Won et al., 2013). The microbial quality in the South African National Standard (SANS 241) is addressed by setting maximal allowed limits of human gastroenteritis bacteria that are total coliforms, *Escherichia coli (E. coli)* and (HPC at 37°C counts.

Coliforms Bacteria

Total coliforms bacteria are a group of bacteria that are present in the environment and in the faeces of all warm-blooded animals and humans. Coliform bacteria will not likely cause illness. However, their presence in drinking water indicates that disease-causing organisms (pathogens) could be in the water system. Most pathogens that can contaminate water supplies come from the faeces of humans or animals. If coliform bacteria are found in a water sample, water system operators work to find the source of contamination and restore safe drinking water. There are three different groups of coliform bacteria and each has a different level of risk. Total coliform, faecal coliform, and *E. coli* are all indicators of microbial water quality. The total coliform group is a large collection of different kinds of bacteria. Faecal coliforms are types of total coliform that mostly exist in faeces. *E. coli* is a sub-group of faecal coliform. Some of these bacteria can grow during decomposition of plant residues in the soil, and some of the plant material in water. Generally, the growth of these bacteria in the soil and water are best at a temperature below 40°C. The analysis of coliform bacteria often takes place at 37°C.

Escherichia coli

Escherichia coli (E. coli) bacteria normally live in the intestines of humans and animals. It is gram-negative, facultatively anaerobic, rod-shaped bacterium that is commonly found in the lower intestine of warm-blooded organisms. Most *E. coli* are harmless and actually are an important part of a healthy human intestinal tract. However, some *E.*

coli are pathogenic, meaning that they can cause illness, either diarrhea or illness outside the intestinal tract. The types of *E. coli* that can cause diarrhea can be transmitted through contaminated water or food, or through contact with animals or persons. Still, other kinds of *E. coli* are used as markers for water contamination, which as said earlier are not themselves harmful, but indicate that the water is contaminated. It is the most appropriate group of coliforms to indicate faecal pollution from warm-blooded animals.

Heterotrophic Plate Counts (HPC at 37°C)

The HPC at 37°C is a procedure used to estimate the number of live heterotrophic bacteria that are present in a water sample. A sample of water is put on a plate that contains nutrients that the bacteria need to survive and grow. Heterotrophic plate counts detect a wide range of bacteria which are omnipresent in nature. Pollution of water can give rise to conditions conducive to bacterial growth, such as high nutrient concentrations and high turbidity and can result in a substantial increase of these naturally-occurring organisms. High heterotrophic plate counts in treated water indicate inadequate treatment of the water, post-treatment contamination or bacterial growth in the distribution system. Therefore, pathogenic micro-organisms, bacteria, viruses or parasites could possibly be present in the water and pose a health risk when the water is used for domestic consumption. The HPC at 37°C results are generally reported as CFU/ml or Colony Forming Units per milliliter. The maximum allowed value of HPC at 37°C is 1000 CFU/mL.

1.4 Objectives

The main objective of the study is to find the statistical methods or techniques to model the rare occurrences of microbiological organisms that exceed the acceptable standards limit at Umgeni Water. A second objective is to compare the performance of different methods for handling both excess zeros and positive counts of microbiological organisms. Sound model adequacy methods will then be used to determine the best method that can handle the data well. The data set will be analyzed to infer on factors and conditions that reduce or increase the occurrence of the microbiological organisms that are of interest to monitoring and evaluation of water quality.

1.5 Thesis Layout

The layout of this study is as follows: Chapter 2 is the exploratory analysis of the data. In Chapters 3 and 4, the methods used to achieve the objectives of the study are discussed. Chapter 5 focuses on the application and the results of the analysis of Umgeni Water quality dataset. Discussion, conclusion and recommendation for further investigations are given in Chapter 6.

Chapter 2

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach for data analysis that employs a variety of techniques (mostly graphical and tabular displays) to extract important variable, test underlying assumptions and help to develop parsimonious models. The EDA approach is not a set of techniques, but a preliminary process about how data analysis may be carried out. Any method of looking at data that does not include formal statistical modeling and inference falls under the term exploratory data analysis. It was promoted by Rosenthal (1995) to encourage statisticians to visually examine the data sets at hand and to formulate hypotheses that could be tested on new datasets. EDA is a critical first step in analyzing any data from an experiment or observational study (Rosenthal, 1995). Here are the main reasons we use EDA:

- Detection of mistakes,
- Checking of assumptions, and
- Determining relationships among the explanatory variables.

2.1 Data Description

In this research, we use water quality data that has been collected over a period of 24 years (1991 - 2015) at Umgeni Water. Umgeni water is one of Africa's most successful organisations involved in water management and is the largest supplier of bulk potable water in the province of KwaZulu-Natal, South Africa. We consider data from two sites at Umgeni Water, namely Midmar (TMM007) and DV Harris (TDV006). The most common microbiological organisms that are detected during routine water quality checks are *E. coli*, total coliform counts and heterotrophic plate counts (HPC at 37 °C). The associated water quality data that influence the response variables are free chlorine, total chlorine, pH, temperature, and turbidity. Table 2.1 displays all the variables, their units of measurements and the acceptable standard limits.

Table 2.1: Variables definition.

Variable	Unit	Standard limits
Total coliform counts	Count per 100 mL	≤ 10
E. coli	Count per 100 mL	$= 0$
HPC at 37°C	Count per ml	≤ 1000
Free chlorine	mg/l	≤ 1.5
Total chlorine	mg/l	≤ 5
Temperature	Degrees celsius	$[10, 25]$ °C
Turbidity	NTU	≤ 1
pH	pH units	$[5, 9.7]$

2.1.1 Descriptive Statistics

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. As shown in the descriptive statistic Table 2.2 below, the highlighted variables of total coliform, *E. coli* and HPC at 37°C counts imply the presence of overdispersion. Overdispersion occurs when the variance of count data exceeds the mean which can lead to the occurrence of extreme values/ outliers. The number of observation is the same for the whole study because the complete case analysis strategy was used. Complete case analysis is used when the analysis is confined to cases or units with complete variable information. However, it should be quick to point out that such a strategy can lead to biased results if the omitted cases are not comparable to those that are used for analysis to estimate the quantities. It should also be noted that HPC at 37°C counts have the highest variability in both sites and it may lead to higher probability of occurrence of extreme values which will skew the mean to higher values (see Table 2.2).

Table 2.2: Descriptive statistics for dependent and independent variables in all sites.

Sites	**Variables**	**N**	**Mean**	**St. Dev.**	**Min**	**Max**
Midmar	Total coliforms	6,853	**0.05**	**2.02**	0	145
	E.coli	6,853	**0.0004**	**0.04**	0	3
	HPC at 37°C	6,853	**1.58**	**23.61**	0	1,000
	Free chlorine	6,853	0.23	0.26	0.05	2.50
	Total chlorine	6,853	2.06	0.53	0.05	6.0
	pH	6,853	8.53	0.37	6.84	9.70
	Temperature	6,853	18.04	3.48	7.0	29.70
	Turbidity	6,853	0.22	0.09	0.01	2.14
DV Harris	Total coliforms	9,011	**0.043**	**0.933**	0	59
	E.coli	9,011	0	0	0	0
	HPC at 37°C	9,011	**1.23**	**17.89**	0	1,000
	Free chlorine	9,011	0.25	0.24	0.050	3.50
	Total chlorine	9,011	1.97	0.48	0.050	4.0
	pH	9,011	8.65	0.33	6.40	9.80
	Temperature	9,011	17.85	3.49	7.0	28.20
	Turbidity	9,011	0.25	0.14	0.01	7.32

2.1.2 Frequency Tables for Total Coliform Counts, *E. coli* and HPC at 37°C.

A frequency table is another way of summarizing data. The table depicts the number of times a data value occurs. Tables 2.3, 2.4 and 2.5 give the distribution of total coliform, *E. coli* and HPC at 37°C for each of the two sites, namely Midmar and DV Harris. The main common feature in all the tables is that most of the samples had more zero counts leading to highly skewed and zero-inflated data which justifies the use of statistical models that can adequately account for such excess zeros in the analysis. The overall compliance for total coliform counts was above 99.0% in all sites. There are many total coliform counts in DV Harris (86 counts) compared to Midmar (38 counts). It can be observed that the compliance for *E. coli* was 100% in DV Harris. *E. coli* was acceptable as it was found to be above 99.9% in Midmar. The overall compliance for HPC at 37°C in Midmar and DV Harris was 100%, according to Umgeni Water limits.

Table 2.3: Frequency table for total coliform counts occurrence in water samples.

Sites	Coliforms	Frequency	Relative frequency (%)
Midmar	0	6815	99.48
	1 - 10	33	0.47
	≥ 11	5	0.08
DV Harris	0	8925	99.05
	1 - 10	76	0.85
	≥ 11	10	0.10

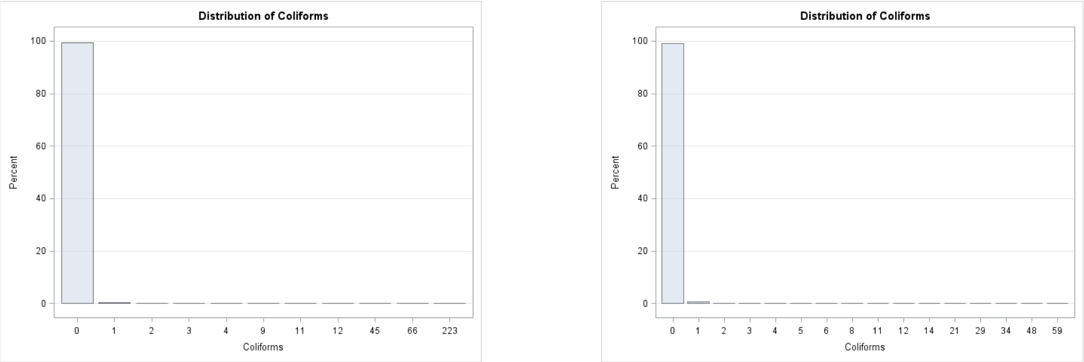
Table 2.4: Frequency table for *E.coli* occurrence in water samples.

Sites	*E. coli*	Frequency	Relative frequency (%)
Midmar	0	6852	99.99
	3	1	0.010
DV Harris	0	9011	100

Table 2.5: Frequency tables for HPC at 37°C occurrence in water samples.

Sites	HPC at 37°C	Frequency	Relative frequency (%)
Midmar	0	5107	74.52
	1 - 100	1737	25.35
	101 - 1000	9	0.13
DV Harris	0	7033	78.05
	1 - 100	1967	21.72
	101 - 1000	17	0.231

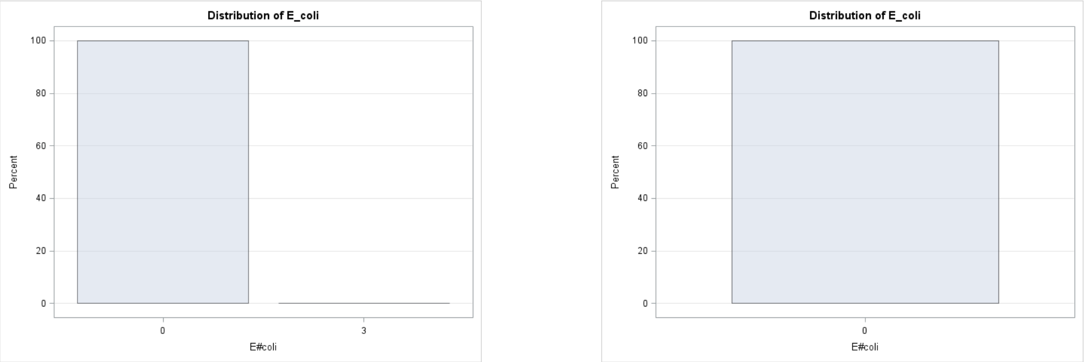
In Figures 2.1 and 2.2, the histograms reveals a high occurrence of zero total coliform and *E. coli* counts, which cannot be accounted for by the variance function of a Poisson or negative binomial distribution. It therefore seems sensible to apply the proposed techniques that account for zero inflation in the data.



(a) Midmar

(b) DV Harris

Figure 2.1: Frequency plot of total coliforms.



