What factors or farm practices affect the susceptibility of dairy cows to sub-clinical vs clinical mastitis?

Seán McMahon

The thesis is submitted to University College Dublin
in part fulfilment of the requirements for the degree of
MA Statistics

School of Mathematics and Statistics
University College Dublin

Supervisor: Dr. Michelle Carey

14/08/2020

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Dr. Michelle Carey for her help and guidance throughout the year. Her swift responses to emails and questions, as well as our weekly meetings, were greatly appreciated.

I would also like to thank Dr Riccardo Rastelli for reading the draft of my dissertation and providing much appreciated feedback.

Abstract

Mastitis is a priority illness in Ireland. The cost of mastitis is approximately €16 million per annum. Clinical mastitis is defined as an infection that is easily detectable due to physical changes associated with the cow or her milk. Subclinical mastitis is much harder to detect. The difficulty in detecting subclinical mastitis can cause it to exist untreated in a herd for long periods of time, negatively impacting the herds output. It is therefore necessary to obtain a better understanding of what factors affect a cow having subclinical vs clinical mastitis. This study looked at 1991 mastitis infections in Ireland from 2007 and 2018, and used a logistic mixed-effects regression to determine what factors influenced a subclinical versus clinical diagnosis. The study found that: infections in the left hind teat was more likely to be a clinical infection than a left fore infection; that a right fore and left hind teat infection was more likely to be a subclinical infection than a left fore infection; that as the severity of an infection increases, so does the likelihood of the infection being clinical; and that as the time gap between the evening and morning milking increases, and the concentration of feed given increases, the likelihood of an infection being clinical increases. This model had a accuracy of 94.25% in predicting clinical mastitis cases, and an accuracy of 85% in predicting subclinical mastitis cases.

Chapter 1

Introduction

1.1 Mastitis in the Irish Dairy Industry

Mastitis in dairy cattle is an inflammation of the mammary gland caused by pathogenic bacteria entering the sterile environment of the udder tissue (Thompson-Crispi et al. 2014). In addition to being an animal welfare issue due to the pain the condition causes (Leslie & Petersson-Wolfe 2012), it accounts for a considerable economic loss to both farmers (Geary et al. 2012) and the processing industry (Geary et al. 2013) in Ireland. Each individual case of mastitis costs an Irish farmer around 190 euro per annum (Gleeson et al. 2009). This cost is due to infected cattle having a decrease in milk yield (Gröhn et al. 2004), reproduction rates (Kumar et al. 2017), and milk quality (Hortet & Seegers 1998), as well as the treatment costs. Overall, it is estimated that mastitis accounts for a loss of 20% of the total dairy revenue in Ireland (Fitzgerald 2019). The rapid detection and removal of infected cattle from the supply chain is crucial in the reduction of these losses.

The dilemma, which the dairy industry now faces, is how to maximise production without compromising welfare, while reducing the impact of carbon emissions from dairy cattle. An important aspect of achieving these aims is to focus on overall lifetime productivity of cattle, rather than on traditional annual targets. To achieve this aim, a reduction in morbidity and mortality due to preventable and prevalent endemic diseases such as mastitis is required. If this is to become a reality, increased efforts are required in the speed and accuracy of detection of mastitis, better targeting of therapy, and improved management and decision making in prevention of exposure to pathogens and avoidance of significant risk factors for disease (Geary et al. 2012).

Mastitis in dairy cattle is divided into clinical cases, and subclinical cases. Clinical cases are those which are easily detectable by the milker. Infected udders can be red,

swollen, and hard, and the milk of infected cattle can be watery and discoloured, with the presence of clots or flakes (Kamphuis et al. 2010). A subclinical mastitis infection effects 20-50% of cattle in a given herd, shows no visible signs of infection, but somatic cell counts are elevated above 200,000 cells/mL (Forsbäck et al. 2009). Subclinical mastitis is often undetected and has the greatest economic consequences because of long term effects on milk yields. Approximately 25-30% of cows with chronic cases of subclinical mastitis may exhibit clinical symptoms that require antibiotic treatments and withholding of milk (Barlow et al. 2009). The failure to treat subclinical mastitis may result in:

- Chronic infections that are unlikely to respond to antibiotic therapy.
- An increase exposure of healthy cows to contagious pathogens as cows with subclinical mastitis maintain a reservoir of infection within the herd.
- A reduction in milk yield (Hortet & Seegers 1998)
- Cows that maintain subclinical mastitis across the dry period (SCC $\geq$ 200,000 cells per ml at the last test of the completed lactation and first test of the subsequent lactation) have been shown to produce 9.1 kgs (20 lbs) less milk at their first test (Pantoja et al. 2008)

1.2 Mastitis detection

The need for detecting clinical mastitis through the analysis of factors related to milking and farm management has become increasingly significant due to the growing popularity of automatic milking systems globally (Tse et al. 2018). Automatic milking systems require no human contact during the milking process and as such, it is not possible to visually detect the signs of clinical mastitis such as discoloured or clotted milk (Hogeveen & Ouweltjes 2003). As such, other methods are required.

The somatic cell count of milk has a well-established relationship with mastitis (Svensson et al. 2006). A somatic cell count above 200,000 cells per ml is the threshold used to indicate a mastitis infection (Schepers et al. 1997). EU directive 92/46/EEC explicitly states that an average somatic cell count above 400,000 cells per ml over a three month period renders the milk unsuitable for human consumption. Potentially mastitis infected milk is unsuitable for consumption due to its significantly reduced shelf life (Santos et al. 2003), and the potential for anti-microbial resistance due to the high anti-microbial usage associated with the treatment of high somatic cell counts (White & McDermott 2001).

A number of other factors have been shown to indicate clinical mastitis infections. Clinical mastitis occurs more frequently early in lactation (Miltenburg et al. 1996, Svensson et al. 2006). The number of offspring also seems to be relevant, with cattle experiencing their first pregnancy being the least likely to contract clinical mastitis (Barkema et al. 1998). Cattle that have previously had clinical mastitis are also more likely to contract it again (Zadoks et al. 2001). Past studies primarily examined the affect of one particular factor on clinical mastitis such as milk flow (Cavero et al. 2006), although more recent research has looked at a combination of multiple cow factors in one multivariate model (Month of lactation, season of lactation, stomatic cell count, infection location and mastitis pathogen), which found that the the infection rate of clinical mastitis had a large difference between different cows depending on these cow factors (Steeneveld et al. 2008).

Mastitis causes a change in the ion concentration of milk, which leads to an effect on its electrical conductivity (Linzell & Peaker 1971). This change in electrical conductivity has become one of the most widely used and studied mastitis detection methods due to its ease of implementation in parlour systems (Hogeveen et al. 2010). However, electrical conductivity of milk can be affected by other factors such as temperature and fat content, leading to potential false positives in mastitis detection (Nielen et al. 1995). Regardless, the majority of research includes electrical conductivity as a factor (Nielen et al. 1995, deMol et al. 1997, Norberg et al. 2004, Sun et al. 2010).

Dry cow therapy, although initially developed to prevent summer mastitis, is used to remove subclinical mastitis infections (Dingwell et al. 2003). Anti-microbials are more effective through this therapy as stronger doses and a constant concentration is maintained due to the lack of lactation and milking (Gruet et al. 2001). How long a cow has had a high somatic cell count has also been shown to impact the effectiveness of dry cow therapy (Barrett et al. 2006).

1.3 Introduction to the Study

We have information on 1991 cows across 10 farms that were diagnosed with either clinical or subclinical Mastitis between 2007 and 2018. The dataset contains observations on 56 different factors related to physiological, and environmental factors related to the cow, lactation stage, bacteria in the udder, milk quality and quantity etc.. A major component of this research will be determining which of these 56 measured factors have a significant effect on the probability of a cow having clinical vs subclinical mastitis. Another challenge is the lack of independence in the data, the same cow can become infected multiple times and therefore appear in the dataset repeatedly. This lack of independence makes the data

unsuitable for a number of traditional modelling techniques such as linear and logistic regression, which both require independence (Yan & Su 2009, Agresti 2013). Mixed effects logistic regression will be used to model the binary outcome variable, subclinical mastitis (1) clinical mastitis (0), in which the log odds of the outcomes are modelled as a linear combination of the predictor variables when observation are not independent and can be grouped/clustered by individual.

The goal of this research is to not only to identify the factors that affect subclinical vs clinical mastitis infections, which can aid better targeting of therapy, improve management and decision making in the prevention of exposure to pathogens. But to also optimise a model for predicting the probability of a cow developing subclinical mastitis, which can be utilized in future research to develop a user-friendly shiny R app which allows farmers to enter new information based on their own observations, and receive a probability score indicating the probability of the cow developing subclinical mastitis.

The first study carried out that applied mixed effects logistic models to mastitis related data, aimed at predicting if a dairy cow had chronic mastitis based on cow factors such as SCC count (Kristulaa et al. 1992), failed at accurately classifying cows with chronic mastitis. A later study aimed to assess poisson, logistic, and linear mixed effects models, and see which were the most accurate in their classification of clinical mastitis in Norwegian dairy cattle (Vazquez et al. 2009). It was found that mixed effect logistic regression had a better predictive ability than the other model types. In this study the random effect was the sire of the dairy cow.

Subclinical mastitis prediction using mixed effects logistic models has been examined in sheep (Vasileiou et al. 2018). The primary factor this study found to be associated with a subclinical infection was the management system employed by the farmer. In this study the random effect was the flock of sheep.

This study will be the first to examine the factors relating to a subclinical versus clinical diagnosis of mastitis in dairy cattle. Past studies have used mixed effects logitic models to predict clinical or subclinical infections, but this will be the first study examining the factors relating to a subclinical versus a clinical mastitis diagnosis. It will also be the first mixed effects logistic model based study to use a cow's identification number as the random effect, encompassing a large dataset with multiple years and infections.

This research has multiple benefits to its completion. Findings will benefit:

- Irish farmers. Understanding what drives the differences between clinical and subclinical mastitis will lead to a reduction in the overall number of subclinical mastitis

cases in Ireland. As the detection and treatment of subclinical mastitis cases are a common financial burden for farmers, reducing their frequency will increase agricultrual revenue.

- The Irish government. The potential reduction in subclinical mastitis infections in Ireland would not only increase agricultural tax revenue received, but would also reduce the need for mastitis relief and support from the department of agriculture and semi-state bodies such as Teagasc. This would reduce their workload and cut expenditure.

- The Irish public. The potential reduction in subclinical mastitis infections will reduce the Irish public's exposure to veterinary antibiotic residue from milk. This will lower the frequency of health issues such as allergic reactions with penicillins and cephalsporins, and the carcinogenic effect associated with sulfamethazine and nitrofurazone.

The remainder of this study is organised as follows. Chapter 2 will discuss classification and provide an overview of logistic regression. Chapter 3 will provide an overview of mixed effects regression. Chapter 4 will discuss the exploratory data analysis carried out on the dataset and provide graphical and numerical summaries of each variable. Chapter 5 will discuss the creation of, and provide the results of the logistic regression. Chapter 6 will discuss the creation of, and provide the results of the mixed effects logistic regression. Chapter 7 will discuss and interpret the results of the mixed effects logistic regression. A full copy of the code used in this study is available in the appendix.

Chapter 2

Classification and Logistic Regression

2.1 Introduction to classification

When a response variable is categorical, the act of predicting which of the categories an individual observation will fall into is known as classification. There are many different techniques available to choose from when carrying out classification, with these techniques being known as classifiers. Three of the most widely-used classifiers are logistic regression, linear discriminant analysis, and K-nearest neighbors. Classifiers, when given a feature vector X, containing values $(X_1, X_2, \ldots X_n)$, and a categorical response Y with values in the set C, creates a function $C(X)$ which uses X to predict the value of Y, with error ϵ.

$$Y = C(X) + \epsilon$$

For example, when predicting if an email is spam or not, a classifier takes X, which may have values for length of the email, the number of "red-flagged" words contained within, and whether or not the recipient has been in contact with the sender before, and uses those to predict Y which is one of the two values of C, spam or not spam.

A linear regression may appear to be a suitable technique to use as a classifier, categorical response variables could be encoded with numerical values for the calculations, for example, 0 for not spam and 1 for spam. An issue arises however when there is no applicable ordering to the responses. When predicting someone's country of birth based on their genetic sequencing, encoding the responses with numbering such as 1 for Ireland, 2 for China, and 3 for Portugal, implies that there is some form of ordering on the outcomes. The implication that there is a measurable difference between being born in each country, and that the difference between being born in China and Ireland is the same as

in China and Portugal, is arbitrary in nature and based purely on what way the classes happened to be ordered when numbered. These encodings could have just as validly been applied as 1 for China, 2 for Portugal, and 3 for Ireland, which would cause the linear regression to imply a different relationship between the classes. "In general there is no natural way to convert a qualitative response variable with more than two levels into a quantitative response that is ready for linear regression" (James et al. 2013).

For binary classes, such as the previously mentioned example of 0 for not spam, and 1 for spam, attempting to use a linear regression as a classifier is more suitable. Each observation would be classified as spam, for example, when $Y > 0.5$. The response variable in this case could be viewed as the conditional probability

$$P(Spam|X)$$

Which can be interpreted as the probability that an observation is a spam email given its feature vector X. There is in issue in the fact that these predicted values from the linear regression can potentially take on values that are below 0 or above 1. This is an impossibility if we are interpreting the response variable as a probability, as probabilities must always fall within the $[0, 1]$ interval.

Therefore, more suitable classification methods are required to make accurate predictions of categorical response variables.

2.2 Logistic Regression

Logistic regression was developed in the 19^{th} century as a way to model the growth of populations. The Belgian astronomer Alphonse Quetelet, found that the currently used model of exponential growth, leads to impossible values as time increases, he asked his pupil Pierre-François Verhulst to look into the problem. Verhulst's findings were published in three papers between 1838 and 1847 (Verhulst 1838, 1845, 1847), in which he revealed a version of the logistic function

$$P(t) = \frac{P(0)e^{rt}}{1 + P(0)(e^{rt} - 1)/K}$$

Where $P(t)$ denotes the population at time t, P(0) is the starting population, and is the population limit. This new model was shown to be very accurate at predicting the actual observed course of population in France, Belgium and Russia at the time. This model was independently rediscovered in 1920 by two American statisticians Raymond Pearl and Lowell J. Reed (Pearl & Reed 1920), who were unaware of Verhulst's findings.

It was this later discovery that popularised the logistic regression and led to the invention of the logit model.

In a logistic regression instead of calculating the probability linearly with

$$P(Y|X) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$$

It uses the logistic function

$$P(Y|X) = \frac{\exp(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n)}{1 + \exp(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n)}$$

The logistic function is the exponential of the linear model, divided by that same number plus one. As the exponential of any number is always positive, and the denominator is always bigger than the numerator due to it having 1 added to it, this function will always return a value between 0 and 1. This feature of the logistic function, always returning values within the $[0, 1]$ interval, makes it suitable for probabilities.

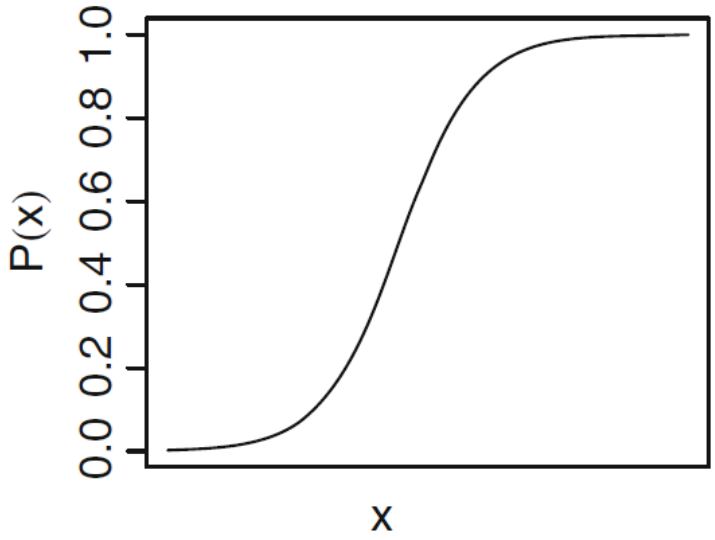


Figure 2.1: The response curve of a logistic regression.

Figure (2.1) shows that in the logistic function, the probabilities, the Y-axis values, are always between 0 and 1. For each variable of the continuous variable X we have a corresponding probability Y.

To determine the odds of a specific outcome Y given the feature/category X, the above logistic function can be rearranged to

$$\frac{p(Y|X)}{1 - p(Y|X)} = \exp(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n)$$

This value, the odds, which can take on any value between 0, which would indicate a low probability of observing outcome Y, and ∞, which would indicate a high probability of observing outcome Y. If a further transformation is carried out, taking the logarithm of both sides,

$$\log(\frac{p(Y|X)}{1-p(Y|X)}) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$$

a model is formed which shows that the log odds is linearly related X. This is referred to as the log-odds or logit transformation. This logit transformation forms the basis for regression modeling, and the origianl logistic function is used to express the outcome in its original scale.

The interpretation of a logistic model with different variable types is as follows:

- Numerical variable: Increasing the value of feature x_j by one unit, will change the estimated odds by a factor of $\exp(\beta_j)$
- Categorical variable: One of the levels is treated as the reference level. Changing the feature x_j from the reference level to the other level changes the estimated odds by a factor of $\exp(\beta_j)$.
- Intercept (β_0: When all numerical variables are zero and the categorical variables are at the reference level, the estimated odds are $\exp(\beta_0)$.

The logistic regression assumes that there isn't any multicollinearity between the independent variables, that the observations are independent, and that the independent variables are linearly related to the log odds. Additionally, a binary logistic regression assumes that the response variable is binary in nature, and a ordinal multiclass logistic regression assumes the response is ordinal.

2.3 Parameter Estimation

The parameters for a logistic regression, are estimated using maximum likelihood. Let θ be a vector of length n containing the parameters $\beta_0, \beta_1, \ldots, \beta_n$. The maximum likelihood finds the set of values in θ for which the probability of the observed data is greatest. The maximum likelihood equation is a derivation of the probability distribution, when there are two possible classes, the binomial distribution is used. This has a joint probability density function of

$$f(y|\beta) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i - y_i)!} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

Where π_i is the probability of success for any given observation in the ith population. In the above probability density function, the values for β are fixed and known, with y as a function of β. The likelihood function is the reverse of this, with the values of y being fixed and known, and β being expressed as a function of these values of y

$$Ł(\beta|y) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i - y_i)!} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

The values of β that are used in the logistic regression are the values that maximise the above function. The maximum of a function occurs when its first derivative equals 0, and the second derivative evaluated at the same point is less than zero. To find the maximum of the above likelihood function its first and second derivative is therefore required, however this is very difficult in its present state.

To simplify the likelihood function for derivation:

- The factorial terms that do not contain π_i are treated as constants and ignored
- a^{b-c} is rewritten as a^b/a^c

From these simplifications we can rewrite the likelihood as

$$Ł(\beta|y) = \prod_{i=1}^{N} (\frac{\pi_i}{1 - \pi_i})^{y_i} (1 - \pi_i)^{n_i}$$

By taking the odds function discussed earlier

$$\frac{\pi_i}{1 - \pi_i} = \exp \sum_{k=0}^{K} X_{ik} \beta_k$$

and solving for π_i, we get

$$\pi_i = \left(\frac{\exp \sum_{n=0}^{N} X_{in} \beta_n}{1 + \exp \sum_{n=0}^{K} X_{in} \beta_n} \right)$$

By substituting in the solved π_i and the odds function into the simplified likelihood we obtain

$$Ł(\beta|y) = \prod_{i=1}^{N} (\exp \sum_{n=0}^{N} X_{in} \beta_n)^{y_i} \left(1 - \frac{\exp \sum_{n=0}^{N} X_{in} \beta_n}{1 + \exp \sum_{n=0}^{K} X_{in} \beta_n} \right)^{ni}$$

This is further simplified to

$$Ł(\beta|y) = \prod_{i=1}^{N}(\exp^{y_i}\sum_{n=0}^{N} X_{in}\beta_n)(1+\exp\sum_{n=0}^{N} x_{in}\beta_n)^{-n_i}$$

Taking the natural log we arrive at the log likelihood function

$$Ł(\beta) = \sum_{i=1}^{N} y_i(\sum_{n=0}^{N} X_{in}\beta_n) - n_i\log(1+\exp\sum_{n=0}^{N} x_{ik}\beta_k) \quad \textbf{(1)}$$

This simplified form can now easily be transformed to the first and second derivatives to obtain the maximum likelihood. The first and second derivatives are

$$\frac{\partial L(\beta)}{\partial beta_k} = \sum_{i=1}^{N} y_i x_{in} - n_i\pi_i x_{in}$$
$$\frac{\partial^2 l(\beta)}{\partial beta_k \partial\beta'_k} = -\sum_{i=1}^{N} n_i x_{in}\pi_i(1-\pi_i)x_{in'}$$

A gradient descent algorithm can be used to minimise equation **(1)** with respect to β , which is based on the Newton-Rhapson method.

2.4 Logistic Model Diagnostics

To assess the quality of a logistic model's fit, the deviance is examined. Deviance is the comparison between the log-likelihood of the fitted model and the log-likelihood of a perfect or saturated model, in which every predicted value is the exact same as its corresponding real value. The smaller the deviance, the closer the fitted value is to the saturated model. A model, known as a null model, where only the intercept term is used for prediction is also used in determining a model's fit. This model can be viewed as the worst possible version for the data provided, as no variables are considered. The difference between the saturated model and the null model is known as the null deviance. By examining at the null deviance and deviance, we can determine just how much more accurate a model becomes when we add in the variables, and if our created model is significantly better at prediction.

McFadden's Pseudo R^2, defined as

$$R^2 = 1 - \frac{\log(L_{model})}{\log(L_{null})}$$

where L_{model} is the likelihood of the fitted model and L_{null} is the likelihood of the null

model, will be used as a measure of the quality of the models fit, with higher values closer to one implying a stronger fit.

To test if the fitted model is significantly more accurate than the null model, a deviance goodness of fit test is used. A Chi-Square test is preformed

$\chi_q^2 = -2[\log(L_{null}) - \log(L_{model})]$

with the following hypothesis

$H_0 : \beta_1 = \beta_2 = ... = 0$
H_a : At least one β_j is not equal to 0

This is checked against a chi-squared distribution with q degrees of freedom. Large chi-square test values lead to small p-values and provide evidence against the intercept-only model in favour of the alternative model. If the corresponding p-value is less than 0.05 then H_0 is rejected in favour of H_a.

To test the significance of each individual variable, Wald χ^2 statistics are used.

$z^2 = (\beta_j)/SE(\beta_j))$

where $SE(\beta_j)$ is the standard error of the β_j parameter, with the following hypothesis

$H_0 : \beta_1, \ldots, \beta_n = 0$
$H_a : \beta_1, \ldots, \beta_n \neq 0$

The Wald statistic is then compared with a χ^2 distribution with 1 degree of freedom. If the value of z^2 (Wald statistic) is bigger than 3.84 and the corresponding p-value is below 0.05 we reject H0 in favour of Ha and conclude that there is a relationship between the log odds and X_j.

To test that the logistic model does not violate the assumption of linearity between the logit transformed response variable and the independent variables, a residuals vs. fitted scatterplot is created. There should be no non linear or variable trendline visible on the scatterplot. If a nonlinear relationship is found, it implies that any inference gained from the model is flawed as a key assumption has been violated.

Logistic models have a key assumption that each observation must be independent

of one another. That is, the values of one particular observation has no impact on the values recorded in another different observation. A violation of independece is usually detected by the presence of autocorrelation, that is, a correlation between values within the same variable across observations. To test that the logistic model does not violate the assumption of independence, a lagged plot is created, which detects negative and positive autocorrelation for a given lag value. Autocorrelation can also be assessed through a Durbin–Watson test. If e_n is taken as the deviance residuals, calculated by

$$\pm\{-2[y_i \log(\hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)]\}^{1/2}$$

where the sign is positive if $y_i \geq \hat{\pi}_i$, and negative otherwise, then Durbin-Watson test can be preformed via

$$d = \frac{\sum_{n=2}^{N}(e_n - e_{n-1})^2}{\sum_{n=2}^{N} e_n^2}$$

where N is the number of observations. A value of 2 on the Durbin–Watson test indicates no autocorrelation, values below 2 indicate a positive autocorrelation, and values above 2 indicate a negative autocorrelation. A value below 2 can be taken as an indication that some observations within the dataset are not independent of eachother, and are taken up of measurements of the same individual at different times.

Overdispersion occurs when there are discrepancies between the observed responses y_i and their predicted values $\hat{\mu}_i = n_i\hat{\pi}_i$ and these values are larger than what the binomial model would predict. It causes the standard errors, test statistics and overall model fit to be inaccurate. If

$$y_i \sim B(n_i, \pi_i)$$

The mean is $\mu_i = n_i\hat{\pi}_i$ and variance is $\mu_i(n_i - \mu_i)/n_i$. Overdispersion is when the data shows evidence of the variance of y_i being above $\mu_i(n_i - \mu_i)/n_i$.

Multicollinearity occurs when two or more independent variables are highly correlated with one another, that is, they describe very similar information. The presence of multicollinearity violates one of the assumptions of logistic regression as it reduces the precision of estimates, making the effect of highly correlated variables on the response hard to determine. To test for the presence of multicollinearity, a correlation matrix of the numerical response variables can be assessed for high positive or negative correlations. The variance inflation factor (VIF) of each variable can also be examined

$$VIF_i = \frac{1}{1 - R_i^2}$$

where R_i is the coefficient of determination from a regression where variable i was the response variable and was predicted by all other response variable. A value of 1 indicates a complete absence of multicollinearity. Values of 10 are considered the absolute maximum to be accepted in a model without interactions (Hair 2006)

Outliers are identified via a Bonferroni Outlier Test through the car package (Fox & Weisberg 2019). This provides a p-value for the largest absolute studentized residual, using the standard-normal distribution, with degrees of freedom one less than the residual degrees of freedom for the model. Observations shown to be significant outliers are removed from the dataset

Chapter 3

Logistic Mixed-Effects Regression

A key requirement of logistic regression is that all the observations are completely independent of one another and are on the same hierarchical level. That each individual observation has no relationship or impact on the values of another observation. This is also known as being a fixed-effects model. It is possible however, for data to not be independent. Data can be clustered in the sense that multiple data points can belong to the same individual. If a dataset measured infections in cows, and cows could have repeated infections recorded, the identification of each individual cow would be considered a level. Each individual cow is considered a level and not a predictor variable because they are randomly sampled from a population, and the identity of each cow has no intrinsic meaning. This dataset could be viewed as containing N participants (the number of infections), that all fall within one of K clusters (the number of cows studied).