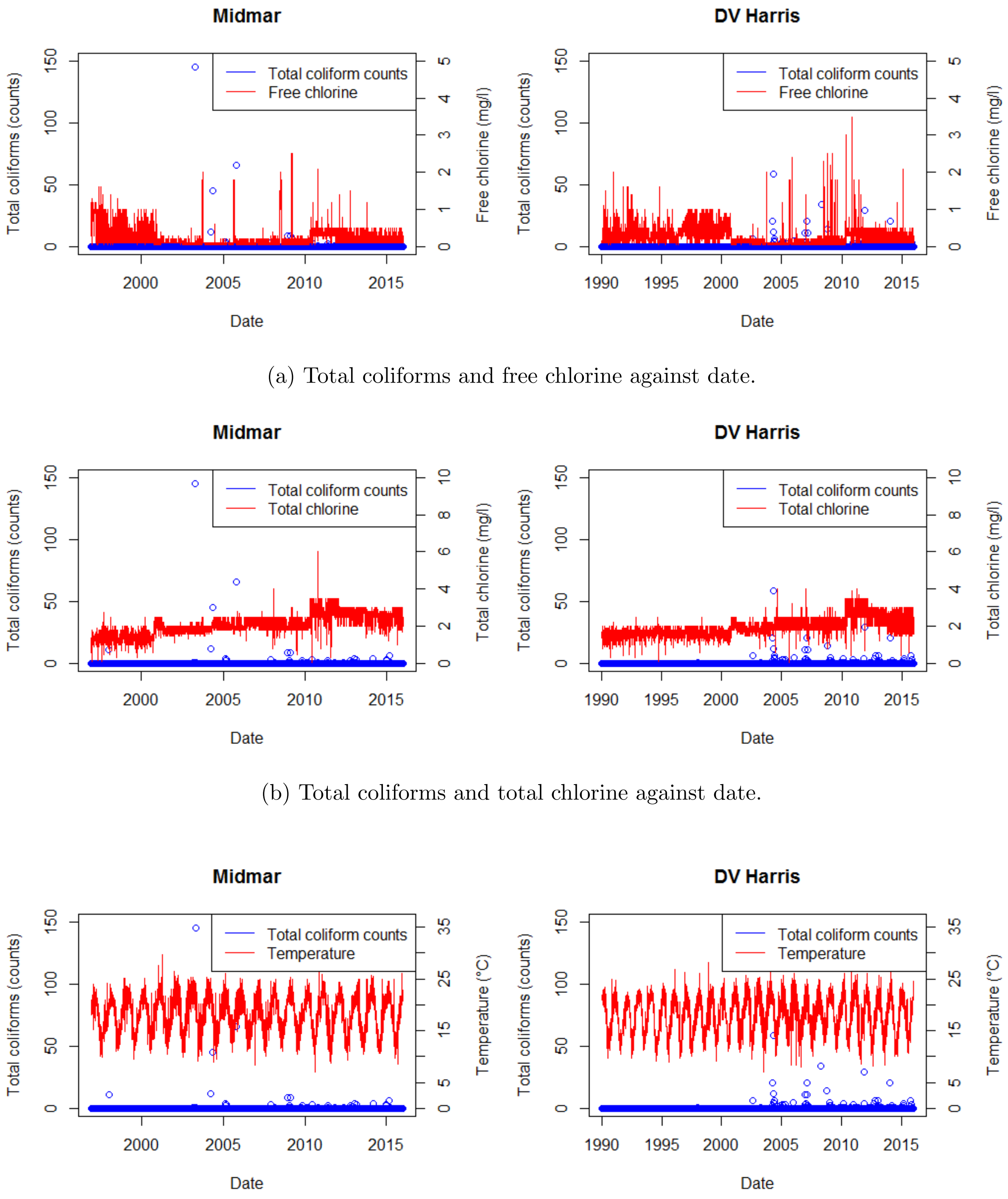
(a) Midmar

(b) DV Harris

Figure 2.2: Frequency plot of *E.coli.*

2.1.3 Plots of Total Coliforms Counts against other Measurable Variables

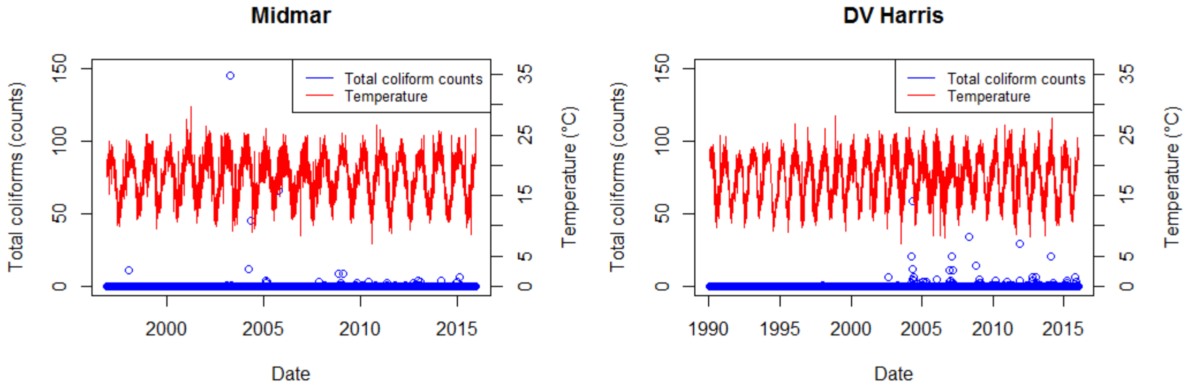
Comparative displays for total coliforms counts against all regressors are shown in the figures below.



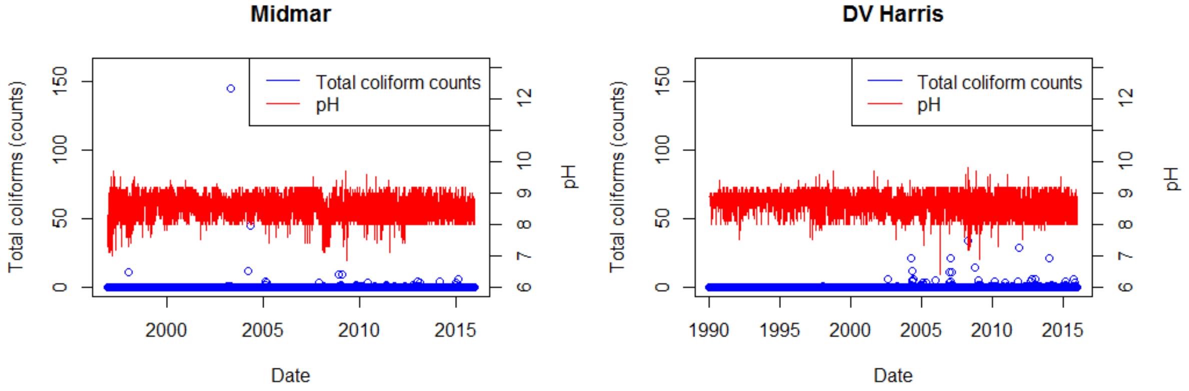
(a) Total coliforms and free chlorine against date.

(b) Total coliforms and total chlorine against date.

(c) Total coliforms and temperature against date.



(d) Total coliforms and turbidity against date.



(e) Total coliforms and pH against date.

All displays show that the number of total coliform counts increases or decreases with the regressors as expected. Total and free chlorine decreases/ inactivates the total coliform counts, the rate of inactivation varies widely but is more rapid when more chlorine is present in the water. The total coliform counts grow very fast at higher temperature as well. Turbidity has a positive effect on positive counts, therefore low (< 1.0 NTU) turbidity measurement is an indication of adequate water treatment. The pH values complies with the accepted limits in both sites, it is a good indication that water is safe to drink.

Chapter 3

Generalized Linear Models (GLM)

Generalized Linear models by Nelder and Baker (1972) and second one by McCullough and Nelder (1989) provide a powerful theoretical and computational framework, including classical linear models and enlarging the scope to a wider class of distributions under the exponential family. Generalized linear models (GLMs) expand the well known linear model to accommodate non-normal response variables in a single unified approach. It is common to find response variables which do not adhere to the standard assumptions of the linear models (normally distributed errors, constant variance), for example count data, dichotomous variables and truncated data. The GLM is based on well developed theory and with the advances in statistical software, these models have become a basic tool for statistical analyses by most researchers. There are two fundamental issues in the notion of generalized linear models, namely the distribution of the response (as we mentioned above), but also the functional form that relates the mean response to the regression variables.

The generalized linear models are an extension of classical linear models so that the latter form a suitable starting point for discussion. A vector of observation $\mathbf{y}$ having n components is assumed to be a realization of a random variable $\mathbf{Y}$ whose components are independently distributed with the vector of the means $\boldsymbol{\mu}$. The systematic part of the model is a specification of the i^{th} component for the vector $\boldsymbol{\mu}$ in terms of a small number of unknown parameters $\beta_0, ..., \beta_q$. In the case of ordinary linear models, this specification takes the form

$$\mu_i = \sum_{j=0}^{q} x_{ij}\beta_j, \tag{3.1}$$

where the β's are the parameter whose values are usually unknown and have to be estimated from the data. If we let i index the observation then the systematic part of the model may generally be written as

$$g(E(Y_i)) = g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j; \quad i = 1, ..., n, \tag{3.2}$$

where x_{ij} is the value of the jth covariate for observation i. In matrix notation where $\boldsymbol{\mu}$ is $n \times 1$, $\mathbf{X}$ is $n \times p$ and $\boldsymbol{\beta}$ is $p \times 1$ we may write

$$g(\boldsymbol{\mu}) = \boldsymbol{X\beta},$$

where $\boldsymbol{X}$ is the model matrix, $\boldsymbol{\beta}$ is the vector of parameters and $p = q+1$. This completes the specification of the systematic part of the model.

3.1 The Generalization

To simplify the transition to generalized linear models, we shall rearrange (3.1) slightly to produce the following three part specification:

1. The *random component*: the components of $\boldsymbol{Y}$ have independent Normal distributions with $\boldsymbol{E(Y)} = \boldsymbol{\mu}$ and constant variance σ^2.

2. The *systematic component*: covariates $\boldsymbol{x_1, x_2, ..., x_p}$ produce a *linear predictor*

$\boldsymbol{\eta}$ given by

$$\boldsymbol{\eta} = \sum_{1}^{p} \boldsymbol{x_j}\beta_j.$$

3. In the case of the identity link we have,

$$\boldsymbol{\mu} = \boldsymbol{\eta}.$$

This generalization introduces a functional relationship between the mean μ and the linear predictor η. If we write

$$\eta_i = g(\mu_i),$$

then $g(\cdot)$ will be called the *link function*. In this formulation, the classical linear models have a Normal (or Gaussian) distribution in component 1 and the identity link function for component 3. Generalized linear models allow two extensions; first the distribution in component 1 may come from an exponential family other than the Normal, and secondly the link function in component 3 may become any monotonic differentiable function.

3.2 The Exponential Family

An important unifying concept underlying the GLM is the exponential family of distributions. The exponential family of distribution was first described by Efron and Hinkley (1978). Members of the exponential family of distributions all have probability density (or probability mass) functions that can be expressed in the form:

$$f(y_i; \theta_i, \phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi)\right\}, \tag{3.3}$$

where θ_i is referred to as a natural or canonical parameter and $a(\phi_i)$, $b(\theta_i)$ and $c(y_i, \phi)$ are known functions. The term $a(\phi_i)$ has the form $a(\phi_i) = \frac{\phi}{w_i}$, where w_i is a known weight depending on whether the data is grouped and ϕ is referred to as the dispersion or scale parameter. It can be shown that if a response Y_i has a distribution belonging to the exponential family, then its mean and variance are

$$E(Y_i) = \mu_i = b'(\theta_i) \tag{3.4}$$

$$Var(Y_i) = b''(\theta_i)a(\phi_i), \tag{3.5}$$

where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of $b(\theta_i)$ with respect to θ_i. $b''(\theta_i)$ is a function of the mean, thus it is referred to as the variance function denoted by $v(\mu_i)$.

The variance of Y_i from the exponential family can be also expressed as

$$\begin{aligned} Var(Y_i) &= a_i(\phi)v(\mu_i) && (3.6) \\ &= \frac{\phi}{w_i}v(\mu_i). && (3.7) \end{aligned}$$

Thus, another property of the GLM is that of a non-constant variance where the variance may vary as a function of the mean. When $a_i(\phi) > 1$ the distribution of Y_i is said to be overdispersed since $Var(Y_i) > v(\mu_i)$. Similarly, the distribution of Y_i will be underdispersed when $a_i(\phi) < 1$. Therefore, standard errors calculated on the assumption $a_i(\phi) = 1$ would be incorrect when $a_i(\phi) \neq 1$.

3.3 Maximum Likelihood Estimation (MLE)

The idea behind MLE is to provide estimates for a given model's parameters. The estimated parameters maximise the likelihood of the sample data. In general, MLEs do not have a closed form for GLMs and therefore one has to rely on approximation methods such as the Newton-Raphson or Fisher scoring to find MLEs. From a statistical point of view, the method of maximum likelihood is considered to be more robust (with some exceptions) and yields estimators with good statistical properties when compared to other methods such as the method of least square estimation. The MLE estimation method is versatile and applies to most models and to different types of data. Due to advances in statistical theory and computer software, this method of estimation has become the most popular technique in applied statistics (Wu, 2005). The estimation of the dispersion parameter ϕ may also become necessary, particularly if it is differed from one implying the case of over-dispersion ($\phi > 1$) or under-dispersion ($\phi < 1$). The estimation of ϕ is imported in order to correctly determine standard errors for parameter estimates.

3.3.1 Parameters Estimation

The log-likelihood function contribution for a single observation is given by

$$\ell_i = ln(f(y_i; \theta_i, \phi)) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi). \tag{3.8}$$

Since Y_i, $i = 1, ..., n$, are independent, the joint log-likelihood function is

$$\ell(\boldsymbol{\beta}, \boldsymbol{y}) \quad = \quad \sum_{i=1}^{n} \ell_i. \tag{3.9}$$

The ML estimate of β_j, $j = 0, ..., p$, is the solution to the score equation

$$\tfrac{\partial \ell_i}{\partial \beta_j} \quad = \quad 0. \tag{3.10}$$

To obtain this solution, we use the chain rule of differentiation as

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}. \tag{3.11}$$

Using Equation (3.8), we get

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi_i)}. \tag{3.12}$$

Since $\mu_i = b'(\theta_i)$, $Var(Y_i) = a(\phi_i)v(\mu_i)$ and $\eta_i = \sum_j \beta_j x_{ij}$,

$$\begin{aligned} \frac{\partial \ell_i}{\partial \theta} &= \frac{y_i - \mu_i}{a(\phi_i)} \\ \frac{\partial \mu_i}{\partial \eta_i} &= b''(\theta_i) = v(\mu_i) \\ \frac{\partial \eta_i}{\partial \beta_j} &= x_{ij}. \end{aligned} \tag{3.13}$$

Thus,

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi_i)} \frac{1}{v(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \\ &= \sum_{i=1}^{n} (y_i - \mu_i) W_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij}, \end{aligned} \tag{3.14}$$

where W_i is referred to as the iterative weights given by

$$\begin{aligned} W_i &= \frac{1}{a(\phi)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 v_i^{-1} \\ &= \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2, \end{aligned} \tag{3.15}$$

and $v_i = v(\mu_i)$ is the variance function. Since $\eta_i = g(\mu_i)$, $\frac{\partial \mu_i}{\partial \eta_i}$ depends on the link function for the model.

Therefore, solving for the score equation below will give the ML estimate of β from

$$\sum_{i=1}^{n} (y_i - \mu_i) W_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij} = 0. \tag{3.16}$$

This score equation is a nonlinear function of β, and therefore requires iterative procedures to be solved. The Newton Raphson and Fisher Score iterative Equations can be used, where the score U is given by the left hand side of Equation (3.16). Thus, the Newton Raphson iterative equation will be

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} - (\boldsymbol{H}^{(t)})^{-1}\boldsymbol{U}^{(t)}, \tag{3.17}$$

and the Fisher Score iterative equation

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} + (\boldsymbol{T}^{(t)})^{-1}\boldsymbol{U}^{(t)}, \tag{3.18}$$

with information matrix

$$\begin{aligned} \boldsymbol{T} &= -E(\boldsymbol{H}) \\ &= -E\left(\tfrac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}}\right) \\ &= \boldsymbol{X'WX}, \end{aligned} \tag{3.19}$$

where $\boldsymbol{W}$ is known as the weight matrix with diagonal elements given in Equation (3.15). Equation (3.18) can also be represented as

$$\boldsymbol{T}^{(t)}\widehat{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{T}^{(t)}\widehat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{U}^{(t)}. \tag{3.20}$$

It can be shown that the right hand side of Equation (3.20) can be written as

$$\boldsymbol{X'W}^{(t)}\boldsymbol{z}^{(t)}, \tag{3.21}$$

where $\boldsymbol{W}^{(t)}$ is weight matrix evaluated at $\widehat{\boldsymbol{\beta}^{(t)}}$, and $\boldsymbol{z}^{(t)}$ has the following elements evaluated at $\widehat{\boldsymbol{\beta}^{(t)}}$

$$z_i = \eta_i + (y_i - \mu_i)\left(\tfrac{\partial \eta_i}{\partial \mu_i}\right). \tag{3.22}$$

This variable z_i is often called the adjusted dependent variable or the working dependent variable. Therefore, we can obtain

$$\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{X'W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X'W}^{(t)}\boldsymbol{z}^{(t)}. \tag{3.23}$$

Thus, each iteration step is the result of a weighted least squares regression of the adjusted variable $\boldsymbol{z}$ on the predictors $\boldsymbol{x}$ with working weight $\boldsymbol{W}$. Fisher scoring can therefore be regarded as an iteratively re-weighted least squares (IRWLS) procedure carried out on a transformed version of the dependent variable (Vazquez et al., 2010).

It follows that the asymptotic variance (also known as the asymptotic covariance) of this estimate of β is the inverse of the information matrix given in Equation (3.19) and can be estimated by

$$\hat{Var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X'}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}, \tag{3.24}$$

where $\widehat{\boldsymbol{W}}$ is $\boldsymbol{W}$ evaluated at $\widehat{\boldsymbol{\beta}}$ and depends on the link function of the model. The dispersion parameter ϕ, in the function $a(\phi_i)$ that is used in the calculation of W_i, gets cancelled out of the IRWLS procedure, thus the value of $\widehat{\boldsymbol{\beta}}$ is the same under any value of ϕ. However, the value of ϕ is required for the calculation of the variance of $\widehat{\boldsymbol{\beta}}$, therefore when ϕ is unknown, it can be estimated using a moment estimator (McCullough and Nelder, 1989), given by

$$\hat{\phi} = \frac{1}{n-p-1}\sum_{i=1}^{n}\frac{w_i(y_i - \hat{\mu_i})^2}{v(\hat{\mu_i})}, \tag{3.25}$$

where w_i is the weight for observation y_i defined in Equation (3.3).

3.4 Measure of Fit

An important step in statistical analysis is to assess the goodness-of-fit of the model of interest. One way in which this could be done is by using the *deviance*, a measure of discrepancy between the predicted values from the fitted model and the actual values from the data set. If, for the fitted model with $p+1$ parameters, $\ell(\widehat{\mu}, \phi, \mathbf{y})$ is the log-likelihood function maximized over $\widehat{\beta}$ for a fixed value of the dispersion parameter ϕ, and $\ell(\mathbf{y}, \phi, \mathbf{y})$ is the maximum log-likelihood achievable under the saturated model where the number of parameters equals the number of observations, the scaled deviance is

$$D^s = \frac{-2[\ell(\widehat{\mu}, \phi, \mathbf{y}) - \ell(\mathbf{y}, \phi, \mathbf{y})]}{\phi}. \tag{3.26}$$

If $\phi = 1$, the deviance is defined as

$$D = -2[\ell(\widehat{\mu}, \phi, \mathbf{y}) - \ell(\mathbf{y}, \phi, \mathbf{y})]. \tag{3.27}$$

The (scaled) deviance converges asymptotically to a χ^2 distribution with $n-p-1$ degrees of freedom. Thus, when testing at a level of significance of α, the fitted model is rejected if the calculated deviance is greater than or equal to $\chi^2_{n-p-1;\alpha}$.

Another commonly used measure of goodness-of-fit is the *generalized Pearsons chi-square statistic* given by

$$\chi^2 = \textstyle\sum_i^n \frac{(y_i - \hat{\mu_i})^n}{v(\hat{\mu_i})}, \tag{3.28}$$

where $v(\widehat{\mu_i})$ is the estimated variance function for the distribution in question. This statistic also asymptotically follows χ^2 distribution with $n-p-1$ degrees of freedom. Similar to the deviance, the smaller the value of the χ^2 statistic, the better the fit of the model. The scaled Pearsons χ^2 statistic is $\frac{\chi^2}{\phi}$ (Wu, 2005). For linear models, the value of the Pearsons χ^2 statistic is the residual sum of squares since $v(\widehat{\mu_i})$ is generally taken as one, and both the deviance and Pearsons χ^2 statistic have exact χ^2 distributions. For other distributions, these measures of goodness-of-fit have asymptotic χ^2 distributions and neither is superior to one another when samples are small. However, the deviance has an advantage over Pearsons χ^2 statistic as it is additive for nested models (Nelder and Baker, 1972).

3.4.1 Wald Test

When a hypothesis test on a single parameter, β_j , is to be carried out, a commonly used method is the Wald test. The test statistic for this test is

$$z_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}. \tag{3.29}$$

The standard error of $\hat{\beta}_j$ is the square root of the diagonal elements in the inverse of the information matrix given in Equation (3.19). This test statistic follows an approximately standard normal distribution. Some software packages square this value of the Wald test statistic and thus compare it to a chi-square distribution with 1 degree of freedom (Heeringa et al., 2010). Thus, for large values of the test statistic, one would reject the null hypothesis $H_0 : \beta_j = 0$ and conclude its corresponding variable is significant to the model.

Chapter 4

Zero-Inflated Count Data Regression Models

Various statistical models have been developed to model count data and zero-inflated count data. When describing count data variables, we note that it is common for many of the units to have never have exhibited or experienced positive counts. The resulting variable distribution, therefore, has many zeros and just a few other values (Atkins et al., 2013). In this chapter, four models are discussed that can deal with the excessive number of zeros, namely, the zero-inflated Poisson (ZIP), the zero-inflated negative binomial (ZINB) models, the Poisson logit hurdle (PLH), and the negative binomial logit hurdle (NBLH) models. There are two main distinctions in these abbreviations, namely zero-inflated (ZI) versus logit hurdle (LH), and Poisson versus negative binomial. The latter pair of Poisson versus negative binomial should be familiar territory with the negative binomial models (ZINB and NBLH) to deal with a certain degree of overdispersion. Furthermore, because a Poisson GLM is nested in a negative binomial GLM, the ZIP is nested in a ZINB, and a PLH is nested in a NBLH. The difference between ZI and LH models is slightly more complicated and is related to the nature of the zeros. Below, a brief outline for each of the models mentioned above is given.

4.1 The Poisson Model

Poisson regression is traditionally conceived as the basic count model upon which a variety of other count models are based (Hilbe, 2011). If events occur randomly over time, without occurrence dependence or duration dependence, the number of events during a unit time interval is Poisson distributed with probability function

$$f(y;\lambda) \quad = P(Y=y) = \tfrac{e^{-\lambda}(\lambda^{y})}{y!}, \quad y = 0,1,2,...;\lambda > 0 \tag{4.1}$$

where the random variable y is the count response and the parameter λ is the mean. One strict assumption about the model is that the mean is equal to the variance. Unlike most other distributions, the Poisson distribution does not have a distinct scale parameter (McCullagh and Nelder, 1989).

The standard Poisson distribution, which assumes equal variance and mean, is not appropriate to fit the observed counts since the variance of the most observed data is much larger than their mean. Violations of equidispersion indicate correlation in the data, which affects both standard errors of the parameter estimates and the further model fit. When such a situation arises, modifications are made to the Poisson model to account for inconsistency in the goodness of fit of the underlying distribution. The negative binomial (NB) distribution is commonly used to model overdispersed count data. A dispersion parameter is included in the NB model to cater for overdispersion by allowing the variance to be greater than the mean and accommodate the unobserved heterogeneity in the count data.

The Poisson regression model is derived from generalized linear models or GLMs (McCullagh and Nelder, 1989) and relates $\lambda, \boldsymbol{\beta}$ and $\boldsymbol{x_i'}$ through:

$$log(\lambda_i) = \boldsymbol{x_i'\beta}. \tag{4.2}$$

Here, $\boldsymbol{x_i'\beta}$ is the linear predictor, which is also symbolized by η within the context of generalized linear models (GLM). In equation (4.1), y denotes dependent variable having the Poisson distribution. The log-likelihood for the Poisson regression model is, (Walhin, 2001)

$$LL\left(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{x}\right) = \sum_{i=1}^{n}\left[y_i x_i'\beta - exp\left(x_i'\beta\right) - lny_i!\right]. \tag{4.3}$$

The vector parameters can be estimated by maximizing the log likelihood function in order to get ML estimates (Yesilova et al., 2012). Methods that can be used including the Newton-Raphson and the Fisher scoring methods which are iterative in nature.

4.2 The Negative Binomial (NB) Model

The negative binomial model is a type of generalized linear model in which the dependent variable Y is a count of the number of times an event occurs (Greene, 2008). The negative binomial distribution is given by:

$$P(Y=y) \quad = \frac{\Gamma(y+1/\alpha)}{\Gamma(y+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y} \tag{4.4}$$

where $\mu > 0$ is the mean of Y, $\alpha > 0$ is the heterogeneity parameter and the random variable Y has a negative binomial distribution with parameters $\tau \geq 0$ and $\lambda \geq 0$. The shape parameter τ quantifies the amount of overdispersion and its mean and variance of the distribution are given by:

$$E(Y) = \tau\lambda \tag{4.5}$$

$$\text{Var}(Y) = \tau\lambda(1+\lambda) = E(Y)(1+\lambda). \tag{4.6}$$

(Greene, 2008) derives this parametrization as a Poisson-gamma mixture, or alternatively as the number of failures before the $(1/\alpha)^{th}$ success, though we will not require $1/\alpha$ to be an integer. The traditional negative binomial regression model, designated the NB2 model in equation 4.4, is

$$ln\mu \quad = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p, \tag{4.7}$$

where the predictor variables $x_1, x_2, ..., x_p$ are given, and the population regression coefficients $\beta_0, \beta_1, \beta_2, ..., \beta_p$ are to be estimated.