It would be incorrect to simply ignore the nature of this dataset and assume that observations that come from the same cow are as different from one another as observations that come from different cows. It is very likely that certain traits will be shared between observations of the same cow. Observations nested in the same cluster (in this example, cow) are more likely to function in the same way than observations nested in different clusters. This correlation among particular observations must be accounted for if we are to receive meaningful results.

In 1918, to account for the correlations between relatives in a genelogical study, Ronald Fisher introduced the random effects model (Fisher 1918). A random effects model is a statistical model where the model parameters are random variables. It is assumed that some type of relationship exists between some of the observations. Logistic linear mixed models (MacCullagh & Nelder 1989, Breslow & Clayton 1993*a*, McCulloch & Searle 2001) were created as an extension to logistic linear models that included these random effects. This can be viewed as a logistic regression where intercepts are allowed to vary among clusters.

Using logistic mixed-effects regression in this way, implies that log-odds of the outcome variable equaling one instead of zero can vary between clusters. In this example, the log-odds of an observation being classified as having a specific type of infection can vary between the different cows examined. A simplified version of the model with only intercept terms highlights the effect.

$$P(Y_{ij} = 1|u_i) = \frac{\exp(\beta_0 + u_{ij})}{1 + \exp(\beta_0 + u_{ij})}$$

It can bee seen that the probability of observation i being the value of 1, when it belongs to cluster j, is based on the shared β_0 coefficient, the intercept that all clusters share, and additionally a unique value for that particular cluster u_{ij}. β_0 is known as the fixed intercept, and corresponds to the overall probability across all clusters. u_{ij} is known as the random intercept variance, and measures the deviation between that particular cluster's probability from the overall fixed intercept β_0. It is typically assumed that u_{ij} follows a normal distribution with mean 0 and variance σ_n^2
The mean of these deviances are assumed to be zero. The higher the random intercept variance, the greater the variation between the probability of different clusters being classified as 1. In the above example, this would mean that certain cows are more prone to a certain infection type than others.

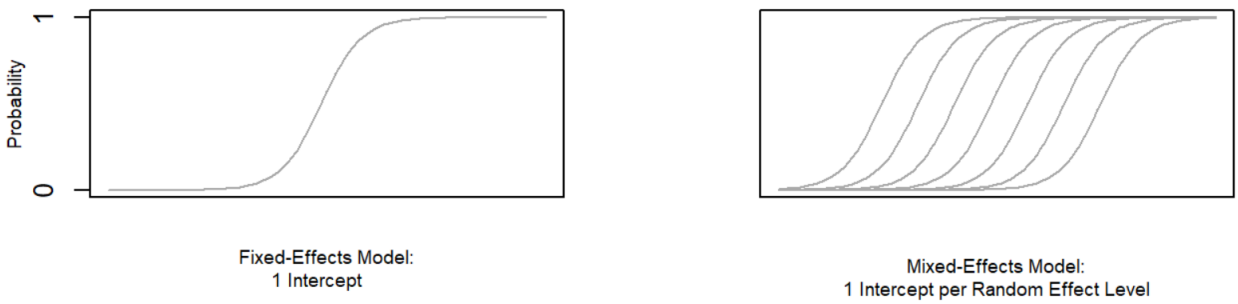


Figure 3.1: A comparison between the logistic curves of a fixed effects model (Left) and a mixed-effects model (Right), with random intercepts

Figure 3.1 shows how in mixed-effects logistic regression, each cluster receiving its own random intercept causes the sigmoid function to shift in position. Each corresponds to a different cluster. Note that the shape remains constant in each cluster, only its position changes. This is because the slope has remained constant. Mixed-effects logistic regression can also account for a random slope based on each cluster, but this is beyond the

scope of this study.

3.1 Parameter estimation

The conditional probability of y_{ij} being 1, which is the i^{th} outcome within the j^{th} cluster, can be expressed as

$$p(y_{ij} = 1|u_{ij}) = \frac{\exp(x'_{ij}\beta + u_j)}{\exp(1 + x'_{ij}\beta + u_j)}$$

$$p(y_{ij} = 0|u_{ij}) = \frac{1}{\exp(1 + x'_{ij}\beta + u_j)}$$

Where x_{ij} is a vector of explanatory variables associated with a vector of fixed-effects parameters β, and u_j is a k-dimensional vector of random effects for the j^{th} cluster. It is assumed that the observations within cluster j are independent given the random effects u_j. The conditional probability of the response vector $y_i = (y_{i1}, ...y_{in})^T$ can be given with

$$g(y_i|u_j; \beta) = \prod_{i=1}^{n_j} p(y_{ij} = 1|u_j)^{y_{ij}} p(y_{ij} = 0|u_j)^{1-y_{ij}}$$

Averaging over the distribution of u_i gives the marginal distribution

$$L_i(\beta, \theta) = \int g(y_i|u_i; \beta) f(u_i; \theta) du_i$$

Where f is the probability distribution for the random effect, a Gaussian distribution. The parameters for logistic mixed effects models are not easily estimated using the maximum likelihood as with logistic regression, this is because the addition of a random effect leads to a likelihood function containing an integral as shown above. This generally does not have an analytical solution (Tapia et al. 2018).

This likelihood function needs to be evaluated numerically or approximated. There have been a number of different methods proposed such as using pseudo-likelihood (Wolfinger & O'Connell 1993) and a penalized quasi-likelihood (Breslow & Clayton 1993*b*). The method employed in this study is Gauss–Hermite quadrature approximation of the marginal likelihood function. In this, the integral is resolved using direct numerical evaluation. The above marginal distribution is rewritten as

$$L_i(\beta, \theta) = \int g(y_i|v_i; \beta, \Gamma) \phi(v_i) dv_i$$

where $u_i = \Gamma v_i, \Gamma\Gamma' = \sum(\theta)$, and v_i has the standard normal density $\phi(v_i)$. The approximated marginal likelihood is then

$L_i(\beta, \theta) \approx \sum_q g(y_i|b_q; \beta, \Gamma)w(b_q)$

where b_q is a vector of quadrature points having the same dimension as u_i and $w(b_g)$ its related weight. The remaining β coefficients are obtained with this marginal likelihood as described in Chapter 2

The random intercept u_j can be interpreted as the effect of being in group j on the log-odds that y =1

3.2 Model Diagnostics

To compare the results of the logistic mixed-effects model to those of the logistic model, the AIC obtained from each will be examined to see if the mixed-effects model has a lower score. If the AIC score of the mixed-effects model is lower than the score for the standard logistic model, it will indicate that the inclusion of random intercepts was justified as it provides better prediction accuracy. To further justify the inclusion of random effects, a test as to whether the mixed-effects model explains significantly more variance will be preformed. This will be done by applying a Model Likelihood Ratio Test to the fixed and the mixed effects models. This will provide a p-value indicating the significance of the addition of random effects. The McFadden's Pseudo R^2 will be calculated using the entire model and using only the fixed effects. This will create another way to see how much the inclusion of random effects increases the model quality.

As with logistic regression, individual fixed effects can have their parameters examined for significance through Wald χ^2 statistics as described in chapter 2. Additionally multi-collinearity can be examined through the variance inflation factor as in logistic regression.

A Wald test will also be carried out on the random-effect intercept to ensure that it is significant. The hypothesis of this test is

$H_0 : \sigma_u^2 = 0$
$H_a : \sigma_u^2 > 0$

The formula used is

$$\text{Wald Statistic} = \left(\frac{\sigma_u^2}{se(\sigma_u)^2}\right)^2$$

Compared with a χ^2 distribution with 1 degree of freedom, we receive a p-value. This value is divided by 2 because the alternative hypothesis is one sided. A value below 0.05 means we reject the null hypothesis in favour of stating that σ_u^2 does not equal 0.

Residual vs fitted and normal Q-Q plots will be created to assess for unusual patterns and non-normality. Additionally, the prediction accuracy of the model will be assessed by forming a ROC (receiver operating characteristic) curve, which plots 1 - specificity on the x-axis and sensitivity on the y-axis at different candidate threshold values between 0 and 1 (Zweig & Campbell 1993). Sensitivity is defined as the proportion of observations with the target condition and were predicted to have positive results. The specificity is defined as the proportion of observations without the target condition and were predicted to have negative results. The formula for these two measurements can be expressed as

$$\text{Sensitivity} = \frac{True\ positives}{True\ positives\ +\ False\ positives}$$
$$\text{Specificity} = \frac{True\ negatives}{True\ negatives\ +\ False\ negatives}$$

The area under this curve is used as a measurement of the prediction accuracy, with 1 indicating a perfect predicting model, and 0 indicating a model that incorrectly predicted every single observation. The optimal threshold τ will then be calculated by maximising the sum of the sensitivity and specificity. If a predicted value exceeds this value, it will be classified as a subclinical mastitis infection, if it is under this value, it will be classified as a clinical mastitis infection.

Chapter 4

Exploratory Data Analysis

The dataset examined is composed of milking parlour information gathered in the Republic of Ireland from 1991 cows across 10 farms that were diagnosed with clinical or subclinical mastitis between 2007 and 2018. A cow was classified as having clinical mastitis when the milk appeared abnormal with the presence of flakes, clots, or strings. The mammary gland also may be warm or hard to the touch and may exhibit increased sensitivity. A cow was classified as having subclinical mastitis when both milk and mammary gland appear normal but SCC was elevated to a level above 200,000 cells/mL. The dataset contains observations on 56 different factors related to physiological, and environmental factors related to the cow, lactation stage, bacteria in the udder, milk quality and quantity etc. Cows are able to contract multiple mastitis infections, and as such, can be entered multiple times into the dataset as individual observations. The total number of infections recorded is 1991, 1205 clinical and 786 subclinical.

In the subsequent section the variables are examined to determine their suitability for the study and examine their relationship with a clinical versus subclinical diagnosis.

4.1 Infection and Treatment Factors

Variables were examined that were based on the severity and location of the mastitis infection as well as the treatments used to combat it. It was hypothesised that since a clinical diagnosis can be viewed as a further developed and therefore stronger infection when compared with subclinical mastitis, variables such as teat and severity may indicate a clinical infection when more teats are infected or the severity of the mastitis is higher. All variables under this category are categorical in nature

Teat

The variable Teat is based on the location of the mastitis infection on the udder. The

possible values this variable can take consist of either the left or right hind udder, the left or right fore udder, or any combination of the four. For example, a cow may have had a subclinical mastitis infection in both her right fore udder and her left hind udder. This would be recorded in the Teat variable as "RF_LH".

There were 7 observations that did not have recorded data for Teat, these incomplete observations were removed from the dataset. One potential outcome, an infection in the right and left fore udder and the left hind udder ("RF_LF_LH") had only 8 observations. This was determined to not be enough to allow for a meaningful contribution to the model, and would instead weaken it due to potentially random variations being interpreted as significant due to the small sample size. This was repeated for categorical variables throughout the dataset, any levels that had less than 10 observations were removed from the dataset.

The frequency of the levels of this variable is as follows

Table 4.1: Clinical and subclinical mastitis infections for each Teat level

Teat	Clinical	Subclinical	Total
LF	224	95	319
LH	334	195	529
RF	254	133	367
RH	320	181	501
LF_LH	7	24	31
RF_LF	8	13	21
RF_LH	10	18	28
RF_RH	10	34	44
RH_LH	17	30	47
RH_LF	2	21	23
RF_RH_LF	3	12	15
RF_RH_LH	5	11	16
RH_LF_LH	1	9	10
RF_RH_LF_LH	3	22	25

Table (4.1) highlights that mastitis infections occurred more frequently in a single udder than in grouped infections. Hind udders are also shown to have become infected more frequently than fore udders. Clinical mastitis infections occurred more often than subclinical infections when only one teat was infected. In contrast, once more than one teat was infected, it was more often subclinical mastitis rather than clinical mastitis.

Drug

Antibiotic therapy via intra-mammary injection is applied at the start of a cows dry period. It is given to deal with any ongoing intra-mammary infections contracted during lactation as well as to provide a level of resistance to further infections during the dry period. Drug is a categorical variable based on the type of antibiotic the cow was placed on. Of the 19 possible drug types, 8 were only found in less than 10 observations each, and as such these observations were removed from the dataset, and the remaining 12 were examined.

Table 4.2: Clinical and subclinical mastitis infections for each Drug type

Drug type	Clinical	Subclinical	Total
BOVACLOX	5	10	15
CEPRAVIN	13	0	13
KANACEF M.C	21	21	42
MILKING TUBE - TERREXINE	6	8	14
MULTIMAST	66	43	109
PATHOCEF	14	46	60
SYNULOX	57	49	106
SYNULOX TUBES	21	4	25
TEREXINE	475	229	704
TERREXINE	409	289	698
TETRA DETLA	90	79	169

Table (4.2) shows that Terrexine and Terexine are the two most frequently used drugs, with double the amount of infections for clinical than subclinical. PATHOCEF was found in subclinical cows more than twice as often than in clinical cows, indicating that the antibiotic may prevent the infection from becoming widespread enough to cause a clinical infection.

Severity

Severity is a measurement of the severity of the mastitis infection, from 1 being the weakest level of infection to 3 being the strongest. 261 observations did not have their infection severity measured and as such were not included in the study. The frequency of the different levels of infection severity among clinical and subclinical infections is as follows

Table (4.3) shows that the most common infection severity for both clinical and subclinical was level 2. The number of cows experiencing an infection of severity 1 is the same for both clinical and subclinical, but clinical infections are more frequent at higher severity levels.

Table 4.3: Clinical and subclinical mastitis infections for each severity level

Severity	Clinical	Subclinical	Total
1	164	164	328
2	808	361	1169
3	170	63	233

Tubes

The variable Tubes signifies the number of antibiotic drying off tubes given to a cow, ranging from 1 to 9. There were 212 observations without information on the number of tubes. These incomplete observations were not considered. Additionally, the application of 7 tubes occurred infrequently, 9 times total throughout all observations. These values were also removed from the dataset. The frequency of the number of tubes used is as follows

Table 4.4: Clinical and subclinical mastitis infection frequency for the number of tubes used

Number of Tubes	Clinical	Subclinical	Total
1	163	75	238
2	377	220	597
3	367	159	526
4	129	92	221
5	32	5	37
6	52	61	113
8	12	10	22
9	7	9	16

Table (4.4) shows that the majority of the observations had between 2 and 3 tubes used, and few had a high number of tubes used. Clinical cases were more likely to have 5 tubes used than subclinical cases. There is a noticeable difference between the number of clinical and subclinical infections when 1 and 3 tubes are used, with far more of the infections being clinical.

Treatment

The variable Treatment signifies the various different treatment types that a cow could be on at the time of infection. Of the 56 different treatment types 43 of them are present in less than 10 different observations individually. Additionally, 387 observations did not have their treatment type recorded at all. These observations were removed from the dataset. The frequency of the clinical and subclinical mastitis for each of the remaining treatment type is as follows

From table 4.5, it can be seen that treatment 1, 2, and 3 are the most frequently used variations. No treatment type was more frequently found in subclinical infected cows than

Table 4.5: Clinical and subclinical mastitis infection frequency for each treatment type

Treatment Type	Clinical	Subclinical	Total
0	16	2	18
1	294	147	411
2	215	169	384
3	173	134	307
4	82	43	125
5	53	29	82
8	12	2	14
10	10	0	10
C150	14	5	19
G250	8	4	12
L602	6	4	12
TMR	15	4	19

clinical. There is a noticeable difference between clinical and subclinical infections when treatment 4 is used, with far more of the infections being clinical than subclinical.

Subtreatment

Subtreatment indicates what subtreatment that cow was on at the time of infection. Of the 16 possible subtreatments, 9 of them are present in less than 10 different observations individually. Additionally, 391 of the observations did not have their subtreatment type recorded. These observations were removed from the dataset. The frequency of the clinical and subclinical mastitis for each remaining subtreatment type is as follows

Table 4.6: Clinical and subclinical mastitis infection frequency for each subtreatment type

Subtreatment Type	Clinical	Subclinical	Total
0	489	299	788
1	239	135	374
2	173	121	294
3	30	38	68
4	12	11	23
A	38	0	38

From table 4.6 it can be seen that subtreatment type 0 is the most common type employed. There are far more clinical than subclinical infections observed when subtreatment 0 and 1 wee used. 3 is the only subtreatment type that had more subclinical than clinical cases. No observations with subclinical mastitis used subtreatment A.

4.2 Cow Factors

Various variables relating to the cow were studied, such as her age, number of offspring, and number of dry off days given during the previous dry period.

Farm

The variable Farm is used to identify which farm a cow comes from and uses a numerical labeling system to preserve anonymity. Marked differences in the clinical versus subclinical frequency of mastitis between certain farms can be viewed as an indicator that these farms have differences in important factors that have not been recorded through the other variables in this dataset. One farm, number 66, was removed from the dataset as it only contained 4 observations. The frequency of the clinical and subclinical mastitis observations on each farm is as follows

Table 4.7: Clinical and subclinical mastitis infections on each farm

Farm Number	Clinical	Subclinical	Total
1	506	439	945
2	202	168	370
3	67	55	122
4	145	21	166
6	59	67	126
7	40	26	66
8	8	4	12
9	145	2	147
13	29	4	33

Table 4.7 shows that the majority of observations were found in farm number 1, which had a roughly even division between clinical and subclinical mastitis cases. Farm 4 in comparison, had nearly 7 times as many clinical than subclinical cases. On all farms but farm 6, clinical cases were more common than subclinical.

Dry Off Days

The variable Dry Off Days indicates how long the cow had been dried off for. Drying off is an important phase in the dairy cycle where milk secretion is abruptly ended to allow for preparation and recovery for the cows next lactation period. This variable measured the number of days between their last milking of the previous lactation cycle, and their next calf is born. In Ireland, the recommended dry of length is approximately 56 days, with greater than 63 days being suggested if a cow has a high stomatic cell count, to increase mastitis cure rates (Glanbia 2017). Research has shown that severely extended dry periods, from 143 to 250 days, increase the odds of a subclinical mastitis infection (Pinedo et al. 2011). To prevent the large values of the variable from overpowering other smaller sized variables in the model, scaling was applied. The values for dry off days were

divided by 30, to roughly convert the variable into the number of months a cow was dried off, if each month had 30 days. 504 observations had no informaton recorded on this variable and were removed. The summary statistics of the variable is as follows

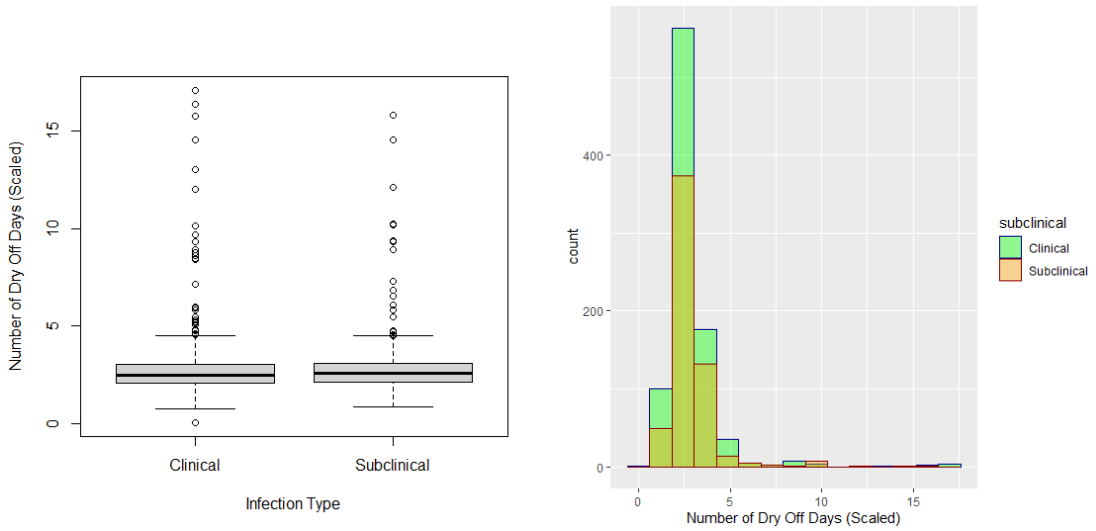
Table 4.8: Summary statistics of the Dry Off Days variable for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
2.8	2.5	2.1	3.1	0.067	17.1

Table 4.9: Summary statistics of the Dry Off Days variable for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
2.843	2.6	2.167	3.1	0.9	15.767

Table 4.8 and 4.9 shows that the average number of scaled dry of days was approximately 2.8 for both clinical and subclinical infections, or 84 days. A clinical case had the shortest dry off period of 0.067, or 2 days. The longest dry off period was also a clinical case, 435 days.



(a) **(b)**

Figure 4.1: (a) A boxplot indicating the difference in the number of dry off days between clinical and subclinical mastisis observations. (b) A histogram shows the distribution of dry off days in clinical and subclinical observations

Figure 4.1a shows that overall the number of dry of days between clinical and subclinical mastitis cases are very similar. Both have a large number of upper outliers. Clinical cases have 43 upper outliers and 1 lower outlier. Subclinical cases have 25 upper outliers. Figure 4.1b shows that for both clinical and subclinical mastitis, the distribution of dry off days is heavily right skewed. The majority of the observations fall in the 3^{rd} or 4^{th}

bin, which signify 61-90 and 91-120 days respectively.

Some of the values recorded for the number of dry off days are incredibly high when the industry standard of 53-63 days is considered. Further examination of the variable was carried out, including analysis of the original milking machine output. This analysis revealed that for 168 observations, there was a error in the recording software where it skipped one of the cow's pregnancies. As such, the dry off length appears abnormally large, because it is encompassing the gap from the previous calf's birth date, rather than the current calf's birth date. As these observations are recorded in error, and not actual outliers, they were removed from the dataset. These values are still included in figure 4.1 and table 4.8/4.9.

Calves

The variable Calves is based on the number of calves the cow has given birth to in her life. The values range from 1 to 10. There was one observation that had 12 calves, but this was removed from the dataset due to insufficient information surrounding that level of calves. The frequency of the clinical and subclinical mastitis observations for each number of calves born is as follows

Table 4.10: Clinical and subclinical mastitis infection frequency for each number of calves born

Calves	Clinical	Subclinical	Total
1	302	198	500
2	278	108	386
3	191	134	325
4	187	138	325
5	105	71	176
6	69	73	142
7	23	34	57
8	30	14	44
9	13	9	22
10	6	7	13

Table (4.10) shows that as the total number of observations decreases as the number of calves born increases. Twice as many observations with 8 calves had clinical rather than subclinical mastitis, and the majority of observations had between 1 and 4 calves. There is a noticeable difference between the number of clinical and subclinical infections when 1 and 2 calves were born, with far more of the infections being clinical than subclinical.

Age

The variable age indicates the age of the cow in years. Past research has shown that

subclinical mastitis cases significant rose in frequency as the age of Swiss dairy cow increased (Busato et al. 2000). The summary statistics of the variable age are as follows

Table 4.11: Summary statistics of the Age variable for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
4.42	4.151	2.981	5.463	1.863	11.479

Table 4.12: Summary statistics of the Age variable for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
4.829	4.521	2.924	6.392	1.841	12.682

Table 4.11 and 4.12 show that the average age of cows is approximately 4 and a half years old for both clinical and subclinical. The mean age for subclinical infections is higher than the age for clinical infections. Subclinical infections contained the lowest and highest observed age at 1.84 and 12.682 years old respectively.

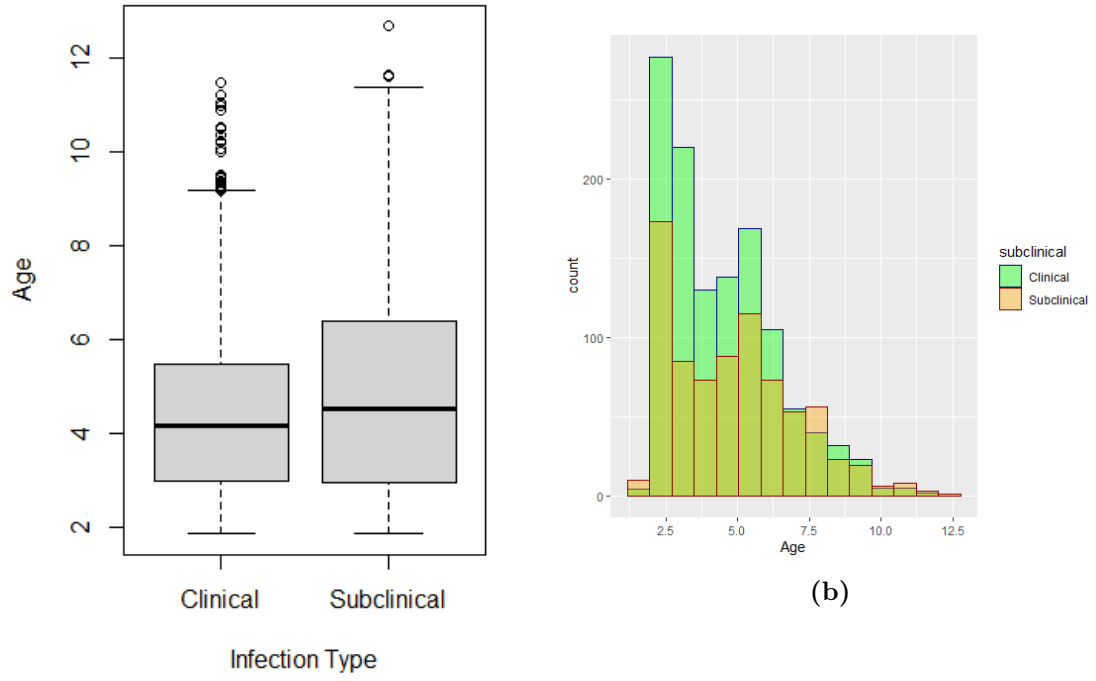
Figure (4.2a) shows that subclinical observations had a higher mean age than clinical observations, as well as a larger upper range. Clinical has a large number of upper outliers. Clinical cases have 27 upper outliers. Subclinical cases have 3 upper outliers. Figure 4.2b shows that for both clinical and subclinical mastitis, the distribution of age is right skewed. respectively. This indicates that there are a small number of observations that had an age much higher than the majority of observations.

Body Condition Score

Body condition score is a measurement of the fat distribution of a cow. Its possible values range from 1 to 5, with 1 indicating an extremely thin cow, and 5 indicating an extremely fat cow. Past research on body condition score and its potential influence on subclinical mastitis infecions did not show any significant relationship (Domecq et al. 1997), but it is still an important variable to examine, as an extremely overweight or underweight cow may be immunocomprimised, therfore increasing their chances of a subclinical mastitis infection. Each cow had her body condition score recorded in this variable, with the exception of 207 observations, that were removed from the dataset

Table 4.13: Summary statistics of the Body Condition Score variable for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
2.941	3	2.75	3.125	1.75	4.75



(a) **(b)**

Figure 4.2: (a) Boxplot indicating the difference in age between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of age in clinical and subclinical observations

Table 4.14: Summary statistics of the Body Condition Score variable for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
2.945	3	2.75	3.125	1.875	4.75

Table 4.13 and 4.14 show that the values for clinical and subclinical infections are very similar. The mean and median body condition scores are very similar, and the majority of the values have a small range of 2.75 to 3.125.

Figure (4.3a) shows that subclinical and clinical observations have similar means and interquartile ranges. Both also have a number of upper and lower outliers. Clinical cases have 32 lower outliers and 19 upper outliers. Subclinical cases have 9 lower outliers and 12 upper outliers. Figure 4.3b shows that for both clinical and subclinical mastitis, the distribution of body condition score is bell shaped around a value of 3.

Weight

The infected cows weight was measured in kilograms. To aid in the convergence of the

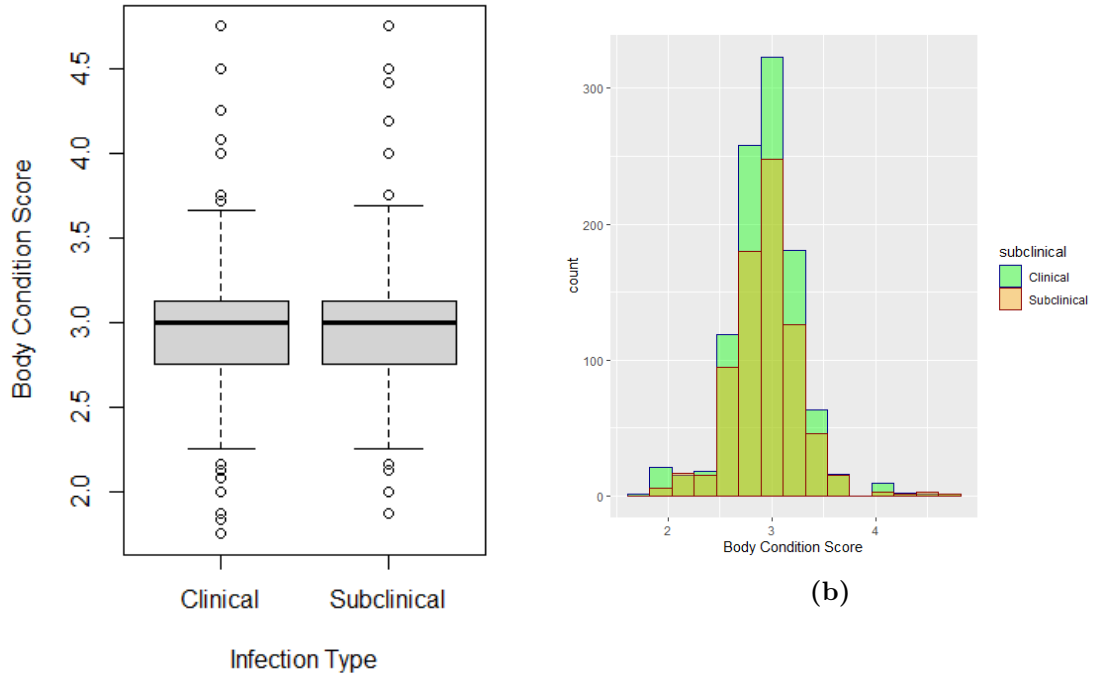


Figure 4.3: (a) Boxplot indicating the difference in body condition score between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of body condition score in clinical and subclinical observations

optimisation scheme used to estimate the parameters, the values were scaled. All values were divided by 100. 41 observations did not have their weight recorded and as such were removed from the dataset

Table 4.15: Summary statistics of the scaled weight variable for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
5.314	5.327	4.76	5.845	3.09	7.66

Table 4.16: Summary statistics of the scaled weight variable for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
5.335	5.328	4.843	5.889	2.42	8.01

Table 4.15 and 4.16 show that the mean and median value for cow weight are very similar indicating a balanced distribution for both clinical and subclinical infections. Subclinical infections contain the lowest and highest observed values, 2.42 and 8.01 respec-

tively.

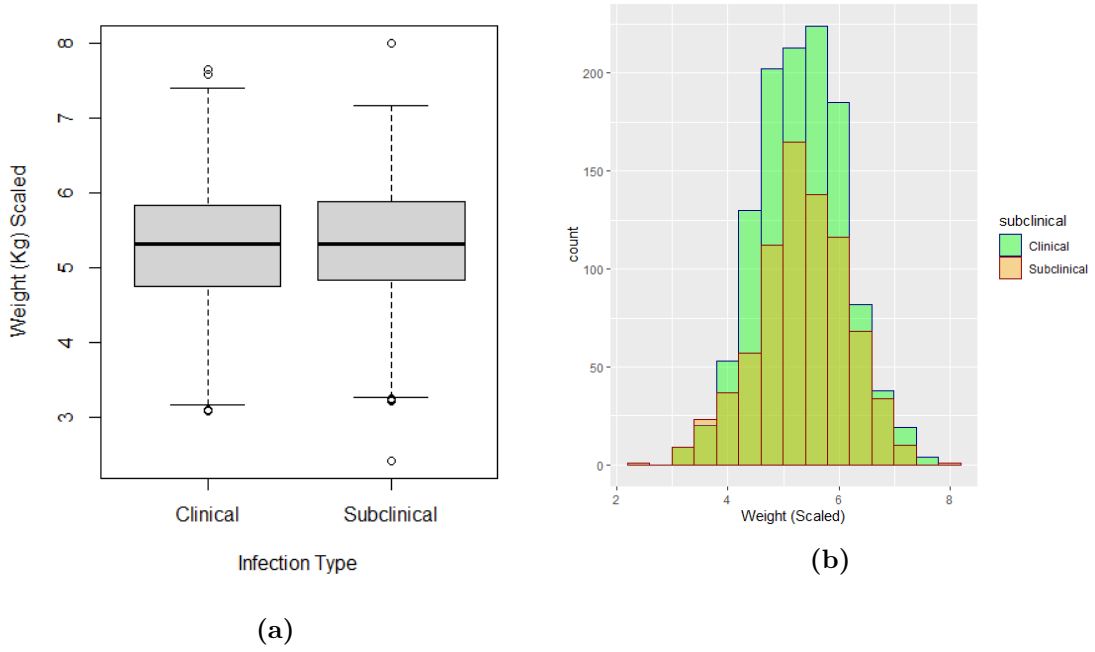


Figure 4.4: (a) Boxplot indicating the difference in scaled weight between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of scaled weight in clinical and subclinical observations

Figure (4.4a) shows that subclinical and clinical observations have similar means and interquartile ranges. Both also have a number of upper and lower outliers. Clinical cases have 2 upper and 2 lower outliers. Subclinical cases have 1 upper outlier and 5 lower outliers. Figure 4.4b shows that for both clinical and subclinical mastitis, the distribution of body condition score is bell shaped around a value of 500.

Cow

The variable Cow is the code used to differentiate each individual cow in the dataset. As it is possible for cows to contract mastitis multiple times during the period of the study, they can be entered as multiple obsersvations in the study. 607 cows were recorded only once in the dataset, having contracted mastitis only once. 435 cows, having contracted mastitis multiple times throughout the study, composed the remainder of the intial 1991 observations in the study. The maximum number of times observed was a cow that was observed 13 seperate times. The average number of times appearing as an observation in the study was 1.911 times.

4.3 Milk Factors

These are variables relating to measurements taken on the observed cows milk. From the values of different nutrients to its stomatic cell count. Variations within milk factors have historically been the primary method of detecting subclinical mastitis, and have been shown to be significantly related to a subclinical diagnosis (Mdegela et al. 2009, Ogola et al. 2007).

Fat

The average fat content of a cow's milk over the month prior to the infection was recorded for both the morning and evening milking. Fat content in milk has been shown to drop by as much as 48% in the presence of subclinical mastitis (Ashworth et al. 1967). 113 observations did not have morning fat recorded and 107 did not have evening fat recorded. These observations were removed from the dataset. The summary statistics of the remaining variables was as follows

Table 4.17: Summary statistics of morning milk fat for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
3.87	3.77	3.23	4.409	0.87	8.09

Table 4.18: Summary statistics of morning milk fat for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
3.932	3.88	3.39	4.4	1.19	7.24

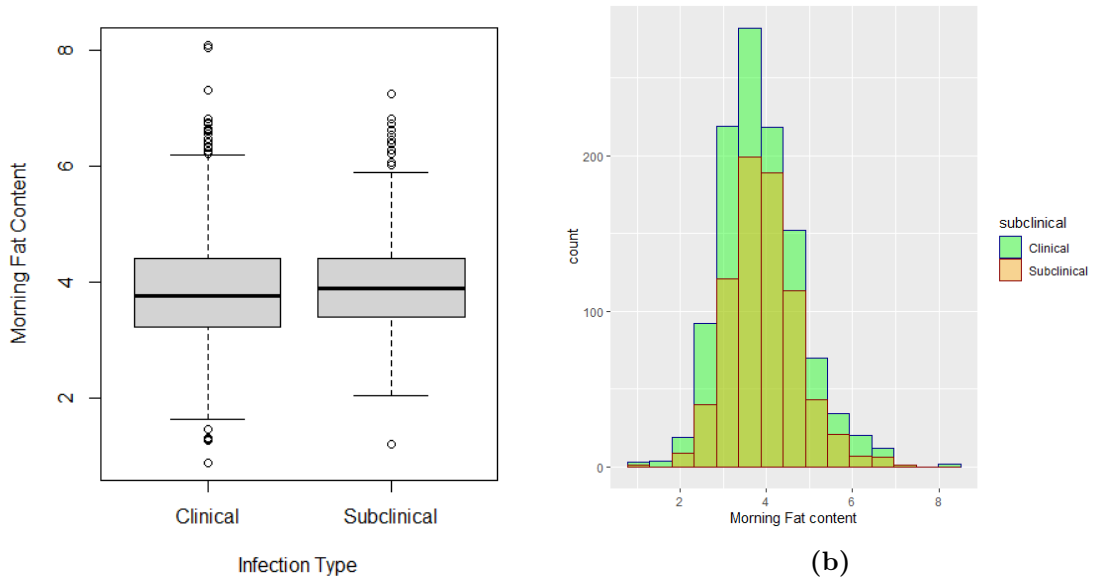
Table 4.19: Summary statistics of evening milk fat in clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
5.42	5.5	4.72	6.14	1.090	12.650

Table 4.20: Summary statistics of evening milk fat in subclinical infections

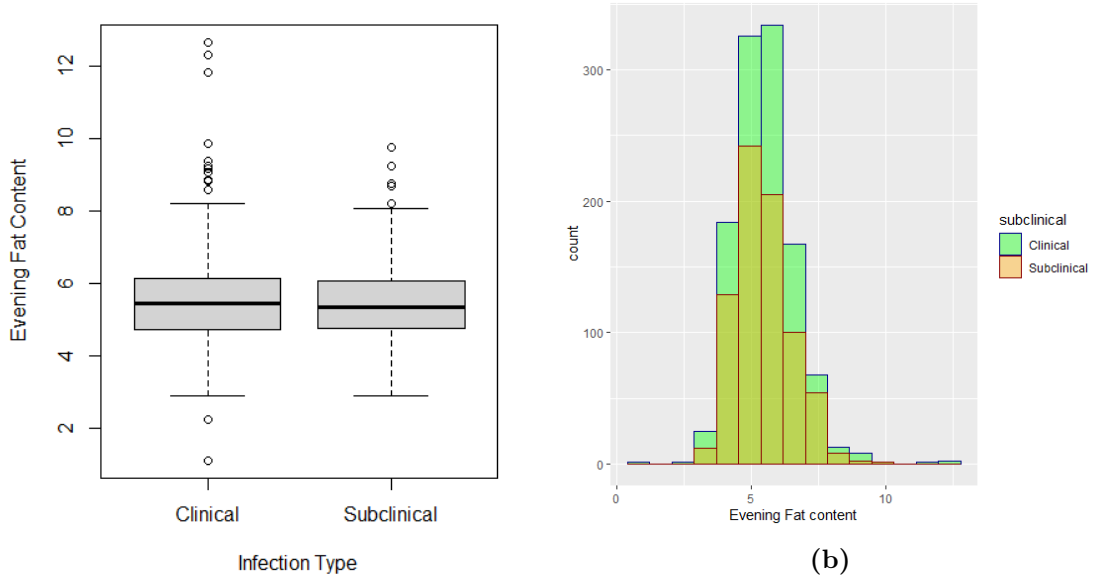
Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
5.47	5.34	4.75	6.1	2.895	9.74

From table 4.17, 4.18, 4.19,and 4.20, an increase in mean fat content from the morning to evening is shown for both clinical and subclinical infections. The values range from 0.87 to 8.09 and 1.09 and 12.65 respectively.



(a) (b)

Figure 4.5: (a) Boxplot indicating the difference in morning fat content between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of morning fat content in clinical and subclinical observations



(a) (b)

Figure 4.6: (a) Boxplot indicating the difference in evening fat content between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of evening fat content in clinical and subclinical observations

Figure (4.5a) indicates that clinical mastitis infections has a slightly lower mean morn-

ing fat content compared to subclinical, and is more variable. Morning fat content for clinical cases have 5 lower outliers and 22 upper outliers. Morning fat content for subclinical cases have 14 upper outliers and 1 lower outlier. Figure (4.5b) shows that morning fat content has a bell shaped distribution for both clinical and subclinical observations. Figure (4.6a) shows that the mean value between both types of infection is approximately equal for evening fat content, with clinical infections having a higher number of outliers. Evening fat content for clinical cases have 2 lower outliers and 14 upper outliers. Evening fat content for subclinical cases have 5 upper outliers. Figure (4.6b) shows that evening fat content has a bell shaped distribution for both clinical and subclinical observations.

Protein

The average protein content of a cow's milk over the month prior to the infection was recorded for both the morning and evening milking. Subclincial mastitis infections cause a rise in the non-caesin (a family of proteins specific to milk) protein concentrations of milk (Ishikawa et al. 1982). 113 observations did not have morning protein recorded and 107 did not have evening protein recorded. These observations were removed from the dataset. The summary statistics of the remaining variables was as follows

Table 4.21: Summary statistics of morning milk protein for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
3.522	3.487	3.254	3.71	2.36	5.96

Table 4.22: Summary statistics of morning milk protein for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
3.538	3.5	3.31	3.745	0.98	5.69

Table 4.23: Summary statistics of evening milk protein for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
3.503	3.542	3.293	3.753	1.54	5.1

Table 4.24: Summary statistics of evening milk protein for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
3.579	3.565	3.34	3.788	2.35	5.02

From table 4.21, 4.22, 4.23,and 4.24 a small increase in mean protein content from the morning to evening is shown for subclinical infections. The values for clinical and subclinical infections are overall very similar for both morning and evening protein content.

The minimum and maximum values for morning protein content are more extreme than those recorded on the final measurement.

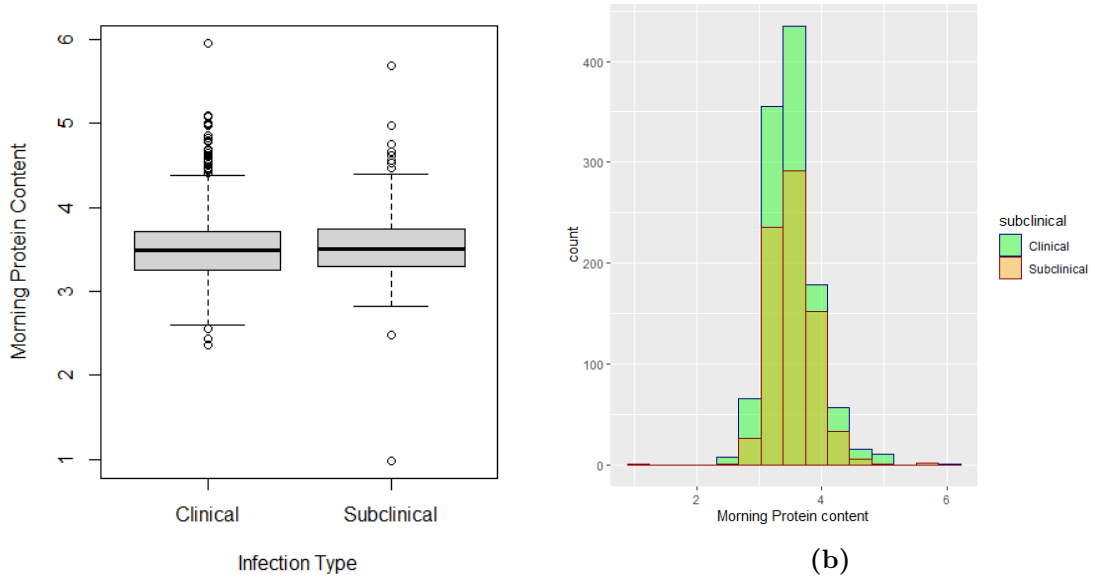


Figure 4.7: (a) Boxplot indicating the difference in morning protein content between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of morning protein content in clinical and subclinical observations

Figure 4.7a indicates that both infection types have approximately equal mean morning protein levels, there is an extreme lower outlier in subclinical however. Morning protein content for clinical cases has 3 lower outliers and 33 upper outliers. Morning protein content for subclinical cases has 2 lower outliers and 9 upper outliers. Figure 4.7b shows that evening protein content has a bell shaped distribution for both clinical and subclinical observations. Figure 4.8a shows that the mean value between both types of infection is approximately equal for final protein content, with clinical infections having an extreme lower outlier. Evening protein content for clinical cases have 3 lower outliers and 29 upper outliers. Evening protein content for subclinical cases have 3 lower outliers and 7 upper outliers. Figure 4.8b shows that evening protein content has a bell shaped distribution for both clinical and subclinical observations.

Lactose

The average lactose content of a cow's milk over the month prior to the infection was recorded for both the morning and evening milking. Lactose concentration in milk has been shown to drop during the presence of a subclinical mastitis infection (M Bruckmaier et al. 2012). 113 observations did not have morning lactose recorded and 107 did not

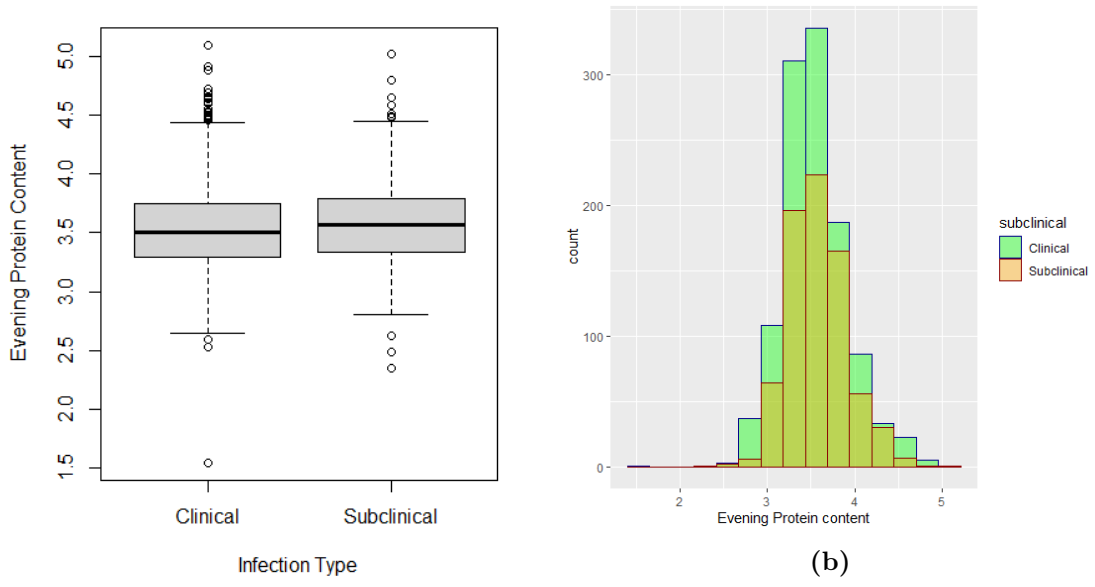


Figure 4.8: (a) Boxplot indicating the difference in evening protein content between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of evening protein content in clinical and subclinical observations

have evening lactose recorded. These observations were removed from the dataset. The summary statistics of the remaining variables was as follows

Table 4.25: Summary statistics of morning milk lactose for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
4.634	4.66	4.51	4.81	2.965	5.297

Table 4.26: Summary statistics of morning milk lactose for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
4.568	4.603	4.43	4.75	0.61	5.22

Table 4.27: Summary statistics of evening milk lactose for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
4.578	4.62	4.446	4.7	1.79	5.421

From table 4.25, 4.26, 4.27,and 4.28 it is shown the the lactose content of milk remains relatively stable between morning and evening milkings, for both clinical and subclinical infections. The minimum value for morning lactose in subclinical infections is much lower

Table 4.28: Summary statistics of evening milk lactose for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
4.523	4.56	4.39	4.7	2.62	5.420

than those observed in the other categories.

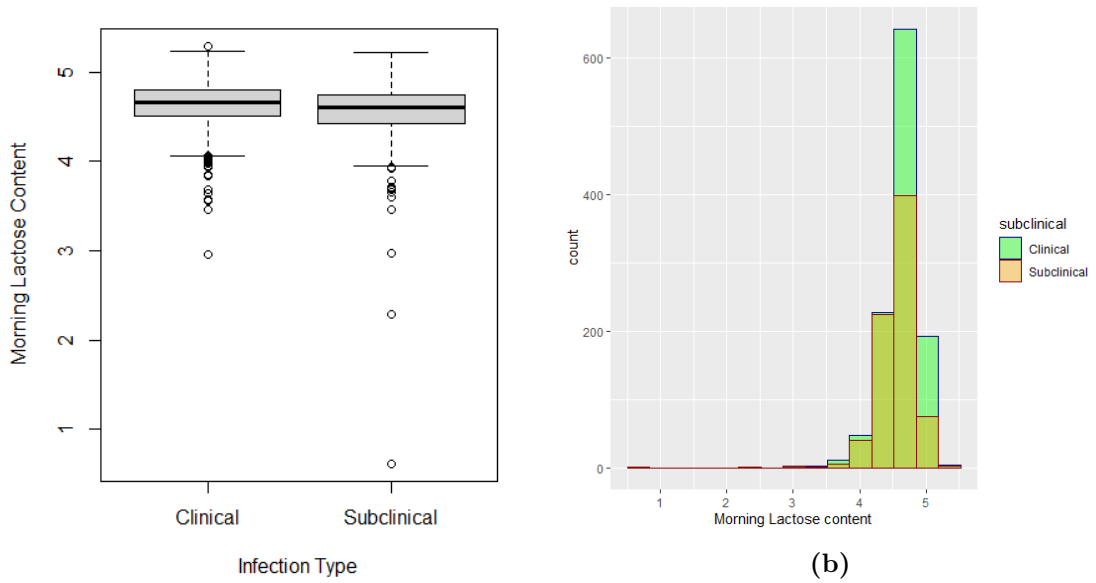


Figure 4.9: (a) Boxplot indicating the difference in morning lactose content between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of morning lactose content in clinical and subclinical observations

Figure 4.9a indicates that both infection types have approximately equal initial mean lactose levels, there is an extreme lower outlier in subclinical however. Morning lactose content for clinical cases have 1 upper outlier and 34 lower outliers. Morning lactose content for subclinical cases have 14 lower outliers. Figure 4.9b shows that both morning lactose values for both clinical and subclinical mastitis have a heavy left skew distribution, with very few observations having lactose values of 2.5-3.5. 4.10a shows that the mean value between both types of infection is approximately equal for final lactose content, with clinical infections having an extreme lower outlier. Evening lactose content for clinical cases have 36 lower outliers. Evening lactose content for sub clinical cases have 1 upper outlier and 24 lower outliers. Figure 4.10b shows that both evening lactose values for both clinical and subclinical mastitis have a heavy left skew distribution, with very few observations having lactose values of 1.5-3.