Given a random sample of n subjects, we observe for subject i the dependent variable y_i and the predictor variables $x_{1i}, x_{2i}, ..., x_{pi}$. Utilizing vector and matrix notation, we let $\boldsymbol{\beta} = (\beta_0 \; \beta_1 \; \beta_2 \; ... \; \beta p)^T$, and we gather the predictor data into the design matrix X as follows:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

Designating the i^{th} row of $\boldsymbol{X}$ to be x_i, and exponentiating equation 4.7, we can then write the distribution in equation 4.4 as

$$P(Y = y_i) \quad = \tfrac{\Gamma(y_i+1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)} \left(\tfrac{1}{1+\alpha e^{x_i\beta}}\right)^{1/\alpha} \left(\tfrac{\alpha e^{x_i\beta}}{1+\alpha e^{x_i\beta}}\right)^{y_i}, \quad i = 1, 2, ..., n. \tag{4.8}$$

We estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ using maximum likelihood estimation. The likelihood function is

$$L\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \prod_{i=1}^{n} p(y_i) = \prod_{i=1}^{n} \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma(1/\alpha)} \left(\frac{1}{1 + \alpha e^{x_i\beta}}\right)^{1/\alpha} \left(\frac{\alpha e^{x_i\beta}}{1 + \alpha e^{x_i\beta}}\right)^{y_i}, \tag{4.9}$$

and the log-likelihood function is

$$lnL\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right)=\sum_{i=1}^{n}\left(y_i ln\alpha+y_i(x_i\beta)-\left(y_i+\tfrac{1}{\alpha}\right)ln(1+\alpha e^{x_i\beta})+ln\Gamma(y_i+\tfrac{1}{\alpha})-ln\Gamma(y_i+1)-ln\Gamma\left(\tfrac{1}{\alpha}\right)\right). \tag{4.10}$$

The values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that maximize $lnL(\boldsymbol{\alpha},\boldsymbol{\beta})$ will be the maximum likelihood estimates we seek, and the estimated variance-covariance matrix of the estimators is $\sum=-H^{-1}$, where H is the Hessian matrix of second derivatives of the log-likelihood function. Then the variance-covariance matrix can be used to find the usual Wald confidence intervals and $p-$values of the coefficient estimates.

Since $\mu\geq 0$, the variance of NB distribution generally exceeds its mean implying overdispersion (Winkelmann, 2008). It has been proposed as one of the distributions to model excessive variation in microbiological organisms that affect water quality. The NB distribution has been widely used for modeling count variables, usually for over-dispersed count outcome variables. However, distributional problems affect both models (Poisson and NB) such as overdispersion resulting from the specification of errors in the systematic part of the regression model, hence NB model themselves may be over-dispersed (Hilbe, 2011). Nevertheless, both models can be extended to accommodate any extra correlation or dispersion in the data that result in a violation of distributional properties of each respective distribution. The enhanced Poisson or NB model can be regarded as a solution to the violation of distributional assumptions of the primary model (4.1). For a better fit, an overdispersed model that incorporates excess zeros should serve as an alternative. Zero modified models such as zero-inflated models and hurdle count models are capable of incorporating excess zeros. They are applied to count data when overdispersion exists and have an excess of zeros.

4.3 Zero-Inflated (ZI) Models

The other problem with Poisson regression model is having far more zeros than expected by the distributional assumption of the Poisson and negative binomial models, hence resulting in incorrect parameter estimates. The use of zero-inflated Poisson or zero-inflated negative binomial models is proposed as a solution for this problem (Loeys et al., 2012).

4.3.1 Zero-Inflated Poisson (ZIP) Regression

This model was proposed by Lambert (1992) to model count data with excess zeros. In ZIP regression, excess zero counts are assumed to occur with probability p_i and follow a Poisson distribution with mean λ_i, with probability $1 - p_i$ where $i = 0, 1, 2, ..., n$. The ZIP model can thus be seen as a mixture of two component distributions, a zero part and non-zero component, given by:

$$P(Y = y_i|\lambda, p_i) = \begin{cases} p_i + (1 - p_i)e^{-\lambda_i}, & y_i = 0 \\ (1 - p_i)\frac{e^{-\lambda}\lambda^{y_i}}{y_i!}, & y_i = 1, 2, ... \end{cases} \tag{4.11}$$

The first part of the equation above is the zero part of the model and the second part is the non-zero count's part of the model. The two components together constitute the zero-inflated model.

The mean and variance of the zero-inflated Poisson model are:

$$E(Y_i) = \lambda_i(1 - p_i) \tag{4.12}$$

$$\mathrm{Var}(Y_i) = \lambda_i(1 - p_i)(1 + \lambda_i p_i), \tag{4.13}$$

where both p_i and λ_i are functions of x_i derived from (4.14) and (4.15) below. Note that this distribution approaches the Poisson distribution as $p_i \to 0$.

Note that, zero observations arise from both the zero-component distribution and the Poisson distribution. The zero-component distribution is therefore related to modeling ‘excess’ or ‘inflated’ zeros that are observed in addition to zeros that are expected to be observed under the assumed Poisson distribution. To assess the impact of covariates on the count distribution in a ZIP model p_i and λ_i can be explicitly expressed as a function of covariates using appropriate link functions. The most natural choice to model the probability of excess zeros is to use a logistic regression model with a logit link specified as:

$$\begin{aligned} logit(p_i) &= \boldsymbol{z_i'\gamma} \\ p_i &= \tfrac{e^{\boldsymbol{z_i'\gamma}}}{1+e^{\boldsymbol{z_i'\gamma}}} \\ 1-p_i &= \tfrac{1}{1+e^{\boldsymbol{z_i'\gamma}}}, \end{aligned} \tag{4.14}$$

where $\boldsymbol{x_i'}$ represents a vector of covariates and $\boldsymbol{\beta}$ a vector of parameters. The effect of covariates count data excluding excess zeros can be modelled through Poisson regression:

$$\begin{aligned} log(\lambda_i) &= \boldsymbol{x_i'\beta} \\ \lambda_i &= \boldsymbol{e^{x_i'\beta}}. \end{aligned} \tag{4.15}$$

Now we have:

$$f(y_i) = \begin{cases} p_i + (1-p_i)e^{-\lambda_i} & = \frac{e^{z_i'\gamma}}{1+e^{z_i'\gamma}} + \frac{e^{-e^{x_i'\beta}}}{1+e^{z_i'\gamma}};\ \text{when } y_i = 0 \\ (1-p_i)\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} & = \frac{e^{-e^{x_i'\beta}}(e^{x_i'\beta})^{y_i}}{(1+e^{z_i'\gamma})y_i!};\ \text{when } y_i = 1, 2, ... \end{cases} \tag{4.16}$$

The likelihood function is :

$$L(\boldsymbol{\gamma}, \boldsymbol{\beta}; \boldsymbol{y_i}) = \prod_{i=1}^{n} \left(\left(\frac{e^{z_i\gamma}}{1+e^{z_i\gamma}} + \frac{e^{-e^{x_i\beta}}}{1+e^{z_i\gamma}} \right) \left(\frac{e^{-e^{x_i\beta}}(e^{x_i\beta})^{y_i}}{(1+e^{z_i\gamma})y_i!} \right) \right), \tag{4.17}$$

and the log-likelihood function to be used to estimate the parameter vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, as well as, λ and p is given by,

$$\ell(\boldsymbol{\gamma}, \boldsymbol{\beta}; \boldsymbol{y_i}) = \sum_{y_i=0} log(e^{z_i\gamma} + e^{-e^{x_i\beta}}) - \sum_{i=1}^{n} log(1 + e^{z_i\gamma}) + \sum_{y_i>0}(y_ix_i\beta - e^{x_i\beta}) - \sum_{y_i>0} log(y_i!). \tag{4.18}$$

The parameter estimation can be carried out by employing EM algorithm or Newton-Raphson algorithm.

4.3.2 Zero-Inflated Negative Binomial (ZINB) Regression

The ZINB distribution is a mixture distribution, similar to the ZIP distribution, where p_i denotes the probability for excess zeros and the probability $(1 - p_i)$ is for the rest of the counts which follow the negative binomial distribution. Note that the negative binomial distribution can be viewed as a mixture of Poisson distributions, which allows the Poisson mean λ_i to be distributed as Gamma, and in this way overdispersion is modeled. The ZINB distribution is given by:

$$P(Y = y_i) = \begin{cases} p_i + (1 - p_i)\left(\dfrac{\tau}{\tau + \lambda_i}\right)^{\tau}, & y_i = 0 \qquad (4.19) \\ (1 - p_i)\dfrac{\Gamma(\tau + y_i)}{y_i!\Gamma(\tau)}\left(\dfrac{\tau}{\tau + \lambda_i}\right)^{\tau}\left(\dfrac{\lambda_i}{\lambda_i + \tau}\right)^{y_i}, & y_i = 1, 2, ... \quad (4.20) \end{cases}$$

The mean and variance of the ZINB distribution are:

$$E(Y_i) = (1 - p_i)\lambda_i \tag{4.21}$$

$$\text{Var}(Y_i) = (1 - p_i)\lambda_i\left(1 + p_i\lambda_i + \frac{\lambda_i}{\tau}\right). \tag{4.22}$$

Observe that this distribution approaches the zero-inflated Poisson distribution and the negative binomial distribution as $\tau \rightarrow \infty$ and $p_i \rightarrow 0$, respectively. If both $\frac{1}{\tau}$ and $p_i \approx 0$ then the ZINB distribution reduces to the Poisson distribution. Now consider a sample of n observation which independently follow the ZINB distribution but not necessary identical. The ZINB regression model relates p_i and λ_i to covariates, through the equations:

$$log(\lambda_i) = \boldsymbol{x_i'\beta}, \tag{4.23}$$

and

$$logit(p_i) = \boldsymbol{z_i'\gamma}, \tag{4.24}$$

where $i = 1, 2, ..., n$ and $\boldsymbol{x_i'}$ and $\boldsymbol{z_i'}$ are p$-$ and q$-$ dimensional vectors of covariates pertaining to the ith subject, and with $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ the corresponding vectors of regression coefficients, respectively.

The ZINB log-likelihood given the observed data is:

$$\begin{aligned} l(\boldsymbol{\beta, \gamma; y}) &= \sum_{i=1}^{n} log(1 + e^{z_i\gamma}) - \sum_{i=1:y_i=0}^{n} log\left(e^{z_i\gamma} + \left(\frac{e^{x_i\beta+\tau}}{\tau}\right)^{-\tau}\right) \\ &+ \sum_{i=1:y_i>0}^{n} \left(\tau log\left(\frac{e^{x_i\beta}+\tau}{\tau}\right) + y_i log(1 + e^{-x_i\beta}\tau)\right) \\ &+ \sum_{i=1:y_i>0}^{n} (log\Gamma(\tau) + log\Gamma(1 + y_i) - log\Gamma(\tau + y_i)). \end{aligned} \tag{4.25}$$

Parameter estimation can be carried out using the quasi-Newton optimization method.

4.4 Hurdle Regression Model

The hurdle model proposed by Mullahy (1986) uses a two-part model where the first part is a binary outcome model, and the second part is a truncated count model. As per

Cameron and Trivedi (1998) "Such a partition permits the interpretation that positive observation arises from crossing the zero hurdle or the zero thresholds. The first part models the probability that the threshold is crossed. In principle, the threshold need not be at zero; it could be any value. The zero value has a special appeal because in many situations it partitions the population into subpopulations in a meaningful way". So, a data set is split into zero and non-zero (positive) values to fit two different models with associated covariates in regression. A variety of population distributions can be considered for zero counts. The frequently used distributions in real-life data are binomial distribution, Poisson distribution, and negative binomial distribution.

4.4.1 Poisson Logit Hurdle (PLH) Model

PLH model is a two-component model comprising of a hurdle component that models zero versus non-zero counts, and a truncated Poisson count component is employed for the non-zero counts:

$$P(Y = y_i|\lambda, p_i) = \begin{cases} p_i, & y_i = 0 \\ \frac{(1-p_i)e^{-\lambda}\lambda^y}{(1-e^{-\lambda})y_i!}, & y_i = 1, 2, ... \end{cases} \tag{4.26}$$

where now p_i models all zeros from the degenerate zero distribution and the standard Poisson count distribution. For PLH model, the most natural choice to model probability of zeros as a function of covariates is to use a logistic regression model:

$$logit(p_i) = z_i'\gamma, \tag{4.27}$$

while the effect of covariates z' on strictly positive (that is, censored) count data are modeled through Poisson regression:

$$log(\lambda_i) = x_i'\beta. \tag{4.28}$$

Note that equations (4.27) and (4.28) allow the covariates in the zero counts and the all non-zero models to be different but in practice the set of covariates in the two components are made to be the same.