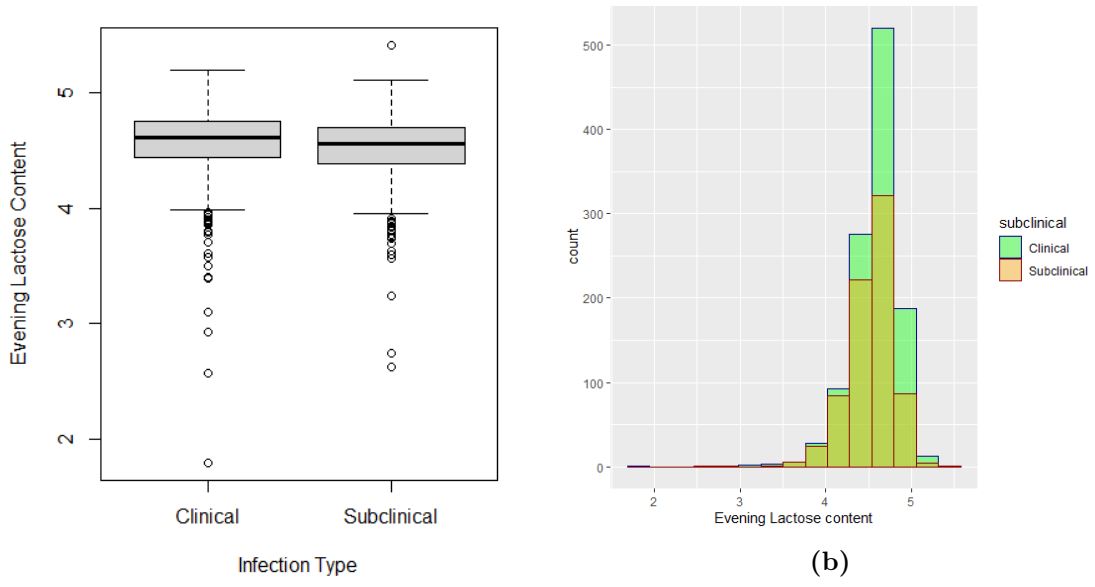


Figure 4.10: (a) Boxplot indicating the difference in evening lactose content between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of evening lactose content in clinical and subclinical observations

Stomatic Cell Count

Stomatic cell count is the number of stomatic cells per millimeter of milk. Inflammation in the udder, caused by the introduction of pathogenic microorgaisms such as in the case of mastitis, causes a large number of stomatic cells to enter the area in an attempt to combat the infection (Harmon 1994). This causes an increase in the number of stomatic cells present in an infected cows milk, and is currently used as the primary detector of mastitis infections in cows (Harmon 1994). For the past 20 years, a stomatic cell count higher than 200,000 cells per ml has been used as the indicator of a mastitis infection (Hillerton 1999).

The average stomatic cell count of a cow's milk over the month prior to the infection was recorded at the morning milking. To stop the very large numbers recorded from dominating the regressions coefficients, a scaling was applied to the values. All stomatic cell count values were divided by 1000 to scale them appropriately. 169 of the observations did not have the stomatic cell count recorded and as such were removed from the data set. The summary statistics of the remaining observations is as follows

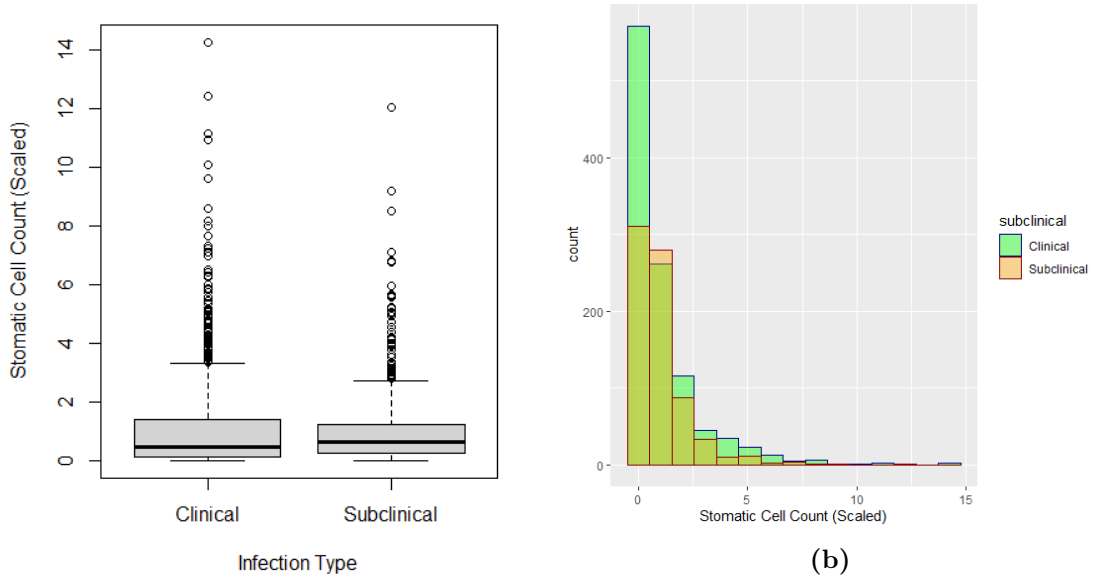
Table 4.29: Summary statistics of milk's stomatic cell count for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
1.12	0.46	0.12	1.41	0.001	14.257

Table 4.30: Summary statistics of milk's stomatic cell count for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
1.006	0.64	0.24	1.26	0.001	12.022

From table 4.29 and 4.30, the mean stomatic cell count is much higher than the median for both clinical and subclinical infections, indicating a right skew. The scaled values range from 0.001 and 14.257



(a) **(b)**

Figure 4.11: (a) Boxplot indicating the difference in scaled stomatic cell count between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of the stomatic cell count of milk in clinical and subclinical observations

Figure 4.11a indicates that clinical infections are more variable in their stomatic cell count value and have more outliers than subclinical infections have. Clinical cases have 97 upper outliers. Subclinical cases have 51 upper outliers. Figure 4.11b shows that scc values are heavily right skewed for both clinical and subclinical observations, with very few values from 10-15.

Yield

The variable yield can be viewed as the slope of the milk yield for the previous month prior to infection. A positive value can be viewed as an increase in milk yields day to day for the previous month, a negative value can be viewed as a decrease in milk yields day to day for the previous month, and a value near 0 indicates that yields were

relatively constant. A decrease in yield has been shown to be associated with subclinical mastitis, causing an average loss in yield of around 6.8% per cow (Gebreyohannes et al. 2010). These values were taken at both morning and evening milkings. To stop the very small numbers recorded from having a minimal impact on the regressions coefficients, a scaling was applied to the values. All yield values were multiplied by 10 to scale them appropriately. 19 observations did not have information gathered on the initial yield and 25 observations did not have information on the final yield. These variables were removed from the dataset

Table 4.31: Summary statistics of the change in previous months morning milk yield for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
-0.0615	-0.15	-0.997	0.744	-67	68

Table 4.32: Summary statistics of the change in previous months morning milk yield for subclinical cases

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
0.136	-0.1524	-0.784	0.625	-35.3	54.2

Table 4.33: Summary statistics of the change in previous months evening milk yield for clinical infections

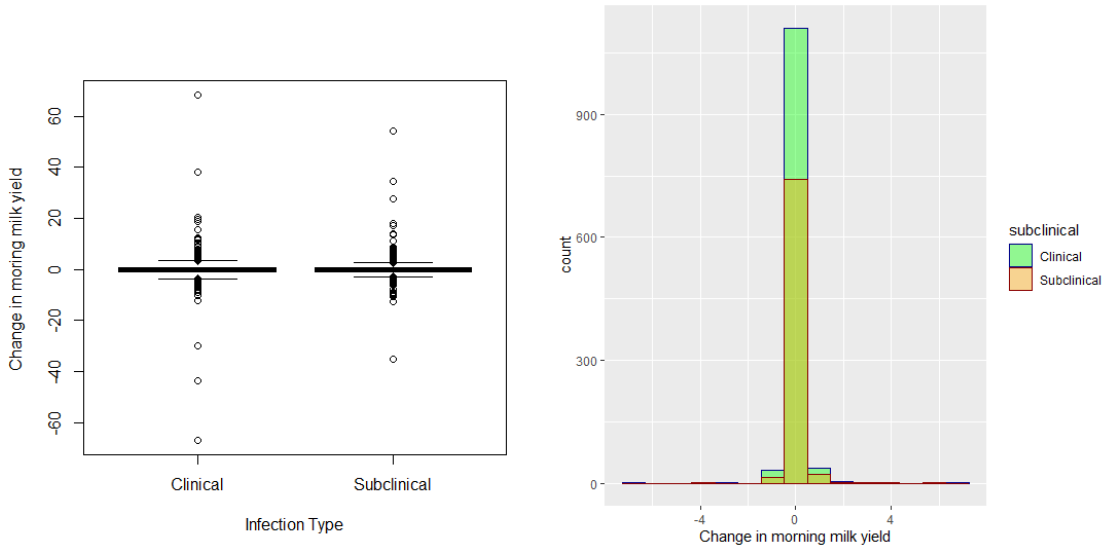
Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
-0.041	-0.134	-0.69	0.53	-15.89	15.5

Table 4.34: Summary statistics of the change in previous months evening milk yield for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
0.0469	-0.1	-0.44	0.36	-23.57	22.5

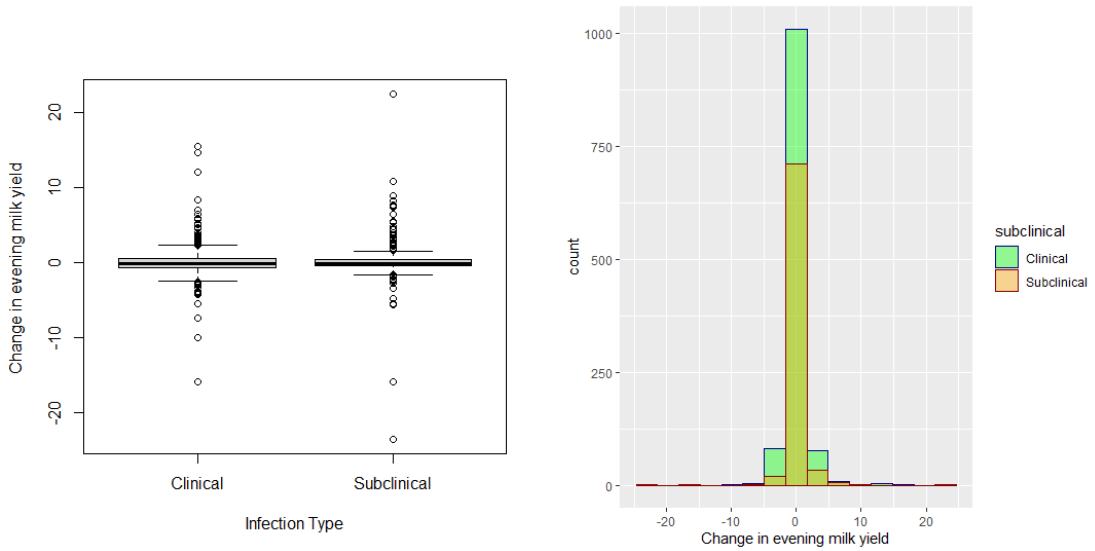
Table 4.31 and 4.32 show that the change in morning yields had a mean value of -0.0615 and 0.136 for clinical and subclinical infections, indicating that they remained relatively constant throughout the previous month. Some were much more variable for example, with one observation having a positive slope of 68, and another a negative slope of -67, with clinical cases containing both of these values. Table 4.33 and 4.34shows that the change in evening yields had a mean value of -0.041 and 0.0469, indicating that they also remained relatively constant throughout the previous month. Some were much more variable for example, with one observation having a positive slope of 22.5, and another a negative slope of -23.57, both these extreme values were found in subclinical infections.

Evening yield was less variable than morning yield.



(a) **(b)**

Figure 4.12: (a) Boxplot indicating the difference in the previous months change in morning milk yield between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of the change in morning milk yield in clinical and subclinical observations



(a) **(b)**

Figure 4.13: (a) Boxplot indicating the difference in the previous months change in evening milk yield between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of the change in evening milk yield in clinical and subclinical observations

Figure 4.12a and figure 4.13a both indicate that clinical and subclinical mastitis infections both have very similar yield values, tightly grouped around 0, with a large number of outliers. Change in yield for clinical infections have 49 and 33 lower outliers and 71 and

54 upper outliers for morning and evening, respectively. Change in yield for subclinical infections have 32 and 25 lower outliers and 53 and 45 upper outliers for morning and evening, respectively. Figure 4.12b and 4.13b show that both morning and evening yield values have distributions that are strongly grouped around the mean value of 0 for both clinical and subclinical mastitis. Evening change in yield appears to have a slightly wider distribution.

Time

The variable time is a measurement of the time in minutes since the previous milking. The gap in time between the evening milking of the previous day, and the morning milking was recorded, as was the gap in time between the morning and evening milking of that same day. This value was averaged over the month prior to the infection. As this time was recorded in minutes, the values are large, to stop these large values from overpowering the regression coefficients, scaling was applied. The variable was divided by 60, thereby transforming the time variable from minutes to hours. The evening to the morning of the next day gap (evening-morning) had 51 missing values and the gap in time from morning to evening of the same day (moring-evening) had 165 missing values. These observations were removed from the dataset

Table 4.35: Summary statistics of the evening-morning milking time gap for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
7.62	7.38	5.92	9.02	0	21.217

Table 4.36: Summary statistics of the evening-morning milking time gap for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
7.35	7.1	6	8.317	2.5	19.02

Table 4.37: Summary statistics of the morning-evening milking time gap for clinical infections

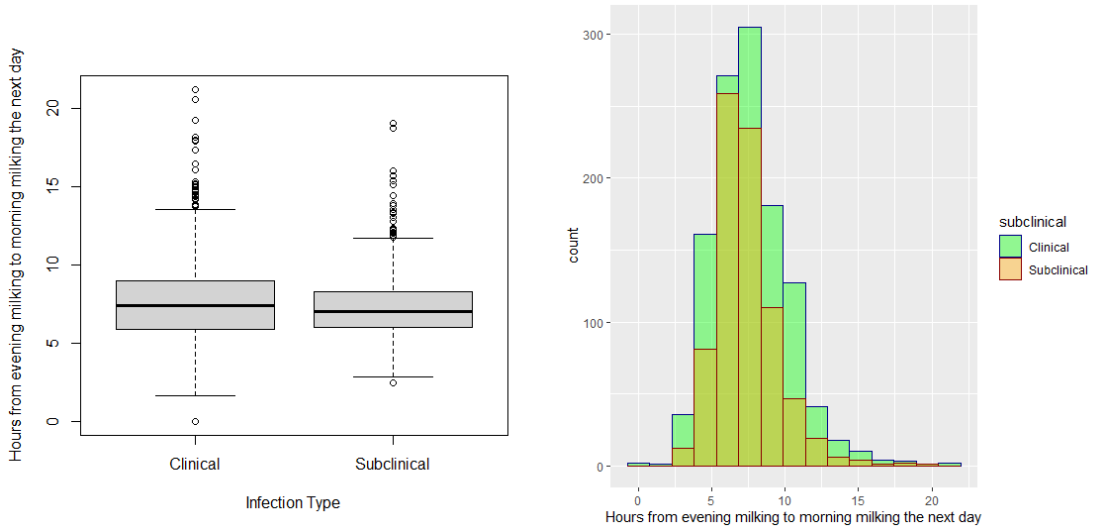
Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
5.899	5.633	4.8	6.8	0	14.4

Table 4.35 and 4.36 show that the mean evening-morning milking time gap was very similar for both clinical and subclinical infections, with a value of rougly 7.5. Table 4.37 and 4.38 show that the mean morning-evening milking time gap was also very similar between clinical and subclinical cases, but was shorter than the evening-morning time gap.

Table 4.38: Summary statistics of the morning-evening milking time gap for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
5.59	5.42	4.75	6.23	2.3	13.32

Clinical cases also contain 0 hour gaps between milkings, with 2 0 hour gaps for evening-morning and 4 0 hour gaps for morning-evening. In comparison, the shortest time gap for subclinical cases is approximately 2.5 for both evening-morning and morning-evening.



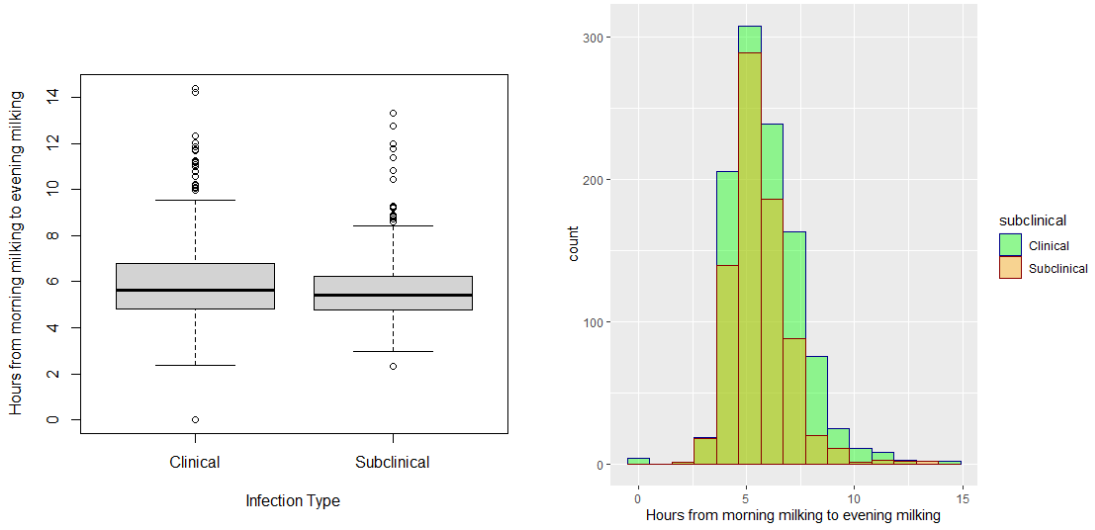
(a) **(b)**

Figure 4.14: (a) Boxplot indicating the difference in the previous month's hours from evening to morning milking of the next day between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of the change in hours from morning to evening milking in clinical and subclinical observations

Figure 4.14a and figure 4.15a both indicate that clinical and subclinical observations have similar values for both evening-morning wait time and morning-evening wait times. Wait time for clinical infections both have 1 lower outlier and 29 and 27 upper outliers for evening-morning wait time and morning-evening wait times, respectively. Wait time for subclinical infections both have 1 lower outlier and 29 and 22 upper outliers for evening-morning wait time and morning-evening wait times, respectively. Figure 4.14b and 4.15b show that both clinical and subclinical observations have evening-morning and morning-evening wait times have a bell-shaped distribution.

Max Flow

Maximum milk flow of a cow's milk over the month prior to the infection was recorded for both the morning and evening milking. The average values were imputed as the Max Flow variable. Milk flow has previously been shown to not change during the presence



(a) **(b)**

Figure 4.15: (a) Boxplot indicating the difference in the previous month's hours from morning to evening milking between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of the change in hours from morning to evening milking in clinical and subclinical observations

of a subclinical mastitis infection (de Felicio Porcionato et al. 2010). However, cows with lower flow rates were more susceptible to mastitis infections, as low flow rates are associated with longer teats, and teat to ground distance influences the chances of contracting mastitis, thereby causing cows with longer teats and thus a smaller max flow value, to potentially become infected more frequently (de Felicio Porcionato et al. 2010). Morning max flow had 51 missing values and evening maximum flow had 165 missing values. These observations were removed from the dataset

Table 4.39: Summary statistics of the morning max milk flow for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
2.865	2.87	2.065	3.71	0	8.341

Table 4.40: Summary statistics of the morning max milk flow for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
2.82	2.786	1.978	3.641	0	8.396

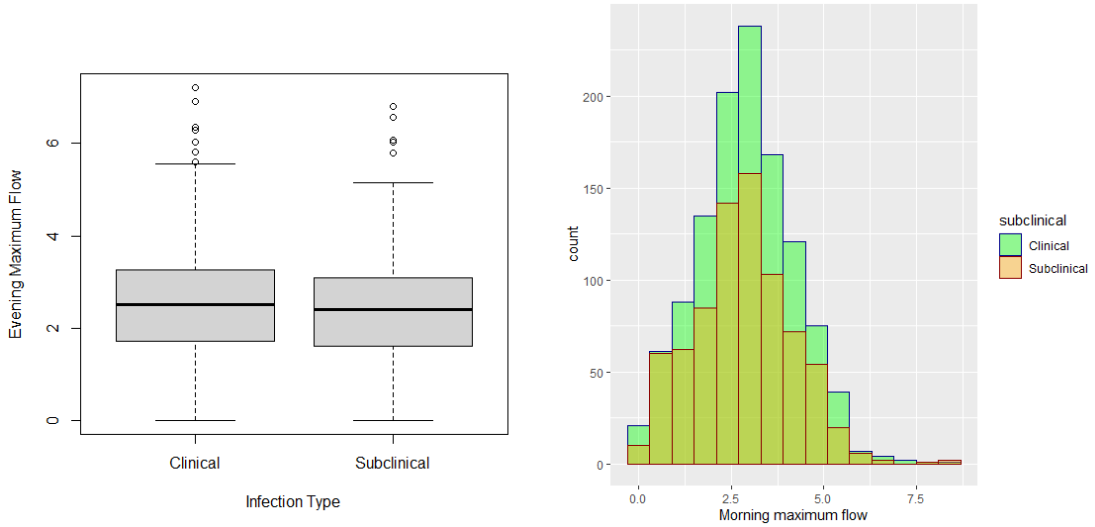
Table 4.41: Summary statistics of the evening max milk flow for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
2.5	2.518	1.72	3.261	0	7.215

Table 4.42: Summary statistics of the evening max milk flow for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
2.369	2.394	1.61	3.1	0	6.81

Table 4.39 and table 4.40 show that the values for morning max flow were similar for both clinical and subclinical infections. Table 4.41 and table 4.42 show that clinical cases had a higher mean and median evening max flow than subclinical cases did. These evening values are also lower than the morning flow values. 16 observations had 0 flow at morning milking, and 17 had 0 flow at evening milking.



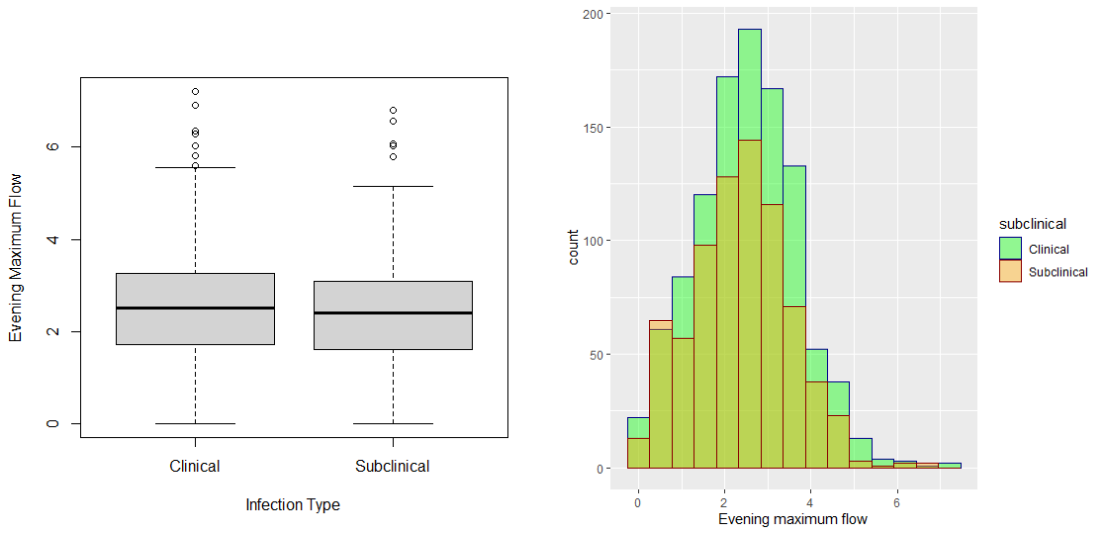
(a) **(b)**

Figure 4.16: (a) Boxplot indicating the difference in the morning maximum milk flow between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of the morning maximum milk flow in clinical and subclinical observations

Figure 4.16a and figure 4.17a both indicate that clinical and subclinical have a very similar range of maximum flow values, with clinical values shifted slightly higher. Maxflow for clinical infections have 9 upper outliers for both morning and evening measurments. Maxflow for subclinical infections have 8 and 5 upper outliers for morning and evening measurements respectively. Figure 4.16b and 4.17b show that the distribution of mornning and evening maximum flow are slightly right skewed for both clinical and subclinical observations indicating that there are a small amount of observations with exteremly high values when compared to the majority.

Conc Fed

The variable conc fed is related to the concentration of feed that the observed cow ate the month prior to the infection. The values were averaged out over the month, and was



(a) (b)

Figure 4.17: (a) Boxplot indicating the difference in evening maximum milk flow between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of evening maximum milk flow in clinical and subclinical observations

recorded for both the morning and evening feeding. Because of the large values of these variables, they were scaled to prevent them from overpowering the regression coefficients. Each value was divided by 10. Morning conc fed had 171 missing values and evening conc fed had 173 missing values. These observations were removed from the dataset

Table 4.43: Summary statistics of the morning concentration fed for clinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
1.214	1.2	0.1	2.2	0.1	4.0

Table 4.44: Summary statistics of the morning concentration fed for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
0.9768	0.6	0.1	1.9	0.1	3.9

Table 4.45: Summary statistics of the evening concentration fed for clinical infecitons

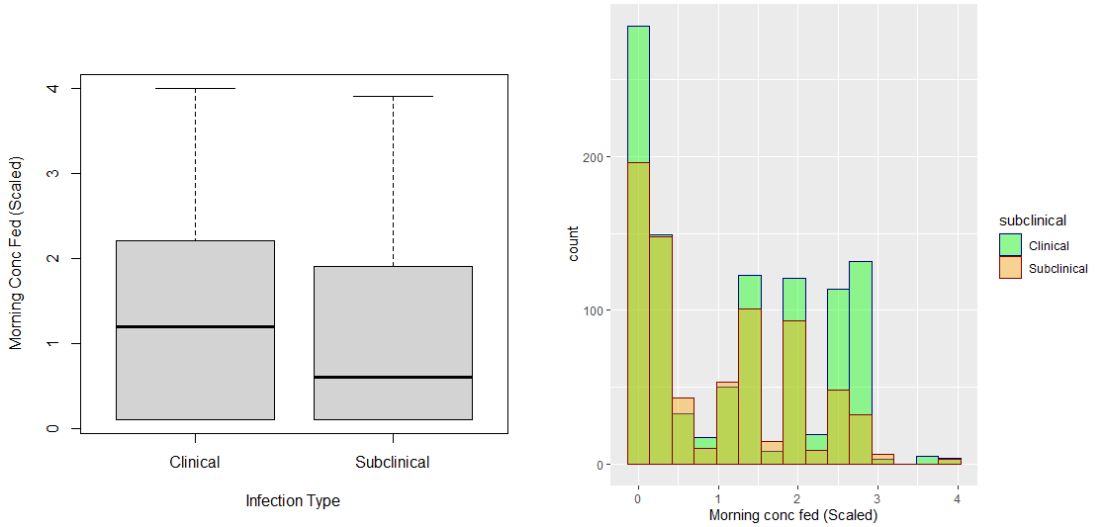
Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
1.21	1.35	0.1	2.1	0.1	4.0

Table 4.43 and 4.43 show that clinical values had a higher mean and median morning conc fed value than subclinical. Table 4.46 and **??** shows that this trend is repeated in the evening conc fed, with clinical cases having higher values than subclinical cases. The

Table 4.46: Summary statistics of the evening concentration fed for subclinical infections

Mean	Median	1^{st} Quartile	3^{rd} Quartile	Min	Max
0.97	0.6	0.1	1.7	0.1	3.9

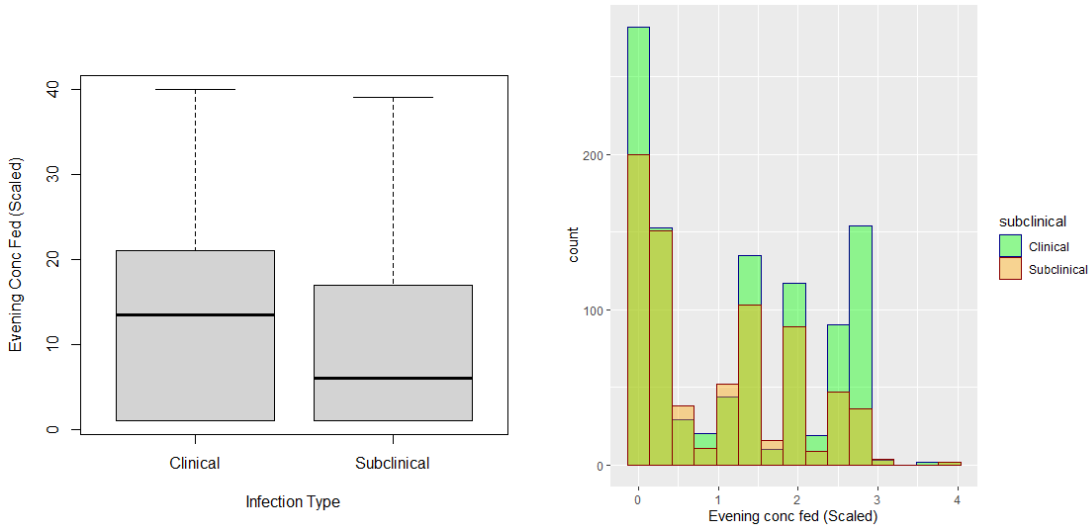
values for morning and evening clinical conc fed are very similar, as are the values for morning and evening subclinical conc fed.



(a) **(b)**

Figure 4.18: (a) Boxplot indicating the difference in the morning conc fed between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of the morning conc fed in clinical and subclinical observations

Figure (4.18a) and figure (4.19a) show that for both morning and evening conc fed, clinical infections have a higher mean conc fed value than subclinical infections do, with the gap being wider for the evening conc fed values. Figure 4.18b and 4.19b show that the distribution of mornning and evening conc fed are right skewed for both clinical and subclinical observations, with the majority of values being around 0. Morning and evening concentration fed are highly correlated (0.94), as such, only morning concentration fed will be used in the regression analysis to limit multicollinearity.



(a) **(b)**

Figure 4.19: (a) Boxplot indicating the difference in evening conc fed between Clinical and Subclinical mastitis observations. (b) A histogram shows the distribution of evening conc fed in clinical and subclinical observations

Chapter 5

Binary Logistic Model

5.1 Model Fitting and Results

A binary logistic model was run with clinical = 0 and subclinical = 1 as the response using the general linear model function of R (R Core Team 2020). All factors listed in the EDA chapter were included as independent variables in the model except for the cow identifier variable Cow. Only full observations were considered by the model, that is, any observations that contained one or more NA value, was ignored by the model. The size of the dataset after the removal of the other variables and the incomplete observations was 356 observations on 28 variables.

This model assumes that each observation is independent of one another, that the logit transformed response variable is linear to each independent variable, and that there is no multicollinearity between the independent variables.

Three observations were identified and removed from the dataset due to being outliers that significantly impacted the models output via the method described in chapter 2. These were observation 1168, 1197, and 1877. The logistic regression was then rerun without these variables.

Certain levels of the categorical variables caused convergence issues in the mixed-effects logistic regression due to their small sample size. To make comparisons between the two models equal, these levels were also dropped in the logistic regression. These levels were as follows:

Teat: LF_LH, RF_LH, RF_RH_LF, RF_RH_LF_LH, RF_RH_LH, RH_LH, RH_LF_LH, RH_LH

Drug: Synulox Tubes

Tubes: 7, 9
Farm: 7, 4, 13
Treatment: 0, 10, 8, L602

The formula obtained by the fitted model is as follows

$$\begin{aligned}Predicted\ logit\ of\ (Subclinical)\ &= 0.54804 + 1.47 * (Teat_{LH}) + 0.46641 * (Teat_{RF}) + \\ &1.05 * (Teat_{RF\_LF}) + 1.003 * (Teat_{RF\_LH}) + 3.41 * (Teat_{RF\_RH}) + 1.17 * (Teat_{RH}) - 0.31 * \\ &(Drug_{KANACEF\ M.C}) + 0.797 * (Drug_{MULTIMAST}) - 0.358 * (Drug_{SYNULOX}) + 0.48 * \\ &(Drug_{TEREXINE}) - 0.113 * (Drug_{TERREXINE}) - 2.47 * (Drug_{TETRA\ DELTA}) - 2.56 * (Severity_2) - \\ &2.85 * (Severity_3) + 0.65 * (Tubes_2) + 0.56 * (Tubes_3) + 1.48 * (Tubes_4) + 0.875 * (Tubes_5) + \\ &1.13 * (Tubes_6) + 2.87 * (Tubes_8) + 1.1 * (Farm_2) - 3.33 * (Farm_3) + 1.89 * (Farm_6) + 0.09 * \\ &(Farm_9) - 0.185 * (Dry\ off\ Days) - 0.136 * (Calf_3) - 1.23 * (Calf_4) - 1.75 * (Calf_5) - \\ &1.1(Calf_6) - 1.35(Calf_7) - 5.1 * (Calf_8) - 2.79 * (Calf_9) - 0.298 * (Treatment_2) - 0.126 * \\ &(Treatment_3) - 0.2 * (Treatment_4) - 0.28 * (Treatment_5) - 1.596 * (Treatment_{C150}) - \\ &1.78 * (Treatment_{G250}) - 2.35 * (Treatment_{TMR}) - 1.077 * (SubTreatment_1) - 0.784 * \\ &(SubTreatment_2) - 1.899 * (SubTreatment_3) + 0.61 * (Age) - 0.5 * (Morning\ Fat) + 0.629 * \\ &(Morning\ Protein) - 2.25 * (Morning\ Lactose) - 0.29 * (Stomatic\ Cell\ Count) + 0.194 * \\ &(MorningYield) - 0.53 * (Evening - Morning\ Time) + 0.139 * (Morning\ MaxFlow) - \\ &0.43 * (Morning\ Conc\ Fed) + 0.55 * (Evening\ Fat) - 0.28 * (Evening\ Protein) + 1.66 * \\ &(Evening\ Lactose) + 1.66 * (EveningYield) + 0.166 * (Morning - Evening\ Time) - 0.52 * \\ &(Evening\ Max\ Flow) - 0.017 * (Weight) - 0.077 * (Body\ Condition\ Score)\end{aligned}$$

Table 5.1: Model coefficients of the binary logistic model with mastitis infection type as the response variable (0 signifying a clinical infection, 1 signifying a subclinical infection). Only the intercept and coefficients with a probability lower than 0.05 are included.

Coefficient	β Estimate	Standard Error	Odds Ratio & CI	P-value
Intercept	0.548	6.6	1.73 $(4.1\times10^{-6}, 7.4\times10^{5})$	0.9
TeatLH	1.469	0.52	4.34 (1.54, 12.2)	0.00542
TeatRF_RH	3.4	1.32	30.3 (2.27, 406.05)	0.0099
TeatRH	1.17	0.538	3.23 (1.12, 9.27)	0.02919
Farm 3	-3.328	1.614	0.036 (0.00152, 0.85)	0.039
Severity 2	-2.56	0.566	0.077 (0.0255, 0.23)	5.9×10^{-6}
Severity 3	-2.85	0.7	0.058 (0.0146, 0.229)	5×10^{-6}
Stomatic Cell Count	-0.29	0.146	0.748 (0.56, 0.996)	0.04759
Evening-Morning Time	-0.53	0.18	0.586 (0.412, 0.833)	0.0029
Evening Fat Content	0.554	0.25	1.74 (1.06, 2.86)	0.029
Morning-Evening Time	-0.57	0.51	1.77 (1.05, 2.97)	0.031

From the above formula and from 5.1 the following interpretation for each significant variable can be gained. The odds of an infection being subclinical based on a particular variable was determined by getting the exponent of the β coefficient. From this odds

ratio, a probability could be obtained through the following formula

$$Probability = \frac{Oddsratio}{1 + Oddsratio}$$

5.2 Significant Variable Interpretation

The significance of each variable was assessed by a Wald test as described in chapter 2. The variables examined in the following section all have p-values below 0.05, signifying that they are significantly different from 0.

Teat

When compared to an infection in the reference teat of left fore, with an infection in the left hind, we expect to see a 334% increase in the odds of that observation having subclinical mastitis, this is due to having an odds ratio value of 4.34. As the confidence interval does not contain 1, which would imply no association between the left hind teat infection and a subclinical diagnosis, we can state that we are 95% confident that there is a relationship between subclincial infections occurring in the left hind relative to the left fore. Using the above formula to produce a probability, we receive a value of 0.81. We can interpret this as an infection in the left hind teat, compared to an infection in her left fore teat, with all other variables held equal, has an increase in the probability of it being a subclinical mastitis infection by 0.81.

When compared to an infection in the reference teat of left fore, with an infection in the right fore and right hind teat, we expect to see a 2930% increase in the odds of that observation being subclinical mastitis, this is due to having an odds ratio value of 30.3. As the confidence interval does not contain 1, which would imply no association between a right fore and hind teat infection and a subclinical diagnosis, we can state that we are 95% confident that there is a relationship between subclincial infections occurring in the right fore and right hind teat relative to the left fore. Using the above formula to produce a probability, we receive a value of 0.96. We can interpret this as an infection in her right hind and fore teat, an infection in her left fore teat, with all other variables held equal, has an increase in the probability of it being a subclinical mastitis infection by 0.96.

When compared to an infection in the reference teat of left fore, to an infection in the right hind teat, we expect to see a 223% increase in the odds of that observation having subclinical mastitis, this is due to having an odds ratio value of 3.23. As the confidence interval does not contain 1, which would imply no association between a right hind infection and a subclinical diagnosis, we can state that we are 95% confident that there is a

relationship between subclincial infections occurring in the right hind relative to the left fore. Using the above formula to produce a probability, we receive a value of 0.764. We can interpret this as an infection in the right hind teat, compared to an infection in the left fore teat, with all other variables held equal, has an increase in the probability of it being a subclinical mastitis infection by 0.764.

Farm

An infection on farm 3 compared to one on the reference farm 1, we expect to see a 96.4% decrease in the odds of that observation being subclinical mastitis, this is due to having an odds ratio value of 0.036. As the confidence interval does not contain 1, which would imply no association between an infection being on farm 3 and a subclinical mastitis diagnosis, we can state that we are 95% confident that there is a relationship between subclincial infections occurring on farm 3 relative to farm 1. Using the above formula to produce a probability, we receive a value of 0.035 . We can interpret this as an infection at farm 3, compared with an infection in farm 1, with all other variables held equal, has an decrease in the probability of it being a subclinical mastitis infection by 0.035.

Severity

An infection of severity level 2 compared to of the reference severity level 1, we expect to see a 92.3% decrease in the odds of that observation being subclinical mastitis, this is due to having an odds ratio value of 0.077. As the confidence interval does not contain 1, which would imply no association between an infection severity level 2 a subclinical mastitis diagnosis, we can state that we are 95% confident that there is a relationship between subclincial infections occurring of severity 2 relative to infections of severity 1. Using the above formula to produce a probability, we receive a value of 0.071 . We can interpret this as an infection of severity level 2, compared to an infection of severity level 1, with all other variables held equal, has an decrease in the probability of it being a subclinical mastitis infection by 0.071.

An infection of severity level 3 compared to of the reference severity level 1, we expect to see a 94.2 % decrease in the odds of that observation having subclinical mastitis, this is due to having an odds ratio value of 0.058. As the confidence interval does not contain 1, which would imply no association between an infection with severity level 3 a subclinical mastitis diagnosis, we can state that we are 95% confident that there is a relationship between subclincial infections occurring of severity 3 relative to infections of severity 1. Using the above formula to produce a probability, we receive a value of 0.055 . We can interpret this as an infection of severity level 3, compared to an infection of severity level 1, with all other variables held equal, has an decrease in the probability of it being a subclinical mastitis infection by 0.0555.

Milk Fat Content

For each unit increase in the average evening milk fat content of a cow a month prior to the infection, we expect to see a 74.8% increase in the odds of that observation having subclinical mastitis, this is due to having an odds ratio value of 1.74. As the confidence interval does not contain 1, which would imply no association between a cows average evening milk fat content and a subclinical mastitis diagnosis, we can state that we are 95% confident that there is a relationship between subclincial infections and the evening milk fat content. Using the above formula to produce a probability, we receive a value of 0.64 . We can interpret this as for every unit increase in the average morning milk fat content of a cow a month prior to the infection, we expect the probability of that observation having subclinical mastitis to decrease by 0.64.

Stomatic Cell Count

For each unit increase in the average Stomatic Cell Count of a cows milk a month prior to the infection, we expect to see a 71% decrease in the odds of that observation having subclinical mastitis, this is due to having an odds ratio value of 0.29. As the confidence interval does not contain 1, which would imply no association between a cows average stomatic cell count and a subclinical mastitis diagnosis, we can state that we are 95% confident that there is a relationship between average stomatic cell count and subclinical mastitis infections. Using the above formula to produce a probability, we receive a value of 0.428. We can interpret this as for every unit increase in the average stomatic cell count of a cow a month prior to the infection, we expect the probability of that observation having subclinical mastitis to decrease by 0.428.

Milking wait time

For each unit increase in the average gap in time between a cows evening milking and the morning miking the next day, a month prior to the infection, we expect to see a 41.4% decrease in the odds of that observation having subclinical mastitis, this is due to having an odds ratio value of 0.586. As the confidence interval does not contain 1, which would imply no association between a cows evening to morning wait time and a subclinical mastitis diagnosis, we can state that we are 95% confident confident that there is a relationship between average evening-morning milking wait time and subclinical mastitis infections. Using the above formula to produce a probability, we receive a value of 0.37 . We can interpret this as for every unit increase in average gap in time between a cows evening milking and the morning miking the next day, a month prior to the infection, we expect the probability of that observation having subclinical mastitis to decrease by 0.37.

For each unit increase in the average gap in time between a cows morning milking

and evening milking, a month prior to the infection, we expect to see a 77% increase in the odds of that observation having subclinical mastitis, this is due to having an odds ratio value of 1.77. As the confidence interval does not contain 1, which would imply no association between a cows morning to evening wait time and a subclinical mastitis diagnosis, we can state that we are 95% confident confident that there is a relationship between average morning-evening milking wait time and subclinical mastitis infections. Using the above formula to produce a probability, we receive a value of 0.63. We can interpret this as for every unit increase in average gap in time between a cows evening morning and evening milking, a month prior to the infection, we expect the probability of that observation having subclinical mastitis to increase by 0.63.

5.3 Model Evaluation

The fitted model had a Null Deviance of 739.75 on 548 degrees of freedom and a Residual Deviance of 397.11 on 470 degrees of freedom

A goodness of fit test was used to assess the models prediction quality compared to a model with just the intercept term included. The intercept only model received a log-likelihood value of -369.64 whereas the fitted model received a value of -198.55. We can state that the fitted model is significantly better at predicting a clinical versus subclinical mastitis diagnosis than the null model, as the probability of this difference in log-likelihood occurring randomly is $<2.2\text{x}10^{-}16$.

The McFadden's Pseudo R^2 value for the model is 0.4628. With McFadden Pseudo R^2 values above 0.4 being considered to "represent excellent fit" (McFadden 1973), we can conclude that the psuedo R^2 value indicates that the model fits the data very well.

Diagnostic plots were also created to assess the models accuracy.

The first plot of 5.1, the residual vs fitted plot shows a small degree of grouping among the residuals pattern indicating that there is some similarities between particular obsevartions. The Q-Q plot indicates that the ordered deviance residuals line up accurately with the standard normal quantities. Plot 3 and 4 indicate no significant influence points or outliers in the data, as they are all within Cook's distance. Cook's distance is the scaled change in fitted values, and shows the influence of each observation on the fitted response values. It is calculated as follows

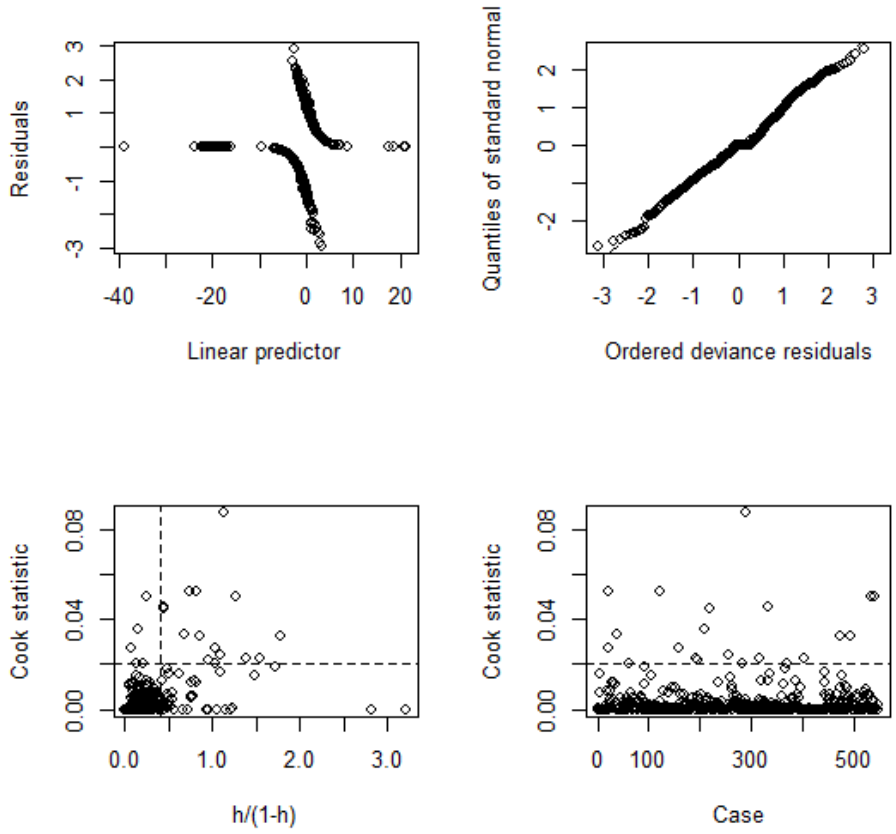


Figure 5.1: Diagnostic plots(Residual vs Fitted, Normal Q-Q, Scale-location, Residuals vs leverage) of the created binary logitsic model predicting clinical vs subclinical mastitis infections using physiological and environmental factors.

$$D_i = \sum_{j=1}^{n} \frac{(\hat{y_j} - y_{\hat{j(i)}})^2}{pMSE}$$

Where $y_{\hat{j}}(i)$ is the fitted response variable not including observation i, MSE is the mean squared error, and p is the number of coefficients in the regression.

An observation with Cook's distance larger than three times the mean Cook's distance would be considered an outlier (Kutner et al. 2004). As no observations bypassed this threshold, it can be concluded that the model includes no outliers.

To ensure that the assumption that none of the independent variables are highly correlated, the variance inflation factor was calculated as described in 2. No variables had a VIF value over 10, the highest was age, which had a value of 7.559807.

To test for independence among the observations, the Duban-Watson test was carried out as explained in chapter2. The test statistic received was 1.23. With a value of 2 indicating no autocorrelation among the variables, the value received for the fitted

model indicates that there is a positive autocorrelation among the variables. This can be explained by the nature of the dataset. The same cow can be recorded as a separate observation each time they contract a mastitis infection. This breaks one of the key assumptions of logistic regression, and invalidates the models accuracy in prediction.

Chapter 6

The Logistic Mixed-Effects Model

6.1 Model Fitting and Results

A mixed-effects logistic model was run with clinical = 0 and subclinical = 1 as the response using the glmer function of the lme4 package in R (Bates et al. 2015). A random intercept was included based on the 'cow' variable, which was the identification code that specified each individual cow. All factors listed in the EDA section, chapter 4, were included in the model as predictor variables. Only full observations were considered by the model, that is, any observations that contained one or more NA value, was ignored by the model. The size of the dataset after the removal of the other variables and the incomplete observations was 356 observations on 28 variables. Of the 354 observations, there were 204 different cow clusters that each could fall into.

The model assumptions are as follows. There is a correlation between observations found within the same cow cluster. Conditional on the radnom effect, the responses are independent and have a Bernoulli distribution.That the residuals are normally distributed. That there is homogeneity of variance.

The formula obtained by the fitted model is as follows

$$\begin{aligned}Predicted\ logit\ of\ (Subclinical_i) &= 1.31439 + RandomIntercept_i + 1.74105 * (Teat_{LH}) + 0.63683 * (Teat_{RF}) + 1.23097 * (Teat_{RF\_LF}) + 1.26967 * (Teat_{RF\_LH}) + \\ &3.98856 * (Teat_{RF\_RH}) + 1.29943 * (Teat_{RH}) - 1.60980 * (Drug_{KANACEF\ M.C}) - \\ &0.05739 * (Drug_{MULTIMAST}) - 1.39149 * (Drug_{SYNULOX}) - 0.13618 * (Drug_{TEREXINE}) - \\ &0.77525 * (Drug_{TERREXINE}) - 3.43382 * (Drug_{TETRA\ DELTA}) - 2.91604 * (Serverity_2) - \\ &3.23056 * (Severity_3) + 0.53749 * (Tubes_2) + 0.11441 * (Tubes_3) + 1.51682 * \\ &(Tubes_4) + 0.50651 * (Tubes_5) + 1.189 * (Tubes_6) + 2.97042 * (Tubes_8) + 1.28127 * \\ &(Farm_2) - 4.20422 * (Farm_3) + 2.09826 * (Farm_6) - 0.31270 * (Farm_9) + 0.03793 *\end{aligned}$$

$(Dry\ Off\ Days) + 0.08092 * (Calves_3) - 1.05136 * (Calves_4) - 1.508 * (Calves_5) - 0.49434 * (Calves_6) - 0.87775 * (Calves_7) - 5.35836 * (Calves_8) - 2.41856\,(Calves_9) - 0.39273 * (Treatment_2) - 0.36926 * (Treatment_3) - 0.0712 * (Treatment_4) - 0.36932 * (Treatment_5) - 2.35774 * (Treatment_{C150}) - 2.31362 * (Treatment_{G250}) - 3.25691 * (Treatment_{TMR}) - 1.21469 * (Subtreatment_1) - 0.75708 * (Subtreatment_2) - 2.30137 * (Subtreatment_3) + 0.59698 * (Age) - 0.55174\,(Morning\ Fat) + 0.72469 * (Morning\ Portein) - 2.05337 * (Morning\ Lactose) - 0.36591 * (SCC) + 0.21841 * (Morning\ Yield) - 0.56585 * (Evening - Morning\ Wait\ Time) + 0.24441 * (Morning\ Max\ Flow) - 0.6767 * (Concentration\ Fed) + 0.60487 * (Evening\ Fat) - 0.0666 * (Evening\ Protein) + 1.49309 * (Evening\ Lactose) + 0.28854 * (Evening\,Yield) + 0.54496 * (Monring - Evening\,Wait\,Time) - 0.76263 * (Evening\ Max\ Flow) - 0.0449 * (Weight) + 0.21336 * (Body\ Condition\ Score)$

Table 6.1: Fixed effects coefficients of the logistic mixed-effects model with mastitis infection type as the response variable (0 signifying a clinical infection, 1 signifying a subsclinical infection), and the cow the infection was observed in as the random effect. Only the intercept and coefficients with a probability lower than 0.05 are included.

Coefficient	β Estimate	Standard Error	Odds Ratio and CI	P-value
Intercept	1.314	9.07	3.72 $(7.1\times10^{-8}, 1.9 \times 10^{8})$	0.884762
TeatLH	1.741	0.715	5.7 (1.4, 23.1)	0.015
TeatRF_RH	3.989	1.72	53.98 (1.86, 155.99)	0.02
Severity2	-2.916	0.75	0.05 (0.012, 0.24)	9.98
Severity3	-3.231	0.919	0.0395 (0.0065,0.24)	0.0004
Evening-Morning Wait	-0.566	0.24	0.568 (0.36, 0.9)	0.0179
Concentration Fed	-0.60487	0.332	0.5 (0.265, 0.975)	0.042

The odds ratio in table 6.1 is be obtained via the same method as used in logistic regression in chapter 5 . The interpretation of the significant variables is based on comparing infection observations within a particular cow, and is as follows

6.2 Significant Variable Interpretation

Teat

When comparing an infection in a specific cow with an infection in the reference teat of left fore, to an infection in her left hind teat, we expect to see a 470% increase in the odds of that observation having subclinical mastitis, this is due to having an odds ratio value of 5.7. As the confidence interval does not contain 1, which would imply no association between a left fore infection and a subclinical diagnosis, we can state that we are 95% confident that there is a relationship between subclincial infections occurring in the left

fore teat relative to the left fore. Converting this to a probability, we receive a value of 0.85. We can interpret this as a specific cow with an infection in her left hind teat, compared to that same cow with an infection in her left fore teat, with all other variables held equal, has an increase in her probability of it being a subclinical mastitis infection by 0.85.