4.4.2 Negative Binomial Logit Hurdle (NBLH) Model

Similarly to the Poisson hurdle model, the NBLH distribution can be used instead of Poisson distribution above in case of over-dispersion. In this case the all-zero and the positive count components are given by

$$P(Y = y_i|\lambda, p_i) = \begin{cases} p_i, & y_i = 0 \\ (1-p_i)\frac{\Gamma(y_i+\tau)}{\Gamma(y_i+1)\Gamma(\tau)}\frac{(1+\tau\lambda)^{-(y_i+\tau)}\tau^{y_i}\lambda^{y}}{1-(1+\tau\lambda)^{\tau}}, & y_i = 1, 2, ... \end{cases} \tag{4.29}$$

Again the most natural choice to model probability of excess zeros is to use a logistic regression model:

$$logit(p_i) = x_i'\gamma. \tag{4.30}$$

Impact of covariates on count data are modelled through NB regression model given by:

$$log(\lambda_i) \;=\; x_i'\beta. \tag{4.31}$$

Given $\pi = P(Y > 0)$, the probability of a non-zero response with $\eta = x_i'\beta$ as given in Equation (4.31), the expected value and the corresponding variance are given by:

$$E(Y) = \eta = \frac{\pi\lambda_i}{1 - p(0;\tau)} \tag{4.32}$$

$$Var(Y_i) = \eta(\lambda_i - \eta_i) + \frac{\pi\sigma^2}{1 - p(0;\tau)}. \tag{4.33}$$

Both zero-inflated and hurdle models need distributional assumptions for their count component. The two classes differ with respect to their dependencies of estimation of parameters of the "zero" component on these assumptions (Loeys et al., 2012). Unlike ZI models, estimation of parameters γ related to p_i in the hurdle model is not dependent on estimation of parameters β related to λ_i. Hence, if assumptions about the (truncated) Poisson/ negative binomial model are violated (for example due to extreme outlying observations), the hurdle model will in contrast to zero-inflated model, still yield consistent estimators for parameters in the logit model (if correctly specified) (Loeys et al., 2012). Much as the hurdle model will be consistent in the absence of a good model for the non-zero counts, one of it's weaknesses is that it assumes all zeros to come from a single degenerate population.

4.5 Model choice between a hurdle model and a zero-inflated model

The choice between a zero-inflated model and a hurdle model is often dependent on the nature of the problem. Although these two models are similar in many aspects, conceptually there is a subtle difference between the two models and depending on the application and the data collection procedures, one may be more appropriate than the other (Hilbe, 2014). Despite differences between the modelling frameworks (the hurdle model includes a mass at zero and a truncated distribution whereas the zero-inflated model is based on a mass at zero and a regular distribution), the inferential results are often very similar. Hurdle models are more general in the sense that they can handle both cases where there are fewer or more zeros than assumed by a regular distribution. Note, the hurdle models does not necessarily have to be set at 0. Zero-inflated models, although less general than hurdle models, are sometimes preferred due to the assumption that two different types of zeros (structural or true zeros, vs. sampling zeros) may exist in the data. Ideally, the hurdle models are more appropriate for cases where a real separation of mechanisms producing the zeros and the positive counts is justified (Hilbe, 2014). Otherwise, zero-inflated models are more appropriate due to lack of information or knowledge regarding the non-existence of overlap between the two potential sources of zeros. In practice, often due to lack of clear evidence regarding the nature of zeros, model selection procedures are implemented in order to arrive at the best model choice.

From a technical standpoint, the two methodologies are substantially different since hurdle models are two-stage models where the algorithms for fitting the model for the binary component and the non-zero data component are implemented separately while zero-inflated models fall under the class of finite mixture models (the parameters for zero and non-zero parts of the models are estimated simultaneously). Consequently, the interpretation of parameter estimation results may be difficult across these two models due to differences in model structures.

4.5.1 Vuong Test

For non-nested models, a comparison between models with p.m.f, $p_1(.)$ and $p_2(.)$ can be performed using Vuong test (Vuong, 1989), $V = \frac{m\sqrt{n}}{sd(m)}$, where m is the mean of m_i, $sd(m)$ is the standard deviation of m_i,n is the sample size and $m_i = ln\left(\frac{p_{1i}(y_i)}{p_{2i}(y_i)}\right)$. The Vuong test statistics follows a standard normal. As an example, for 0.05 significance level, the first model is “closer” to the actual model if V is larger than 1.96. On the other hand, the second model is “closer” to the actual model if V is smaller than -1.96. Otherwise, neither model is “closer” to the actual model and there is no difference between using the first or the second model.

For models with unequal number of parameters, the equation for m_i in the Vuong test is slightly modified to account for the difference in the number of parameters, $m_i = \ln\left(\frac{p_{1i}(y_i)}{p_{2i}(y_i)}\right) - \frac{k_1-k_2}{2}\ln(n)$, where k_1 and k_2 are the number of parameters in model 1 and model 2 respectively.

4.6 Model Comparison

Akaike's information criterion was developed by Hirotsugu Akaike under the name of "an information criterion" (AIC) in 1971 and proposed in Posada and Crandall (1998). For comparison of non-nested models based on maximum likelihood, to choose the best fitting model, Akaike's information criterion (AIC) has been proposed for model selection criteria based on the fitted log-likelihood function. As a measure of the relative goodness of fit of a statistical model, AIC not only rewards goodness of fit but also includes a penalty that is an increasing function of the number of estimated parameters. Since the log-likelihood is expected to increase as parameters are added to a model, the AIC criteria penalize models with larger q. This penalty function may also be a function of n, the number of observations. This penalty discourages over fitting. Thus the AIC is specified as

$$AIC = -2log(L) + 2q, \tag{4.34}$$

where L is the maximized value of the likelihood function for the estimated model, with q being equal to the number of degrees of freedom used in the model and 2 is a tuning parameter meant to balance the information in the model based on the degrees of freedom with information in the residuals. A model with lowest AIC is preferred. Several alternatives of AIC also exist, viz Bayesian information criteria (BIC) and Consistent Akaike's information criterion (CAIC). AIC is optimal in selecting the model with the least mean squared error while BIC is not asymptotically optimal. An AIC, CAIC or BIC difference of less than 4 indicates that the two competing models are indistinguishable, while a value difference of 4 to 10 suggests moderate superiority of one model against the other, and an AIC, CAIC or BIC difference of greater than 10 implies that for two competing models, one model is better than the other.

Chapter 5

Application and Results

In the following, we illustrate all models described above by applying them to the data set from Umgeni Water. At the end of this chapter, all fitted models are compared highlighting that the mean function is similar across models and that the fitted likelihoods are different. The models differ with respect to explaining overdispersion and the excess zeros in the data.

5.1 Fitting the Poisson Model

As a first attempt to capture the relationship between the total coliform counts and all regressors in a parametric regression model, we fitted the basic Poisson regression model and obtain the coefficient estimates along with associated partial Wald tests. Note, *E. coli* could not be modelled because it had very few positive counts. The level of significant used for the statistical tests is 0.05. Results in Tables 5.1 and 5.2 are statistical test results for the two sites after fitting the Poisson model.

Table 5.1: Results of Poisson model estimates for total coliform counts at Midmar site.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.1699	1.4160	-7.18	<0.0010***
Free chlorine	-5.8140	0.7033	-8.27	<0.0010***
Total chlorine	-1.8879	0.1367	-13.81	<0.0010***
pH	0.6114	0.1509	4.05	<0.0010***
Temperature	0.2544	0.0192	13.27	<0.0010***
Turbidity	3.1501	0.3622	8.70	<0.0010***
Time	0.0741	0.0130	5.70	<0.0010***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.2: Results of Poisson model estimates for total coliform counts at DV Harris site.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.0724	1.4562	-6.23	<0.0010***
Free chlorine	-6.6724	0.6759	-9.87	<0.0010***
Total chlorine	-1.7079	0.1225	-13.94	<0.0010***
pH	0.7064	0.1555	4.54	<0.0010***
Temperature	0.1141	0.0163	7.00	<0.0010***
Turbidity	-2.2525	0.5660	-3.98	<0.0010***
Time	0.1632	0.0098	16.58	<0.0010***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

From both Tables 5.1 and 5.2, we notice that all coefficients are highly significant on both sites. However, the Wald test results might be too optimistic due to a misspecification of the likelihood. From the exploratory analysis in Chapter 2, it was clear that overdispersion is present in this data set. Thus, a first remedial action is to re-compute the Wald tests using sandwich standard errors and the R-statement. Sandwich standard errors can be used to estimate the variance of MLE when underlying model is incorrect.

`coeftest(Poisson, vcov = sandwich)` .

Table 5.3: Results of Wald test for total coliform counts at Midmar site.

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-10.1699	4.6499	-2.19	0.0287*
Free chlorine	-5.8140	2.7248	-2.13	0.0328*
Total chlorine	-1.8879	0.7414	-2.55	0.0109*
pH	0.6114	0.4871	1.26	0.2094
Temperature	0.2544	0.1026	2.48	0.0132*
Turbidity	3.1501	1.3526	2.33	0.0199*
Time	0.0741	0.0339	2.19	0.0286*

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.4: Results of Wald test for total coliform counts at DV Harris site.

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-9.0724	6.2263	-1.46	0.1451
Free chlorine	-6.6724	1.4357	-4.65	< 0.0010***
Total chlorine	-1.7079	0.6197	-2.76	0.0058**
pH	0.7064	0.7046	1.00	0.3161
Temperature	0.1141	0.0537	2.13	0.0335*
Turbidity	-2.2525	2.3467	-0.96	0.3371
Time	0.1632	0.0430	3.78	0.0002***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

The only difference we noticed between the two pairs of tables (Tables 5.1, 5.2 versus 5.3, 5.4), is that in the Tables 5.3 and 5.4 the standard errors are large but the coefficient estimates remain the same. All regressors are still significant ($p < 0.05$) except pH in the Midmar site, pH and temperature in the DV Harris site. Next we consider more superior models that deal with overdispersion and excess zeros in a more formal way.

5.1.1 Fitting the Negative Binomial Model

A more formal way to accommodate overdispersion in a count data regression model is to use a negative binomial model. This model was fitted to the data and the results are presented in Tables 5.5 and 5.6 for Midmar and DV Harris sites respectively.

Table 5.5: Results of negative binomial model estimates for total coliform counts at Midmar site.

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-19.1488	7.4805	-2.56	0.0105*
Free chlorine	-3.4536	1.7479	-1.98	0.0482*
Total chlorine	-0.7836	0.8756	-0.89	0.3709
pH	0.4865	0.8042	0.60	0.5452
Temperature	0.5778	0.1008	5.73	< 0.0010***
Turbidity	7.8449	2.7783	2.82	0.0048**
Time	0.0714	0.0824	0.87	0.3865

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.6: Results of negative binomial model estimates for total coliform counts at DV Harris site.

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	0.1908	4.6746	0.04	0.9674
Free chlorine	-7.5007	1.5617	-4.80	< 0.0010***
Total chlorine	-1.3974	0.4806	-2.91	0.0036**
pH	-0.6214	0.5080	-1.22	0.2213
Temperature	0.1248	0.0501	2.49	0.0128*
Turbidity	-2.3980	1.7119	-1.40	0.1613
Time	0.2511	0.0360	6.97	< 0.0010***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

As shown in Tables 5.5 and 5.6, both regression coefficients and standard errors are not quite similar to the results in previous models (Tables 5.1, 5.2, 5.3 and 5.4). We note that the standard errors of negative binomial model are also generally larger than those from the standard Poisson model as expected. One advantage of the negative binomial model is that it is associated with a formal likelihood so that information criterion are readily available. The non-significant variables are total chlorine ($p = 0.3709$), pH ($p = 0.5452$), and time ($p = 0.3865$) at the Midmar site while for the DV Harris site it is pH ($p = 0.2213$) and turbidity ($p = 0.1613$) that are non-significant. Free chlorine has a significant negative effect on total coliform counts on both sites as expected. Total chlorine also has a significant negative effect on total coliform counts at the DV Harris site as expected. Temperature ($p < 0.001$) and turbidity ($p < 0.01$) have significant positive effects on total coliform counts in the Midmar site while turbidity ($p = 0.1613$) is not significant at the DV Harris site but the temperature ($p < 0.01$) is significant. The time effect is not significant at the Midmar site but it is significant at the DV Harris site. The model seems to indicate an increasing occurrence of total coliform counts over time but caution is needed because the excess zeros are not well accounted in the model.