When comparing an infection in a specific cow with an infection in the reference teat of left fore, to an infection in her right fore and hind teat, we expect to see a 5298% increase in the odds of that observation having subclinical mastitis, this is due to having an odds ratio value of 53.98. As the confidence interval does not contain 1, which would imply no association between a right hind and fore infection and a subclinical diagnosis, we can state that we are 95% confident that there is a relationship between subclincial infections occurring in the right hind and fore teat relative to the left fore. Converting this to a probability, we receive a value of 0.98. We can interpret this as a specific cow with an infection in her right hind and fore teat, compared to that same cow with an infection in her left fore teat, with all other variables held equal, has an increase in her probability of it being a subclinical mastitis infection by 0.98.

Severity

When comparing an infection in a specific cow with an infection in the reference severity level of 1, to an infection of severity 2, we expect to see a 95% decrease in the odds of that observation being subclinical mastitis, this is due to having an odds ratio value of 0.05. As the confidence interval does not contain 1, which would imply no association between a infection of severity 2 and a subclinical diagnosis, we can state that we are 95% confident that there is a relationship between subclincial infections occurring in infecions of severity 2 relative to severity 1. Converting this to a probability, we receive a value of 0.048. We can interpret this as a specific cow with an infection of severity 2 compared to that same cow with an infection of severity 1, with all other variables held equal, has an decrease in her probability of it being a subclinical mastitis infection by 0.048.

When comparing an infection in a specific cow with an infection in the reference severity level of 1, to an infection of severity 3, we expect to see a 96.1% decrease in the odds of that observation being subclinical mastitis, this is due to having an odds ratio value of 0.0395. As the confidence interval does not contain 1, which would imply no association between a infection of severity 3 and a subclinical diagnosis, we can state that we are 95% confident that there is a relationship between subclincial infections occurring in infecions of severity 3 relative to severity 1. Converting this to a probability, we receive a value of 0.038. We can interpret this as a specific cow with an infection of severity 3 compared to that same cow with an infection of severity 1, with all other variables held equal, has an decrease in her probability of it being a subclinical mastitis infection by 0.038.

Wait time for milking

For each unit increase in the average evening to morning wait time for milking in a specific cow with an infection, we expect to see a 43.2% decrease in the odds of that observation being subclinical mastitis, this is due to having an odds ratio value of 0.568. As the confidence interval does not contain 1, which would imply no association between a cows average evening to morning wait time for milking and a subclinical diagnosis, we can state that we are 95% confident that there is a relationship between a cows average evening to morning wait time for milking and a subclinical diagnosis. Converting this to a probability, we receive a value of 0.36. We can interpret this as for every unit increase in the evening-morning wait time in a cow, with all other variables held equal, we expect to see a decrease in her probability of her mastitis infection being subclinical by 0.36.

Concentration Fed

For each unit increase in the average concentration fed to a specific cow with an infection, we expect to see a 50% decrease in the odds of that observation being subclinical mastitis, this is due to having an odds ratio value of 0.5. As the confidence interval does not contain 1, which would imply no association between a cows average concentration fed and a subclinical diagnosis, we can state that we are 95% confident that there is a relationship between a cows average concentration fed and a subclinical diagnosis. Converting this to a probability, we receive a value of 0.33. We can interpret this as for every unit increase in the average concentration fed to a cow, with all other variables held equal, we expect to see a decrease in her probability of her mastitis infection being subclinical by 0.33.

6.3 Model Evaluation

Table 6.2: Comparison between logistic regression model and logistic mixed-effects model.

Model	AIC	log Likelihood	Deviance
Logistic Model	415.53	-147.7668	295.53
Mixed-Effects Model	412.9	-145.4	290.9

Table 6.2 shows that the AIC score for the mixed-effects model is lower than that for the logistic model. A model likelihood ratio test shows that the mixed-effects model is significantly better than the logistic model, with a p-value of 0.03. The mixed-effects model receives a pseudo RR^2 value of 0.59 when only the fixed effects are considered. Once the random effects are also considered, that value increases to 0.75. From these results we can conclude that the inclusion of random effects was justified, and that the

mixed effect model fits the data better than the logistic model.

Table 6.3: Random effects of the logistic mixed-effects model with mastitis infection type as the response variable (0 signifying a clinical infection, 1 signifying a subclinical infection), and the cow the infection was observed in as the random effect.

Group	Effect	Number of clusters	Variance	Standard Deviation	p-value
Cow Identification	Intercept	204	2.014	1.419	0.0779

The 204 random intercepts accounted for each cow in the dataset. Table 6.3 shows that the random intercept values had a range of 0.717-1.419, a variance of 2.014, and a standard deviation of 1.419. Calculating the p-value for significance as described in chapter 5 we receive a value of 0.0779. This shows that we fail to reject the null hypothesis that the random intercepts are equal to 0. This is an interesting result when compared with the significance tests that showed how the mixed-effects model is significantly better than the logistic model. This is possibly caused by the low number of cows that were infected multiple times in comparison to the overall number of infections. Only 42.7% of the cows had more than one infection. This smaller sample size for cows with multiple infections makes the standard deviation hard to quantify, and is causing the random effect to be considered insignificant. This could be improved by using a larger dataset where more cows were recording having more than 1 infection, so provide a more accurate estimate of the variance of the random effect.

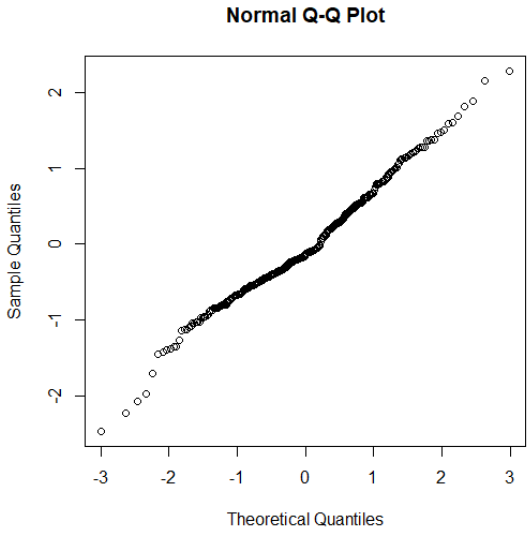


Figure 6.1: Quantile-Quantile plot of the created logistic mixed-effects model predicting clinical vs subclinical mastitis infections using physiological and environmental factors.

Figure 6.1 shows that the quantiles follow a normal distribution and that the assumption of normality is not broken. Figure 6.2 shows that there are no irregularities among

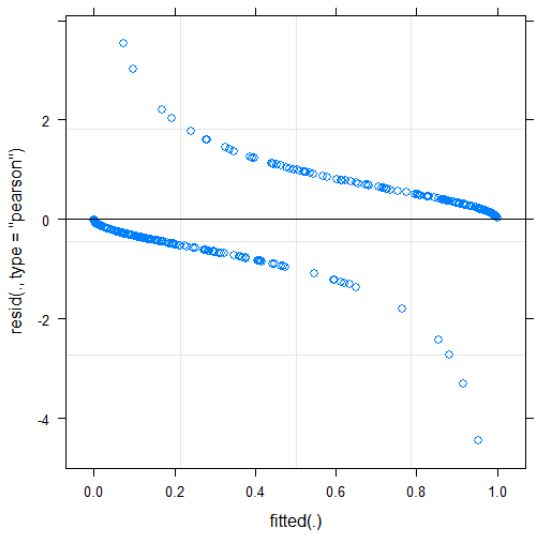


Figure 6.2: Residual vs Fitted plot of the created logistic mixed-effects model predicting clinical vs subclinical mastitis infections using physiological and environmental factors.

the residuals that would indicate that there is heterogeneity of variance, and the grouping among the residuals that was visible in the logistic regression plot (Figure 5.1) is no longer present.

6.4 Model Prediction Accuracy

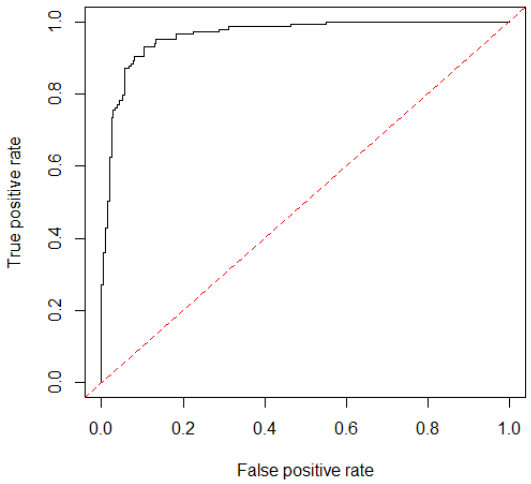


Figure 6.3: Receiver operating characteristic curve of the logistic mixed-effects model predicting clinical vs subclinical mastitis infections using physiological and environmental factors.

Figure 6.3 shows that the true positive and false positive rate is very different from the straight red line. This red line shows what the ROC curve would look like if the model was no better at predicting than random guessing. The total area underneath the curve can be used as a measure of the overall accuracy of the classification. The value received, AUC= 0.96 indicates a very high level of accuracy, as a value of 1 indicates perfect classification, and a value of 0 indicates all wrong classification. The optimal threshold τ was shown to be 0.387. That is, an observation with a probability above 0.387 should be classified as subclinical, and below should be classified as clinical. We can tabulate the predicted versus actual infection types using this optimal threshold value.

Table 6.4: Prediction accuracy of the mixed-effects logistic model at classifying clinical versus subclinical mastitis when compared to the actual diagnoses.

Infection Type	Predicted Clinical Infection	Predicted Subclinical Infection
Clinical Infection	197	12
Subclinical Infection	22	125

Table 6.4 shows that the mixed-effects logistic model correctly classified 197 of the clinical infections, and 125 of the subclinical infections. It incorrectly identified 22 of the subclinical infections as clinical, and 12 of the clinical infections as subclinical. This can be expressed as a 94.25% accuracy in predicting clinical infections, and a 85% accuracy in predicting subclinical infections.

Chapter 7

Discussion

Mastitis infections cause significant financial strain on farmers in Ireland. With subclinical mastitis not easily detected by farmers, understanding the causes lead to a mastsitis infection becoming subclinical or clinical, and then predicting the infection type, is of great importance to the agricultural industry. The goal of this research was to identify what factors are influencing whether a cow gets subclinical or clinincal mastitis. Additionally, this researched aimed to predict whether a cow would contract subclinical or clinical mastitis.

7.1 Logistic Model

A logistic regression was created using the dataset. This regression found that mastitis infections in the left hind teat, in the right fore and hind teat, and in the right hind teat, are all more likely to be subclinical mastitis infections than mastitis infections that are in the left fore teat. Increases in the evening fat content of milk, and in the wait time between morning and evening milking, we also shown to increase the likelihood of an infection being subclinical mastitis. An infection occurring on farm 3 rather than on farm 1, or of severity 2 or 3 rather than of severity 1, were shown to decrease the likelihood of an infection being subclinical mastitis. Additionally, the likelihood of an infection being subclinical drops as the average stomatic cell count rises.

The McFadden's pseudo R^2 value was 0.4628 which indicated an excellent fit. The diagnostic plots did not indicate any issues with outliers or non-normality. However, the residual vs fitted plot showed a small degree of grouping among the residuals.

This grouping in the residuals vs fitted plot is explained by the nature of the dataset, that cows can become infected multiple times. The Durban-Watson test preformed returned a value of 1.23, indicating a positive autocorrelation, also confirmed this. It was therefore necessary to fit a more accurate model that could account for the fact that a

cow could become infected multiple times.

7.2 Mixed-Effects Model

A mixed effects model was created with each cow's identification used as the random effect. This model had coefficients that agreed with the logistic model for certain variables in their significance and effect, but provided different interpretations for others.

An infection in the right fore and hind teat rather than the left fore, was shown again to increase the likelihood of it being a subclinical infection. This mixed-effects model also agreed with the logistic model in determining that infections of severity 2 or 3 compared to severity 1, are less likely to be subclinical infections.

The mixed-effects model found that an infection in the left hind teat has a decreased probability of being a subclinical mastitis infection when compared with one in the left fore teat. This is in direct contrast with the logistic regression which found that it increased the probability. An increase in the average time gap between a cows evening to morning milking was shown to make a diagnosis less likely to be subclinical. As was an increase in the average concentration fed to a cow.

No previous studies have examined the location of the mastitis infection in regards to it being a clinical vs subclinical infection. One study has shown that lesions on the left hind teat were more common than any other teat in buffalo, and that this lead to an increase in mastitis infections (Sharma et al. 2005). It can be hypothesised that this increase in the frequency of lesions in the left hind teat leads to stronger, clinical, mastitis infections rather than subclinical. This would corroborate the results from the mixed-effects regression.

Although no previous studies have confirmed this, it can be hypothesised that the decrease in probability of an infection being subclinical as the severity of the infection increases, is related to the nature of mastitis infections. As the infection becomes more established and therefore of a higher severity, it beings to cause visual symptoms in the infected cow. This then causes the infection to be classified as clinical. This would explain the negative relationship between higher severity infections and a subclinical diagnosis.

This is the first study to find a significant impact of an increased time gap between evening and morning milking the next day. There is similarly little past studies relating the concentration of feed given to a cow and a clinical or subclinical mastitis infection. One study has shown that cows fed $\geq$ 13kg of concentrated feed 15 days after calving

had a higher frequency of clinical mastitis infections than those fed < 13kg of concentrated feed (Nyman et al. 2007). This would agree with the findings of this study, than an increase in the concentration fed lowers the subclinical mastitis chance, and therefore increases the clinical.

The mixed-effects model received a pseudo R^2 value of 0.75, indicating that the model fits the data better than the logistic model, this is also confirmed by the likelihood ratio test showing that the mixed-effects model is significantly better. The residual vs fitted plot no longer shows the grouping structure that was present in the logistic model, as the group structure of different observations from the same cow has been accounted for in this model.

As the mixed-effects model was shown to be the superior model in terms of fit, it was chosen as the method of testing the prediction accuracy for clinical and subclinical mastitis cases. The model recieved a AUC value of 0.96 indicating a high degree of accuracy. 91.25% of the clinical cases were correctly classified, and 85% of the subclinical cases were correctly classified.

7.3 Conclusion

This study has found that the location of a mastitis infection in dairy cows is significantly relevant to that infection being clinical or subclinical, with left hind infections being less likely to be subclinical than left fore infections, and right fore and hind infections being more likely to be subclinical than left fore infections. This study also found that the severity of a mastitis infection is significantly relevant to that infection being clinical or subclinical, with subclinical infections becoming less likely as the severity increases. An increase in the wait period between evening and morning milkings, as well as an increase in the concentration fed, has also been shown to significantly lower the probability of observing a subclinical rather than clinical infection.

This information can be employed by farmers to reduce the number of subclinical mastitis infections by monitoring the individual teats for infection, as the location is relevant, by increasing the gap between the evening and morning milking, and by altering the concentration of feed given to the cows.

7.4 Future Work

To extend this research, an L1-penalised estimation term could be added to the mixed-effects model. This is due to the large number of predictor variables included in the model, in comparison to the amount that were found to be statistically significant. This term would detect variables that do not significantly contribute to the models fit, and reduce their coefficient value to 0, effectively removing them from the regression. This would produce a simpler, more sparse model, and could potentially enhance the prediction accuracy.

Additionally, a study where model selection is chosen by the accuracy of assessing a validation subset of the dataset, that was not used as part of the model fitting process, and where a validation subset used is for determining prediction accuracy, is suggested. This would produce more unbiased accuracy value, as the model would not be predicting outcomes based on the same data that it was built on.