5.2 Zero-Inflated and Hurdle Regression Estimation Results

The exploratory analysis in Chapter 2 conveyed the impression that there might be more zero observations than explained by the basic count data distributions, hence in this section the zero-inflated and hurdle models are used to model both positive counts and excess zeros. Tables 5.7 and 5.8 show the parameter estimates of zero-inflated and hurdle models with associated standard error under both the Poisson and negative binomial distributions at Midmar and DV Harris sites.

Table 5.7: Estimation of coefficients using zero-inflated and hurdle models for total coliform counts at Midmar site.

	ZIP	ZINB	PLH	NBLH
Count model coefficients				
(Intercept)	3.33	-11.29	3.67	-3.82
	(2.95)	(10.73)	(3.00)	(11.53)
Free chlorine	-6.39^{***}	-4.18	-6.46^{***}	-6.09
	(0.72)	(3.64)	(0.73)	(3.75)
Total chlorine	-0.48^{***}	-0.65	-0.53^{***}	-0.59
	(0.13)	(1.05)	(0.14)	(1.72)
pH	0.45	1.30	0.39	-0.19
	(0.29)	(1.07)	(0.30)	(1.64)
Temperature	-0.05	0.16	-0.04	0.03
	(0.03)	(0.15)	(0.03)	(0.31)
Turbidity	6.56^{***}	3.59	6.84^{***}	6.48
	(0.77)	(2.57)	(0.80)	(4.01)
Time	-0.35^{***}	-0.29^{**}	-0.35^{***}	-0.29^{*}
	(0.02)	(0.09)	(0.02)	(0.13)
Zero model coefficients				
(Intercept)	13.45^{**}	4.53	-12.92^{**}	-12.92^{**}
	(4.53)	(12.51)	(4.44)	(4.44)
Free chlorine	-1.35	-2.60	-1.56	-1.56^{**}
	(1.35)	(2.62)	(1.35)	(1.35)
Total chlorine	1.05^{*}	1.57	-1.18^{**}	-1.18^{**}
	(0.49)	(1.01)	(0.46)	(0.46)
pH	-0.28	0.88	0.42	0.42
	(0.48)	(1.20)	(0.47)	(0.47)
Temperature	-0.27^{***}	-0.37	0.25^{***}	0.25^{***}
	(0.06)	(0.20)	(0.06)	(0.06)
Turbidity	2.28	0.84	-0.65	-0.65
	(1.75)	(3.83)	(1.72)	(1.72)
Time	-0.29^{***}	-0.58^{**}	0.19^{***}	0.19^{***}
	(0.05)	(0.20)	(0.04)	(0.04)
Log(theta)		-4.31^{***}		-11.80
		(0.56)		(15.89)
AIC	1020.63	624.77	1021.75	618.33
Log Likelihood	-496.32	-297.39	-496.87	-294.16
Number of observations	6853	6853	6853	6853

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.8: Estimation of coefficients using zero-inflated and hurdle models for total coliform counts at Midmar site.

	ZIP	ZINB	PLH	NBLH
Count model coefficients				
(Intercept)	13.26^{***}	8.65	13.08^{***}	6.78
	(1.68)	(7.16)	(1.70)	(24.72)
Free chlorine	-2.71^{**}	-1.58	-2.78^{**}	-0.92
	(0.84)	(2.36)	(0.86)	(2.82)
Total chlorine	-0.90^{***}	-1.83^{***}	-0.90^{***}	-2.54^{***}
	(0.12)	(0.50)	(0.13)	(0.77)
pH	-0.69^{***}	-0.19	-0.67^{***}	-0.30
	(0.17)	(0.69)	(0.17)	(0.84)
Temperature	-0.07^{***}	-0.02	-0.07^{***}	-0.08
	(0.02)	(0.11)	(0.02)	(0.10)
Turbidity	-3.81^{***}	-3.05	-3.89^{***}	-5.45^{*}
	(0.76)	(1.99)	(0.78)	(2.60)
Time	-0.07^{***}	-0.15^{*}	-0.07^{***}	-0.12
	(0.02)	(0.07)	(0.02)	(0.07)
Zero model coefficients				
(Intercept)	18.77^{***}	24.54^{***}	-15.88^{***}	-15.88^{***}
	(3.24)	(6.87)	(3.07)	(3.07)
Free chlorine	2.64^{*}	2.27	-3.54^{**}	-3.54^{**}
	(1.13)	(2.16)	(1.09)	(1.09)
Total chlorine	0.43	-0.02	-0.65^{*}	-0.65^{*}
	(0.30)	(0.57)	(0.29)	(0.29)
pH	-0.93^{**}	-1.27	0.76^{*}	0.76^{*}
	(0.34)	(0.66)	(0.33)	(0.33)
Temperature	-0.20^{***}	-0.27^{**}	0.18^{***}	0.18^{***}
	(0.04)	(0.09)	(0.04)	(0.04)
Turbidity	-1.50	-2.71	0.07	0.07
	(1.17)	(2.13)	(0.64)	(0.64)
Time	-0.20^{***}	-0.36^{***}	0.19^{***}	0.19^{***}
	(0.02)	(0.08)	(0.02)	(0.02)
Log(theta)		-3.14^{***}		-8.55
		(0.50)		(23.54)
AIC	1578.91	1191.54	1580.68	1187.58
Log Likelihood	-775.45	-580.77	-776.34	-578.79
Number of observations	9011	9011	9011	9011

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

5.2.1 Heterotrophic Plate Counts (HPC at 37 °C)

Tables 5.9 and 5.10 show results of HPC at 37 °C for the zero-inflated and hurdle models under both the Poisson and negative binomial distributions at Midmar and DV Harris sites.

Table 5.9: Estimation of coefficients using zero-inflated and hurdle models for HPC at 37 °C at Midmar site.

	ZIP	ZINB	PLH	NBLH
Count model coefficients				
(Intercept)	2.77^{***}	-1.62	2.79^{***}	-9.41
	(0.24)	(1.01)	(0.24)	(18.79)
Free chlorine	-0.28^{***}	-0.31^{*}	-0.30^{***}	-0.47^{**}
	(0.05)	(0.14)	(0.05)	(0.18)
Total chlorine	-1.42^{***}	-0.78^{***}	-1.43^{***}	-0.88^{***}
	(0.02)	(0.09)	(0.02)	(0.11)
pH	-0.08^{**}	0.03	-0.08^{**}	-0.19
	(0.03)	(0.11)	(0.03)	(0.14)
Temperature	0.06^{***}	0.12^{***}	0.06^{***}	0.08^{***}
	(0.00)	(0.01)	(0.00)	(0.02)
Turbidity	1.12^{***}	1.73^{***}	1.12^{***}	1.30^{*}
	(0.07)	(0.49)	(0.07)	(0.59)
Time	0.11^{***}	0.09^{***}	0.11^{***}	0.11^{***}
	(0.00)	(0.01)	(0.00)	(0.01)
Zero model coefficients				
(Intercept)	5.44^{***}	1.28	-5.34^{***}	-5.34^{***}
	(0.75)	(2.71)	(0.74)	(0.74)
Free chlorine	0.13	-0.18	-0.18	-0.18
	(0.12)	(0.41)	(0.12)	(0.12)
Total chlorine	-0.03	0.03	-0.13	-0.13
	(0.09)	(0.27)	(0.09)	(0.09)
pH	-0.08	0.32	0.08	0.08
	(0.08)	(0.29)	(0.08)	(0.08)
Temperature	-0.19^{***}	-0.32^{***}	0.20^{***}	0.20^{***}
	(0.01)	(0.03)	(0.01)	(0.01)
Turbidity	-1.51^{***}	-1.95	1.59^{***}	1.59^{***}
	(0.29)	(1.05)	(0.28)	(0.28)
Time	0.02^{**}	0.08^{**}	-0.01	-0.01
	(0.01)	(0.03)	(0.01)	(0.01)
Log(theta)		-1.88^{***}		-12.61
		(0.05)		(18.75)
AIC	55667.45	14912.48	55658.46	14508.39
Log Likelihood	-27819.73	-7441.24	-27815.23	-7239.20
Number of observations	6853	6853	6853	6853

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.10: Estimation of coefficients using zero-inflated and hurdle models for HPC at 37 °C at DV Harris site.

	ZIP	ZINB	PLH	NBLH
Count model coefficients				
(Intercept)	-5.09^{***}	-8.88^{***}	-5.11^{***}	-22.87
	(0.28)	(1.12)	(0.28)	(64.15)
Free chlorine	0.10^{**}	-0.04	0.10^{**}	0.04
	(0.04)	(0.13)	(0.04)	(0.16)
Total chlorine	-0.39^{***}	-0.01	-0.39^{***}	-0.20
	(0.03)	(0.14)	(0.03)	(0.19)
pH	0.72^{***}	0.72^{***}	0.72^{***}	1.04^{***}
	(0.03)	(0.12)	(0.03)	(0.16)
Temperature	0.02^{***}	0.11^{***}	0.02^{***}	0.03
	(0.00)	(0.01)	(0.00)	(0.02)
Turbidity	0.00	0.39	-0.00	0.15
	(0.04)	(0.26)	(0.04)	(0.16)
Time	0.07^{***}	0.06^{***}	0.07^{***}	0.06^{***}
	(0.00)	(0.01)	(0.00)	(0.01)
Zero model coefficients				
(Intercept)	4.58^{***}	-14.42^{**}	-5.04^{***}	-5.04^{***}
	(0.77)	(5.16)	(0.76)	(0.76)
Free chlorine	-0.05	-0.76	0.06	0.06
	(0.11)	(0.75)	(0.11)	(0.11)
Total chlorine	0.09	0.78	-0.11	-0.11
	(0.08)	(0.43)	(0.08)	(0.08)
pH	-0.11	0.58	0.15	0.15
	(0.08)	(0.48)	(0.08)	(0.08)
Temperature	-0.11^{***}	-0.23^{***}	0.12^{***}	0.12^{***}
	(0.01)	(0.04)	(0.01)	(0.01)
Turbidity	-1.32^{***}	-6.29^{**}	1.32^{***}	1.32^{***}
	(0.23)	(2.22)	(0.23)	(0.23)
Time	-0.01	0.52^{***}	0.01^{*}	0.01^{*}
	(0.01)	(0.10)	(0.01)	(0.01)
Log(theta)		-2.30^{***}		-15.55
		(0.03)		(64.13)
AIC	61066.33	17011.13	61067.78	16669.30
Log Likelihood	-30519.16	-8490.57	-30519.89	-8319.65
Number of observations	9011	9011	9011	9011

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Poisson regression was not used for this model because there was clear indication that the distribution of the dependent variable (total coliform counts) had a variance considerably larger than the mean. The Poisson goodness of fit analysis indicated that the Poisson distribution should not be used ($p < 0.001$). Instead, a negative binomial model was preferred with the total coliform counts as the dependent variable. A test of the overdispersion parameter alpha revealed that alpha was significantly different from zero ($p < 0.001$), indicating that the negative binomial distribution was superior to the Poisson distribution. Zero-inflated and hurdle models were used because of the prevalence of zero counts. The negative binomial hurdle model was considered as the best model because it had smaller AIC values (618.33 and 1187.58). Tables 5.7, 5.8, 5.9 and 5.10 show the summary of fitted count regression for Umgeni Water data: coefficient estimates from count model, zero-inflation model (both with standard errors), the number of estimated parameters, maximized log-likelihood and AIC. All coefficient estimates confirm the results from the exploratory analysis in Figure 2.3a to Figure 2.3e. Tables 5.7 and 5.8 show the results for total coliform counts at the Midmar and DV Harris sites. The count model parts at the Midmar site shows that time has a negative effect with 0.29 units on the total coliform counts, meaning that total coliform counts are significantly decreasing over time. Total chlorine is significant ($p < 0.001$) at DV Harris site and that means it is reducing the total coliform counts. Turbidity at DV Harris site has a negative effect with 5.45 units on total coliform counts. The low turbidity measurements are indications of adequate water treatment. The zero model parts in Tables 5.7 and 5.8 show that free and total chlorine have a positive significant effect ($p < 0.01$) on zero counts at the Midmar and DV Harris sites, therefore it increases the number of zeros (total coliform counts meet the accepted standards limits). Free chlorine was not significant ($p > 0.05$). The model also showed that at low temperature, the total coliform counts would not increase significantly ($p < 0.001$).

Tables 5.9 and 5.10 show the results for HPC at 37 °C at the Midmar and DV Harris sites. It can be observed that from count model, free and total chlorine are significantly

($p < 0.01$) reducing the HPC at 37 °C at Midmar site. The model further shows that high temperature significantly ($p < 0.01$) increases bacteria growth and this is also supported by previous findings. Time has a negative effect on HPC at 37 °C, that is, the counts are decreasing over time at Midmar and DV Harris sites. The pH has a positive significant ($p < 0.01$) relationship with HPC at 37 °C at DV Harris site. The zero model parts in Table 5.9 and 5.10 shows that temperature and turbidity significantly ($p < 0.001$) affect HPC at 37 °C on both sites.