Appendix A

Code

```
setwd("~//Program Files/RStudio/Statistics/thesis")
clinical = read.csv("Subclinical.csv")
library(psych)
library(ggplot2)

summary(clinical)

#####Labelling the clinical levels as text
clinical$subclinical = factor(clinical$subclinical, labels = c("Clinical
    ", "Subclinical"))
#####


#Factors

clinical$subclinical = as.factor(clinical$subclinical)
clinical$Teat = as.factor(clinical$Teat)
clinical$drug = as.factor(clinical$drug)
clinical$serverity = as.factor(clinical$serverity)
clinical$farm = as.factor(clinical$farm)
clinical$treatement = as.factor(clinical$treatement)
clinical$subtreatment = as.factor(clinical$subtreatment)
clinical$tubes = as.factor(clinical$tubes)
clinical$Calfs = as.factor(clinical$Calfs)


##############################
#EDA TILL LINE 675############
##############################

clinicals = subset(clinical, subclinical == "Subclinical")
clinicalc = subset(clinical, subclinical == "Clinical")
```

```
#Teat



#describeBy(clinical$Teat, as.factor(clinical$subclinical))

clinical$Teat = as.factor(clinical$Teat)

boxplot(as.numeric(clinical$subclinical) ~ clinical$Teat)
cdplot(as.factor(subclinical) ~ Teat, data=clinical)

table(subset(clinical, Teat == "LF_")$subclinical)
table(subset(clinical, Teat == "RF_")$subclinical)
table(subset(clinical, Teat == "RH_")$subclinical)
table(subset(clinical, Teat == "LH_")$subclinical)

table(subset(clinical, Teat == "LF_LH_")$subclinical) #More 1 than 0
table(subset(clinical, Teat == "RF_LF_")$subclinical) #More 1 than 0
table(subset(clinical, Teat == "RF_LH_")$subclinical) #More 1 than 0
table(subset(clinical, Teat == "RF_RH_")$subclinical) #More 1 than 0
table(subset(clinical, Teat == "RH_LH_")$subclinical) #More 1 than 0
table(subset(clinical, Teat == "RH_LF_")$subclinical) #More 1 than 0

table(subset(clinical, Teat == "RF_LF_LH_")$subclinical) #LESS THAN 10
table(subset(clinical, Teat == "RF_RH_LF_")$subclinical) #More 1 than 0
table(subset(clinical, Teat == "RF_RH_LH_")$subclinical) #More 1 than 0
table(subset(clinical, Teat == "RH_LF_LH_")$subclinical) #More 1 than 0

table(subset(clinical, Teat == "RF_RH_LF_LH_")$subclinical) #More 1 than
    0



#drug
clinical$drug = as.factor(clinical$drug)
summary(as.factor(clinical$drug))
cdplot(as.factor(subclinical) ~drug, data=clinical)

table(subset(clinical, drug == "BOVACLOX")$subclinical)
table(subset(clinical, drug == "BOVACLOX MC")$subclinical)
table(subset(clinical, drug == "CEPHAGARD")$subclinical)
table(subset(clinical, drug == "CEPRAVIN")$subclinical)
table(subset(clinical, drug == "CEPRAVIN DRY COW")$subclinical)
table(subset(clinical, drug == "DOVACLOX")$subclinical)
```

```
table(subset(clinical, drug == "DUOFAST")$subclinical)
table(subset(clinical, drug == "KANACEF M.C")$subclinical)
table(subset(clinical, drug == "LEO YELLOW")$subclinical)
table(subset(clinical, drug == "MILKING TUBE - TERREXINE")$subclinical)
table(subset(clinical, drug == "MULTIMAST")$subclinical)
table(subset(clinical, drug == "NO DRUGS")$subclinical)
table(subset(clinical, drug == "PATHOCEF")$subclinical)
table(subset(clinical, drug == "SYNULOX")$subclinical)
table(subset(clinical, drug == "SYNULOX TUBES")$subclinical)
table(subset(clinical, drug == "TEREXINE")$subclinical)
table(subset(clinical, drug == "TEREXINS")$subclinical)
table(subset(clinical, drug == "TERREXINE")$subclinical)
table(subset(clinical, drug == "TETRA DELTA")$subclinical)

#severity
summary(clinical$serverity)
boxplot(clinical$serverity ~ clinical$subclinical)

describeBy(clinical$serverity, clinical$subclinical) #Mean higher in 0
    than in 1

cdplot(as.factor(subclinical) ~serverity, data=clinical)

table(subset(clinical, serverity == 1)$subclinical)
table(subset(clinical, serverity == 2)$subclinical)
table(subset(clinical, serverity == 3)$subclinical)

#tubes
summary(clinical$tubes)

describeBy(clinical$tubes, clinical$subclinical) #Mean higher in 1 than
    in 0

cdplot(as.factor(subclinical) ~ tubes, data=clinical)

table(subset(clinical, tubes == "1")$subclinical)
table(subset(clinical, tubes == "2")$subclinical)
table(subset(clinical, tubes == "3")$subclinical)
table(subset(clinical, tubes == "4")$subclinical)
table(subset(clinical, tubes == "5")$subclinical)
table(subset(clinical, tubes == "6")$subclinical)
table(subset(clinical, tubes == "7")$subclinical)
table(subset(clinical, tubes == "8")$subclinical)
table(subset(clinical, tubes == "9")$subclinical)
```

```
boxplot(clinical$tubes ~ clinical$subclinical, xlab = "Infection Type",
    ylab = "Tubes")


#D_calf
boxplot(clinical$D_Calf ~ clinical$subclinical)

describeBy(clinical$D_Calf, clinical$subclinical) #Mean higher in 1 than
     0

#D_Dry
boxplot(clinical$D_Dry ~ clinical$subclinical)

describeBy(clinical$D_Dry, clinical$subclinical) #Mean higher in 1 than
    0

#Calfs
boxplot(clinical$Calfs ~ clinical$subclinical)
cdplot(as.factor(subclinical) ~Calfs, data=clinical)

describeBy(clinical$Calfs, clinical$subclinical) #Mean higher in 1 than
    0
boxplot(clinical$Calfs~ clinical$subclinical, xlab = "Infection Type",
    ylab = "Number of Calves")

table(subset(clinical, Calfs == "1")$subclinical)
table(subset(clinical, Calfs == "2")$subclinical)
table(subset(clinical, Calfs == "3")$subclinical)
table(subset(clinical, Calfs == "4")$subclinical)
table(subset(clinical, Calfs == "5")$subclinical)
table(subset(clinical, Calfs == "6")$subclinical)
table(subset(clinical, Calfs == "7")$subclinical)
table(subset(clinical, Calfs == "8")$subclinical)
table(subset(clinical, Calfs == "9")$subclinical)
table(subset(clinical, Calfs == "10")$subclinical)
summary(clinical$Calfs)



#Dry off Days

summa


clinical$Dry_Off_Days = clinical$Dry_Off_Days/30
```

```
clinical2$Dry_Off_Days = clinical2$Dry_Off_Days/30


boxplot(clinical$Dry_Off_Days ~ clinical$subclinical)

cdplot(as.factor(subclinical) ~Dry_Off_Days, data=clinical)

describeBy(clinical$Dry_Off_Days, clinical$subclinical) #Similar values
summary(clinical2$Dry_Off_Days)

boxplot(clinical2$Dry_Off_Days ~ clinical2$subclinical, xlab = "
    Infection Type", ylab = "Number of Dry Off Days (Scaled)")
hist(clinical2$Dry_Off_Days, xlab = "Number of Dry Off Days (Scaled)",
    main = "")

ggplot(clinical, aes(x = Dry_Off_Days)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Number of Dry Off Days (Scaled)")

summary(clinicalc$Dry_Off_Days)


#Treatment
#Numerical values then also letter codes?
levels(as.factor(clinical$treatement))
summary(as.factor(clinical$treatement))
#NA values are just blank not NA

boxplot(as.factor(clinical$subclinical) ~ as.factor(clinical$treatement)
    )


table(subset(clinical2, treatement == "")$subclinical)
table(subset(clinical2, treatement == "0")$subclinical)
table(subset(clinical2, treatement == "1")$subclinical)
table(subset(clinical2, treatement == "10")$subclinical)
table(subset(clinical2, treatement == "2")$subclinical)
table(subset(clinical2, treatement == "3")$subclinical)
table(subset(clinical2, treatement == "4")$subclinical)
table(subset(clinical2, treatement == "5")$subclinical)
table(subset(clinical2, treatement == "8")$subclinical)
table(subset(clinical2, treatement == "C150")$subclinical)
table(subset(clinical2, treatement == "G250")$subclinical)
table(subset(clinical2, treatement == "TMR")$subclinical)
```

```
table(subset(clinical2, treatement == "L602")$subclinical)




#subtreatment
levels(as.factor(clinical$subtreatment))
summary(as.factor(clinical$subtreatment))

table(subset(clinical, subtreatment == "")$subclinical)
table(subset(clinical, subtreatment == "0")$subclinical)
table(subset(clinical, subtreatment == "1")$subclinical)
table(subset(clinical, subtreatment == "2")$subclinical)
table(subset(clinical, subtreatment == "3")$subclinical)
table(subset(clinical, subtreatment == "4")$subclinical)
table(subset(clinical, subtreatment == "A")$subclinical)
table(subset(clinical, subtreatment == "")$subclinical)


#age

summary(clinical$age)
boxplot(clinical$age ~ clinical$subclinical)
describeBy(clinical$age, clinical$subclinical) #Group 1 older

boxplot(clinical$age ~ clinical$subclinical, xlab = "Infection Type",
    ylab = "Age")

ggplot(clinical, aes(x = age)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Age")


summary(clinicalc$age)
summary(clinicals$age)

#farm
summary(as.factor(clinical$farm))
describeBy(clinical$farm, clinical$subclinical)
 #Not sure how to visualise this one, its categorical but with a lot of
    categories
table(subset(clinical, farm == 1)$subclinical)
table(subset(clinical, farm == 2)$subclinical)
```

```
table(subset(clinical, farm ==3)$subclinical)
table(subset(clinical, farm == 4)$subclinical)
table(subset(clinical, farm == 5)$subclinical)
table(subset(clinical, farm == 6)$subclinical)
table(subset(clinical, farm == 7)$subclinical)
table(subset(clinical, farm == 8)$subclinical)
table(subset(clinical, farm == 9)$subclinical)
table(subset(clinical, farm == 13)$subclinical)
table(subset(clinical, farm == 66)$subclinical)

#NoofCalvesT
summary(clinical$NoOfCalvesT)
boxplot(clinical$NoOfCalvesT ~ clinical$subclinical)
cdplot(as.factor(subclinical) ~NoOfCalvesT, data=clinical)
hist(clinical2$Dry_Off_Days)

#fat_1
summary(clinical$fat_1)
boxplot(clinical$fat_1 ~ clinical$subclinical)
describeBy(clinical$fat_1, clinical$subclinical)   #Similar but 1 is
    higher

boxplot(clinical$fat_1 ~ clinical$subclinical, xlab = "Infection Type",
    ylab = "Morning Fat Content")

ggplot(clinical, aes(x = fat_1)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Morning Fat content")



#fat_2
summary(clinical$fat_2)
boxplot(clinical$fat_2 ~ clinical$subclinical)
describeBy(clinical$fat_2, clinical$subclinical)   #very similar, 0 more
    variable

boxplot(clinical$fat_2 ~ clinical$subclinical, xlab = "Infection Type",
    ylab = "Evening Fat Content")

ggplot(clinical, aes(x = fat_2)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
```

```
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Evening Fat content")



#protein_1
summary(clinical$protein_1)
boxplot(clinical$protein_1 ~ clinical$subclinical)
describeBy(clinical$protein_1, clinical$subclinical)   #very similar, 0
   more variable

boxplot(clinical$protein_1 ~ clinical$subclinical, xlab = "Infection
   Type", ylab = "Morning Protein Content")


ggplot(clinical, aes(x = protein_1)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Morning Protein content")

#protein_2
summary(clinical$protein_2)
boxplot(clinical$protein_2 ~ clinical$subclinical)
describeBy(clinical$protein_2, clinical$subclinical)   #very similar, 0
   more variable

boxplot(clinical$protein_2 ~ clinical$subclinical, xlab = "Infection
   Type", ylab = "Evening Protein Content")

ggplot(clinical, aes(x = protein_2)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Evening Protein content")

#lactose_1
summary(clinical$lactose_1)
boxplot(clinical$lactose_1 ~ clinical$subclinical)
describeBy(clinical$lactose_1, clinical$subclinical)   #very similar

boxplot(clinical$lactose_1 ~ clinical$subclinical, xlab = "Infection
   Type", ylab = "Morning Lactose Content")

ggplot(clinical, aes(x = lactose_1)) +
```

```
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Morning Lactose content")


#lactose_2
summary(clinical$lactose_2)
boxplot(clinical$lactose_2 ~ clinical$subclinical)
describeBy(clinical$lactose_2, clinical$subclinical)   #very similar, 0
   more variable
boxplot(clinical$lactose_2 ~ clinical$subclinical, xlab = "Infection
   Type", ylab = "Evening Lactose Content")

ggplot(clinical, aes(x = lactose_2)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Evening Lactose content")


#casein_1
summary(clinical$casein_1) #561 NA's
boxplot(clinical$casein_1 ~ clinical$subclinical)
describeBy(clinical$casein_1, clinical$subclinical)   #same mean, 0 more
    variable - due to 1 outlier

#casein_2
summary(clinical$casein_2) #561 NA's
boxplot(clinical$casein_2 ~ clinical$subclinical)
describeBy(clinical$casein_2, clinical$subclinical)   #very similar, 0
   more variable

#ffa_1
summary(clinical$ffa_1) #561 NA's
boxplot(clinical$ffa_1 ~ clinical$subclinical)
describeBy(clinical$ffa_1, clinical$subclinical)   #0 higher than 1

#ffa_2
summary(clinical$ffa_2) #561 NA's
boxplot(clinical$ffa_2 ~ clinical$subclinical)
describeBy(clinical$ffa_2, clinical$subclinical)   #0 higher than 1

#ts_1
summary(clinical$ts_1) #561 NA's
```

```
boxplot(clinical$ts_1 ~ clinical$subclinical)
describeBy(clinical$ts_1, clinical$subclinical)   #Same mean, 1 more
    variable

#ts_2
summary(clinical$ts_2) #561 NA's
boxplot(clinical$ts_2 ~ clinical$subclinical)
describeBy(clinical$ts_2, clinical$subclinical)   #similar mean, 0
    higher and more variable

#urea_1
summary(clinical$urea_1) #561 NA's
boxplot(clinical$urea_1 ~ clinical$subclinical)
describeBy(clinical$urea_1, clinical$subclinical)   #1 higher mean, 0
    has low value outliers

#urea_2
summary(clinical$urea_2) #561 NA's
boxplot(clinical$urea_2 ~ clinical$subclinical)
describeBy(clinical$urea_2, clinical$subclinical)   #1 higher mean, 0
    has lowe value outliers

#scc_1

clinical$scc_1 = clinical$scc_1 / 1000


summary(clinical$scc_1) #169 NA's
boxplot(clinical$scc_1 ~ clinical$subclinical)
describeBy(clinical$scc_1, clinical$subclinical)   #0 higher than 1

boxplot(clinical$scc_1 ~ clinical$subclinical, xlab = "Infection Type",
    ylab = "Stomatic Cell Count (Scaled)")

ggplot(clinical, aes(x = scc_1)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Stomatic Cell Count (Scaled)")


#scc_2
summary(clinical$scc_2) #1934!!
boxplot(clinical$scc_2 ~ clinical$subclinical)
describeBy(clinical$scc_2, clinical$subclinical)   #0 higher and more
    varialbe
```

```
#yield_1
summary(clinical$yield_1) #19 NA's
boxplot(clinical$yield_1 ~ clinical$subclinical)
describeBy(clinical$yield_1, clinical$subclinical)   #means very similar
    0 more variable

boxplot(clinical$yield_1 ~ clinical$subclinical, xlab = "Infection Type"
   , ylab = "Change in moring milk yield ")

ggplot(clinical, aes(x = yield_1)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Change in morning milk yield")


#yield_2
summary(clinical$yield_2) #25 NA's
boxplot(clinical$yield_2 ~ clinical$subclinical)
describeBy(clinical$yield_2, clinical$subclinical)   #mean of 0 - does
   this make sense for yield?
boxplot(clinical$yield_2 ~ clinical$subclinical, xlab = "Infection Type"
   , ylab = "Change in evening milk yield ")

ggplot(clinical, aes(x = yield_2)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Change in evening milk yield")

#time_1

clinical$time_1 = clinical$time_1 /60


summary(clinical$time_1) #51 NA's
boxplot(clinical$time_1 ~ clinical$subclinical)
describeBy(clinical$time_1, clinical$subclinical)   #0 higher mean also
   more variable

boxplot(clinical$time_1 ~ clinical$subclinical, xlab = "Infection Type",
    ylab = "Hours from evening milking to morning milking the next day")
```

```
ggplot(clinical, aes(x = time_1)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Hours from evening milking to morning milking the next day")


#time_2

clinical$time_2 = clinical$time_2 /60
summary(clinical$time_2) #165 NA's
boxplot(clinical$time_2 ~ clinical$subclinical)
describeBy(clinical$time_2, clinical$subclinical)   #0 higher mean also
   more variable
boxplot(clinical$time_2 ~ clinical$subclinical, xlab = "Infection Type",
    ylab = "Hours from morning milking to evening milking")

ggplot(clinical, aes(x = time_2)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Hours from morning milking to evening milking")

#max_flow_1
summary(clinical$max_flow_1) #51 NA's
boxplot(clinical$max_flow_1 ~ clinical$subclinical)
describeBy(clinical$max_flow_1, clinical$subclinical)   #very similar
boxplot(clinical$max_flow_1 ~ clinical$subclinical, xlab = "Infection
   Type", ylab = "Starting Maximum Flow")

ggplot(clinical, aes(x = max_flow_1)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Morning maximum flow")



#max_flow_2
summary(clinical$max_flow_2) #165 NA's
boxplot(clinical$max_flow_2 ~ clinical$subclinical)
describeBy(clinical$max_flow_2, clinical$subclinical)   #very similar
```

```
boxplot(clinical$max_flow_2 ~ clinical$subclinical, xlab = "Infection
    Type", ylab = "Final Maximum Flow")

ggplot(clinical, aes(x = max_flow_2)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Evening maximum flow")


#conc_fed_1
summary(clinical$conc_fed_1) #171 NA's
boxplot(clinical$conc_fed_1 ~ clinical$subclinical)
describeBy(clinical$conc_fed_1, clinical$subclinical)   #mean in 0 quite
    higher
boxplot(clinical$conc_fed_1 ~ clinical$subclinical, xlab = "Infection
    Type", ylab = "Morning Conc Fed (Scaled)")

clinical$conc_fed_1 = clinical$conc_fed_1 / 10


ggplot(clinical, aes(x = conc_fed_1)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Morning conc fed (Scaled)")


#conc_fed_2
summary(clinical$conc_fed_2) #173 NA's
boxplot(clinical$conc_fed_2 ~ clinical$subclinical)
describeBy(clinical$conc_fed_2, clinical$subclinical)   #mean in 0 quite
    higher
boxplot(clinical$conc_fed_2 ~ clinical$subclinical, xlab = "Infection
    Type", ylab = "Evening Conc Fed (Scaled)")

clinical$conc_fed_2 = clinical$conc_fed_2 / 10


ggplot(clinical, aes(x = conc_fed_2)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Evening conc fed (Scaled)")
```

```
#row_1
summary(clinical$row_1) #1952 NA's
boxplot(clinical$row_1 ~ clinical$subclinical)
describeBy(clinical$row_1, clinical$subclinical)   #mean in 0

#row_2
summary(clinical$row_2) #1952NA's
boxplot(clinical$row_2 ~ clinical$subclinical)
describeBy(clinical$row_2, clinical$subclinical)

#side_1
summary(clinical$side_1) #1952 NA's
boxplot(clinical$side_1 ~ clinical$subclinical)
describeBy(clinical$side_1, clinical$subclinical)

#side_2
summary(clinical$side_2) #1952NA's
boxplot(clinical$side_2 ~ clinical$subclinical)
describeBy(clinical$side_2, clinical$subclinical)


#unit_1
summary(clinical$unit_1) #1953 NA's
boxplot(clinical$unit_1 ~ clinical$subclinical)
describeBy(clinical$unit_1, clinical$subclinical)

#unit_2
summary(clinical$unit_2) #1952NA's
boxplot(clinical$unit_2 ~ clinical$subclinical)
describeBy(clinical$unit_2, clinical$subclinical)


#qtr_B

clinical$qtr_B = as.factor(clinical$qtr_B)
summary(clinical$qtr_B)
levels(as.factor(clinical$qtr_B))
describeBy(clinical$subclinical, clinical$qtr_B)
cdplot(as.factor(subclinical) ~qtr_B, data=clinical)


#num_of_bugs

summary(as.factor(clinical$num_of_bugs))
levels(as.factor(clinical$num_of_bugs))   #122 potential levels, what is
    the best way to go about this
```

```
#type_bugs

summary(as.factor(clinical$type_bugs))   #41 potential levels


#cmt_RF

summary(clinical$cmt_RF)   #1636 NA's

#cmt_RH

summary(clinical$cmt_RH)   #1603 NA's


#cmt_LF

summary(clinical$cmt_LF)   #1687 NA's

#cmt_LH

summary(clinical$cmt_LH)   #1610 NA's

#SsC_RF

summary(clinical$SSC_RF)   #1022 NA's

#SSC_RH

summary(clinical$SSC_RH)   #1018 NA's


#SSC_LF

summary(clinical$SSC_LF)   #1017 NA's

#SSC_LH

summary(clinical$SSC_LH)   #1013 NA's


#weight

clinical$weight = clinical$weight / 100
```

```
summary(clinical$weight) #41 NA's
boxplot(clinical$weight~ clinical$subclinical)
describeBy(clinical$weight, clinical$subclinical) #1 slightly more
   variable

boxplot(clinical$weight ~ clinical$subclinical, xlab = "Infection Type",
    ylab = "Weight (Kg) Scaled")

ggplot(clinical, aes(x = weight)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Weight (Scaled)")


boxplot(clinical$weight ~ clinical$subclinical, xlab = "Infection Type",
    ylab = "Weight (Kg) Scaled")$out

clinical[clinical$weight == 7.660000,]


#BCS

summary(clinical$BCS) #207 NA's
boxplot(clinical$BCS ~  clinical$subclinical)
describeBy(clinical$BCS, clinical$subclinical) #near identical

boxplot(clinical$BCS ~ clinical$subclinical, xlab = "Infection Type",
   ylab = "Body Condition Score")

ggplot(clinical, aes(x = BCS)) +
  geom_histogram(aes(color = subclinical, fill = subclinical),
                 position = "identity", bins = 15, alpha = 0.4) +
  scale_color_manual(values = c("Dark Blue", "Dark Red")) +
  scale_fill_manual(values = c("Green", "Orange")) +
  xlab("Body Condition Score")



#COW

n_occur <- data.frame(table(clinical$Cow))
mean(n_occur[n_occur$Freq > 1,]$Freq)
summary(n_occur$Freq)
```

```
#Making a new dataset without the high NA value variables

clinical2 = subset(clinical, select=-c(scc_2,row_1, row_2,side_1,side_2,
   unit_1,unit_2,cmt_RF,cmt_RH,cmt_LF,cmt_LH, SSC_RF,SSC_RH,SSC_LF,SSC_
   LH))

#Removing all the levels of variables that have less than 10
   observations

clinical2 = clinical2[!clinical2$Teat == "RF_LF_LH_",]
clinical2 = clinical2[!clinical2$drug == "BOVACLOX MC",]
clinical2 = clinical2[!clinical2$drug == "CEPHAGARD",]
clinical2 = clinical2[!clinical2$drug == "CEPRAVIN DRY COW",]
clinical2 = clinical2[!clinical2$drug == "DOVACLOX",]
clinical2 = clinical2[!clinical2$drug == "DUOFAST",]
clinical2 = clinical2[!clinical2$drug == "LEO YELLOW",]
clinical2 = clinical2[!clinical2$drug == "NO DRUGS",]
clinical2 = clinical2[!clinical2$drug == "TEREXINS",]
clinical2 = clinical2[!clinical2$treatement == "1002",]
clinical2 = clinical2[!clinical2$treatement == "1202",]
clinical2 = clinical2[!clinical2$treatement == "",]
clinical2 = clinical2[!clinical2$treatement == "1206",]
clinical2 = clinical2[!clinical2$treatement == "2.5",]
clinical2 = clinical2[!clinical2$treatement == "23",]
clinical2 = clinical2[!clinical2$treatement == "24",]
clinical2 = clinical2[!clinical2$treatement == "3.5",]
clinical2 = clinical2[!clinical2$treatement == "45",]
clinical2 = clinical2[!clinical2$treatement == "6",]
clinical2 = clinical2[!clinical2$treatement == "60x2",]

clinical2 = clinical2[!clinical2$treatement == "60x6",]
clinical2 = clinical2[!clinical2$treatement == "4.5",]
clinical2 = clinical2[!clinical2$treatement == "GR",]
clinical2 = clinical2[!clinical2$treatement == "Gr25",]

clinical2 = clinical2[!clinical2$treatement == "7",]
clinical2 = clinical2[!clinical2$treatement == "80x2",]
clinical2 = clinical2[!clinical2$treatement == "80x6",]
clinical2 = clinical2[!clinical2$treatement == "9",]
clinical2 = clinical2[!clinical2$treatement == "A",]
clinical2 = clinical2[!clinical2$treatement == "B",]
clinical2 = clinical2[!clinical2$treatement == "C100",]
clinical2 = clinical2[!clinical2$treatement == "C250",]
clinical2 = clinical2[!clinical2$treatement == "Cl15",]
clinical2 = clinical2[!clinical2$treatement == "Cl25",]
clinical2 = clinical2[!clinical2$treatement == "CONT",]
clinical2 = clinical2[!clinical2$treatement == "E602",]
```

```
clinical2 = clinical2[!clinical2$treatement == "E606",]
clinical2 = clinical2[!clinical2$treatement == "EF",]
clinical2 = clinical2[!clinical2$treatement == "ES",]
clinical2 = clinical2[!clinical2$treatement == "GG",]
clinical2 = clinical2[!clinical2$treatement == "GLU",]
clinical2 = clinical2[!clinical2$treatement == "GR25",]
clinical2 = clinical2[!clinical2$treatement == "HiF",]
clinical2 = clinical2[!clinical2$treatement == "L3.5",]

clinical2 = clinical2[!clinical2$treatement == "L606",]
clinical2 = clinical2[!clinical2$treatement == "LF",]
clinical2 = clinical2[!clinical2$treatement == "LoF",]
clinical2 = clinical2[!clinical2$treatement == "LoV",]
clinical2 = clinical2[!clinical2$treatement == "LS",]
clinical2 = clinical2[!clinical2$treatement == "M602",]
clinical2 = clinical2[!clinical2$treatement == "M606",]
clinical2 = clinical2[!clinical2$treatement == "NF",]
clinical2 = clinical2[!clinical2$treatement == "SAL",]
clinical2 = clinical2[!clinical2$treatement == "L3.5",]

clinical2 = clinical2[!clinical2$subtreatment == "",]
clinical2 = clinical2[!clinical2$subtreatment == "2.7",]
clinical2 = clinical2[!clinical2$subtreatment == "2.7L",]
clinical2 = clinical2[!clinical2$subtreatment == "3.5",]
clinical2 = clinical2[!clinical2$subtreatment == "3.5s",]
clinical2 = clinical2[!clinical2$subtreatment == "4",]
clinical2 = clinical2[!clinical2$subtreatment == "5",]
clinical2 = clinical2[!clinical2$subtreatment == "HF",]
clinical2 = clinical2[!clinical2$subtreatment == "HM",]
clinical2 = clinical2[!clinical2$subtreatment == "LF",]
clinical2 = clinical2[!clinical2$subtreatment == "LH",]
clinical2 = clinical2[!clinical2$farm == "66",]
clinical2 = clinical2[!clinical2$farm == "8",]

clinical2$Teat = factor(clinical2$Teat)
clinical2$drug = factor(clinical2$drug)
clinical2$serverity = factor(clinical2$serverity)
clinical2$farm = factor(clinical2$farm)
clinical2$treatement = factor(clinical2$treatement)
clinical2$subtreatment = factor(clinical2$subtreatment)
```

```
summary(as.factor(clinical2$Teat))
summary(as.factor(clinical2$serverity))
summary(as.factor(clinical2$drug))
```

```
summary(as.factor(clinical2$treatement))
summary(as.factor(clinical2$subtreatment))
summary(as.factor(clinical2$farm))



#also removing dates + additional variables that have very high na count
##########
clinical2 = subset(clinical2, select = -c(qtr_B,casein_1, casein_2, ffa_
    1,ffa_2, ts_1,ts_2,urea_1,urea_2,D_Calf, D_Dry, X, exit_Date, date_of
    _infection,date_of_closest_milking,date_of_bacteria_test, date_of_CM_
    test, date_of_SSC_test, date_of_weight, num_of_bugs,type_bugs))

#SCALING

clinical2$yield_1 = clinical2$yield_1 * 10
clinical2$yield_2 = clinical2$yield_2 * 10
clinical2$time_1 = clinical2$time_1 /60
clinical2$time_2 = clinical2$time_2 /60
clinical2$weight = clinical2$weight / 100
clinical2$conc_fed_1 = clinical2$conc_fed_1 / 10
clinical2$conc_fed_2 = clinical2$conc_fed_2 / 10
clinical2$scc_1 = clinical2$scc_1 / 1000
clinical2$Dry_Off_Days = clinical2$Dry_Off_Days/30



clinical2 = na.omit(clinical2)
#Removal of incorrectly recorded dry off observations

clinical2 = clinical2[!clinical2$Cow == "IE151828760007",]
clinical2 = clinical2[!clinical2$Cow == "IE151549212574",]
clinical2 = clinical2[!clinical2$Cow == "IE151549281970",]
clinical2 = clinical2[!clinical2$Cow == "IE151549281814",]
clinical2 = clinical2[!clinical2$Cow == "IE241103420377",]
clinical2 = clinical2[!clinical2$Cow == "IE151549253527",]
clinical2 = clinical2[!clinical2$Cow == "IE141650520781",]
clinical2 = clinical2[!clinical2$Cow == "IE151549282903",]
clinical2 = clinical2[!clinical2$Cow == "IE151549283414",]
clinical2 = clinical2[!clinical2$Cow == "IE151549283001",]
clinical2 = clinical2[!clinical2$Cow == "IE151549282787",]
clinical2 = clinical2[!clinical2$Cow == "IE151549273611",]
clinical2 = clinical2[!clinical2$Cow == "IE151549283001",]
clinical2 = clinical2[!clinical2$Cow == "IE151549282787",]
clinical2 = clinical2[!clinical2$Cow == "IE151549273611",]
clinical2 = clinical2[!clinical2$Cow == "IE151549233475",]
```

```
clinical2 = clinical2[!clinical2$Cow == "IE151549282382",]
clinical2 = clinical2[!clinical2$Cow == "IE151549273272",]
clinical2 = clinical2[!clinical2$Cow == "IE999999921438",]
clinical2 = clinical2[!clinical2$Cow == "IE151828730689",]
clinical2 = clinical2[!clinical2$Cow == "IE151549253790",]
clinical2 = clinical2[!clinical2$Cow == "IE151828740673",]
clinical2 = clinical2[!clinical2$Cow == "IE151549273859",]
clinical2 = clinical2[!clinical2$Cow == "IE151570480836",]
clinical2 = clinical2[!clinical2$Cow == "IE151549275632",]
clinical2 = clinical2[!clinical2$Cow == "IE151549275938",]
clinical2 = clinical2[!clinical2$Cow == "IE151549222633",]
clinical2 = clinical2[!clinical2$Cow == "IE151549244565",]
clinical2 = clinical2[!clinical2$Cow == "IE151549293448",]
clinical2 = clinical2[!clinical2$Cow == "IE151549245720",]
clinical2 = clinical2[!clinical2$Cow == "IE151549297341",]
clinical2 = clinical2[!clinical2$Cow == "IE141425633047",]
clinical2 = clinical2[!clinical2$Cow == "IE151549297069",]
clinical2 = clinical2[!clinical2$Cow == "IE151549265937",]
clinical2 = clinical2[!clinical2$Cow == "IE141425612682",]
clinical2 = clinical2[!clinical2$Cow == "IE151549296962",]
clinical2 = clinical2[!clinical2$Cow == "IE151549296319",]
clinical2 = clinical2[!clinical2$Cow == "IE341381821173",]
clinical2 = clinical2[!clinical2$Cow == "IE151753131486",]
clinical2 = clinical2[!clinical2$Cow == "IE281436292204",]
clinical2 = clinical2[!clinical2$Cow == "IE151828772103",]
clinical2 = clinical2[!clinical2$Cow == "IE201147970456",]
clinical2 = clinical2[!clinical2$Cow == "IE151549271549",]
clinical2 = clinical2[!clinical2$Cow == "IE151549261119",]
clinical2 = clinical2[!clinical2$Cow == "IE151549251869",]
clinical2 = clinical2[!clinical2$Cow == "IE151549293514",]
clinical2 = clinical2[!clinical2$Cow == "IE151549221585",]
clinical2 = clinical2[!clinical2$Cow == "IE151549231900",]
clinical2 = clinical2[!clinical2$Cow == "IE151549231355",]
clinical2 = clinical2[!clinical2$Cow == "IE151549231867",]
clinical2 = clinical2[!clinical2$Cow == "IE151549271499",]
clinical2 = clinical2[!clinical2$Cow == "IE141482260059",]
clinical2 = clinical2[!clinical2$Cow == "IE151549232659",]
clinical2 = clinical2[!clinical2$Cow == "IE151549222245",]
clinical2 = clinical2[!clinical2$Cow == "IE151549231883",]
clinical2 = clinical2[!clinical2$Cow == "IE141407260676",]
clinical2 = clinical2[!clinical2$Cow == "IE141425691881",]
clinical2 = clinical2[!clinical2$Cow == "IE141482240057",]
clinical2 = clinical2[!clinical2$Cow == "IE151549223020",]
clinical2 = clinical2[!clinical2$Cow == "IE151549262447",]
clinical2 = clinical2[!clinical2$Cow == "IE151941660009",]
clinical2 = clinical2[!clinical2$Cow == "IE151549213432",]
clinical2 = clinical2[!clinical2$Cow == "IE151549243848",]
```

```
clinical2 = clinical2[!clinical2$Cow == "IE151549297861",]
clinical2 = clinical2[!clinical2$Cow == "IE151549217193",]
clinical2 = clinical2[!clinical2$Cow == "IE151549253782",]
clinical2 = clinical2[!clinical2$Cow == "IE151549254905",]
clinical2 = clinical2[!clinical2$Cow == "IE151549252413",]
clinical2 = clinical2[!clinical2$Cow == "IE141425691816",]
clinical2 = clinical2[!clinical2$Cow == "IE151549217598",]
clinical2 = clinical2[!clinical2$Cow == "IE151549253056",]
clinical2 = clinical2[!clinical2$Cow == "IE151549284619",]
clinical2 = clinical2[!clinical2$Cow == "IE151549265400",]
clinical2 = clinical2[!clinical2$Cow == "IE151549234011",]
clinical2 = clinical2[!clinical2$Cow == "IE151549283505",]
clinical2 = clinical2[!clinical2$Cow == "IE151549266282",]
clinical2 = clinical2[!clinical2$Cow == "IE151549237063",]
clinical2 = clinical2[!clinical2$Cow == "IE141425662217",]
clinical2 = clinical2[!clinical2$Cow == "IE141425622634",]
clinical2 = clinical2[!clinical2$Cow == "IE151549245407",]



rownames(clinical2)<-1:nrow(clinical2)   # resets numbering so that
   problem observations are easily found



library(boot)
library(car)

logistic1 = (glm(subclinical ~ Teat + drug + serverity + tubes  + farm +
                   Dry_Off_Days + Calfs + treatement + subtreatment +
                      age +
                   fat_1 + protein_1 + lactose_1 + scc_1 + yield_1 +
                      time_1 +
                   max_flow_1 + conc_fed_1 + fat_2 + protein_2 + lactose
                      _2 +
                   yield_2 + time_2 + max_flow_2 + conc_fed_2 + weight +
                       BCS,
                   data = clinical2, family = binomial))

glm.diag.plots(logistic1)
outlierTest(logistic1)
avPlots(logistic1)

dfbetaPlots(logistic1,id.n=1)
```

```
glm.diag(logistic1)$cook[glm.diag(logistic1)$cook > 5000]  #ObservatION
    320 has a very large cooks distance


clinical3 = clinical2[-c(320),]

logistic2 = (glm(subclinical ~ Teat + drug + serverity + tubes  + farm +
                    Dry_Off_Days + Calfs + treatement + subtreatment +
                       age +
                    fat_1 + protein_1 + lactose_1 + scc_1 + yield_1 +
                       time_1 +
                    max_flow_1 + conc_fed_1 + fat_2 + protein_2 + lactose
                       _2 +
                    yield_2 + time_2 + max_flow_2 + conc_fed_2 + weight +
                        BCS,
                    data = clinical3, family = binomial))


glm.diag.plots(logistic2)

outlierTest(logistic2)
avPlots(logistic2)
#remove 248

dfbetaPlots(logistic2,id.n=1)
#remove 542   20 210 420


clinical3 = clinical2[-c(320,310,420),]

logistic2 = (glm(subclinical ~ Teat + drug + serverity + tubes  + farm +
                    Dry_Off_Days + Calfs + treatement + subtreatment +
                       age +
                    fat_1 + protein_1 + lactose_1 + scc_1 + yield_1 +
                       time_1 +
                    max_flow_1 + conc_fed_1 + fat_2 + protein_2 + lactose
                       _2 +
                    yield_2 + time_2 + max_flow_2 + conc_fed_2 + weight +
                        BCS,
                  data = clinical3, family = binomial))

outlierTest(logistic2)
avPlots(logistic2)
dfbetaPlots(logistic2,id.n=1)

library(lmtest)
```

```
lrtest(logistic2)



exp(cbind("Odds ratio" = coef(logistic2), confint.default(logistic2,
    level = 0.95))) # The CI for the odds ratio


#Independence
plot(logistic2$residuals, type = "o")
lag.plot(logistic2$residuals, lags=1, do.lines=FALSE)
car::durbinWatsonTest(logistic2) #Value of 2 is no autocor, greater than
     1 indicates positive autocor which is common in time-series data (
    this)

#Multicollinearity
car::vif(logistic2)


ll.null = logistic2$null.deviance/-2
ll.proposed = logistic2$deviance/-2
(ll.null - ll.proposed)/ ll.null

1-pchisq(2*(ll.proposed - ll.null),df=length(logistic2$coefficients)-1)



#Mixed effects Model


library(lme4)


#This model shows that additional levels need to be removed
mixedeff = glmer(subclinical ~ Teat + drug + serverity + tubes  + farm +
                         Dry_Off_Days + Calfs + subtreatment + age +
                         fat_1 + protein_1 + lactose_1 + scc_1 + yield_1
                            + time_1 +
                         max_flow_1 + conc_fed_1 + fat_2 + protein_2 +
                            lactose_2 +
                         yield_2 + time_2 + max_flow_2 + conc_fed_2 +
                            weight + BCS + (1 | Cow), data = clinical3,
                            family = binomial)

print(mixedeff)
```

```
se <- sqrt(diag(vcov(mixedeff)))

(tab <- cbind(Est = fixef(mixedeff), LL = fixef(mixedeff) - 1.96 * se,
   UL = fixef(mixedeff) + 1.96 *
                se))

exp(tab)

#Issue with Teat two few observations to achieve meanful estimators inf
   in confidence intervals
table(clinical3$Teat)
clinical3 = clinical3[!clinical3$Teat == "LF_LH_",]
clinical3 = clinical3[!clinical3$Teat == "RF_LH",]
clinical3 = clinical3[!clinical3$Teat == "RF_RH_LF_",]
clinical3 = clinical3[!clinical3$Teat == "RF_RH_LF_LH_",]
clinical3 = clinical3[!clinical3$Teat == "RF_RH_LH_",]
clinical3 = clinical3[!clinical3$Teat == "RH_LF_",]
clinical3 = clinical3[!clinical3$Teat == "RH_LF_LH_",]
clinical3 = clinical3[!clinical3$Teat == "RH_LH_",]

clinical3$Teat=factor(clinical3$Teat)

table(clinical3$drug) #remove the ones with no observations
clinical3$drug=factor(clinical3$drug)
clinical3=clinical3[-which(clinical3$drug=="SYNULOX TUBES"),]
clinical3$drug=factor(clinical3$drug)

table(clinical3$tubes)
ind=which(clinical3$tubes==7)
clinical3=clinical3[-ind,]
ind=which(clinical3$tubes==9)
clinical3=clinical3[-ind,]

clinical3$tubes=factor(clinical3$tubes)

table(clinical3$farm)
ind=which(clinical3$farm==7)
clinical3=clinical3[-ind,]
ind=which(clinical3$farm==4)
clinical3=clinical3[-ind,]
ind=which(clinical3$farm==13)
clinical3=clinical3[-ind,]
clinical3$farm=factor(clinical3$farm)

table(clinical3$subtreatment)
ind=which(clinical3$subtreatment=="A")
clinical3=clinical3[-ind,]
```

```
clinical3$subtreatment=factor(clinical3$subtreatment)

table(clinical3$treatement)
ind=which(clinical3$treatement=="0")
clinical3=clinical3[-ind,]
ind=which(clinical3$treatement=="10")
clinical3=clinical3[-ind,]
ind=which(clinical3$treatement=="8")
clinical3=clinical3[-ind,]
ind=which(clinical3$treatement=="L602")
clinical3=clinical3[-ind,]

clinical3$treatement=factor(clinical3$treatement)

mixedeff = glmer(subclinical ~ Teat + drug + serverity + tubes  + farm +
                   Dry_Off_Days + Calfs  +treatement + subtreatment +
                      age +
                   fat_1 + protein_1 + lactose_1 + scc_1 + yield_1 +
                      time_1 +
                   max_flow_1 + conc_fed_1 + fat_2 + protein_2 + lactose
                      _2 +
                   yield_2 + time_2 + max_flow_2 + conc_fed_2 + weight +
                       BCS + (1 | Cow), data = clinical3, family =
                      binomial)

print(mixedeff)

se <- sqrt(diag(vcov(mixedeff)))

(tab <- cbind(Est = fixef(mixedeff), LL = fixef(mixedeff) - 1.96 * se,
   UL = fixef(mixedeff) + 1.96 *
                 se))

exp(tab)



#We now have a dataset that can correctly be modelled for both logistic
   and mixed effects. The final versions of each using this dataset
   follows

#Logistic

logisticfinal = (glm(subclinical ~ Teat + drug + serverity + tubes  +
   farm +
                   Dry_Off_Days + Calfs + treatement + subtreatment +
                      age +
```

```
                    fat_1 + protein_1 + lactose_1 + scc_1 + yield_1 +
                       time_1 +
                    max_flow_1 + conc_fed_1 + fat_2 + protein_2 + lactose
                       _2 +
                    yield_2 + time_2 + max_flow_2 + weight + BCS,
                  data = clinical3, family = binomial))

summary(logisticfinal)


outlierTest(logisticfinal)
avPlots(logisticfinal)
dfbetaPlots(logisticfinal,id.n=1)
glm.diag.plots(logisticfinal)


BIC(logisticfinal)

library(lmtest)

lrtest(logisticfinal)



exp(cbind("Odds ratio" = coef(logisticfinal), confint.default(
   logisticfinal, level = 0.95))) # The CI for the odds ratio


#Independence
plot(logisticfinal$residuals, type = "o")
lag.plot(logisticfinal$residuals, lags=1, do.lines=FALSE)
car::durbinWatsonTest(logisticfinal) #Value of 2 is no autocor, greater
   than 1 indicates positive autocor which is common in time-series data
    (this)

#Multicollinearity
car::vif(logisticfinal)


ll.null = logisticfinal$null.deviance/-2
ll.proposed = logisticfinal$deviance/-2
(ll.null - ll.proposed)/ ll.null

1-pchisq(2*(ll.proposed - ll.null),df=length(logisticfinal$coefficients)
   -1)
```

```
#Mixed Effects

scaled = clinical3
ind <- sapply(scaled, is.numeric)
scaled[ind] <- lapply(scaled[ind], scale)



mixedeff = glmer(subclinical ~ Teat + drug + serverity + tubes  + farm +
                   Dry_Off_Days + Calfs  +treatement + subtreatment +
                      age +
                   fat_1 + protein_1 + lactose_1 + scc_1 + yield_1 +
                      time_1 +
                   max_flow_1 + conc_fed_1 + fat_2 + protein_2 + lactose
                      _2 +
                   yield_2 + time_2 + max_flow_2  + weight + BCS + (1 |
                      Cow), data = clinical3, family = binomial, nAGQ =
                      25)

 summary(mixedeff)
print(mixedeff)

se <- sqrt(diag(vcov(mixedeff)))

(tab <- cbind(Est = fixef(mixedeff), LL = fixef(mixedeff) - 1.96 * se,
   UL = fixef(mixedeff) + 1.96 *
                se))

exp(tab)

aic.glmer <- AIC(logLik(mixedeff))
aic.glm <- AIC(logLik(logisticfinal))
aic.glmer; aic.glm


null.id = -2 * logLik(logisticfinal) + 2 * logLik(mixedeff)
pchisq(as.numeric(null.id), df=1, lower.tail=F)


qqnorm(residuals(mixedeff))



library(arm)
nrow(se.ranef(mixedeff)$Cow)
```

```
max(se.ranef(mixedeff)$Cow)
min(se.ranef(mixedeff)$Cow)

sjPlot(mixedeff)


install.packages("Hmisc")
library(Hmisc)

probs = 1/(1+exp(-fitted(mixedeff)))
probs = binomial()$linkinv(fitted(mixedeff))
somers2(probs, (as.numeric(clinical3$subclinical) -1))


library(ROCR)


predobj = prediction(fitted(mixedeff), clinical3$subclinical)
perf = performance(predobj, "tpr","fpr")
plot(perf)
abline(0,1,col ="red",lty = 2)
auc=performance(predobj,"auc")
auc@y.values

sens = performance(predobj, "sens")
spec = performance(predobj, "spec")
tau = sens@x.values[[1]]
sensSpec = sens@y.values[[1]] + spec@y.values[[1]]
best = which.max(sensSpec)
plot(tau, sensSpec, type = "l")
points(tau[best],sensSpec[best], pch=19)
tau[best]


table(clinical3$subclinical, round(exp(predict(mixedeff))/(1+exp(predict
   (mixedeff)))))


summary(rePCA(mixedeff))

library(jtools)

summ(mixedeff)
```

Bibliography

Agresti, A. (2013), *Categorical data analysis*, John Wiley Sons.

Ashworth, U. S., Forster, T. L. & Luedecke, L. O. (1967), 'Relationship between california mastitis test reaction and composition of milk from opposite quarters.', *Journal of dairy science* 50(7), 1078–82.

Barkema, H. W., Schukken, Y. H., Lam, T. J., Beiboer, M. L., Wilmink, H., Benedictus, G. & Brand, A. (1998), 'Incidence of clinical mastitis in dairy herds grouped in three categories by bulk milk somatic cell counts.', *Journal of dairy science* 81(2), 411–9.

Barlow, W, J., White, L., Zadocks, R. & Schukken, Y. (2009), 'A mathematical model demonstrating indirect and overall effects of lactation therapy targeting subclinical mastitis in dairy herds .', *Preventive Veterinary Medicine* 90(1), 31–42.

Barrett, D. J., Clegg, T., Healy, A. M. & Doherty, M. L. (2006), 'A study of dry cow therapy and effects on scc in 10 irish dairy herds.', *Journal of veterinary medicine. A, Physiology, pathology, clinical medicine* 53(3), 140–4.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015), 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software* 67(1), 1–48.

Breslow, N. E. & Clayton, D. G. (1993*a*), 'Approximate inference in generalized linear mixed models (in applications and case studies)', *Journal of the American Statistical Association* 88(421), 9–25.

Breslow, N. E. & Clayton, D. G. (1993*b*), 'Approximate inference in generalized linear mixed models (in applications and case studies)', *Journal of the American Statistical Association* 88(421), 9–25.

Busato, A., Trachsel, P., Schällibaum, M. & Blum, J. W. (2000), 'Udder health and risk factors for subclinical mastitis in organic dairy farms in switzerland.', *Preventive veterinary medicine* 44(3-4), 205–20.

Cavero, D., Tölle, K., Buxade, C. & Krieter, J. (2006), 'Mastitis detection in dairy cows by application of fuzzy logic.', *Livestock Science* 105, 207–213.

de Felicio Porcionato, M. A., Soares, W. V. B., dos Reis, C. B. M., Cortinhas, C. S., Mestieri, L. & dos Santos, M. V. (2010), 'Milk flow, teat morphology and subclinical mastitis prevalence in gir cows.', *Pesquisa Agropecuária Brasileira* 45(12), 1507–1512.

deMol, R., Kroeze, G., Achten, J., Maatje, K. & Rossing, W. (1997), 'Results of a multivariate approach to automated oestrus and mastitis detection', *Livestock Production Science* 48(3), 219–227.

Dingwell, R. T., Kelton, D. F. & Leslie, K. E. (2003), 'Management of the dry cow in control of peripartum disease and mastitis.', *The Veterinary clinics of North America. Food animal practice* 19(1), 235–65.

Domecq, J. J., Skidmore, A. L., Lloyd, J. W. & Kaneene, J. B. (1997), 'Relationship between body condition scores and conception at first artificial insemination in a large dairy herd of high yielding holstein cows.', *Journal of dairy science* 80(1), 113–20.

Fisher, R. (1918), 'The correlation between relatives on the supposition of mendelian inheritance.', *Transactions of the Royal Society of Edinburgh* 52(2), 399–433.

Fitzgerald, C. (2019), Dairy in the irish economy, Technical report, Teagasc.

Forsbäck, L., Lindmark-M, H., Andrén, A., Akerstedt, M. & Svennersten-Sjaunja, K. (2009), 'Udder quarter milk composition at different levels of somatic cell count in cow composite milk.', *Animal : an international journal of animal bioscience* 3(5), 710–7.

Fox, J. & Weisberg, S. (2019), *An R Companion to Applied Regression*, third edn, Sage, Thousand Oaks CA.
URL: *https://socialsciences.mcmaster.ca/jfox/Books/Companion/*

Geary, U., Lopez-Villalobos, N., Begley, N., McCoy, F., O'Brien, B., O'Grady, L. & Shalloo, L. (2012), 'Estimating the effect of mastitis on the profitability of irish dairy farms.', *Journal of dairy science* 95(7), 3662–73.

Geary, U., Lopez-Villalobos, N., O'Brien, B., Garrick, D. & Shalloo, L. (2013), 'Examining the impact of mastitis on the profitability of the irish dairy industry.', *Irish Journal of Agricultural and Food Research* 525(2), 135–149.

Gebreyohannes, Y., Fekadu, G. R. & B., K. (2010), 'Milk yield and associated economic losses in quarters with subclinical mastitis due to staphylococcus aureus in ethiopian crossbred dairy cows.', *Tropical Animal Health and Production* 42, 925–931.

Glanbia (2017), 'Dry cow management'.
URL: *https://www.glanbiaconnect.com/medias/6468-16-GIN-Dry-Cow-Management-Leaflet-FA-screen.pdf?context=bWFzdGVyfGltYWdlc3wyMjQ0MDQxfGFwcGxpY2F0aW9uL3Bk2*

Gleeson, D., O'Brien, B., Flynn, J., O'Callaghan, E. & Galli, F. (2009), 'Effect of pre-milking teat preparation procedures on the microbial count on teats prior to cluster application.', *Irish veterinary journal* 62(7), 461–7.

Gruet, P., Maincent, P., Berthelot, X. & Kaltsatos, V. (2001), 'Bovine mastitis and intramammary drug delivery: review and perspectives.', *Advanced drug delivery reviews* 50(3), 245–59.

Gröhn, Y. T., Wilson, D. J., González, R. N., Hertl, J. A., Schulte, H., Bennett, G. & Schukken, Y. H. (2004), 'Effect of pathogen-specific clinical mastitis on milk yield in dairy cows.', *Journal of dairy science* 87(10), 3358–74.

Hair, J. F. (2006), *Multivariate data analysis*, Prentice hall.

Harmon, R. J. (1994), 'Physiology of mastitis and factors affecting somatic cell counts.', *Journal of dairy science* 77(7), 2103–12.

Hillerton, J. (1999), 'Redefining mastitis based on somatic cell count.', *Institute for Animal Health* 345(7), 4–6.

Hogeveen, H., Kamphuis, C., Steeneveld, W. & Mollenhorst, H. (2010), 'Sensors and clinical mastitis–the quest for the perfect alert.', *Sensors (Basel, Switzerland)* 10(9), 7991–8009.

Hogeveen, H. & Ouweltjes, W. (2003), 'Sensors and management support in high-technology milking.', *Journal of animal science* 81 Suppl 3, 1–10.

Hortet, P. & Seegers, H. (1998), 'Loss in milk yield and related composition changes resulting from clinical mastitis in dairy cows.', *Preventive veterinary medicine* 37(1-4), 1–20.

Ishikawa, H., Shimizu, T., Hirano, H., Saito, N. & Nakano, T. (1982), 'Protein composition of whey from subclinical mastitis and effect of treatment with levamisole.', *Journal of dairy science* 65(4), 653–8.

James, G., Hastie, T., Witten, D. & Tibshirani, R. (2013), *An Introduction to Statistical Learning: With Applications in R*, Springer Texts in Statistics, Springer London, Limited.

Kamphuis, C., Mollenhorst, H., Heesterbeek, J. A. P. & Hogeveen, H. (2010), 'Detection of clinical mastitis with sensor data from automatic milking systems is improved by using decision-tree induction.', *Journal of dairy science* 93(8), 3616–27.

Kristulaa, M., Curtis, C., Galligan, D. & Bartholomew, C. (1992), 'Use of a repeated-measures logistic regression model to predict chronic mastitis in dairy cows', *Preventive Veterinary Medicine* 14(1), 57–68.

Kumar, N., Manimaran, A., Kumaresan, A., Sakthivel, J., Sreela, L., Mooventhan, p. & Sivaram, M. (2017), 'Mastitis effects on reproductive performance in dairy cattle: a review.', *Tropical Animal Health and Production* 49(1).

Kutner, M. H., Nachtsheim, C. J., Neter, J. & Li, W. (2004), *Applied Linear Statistical Models with Student CD*, Irwin.

Leslie, K. E. & Petersson-Wolfe, C. S. (2012), 'Assessment and management of pain in dairy cows with clinical mastitis.', *The Veterinary clinics of North America. Food animal practice* 28(2), 289–305.

Linzell, J. L. & Peaker, M. (1971), 'Mechanism of milk secretion.', *Physiological reviews* 51(3), 564–97.

M Bruckmaier, R., E Ontsouka, C. & Blum, W. (2012), 'Fractionized milk composition in dairy cows with subclinical mastitis.', *Veterinární medicína* 49(8), 283–290.

MacCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, Chapman Hall, London.

McCulloch, C. E. & Searle, S. R. (2001), *Generalized, Linear, and Mixed Models*, 1 edn, Wiley-Interscience.

McFadden, D. (1973), 'Conditional logit analysis of qualitative choice behavior'.

Mdegela, R. H., Ryoba, R., Karimuribo, E. D., Phiri, E. J., L, T., Reksen, O., Mtengeti, E. & Urio, N. A. (2009), 'Prevalence of clinical and subclinical mastitis and quality of milk on smallholder dairy farms in tanzania.', *Journal of the South African Veterinary Association* 80(3), 163–8.

Miltenburg, J. D., de Lange, D., Crauwels, A. P., Bongers, J. H., Tielen, M. J., Schukken, Y. H. & Elbers, A. R. (1996), 'Incidence of clinical mastitis in a random sample of dairy herds in the southern netherlands.', *The Veterinary record* 139(9), 204–7.

Nielen, M., Schukken, Y. H., Brand, A., Haring, S. & Ferwerda-van Zonneveld, R. T. (1995), 'Comparison of analysis techniques for on-line detection of clinical mastitis.', *Journal of dairy science* 78(5), 1050–61.

Norberg, E., Hogeveen, H., Korsgaard, I. R., Friggens, N. C., Sloth, K. H. M. N. & L, P. (2004), 'Electrical conductivity of milk: ability to predict mastitis status.', *Journal of dairy science* 87(4), 1099–107.

Nyman, A., T, E., Emanuelson, U., Gustafsson, H., Holteinus, K., K, P. W. & Hallen-Sandgren, C. (2007), 'Risk factors associated with the incidence of veterinary-treated clinical mastitis in swedish dairy herds with a high milk yield and a low prevalence of subclinical mastitis', *Preventive Veterinary Medicine* 78(2), 142–160.

Ogola, H., Shitandi, A. & Nanua, J. (2007), 'Effect of mastitis on raw milk compositional quality.', *Journal of veterinary science* 8(3), 237–42.

Pantoja, J., Hulland, C. & Ruegg, P. (2008), 'Somatic cell count status across the dry period as a risk factor for the development of clinical mastitis in the subsequent lactation.', *Journal of Dairy Science* 92(1), 139–148.

Pearl, R. & Reed, L. (1920), 'On the rate of growth of the population of the united states since 1790 and its mathematical representation', *Proceedings of the National Academy of Sciences of the United States of America* 6(6), 275.

Pinedo, P., Risco, C. & Melendez, P. (2011), 'A retrospective study on the association between different lengths of the dry period and subclinical mastitis, milk yield, reproductive performance, and culling in chilean dairy cows', *Journal of Dairy Science* 94(1), 106–115.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Santos, M. V., Ma, Y. & Barbano, D. M. (2003), 'Effect of somatic cell count on proteolysis and lipolysis in pasteurized fluid milk during shelf-life storage.', *Journal of dairy science* 86(8), 2491–503.

Schepers, A. J., Lam, T. J., Schukken, Y. H., Wilmink, J. B. & Hanekamp, W. J. (1997), 'Estimation of variance components for somatic cell counts to determine thresholds for uninfected quarters.', *Journal of dairy science* 80(8), 1833–40.

Sharma, S., Singh, K., Bansal, B. & Sharma, D. (2005), 'Clinical symptomatology and epidemiological observations on teat skin lesions in buffaloes', *Buffalo Bulletin* 24(1), 12–16.

Steeneveld, W., Hogeveen, H., Barkema, H. W., van den Broek, J. & Huirne, R. B. M. (2008), 'The influence of cow factors on the incidence of clinical mastitis in dairy cows.', *Journal of dairy science* 91(4), 1391–402.

Sun, Z., Samarasinghe, S. & Jago, J. (2010), 'Detection of mastitis and its stage of progression by automatic milking systems using artificial neural networks.', *The Journal of dairy research* 77(2), 168–75.

Svensson, C., Nyman, A.-K., Persson Waller, K. & Emanuelson, U. (2006), 'Effects of housing, management, and health of dairy heifers on first-lactation udder health in southwest sweden.', *Journal of dairy science* 89(6), 1990–9.

Tapia, A., Leiva, V. & del Pilar Diaz, M, G. V. (2018), 'Influence diagnostics in mixed effects logistic regression models', *TEST* 28, 920–942.

Thompson-Crispi, K., Atalla, H., Miglior, F. & Mallard, B. A. (2014), 'Bovine mastitis: frontiers in immunogenetics.', *Frontiers in immunology* 5, 493.

Tse, C., Barkema, H., DeVries, T., Rushen, J. & Pajor, E. (2018), 'Impact of automatic milking systems on dairy cattle producers' reports of milking labour management, milk production and milk quality .', *Animal* 12(12), 2649–2656.

Vasileiou, N., Cripps, P., Ioannidi, K., Chatzopoulos, D., Gougolis, D. & Sarrou, S. (2018), 'Extensive countrywide field investigation of subclinical mastitis in sheep in greece', *Journal of Dairy Science* 101(8), 7297–7310.

Vazquez, A. I., Gianola, D., Bates, D., Weigel, K. A. & Heringstad, B. (2009), 'Assessment of poisson, logit, and linear models for genetic analysis of clinical mastitis in norwegian red cows.', *Journal of dairy science* 92(2), 739–48.

Verhulst, P.-F. (1838), 'Notice sur la loi que la population poursuit dans son accroissement', *Corresp. Math. Phys.* 10, 113–126.

Verhulst, P.-F. (1845), 'La loi d'accroissement de la population', *Nouv. Mem. Acad. Roy. Soc. Belle-lettr. Bruxelles* 8(1).

Verhulst, P.-F. (1847), 'Deuxième mémoire sur la loi d'accroissement de la population. mém. acad. r', *Sci. Lett. B.-Arts Belg* 20, 142–173.

White, D. & McDermott, P. (2001), 'Emergence and transfer of antibacterial resistance.', *Journal of dairy science* 84, 151–155.

Wolfinger, R. & O'Connell, M. (1993), 'Generalized linear mixed models a pseudo-likelihood approach', *Journal of Statistical Computation and Simulation* 49, 233–243.

Yan, X. & Su, X. G. (2009), *Linear regression analysis: theory and computing*, World Scientific Publishing Company.

Zadoks, R. N., Allore, H. G., Barkema, H. W., Sampimon, O. C., Wellenberg, G. J., Gröhn, Y. T. & Schukkent, Y. H. (2001), 'Cow- and quarter-level risk factors for streptococcus uberis and staphylococcus aureus mastitis.', *Journal of dairy science* 84(12), 2649–63.

Zweig, M. H. & Campbell, G. (1993), 'Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine.', *Clinical chemistry* 39(4), 561–577.