5.3 Model Comparison

Note that the model output above does not indicate in any way if the zero-inflated and hurdle models are an improvement over a standard Poisson and negative binomial model. This can be determined by running the corresponding standard Poisson/ negative binomial models and then performing a Vuong test of the two models. To compare the performance of each model we may use the Vuong test as the models are non-nested. Tables 5.11 and 5.12 show that the hurdle model is the most superior to the rests. The ZIP and ZINB have almost similar performance and the Vuong test indicated that the two models do not have a significant difference. The reason might be due to the statistical non-significance of the estimate of the dispersion parameter, that is log(theta).

Table 5.11: Vuong test for comparisons of different models at Midmar site.

	ZIP	ZINB	PLH	NBLH	P	NB	Best model
ZIP		0.02^{*}	0.398	0.000^{***}	0.01^{*}	0.021^{*}	Hurdle
ZINB			0.02^{*}	0.000^{***}	0.008^{**}	0.144	Hurdle
HLP				0.009^{**}	0.014^{*}	0.022^{*}	Hurdle
HLNB					0.008^{**}	0.061^{*}	Hurdle
POI						0.009^{**}	NB
NB							

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.12: Vuong test for comparisons of different models at DV Harris site.

	ZIP	ZINB	HLP	HLNB	Poi	NB	Best model
ZIP		0.005**	0.262	0.000***	0.000***	0.011*	Hurdle
ZINB			0.004**	0.000***	0.000***	0.001**	Hurdle
HLP				0.000***	0.000***	0.011*	Hurdle
HLNB					0.000***	0.000***	Hurdle
POI						0.000***	NB
NB							

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

5.3.1 Akaike information criterion (AIC)

In this section we compare all the fitted models for total coliform counts using Akaike information criteria. The AIC values to select the best model that fits the data set are presented in Table 5.13.

Table 5.13: The comparison of different models using Akaike Information Criterion (AIC).

Models	AIC (Midmar)	AIC (DV Harris)
Poisson	4304.23	3828.76
Negative Binomial	646.75	1241.81
Zero-Inflated Poisson	1020.63	1578.91
Zero-Inflated Negative Binomial	624.77	1191.54
Poisson Logit Hurdle	1021.75	1580.68
Negative Binomial Logit Hurdle	**618.33**	**1187.58**

According to the Table 5.13, the Poisson model is not the best performing model because it yields the largest AIC values in both sites. Since the negative binomial model has the small AIC value, one can say that it is the best fitting model for the data set. However, the dispersion parameter for the negative binomial model is $1,0405$ and the dispersion parameter for the Poisson model is $1,201623$ which indicate that the dependent variable (total coliform counts data) is overdispersed. On the other hand, we used the Vuong test to check if the zero-inflated model is better than the Poisson model, and zero-inflated negative binomial model is better than the negative binomial model. For the Poisson part, the computed Vuong test statistic is $V = -6,722759$ (p-value < 0.001) which indicates that zero-inflated Poisson model fits better than the standard Poisson model, and for the negative binomial part, the computed Vuong's test statistic is $V = -1,569072$ (p-value $= 0,0583156$) which indicates that the zero-inflated negative binomial model fits the data better than the standard negative binomial model because we test significance at the 0.05 level. We can also state that zero-inflated Poisson model is better than the standard Poisson model, and that the zero-inflated negative binomial model is better than the negative binomial model. Since the zero-inflated negative binomial model and the hurdle negative binomial model have the closest AIC values, we can say that these models perform best for our data set.

Table 5.14: The zero counts capturing in Midmar site.

Observed	P	NB	ZIP	ZINB	PLH	NBLH
6815	6524	6815	6815	6815	6815	6815

Table 5.15: The zero counts capturing in DV Harris.

Observed	P	NB	ZIP	ZINB	PLH	NBLH
8925	8643	8925	8925	8926	8925	8925

One may conclude that the Poisson model is again not appropriate as it accounted for the least number of zeros compared to the other models. The NBLH, ZINB, PLH, and ZIP models captured 6815 and 8926 zeros which are equal to the observed (Tables 5.14 and 5.15). The modified NB based models (ZINB and NBLH) offered the best fit to zero-inflated microbial data in terms of the AIC (minimum value for all the models fitted).

In summary, the hurdle and zero-inflation models lead to the best results (in terms of likelihood) on this data set. Above, their mean function for the count component of the model was already shown to be similar, below look at the fitted zero components.

Table 5.16: The fitted zero components in Midmar site.

	Intercept	Total chlorine	Free chlorine	Temperature	pH	Turbidity
ZINB	4.53	-2.60	1.57	0.88	-0.37	0.84
Hurdle-NB	-12.92	-1.56	-1.18	0.42	0.25	-0.65

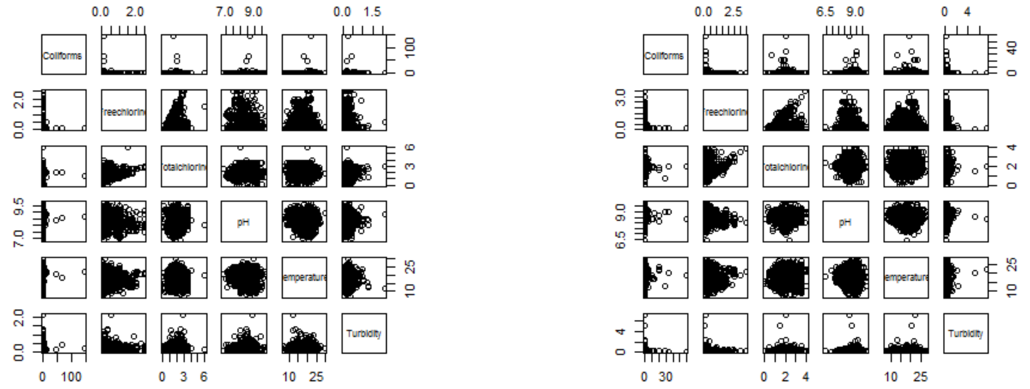
Table 5.17: The fitted zero component in DV Harris site.

	Intercept	Total chlorine	Free chlorine	Temperature	pH	Turbidity
ZINB	24.54	2.27	-0.02	-1.27	-0.27	-2.71
Hurdle-NB	-15.88	-3.54	-0.65	0.76	0.18	0.07

This shows that the absolute values are different, which is not surprising as they pertain to slightly different ways of modeling zero counts but the signs of the coefficients match, i.e., they oppose each other. For the hurdle model, the zero hurdle component describes the probability of observing a positive count whereas, for the ZINB model, the zero-inflation component predicts the probability of observing a zero count from the point mass component. Overall, both models lead to the same qualitative results and very similar model fits. Perhaps the hurdle model is preferable because it has the nicer interpretation: there is one process that controls the non-occurrence of microbiological organisms, and a second process that determines how many microbiological-organisms (positive counts) have been detected/ occurred.

5.4 Model Validation

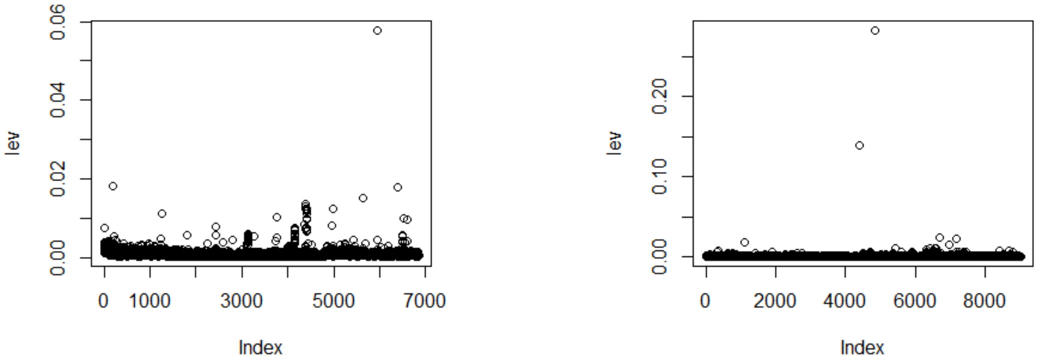
An important part of all regression analyses is to examine residual diagnostics, influential data points, and non-linearity in the predictors (Andersen, 2012). After fitting a regression model it is important to determine whether all the necessary model assumptions are valid before performing inference. If there are any violations, subsequent inferential procedures may be invalid resulting in faulty conclusions. Therefore, it is crucial to perform appropriate model diagnostics.



(a) Midmar site

(b) DV Harris site

Figure 5.1: Correlation between variables.

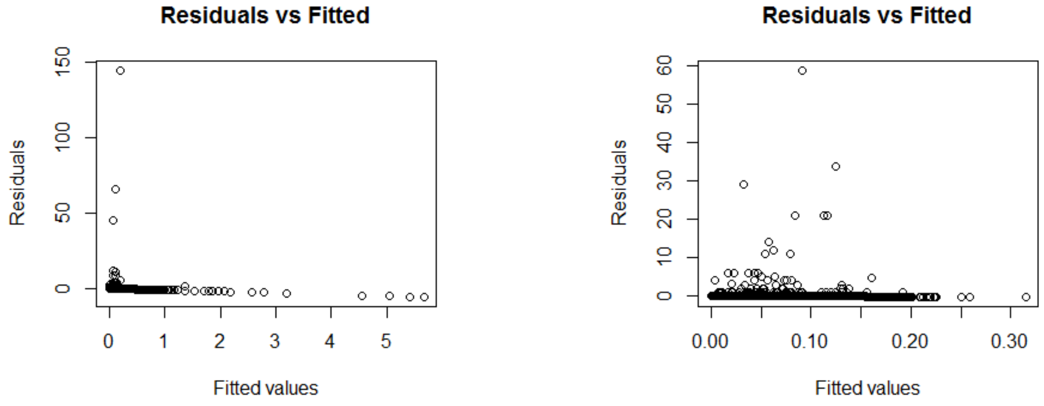


(a) Midmar site

(b) DV Harris site

Figure 5.2: The leverage plots for the model fitted to total coliform counts data.

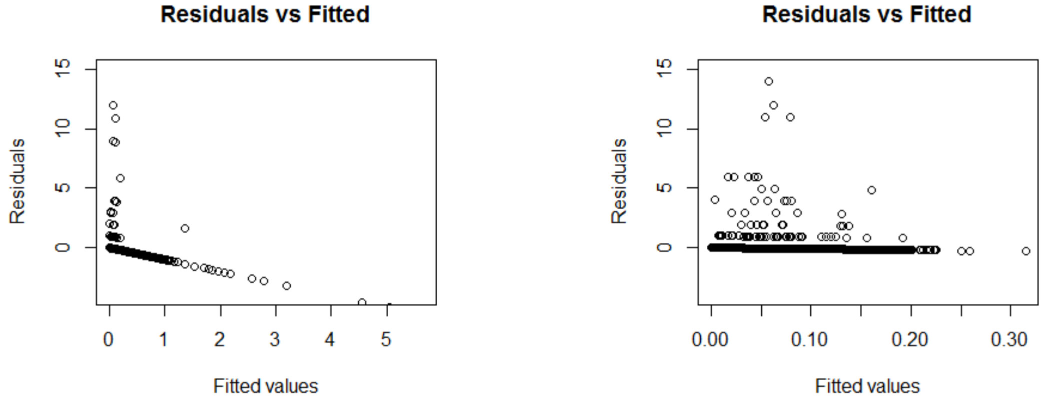
Figure 5.2 helps us to find influential cases (i.e. subjects) if any. Not all outliers are influential in linear regression analysis. Even though data might have extreme values, they might not be influential in determining a regression line. That means, the results would not be much different if we either include or exclude them from analysis. They follow the trend in the majority of cases and they do not really matter; they are not influential. On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values. They could be extreme cases against a regression line and can alter the results if we exclude them from analysis. Another way to put it is that they do not get along with the trend in the majority of the cases. In both Figures 5.3 and 5.4 the we have plots for residuals versus the fitted values ($\hat{y}$) for hurdle model



(a) Midmar site

(b) DV Harris site

Figure 5.3: Residuals versus fitted values for the hurdle model.



(a) Midmar site

(b) DV Harris site

Figure 5.4: Residuals versus fitted values for the hurdle model.

and the scale of the residuals have been decreased to make pattern visible. The typical regression model shows a slight pattern in the results as the line slated down. There is no systematic trends thus the model's predictions are very good. However, the plot of the fitted versus residuals (DV Harris site) seems to have more variation at low-level values compared with the high fitted values.

Chapter 6

Discussions and Conclusions

The primary objective of this study was to find the statistical methods or techniques to model the rare microbiological organisms that exceed the acceptable standards limit at Umgeni Water. Analyzing the distribution of the data set is crucial, especially the microbiological data set which has more zeros than expected. This drives analyst to work with zero modified models such as zero-inflated and hurdle models. In this study, the Poisson and negative binomial models which are traditional methods to analyze count data, the zero-inflated Poisson, zero-inflated negative binomial, hurdle model with hurdle negative binomial model are applied to microbiological data. There are five independent variables, which are free chlorine, total chlorine, temperature, turbidity, pH and time. Plot analysis indicated that the datasets contained a very large proportion of zeros which leads to overdispersion (see Figures 2.1 and 2.2). The dispersion parameter is used to see if there is overdispersion in the data set and the Vuong test is used to compare non-nested models. The AIC and log-likelihood values are used to compare different models.

On average, the count model shows that free and total chlorine reduce the total coliform counts. Turbidity has a negative effect on total coliform counts, that is, as turbidity increases the total coliform counts decreases. Time has a negative effect on total coliform counts and that means the total coliform counts are decreasing over time. The study has demonstrated that temperatures of 15, 25 and 35°C (Figure 2.3c), generally have negative impact on coliform, *E. coli* and HPC at 37 °C bacteria levels, resulting in a decrease in their counts in the water phase. The results indicate that, in the absence of nutrients, high temperature could be an important factor in reducing the survival and growth of bacteria. The zero part model shows that the free and total chlorine are significantly increasing the number of zero counts, meaning very few positive counts are found in water with higher levels of free and total chlorine. The temperature has a negative effect on zeros counts, the positive counts would not achieve any significant growth at low temperature.

From our simulation experiment, we see that the zero-inflated models, ZIP, ZINB, and hurdle, are consistent with the changes of the model parameters. The specification of the correct model is very important. Based on the AIC and Vuong tests, the hurdle model has higher flexibility to fit a model with a mixture of distribution for zeros and positive counts and it performs in a competitive way with ZIP and ZINB. The ZIP model is a very good fit over the standard Poisson model and the ZINB is the better statistical fit compared to the negative binomial model. The zero-inflated binomial model is a better fit over the zero-inflated Poisson model for modeling the microbiological data. The zero-inflated models fit better than their corresponding non-zero inflated counterparts; this suggests the best fitting model needs to account for both overdispersion and zero-inflation in the observed data. Sometimes overdispersion of a data set may not be significant if the percentage of zeros is too high (might be 80% or more) and in such a case, the ZIP and ZINB have nearly identical estimate of the parameters. In most cases, ZIP does not fit the data well if there is overdispersion with a moderate percentage of zeros.

6.1 Future Research

It is suggested that future research considers other proportions of zeros and event stage distributions, underdispersion adjustments, different optimization procedures. This research should also be extended to an approach for direct marginal inference. The aim is to develop a marginalized model for zero-inflated univariate count outcome in the presence of overdispersion.

Bibliography

Andersen, R. (2012). Methods for detecting badly behaved data: Distributions, linear models, and beyond. *American Psychological Association*, 3:5–26.

Atkins, D. C., Baldwin, S. A., Zheng, C., Gallop, R. J., and Neighbors, C. (2013). A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors*, 27(1):166.

Barry, S. C. and Welsh, A. H. (2002). Generalized additive modelling and zero inflated count data. *Ecological Modelling*, 157(2):179–188.

Bermúdez, L. and Karlis, D. (2011). Bayesian multivariate poisson models for insurance ratemaking. *Insurance: Mathematics and Economics*, 48(2):226–236.

Bolker, B. M. (2008). *Ecological models and data in R.* Princeton: Princeton University Press.

Boucher, J. P., Denuit, M., and Guillen, M. (2009). Number of accidents or number of claims an approach with zero-inflated poisson models for panel data. *Journal of Risk and Insurance*, 76(4):821–846.

Cameron, C. and Trivedi, P. (1998). *Models for Count Data*, volume 53. Cambridge: Cambridge University Press.

Dalrymple, M. L., Hudson, I., and Ford, R. P. K. (2003). Finite mixture, zero-inflated poisson and hurdle models with application to sids. *Computational Statistics & Data Analysis*, 41(3):491–504.

Dobbie, M. J. and Welsh, A. H. (2001). Theory & methods: Modelling correlated zero-inflated count data. *Australian & New Zealand Journal of Statistics*, 43(4):431–444.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3):457–483.

Fahrmeir, L. and Osuna Echavarría, L. (2006). Structured additive regression for overdispersed and zero-inflated count data. *Applied Stochastic Models in Business and Industry*, 22(4):351–369.

Greene, W. (2008). Functional forms for the negative binomial model for count data. *Economics Letters*, 99(3):585–590.

Gurmu, S. (1998). Generalized hurdle count data regression models. *Economics Letters*, 58(3):263–268.

Heeringa, S. G., West, B. T., and Berglund, P. A. (2010). *Applied survey data analysis.* Boca Raton: CRC Press.

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, 36(5):531–547.

Hilbe, J. M. (2011). *Negative binomial regression.* Cambridge: Cambridge University Press.

Hilbe, J. M. (2014). *Modeling count data.* Cambridge: Cambridge University Press.

Hua, H., Wan, T., Wenjuan, W., and Paul, C.-C. (2014). Structural zeroes and zero-inflated models. *Shanghai archives of psychiatry*, 26(4):236.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Loeys, T., Moerkerke, B., De Smet, O., and Buysse, A. (2012). The analysis of zero-inflated count data: Beyond zero-inflated poisson regression. *British Journal of Mathematical and Statistical Psychology*, 65(1):163–180.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., and Possingham, H. P. (2005). Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology letters*, 8(11):1235–1246.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. Boca Raton: CRC press.

McCullough, P. and Nelder, J. (1989). Generalized linear models. *Chapman and Hall*, 11(8):35–46.

Mouatassim, Y. and Ezzahid, E. H. (2012). Poisson regression and zero-inflated poisson regression: application to private health insurance data. *European Actuarial Journal*, 2(2):187–204.

Mouatassim, Y., Ezzahid, E. H., and Belasri, Y. (2012). Operational value-at-risk in case of zero-inflated frequency. *International Journal of Economics and Finance*, 4(6):70.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365.

Nelder, J. A. and Baker, R. J. (1972). Generalized linear models. *Encyclopedia of statistical sciences*, 135(3):370–384.

Posada, D. and Crandall, K. A. (1998). Modeltest: testing the model of dna substitution. *Bioinformatics*, 14(9):817–818.

Ridout, M., Demétrio, C. G., and Hinde, J. (1998). Models for count data with many zeros. *Proceedings of the XIXth international biometric conference*, 19(1):179–192.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological bulletin*, 118(2):183.

Saffari, S. E., Adnan, R., and Greene, W. (2012). Hurdle negative binomial regression model with right censored count data. *SORT-Statistics and Operations Research Transactions*, 36(2):181–194.

Sharp, E., Parson, S., and Jefferson, B. (2006). Coagulation of nom: linking character to treatment. *Water Science & Technology*, 53(7):67–76.

Vazquez, A., Bates, D., Rosa, G., Gianola, D., and Weigel, K. (2010). Technical note: an r package for fitting generalized linear mixed models in animal breeding. *Journal of animal science*, 88(2):497–504.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 57(2):307–333.

Walhin, J. F. (2001). Bivariate zip models. *Biometrical Journal*, 43(2):147–160.

Warton, D. I. (2005). Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics*, 16(3):275–289.

Winkelmann, R. (2008). *Econometric analysis of count data*. Verlag: Springer Science & Business Media.

Won, G., Kline, T. R., and LeJeune, J. T. (2013). Spatial-temporal variations of microbial water quality in surface reservoirs and canals used for irrigation. *Agricultural water management*, 116:73–78.

Wu, Z. (2005). Generalized linear models in family studies. *Journal of Marriage and Family*, 67(4):1029–1047.

Yesilova, A., Ozgokce, M. S., and Kaya, Y. (2012). Zero-inflated regression methods for insecticides. *University of Exeter*, 12(2):259–276.

Appendix A

R Syntax

Import Data

```
Midmar <-is data sets from Midmar site
DV Harris <- is data sets from DV Harris site

# import data
Midmar <-read.spss("C:/Users/Zibusiso Hlongwane/Documents/R/Analysis/New Data
/U.sav",use.value.labels = T,to.data.frame = T)
DV Harris <-read.spss("C:/Users/Zibusiso Hlongwane/Documents/R/Analysis
/New Data/TDVHarris.sav",use.value.labels = T,to.data.frame = T)

# Create new data sets without missing data
Midmar <-na.exclude(Midmar)
DV Harris <-na.exclude(DV Harris)



# Descriptive statistics
stargazer(Midmar)
stargazer(DV Harris)


# Frequency table
#Midmar site
A1 <-cbind(Freq=table(Coliforms),Cumul=cumsum(table(Midmar$Coliforms)),
relative=100*prop.table(table(Coliforms)))
A2 <-cbind(Freq=table(\textit{E. coli}),Cumul=cumsum(table(
\textit{E. coli})),relative=100*prop.table(table(\textit{E. coli})))
A3 <-cbind(Freq=table(HPC at 37 $^{\circ}$C),Cumul=cumsum(table(
HPC at 37 $^{\circ}$C)),relative=100*prop.table(table(HPC at 37 $^{\circ}$C)))


# DV Harris site
B1 <-cbind(Freq=table(Coliforms),Cumul=cumsum(table(Coliforms)),
relative=100*prop.table(table(Coliforms)))
B2 <-cbind(Freq=table(\textit{E. coli}),Cumul=cumsum(table
(\textit{E. coli})),relative=100*prop.table(table(\textit{E. coli})))
B3 <-cbind(Freq=table(HPC at 37 $^{\circ}$C),Cumul=cumsum(table
(HPC at 37 $^{\circ}$C)),relative=100*prop.table(table(HPC at 37 $^{\circ}$C)))


# Plots
date1 <-as.Date(Midmar$Date)
date2 <-as.Date(DV Harris$Date)
par(mar=c(5,4,4,5)+.1)
plot(date1,Coliforms,ylim =c(0,150),main="Midmar",col="blue",
ylab="Total coliforms (counts)",xlab="Date",data=Midmar)
par(new=T)
plot(date1,pH,type="l",col="red",xaxt="n",ylim =c(6,10),
yaxt="n",xlab="",ylab="")
```

```
axis(4)

mtext("pH",side=4,line=3)

############################################################################

par(mar=c(5,4,4,5)+.1)

plot(date2,Coliforms,ylim =c(0,150),main="Midmar",col="blue",

 ylab="Total coliforms (counts)",xlab="Date",data=DV Harris)

par(new=T)

plot(date2,pH,type="l",col="red",xaxt="n",ylim =c(6,10),yaxt="n",xlab="",ylab="")

axis(4)

mtext("pH",side=4,line=3)
```

Fit Regression Models

```
/********************* Model Fitting ********************/ \\

# Midmar site

# Standard regression models

summary(m1 <-glm(Colifor0ms~freechlorine+totalchlorine+pH+Temperature+Turbidity+

Time,family="poisson",data=Midmar))

summary(m2 <-glm.nb(Coliforms~freechlorine+tchlorine+pH+Temp+Turb+

Time,data=Midmar))


# Zero modified and hurdle models

summary(m3 <-zeroinfl(Coliforms~freechlorine+totalchlorine+pH+Temperature+Turbidity+

Time,data=Midmar,dist="poisson"))

summary(m4 <-zeroinfl(Coliforms~freechlorine+totalchlorine+pH+Temperature+Turbidity+

Time,data=Midmar,dist="negbin"))

summary(m5 <-hurdle(Coliforms~freechlorine+totalchlorine+pH+Temperature+Turbidity+

Time,data=Midmar,dist="poisson"))
```

```
summary(m6 <-hurdle(Coliforms~freechlorine+totalchlorine+pH+Temperature+Turbidity+
Time,data=Midmar,dist="negbin"))


# DV Harris site
# Standard regression models
summary(m1 <-glm(Coliforms~freechlorine+totalchlorine+pH+Temperature+
Turbidity+Time,family="poisson",data=DV Harris))
summary(m2 <-glm.nb(Coliforms~freechlorine+totalchlorine+pH+Temperature+
Turbidity+Time,data=DV Harris))


# Zero modified and hurdle models
summary(m3 <-zeroinfl(D1$Coliforms~freechlorine+totalchlorine+pH+Temperature+
Turbidity+Time,data=DV Harris,dist="poisson"))
summary(m4 <-zeroinfl(D1$Coliforms~freechlorine+totalchlorine+pH+Temperature+
Turbidity+Time,data=DV Harris,dist="negbin"))
summary(m5 <-hurdle(D1$Coliforms~freechlorine+totalchlorine+pH+Temperature+
Turbidity+Time,data=DV Harris,dist="poisson"))
summary(m6 <-hurdle(D1$Coliforms~freechlorine+totalchlorine+pH+Temperature+
Turbidity+Time,data=DV Harris,dist="negbin"))
```