Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Ingrid Langevei Mæland

A study of regression models for count data, with applications to salmon lice data

Master's thesis in Applied Physics and Mathematics
Supervisor: Thea Bjørnland
June 2023

NTNU
Norwegian University of
Science and Technology

Ingrid Langevei Mæland

A study of regression models for count data, with applications to salmon lice data

Master's thesis in Applied Physics and Mathematics
Supervisor: Thea Bjørnland
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Preface

During the spring of 2023, I completed my master's thesis, which also marks the end of my studies at Norwegian University of Science and Technology (NTNU) in Trondheim. My five-year journey as a student have been a wonderful experience, and I am grateful for all the knowledge I have gained and the friends I have made. I wrote this thesis in collaboration with *Taskforce Salmon Lice*, on an R&D project named *RegLus*.

Having the chance to finalize my studies through the composition of a thesis centred around the salmon farming industry has been truly interesting. Working on this thesis has not only enhanced my comprehension of theoretical statistics but also illuminated its practical applications in various fields. Additionally, it has greatly improved my skill in the `R` programming language. Furthermore, I have acquired extensive knowledge about the intricacies of the salmon farming industry.

A special thanks goes to my supervisor Thea Bjørnland, at the Department of Mathematical Science at NTNU. Thank you all for sharing your knowledge and your patience with me. The master thesis served as a continuation of my previous project assignment (Mæland 2022), which also involved the guidance of Thea Bjørnland as my supervisor. Finally, I would like to thank my friends and family for supporting me throughout my studies.

Abstract

The Norwegian aquaculture industry faces a significant challenge with the prevalence of salmon lice. The salmon louse *Lepeophtheirus salmonis* harm both wild and farmed salmon by reducing their growth rate and ultimately causing their death. To address this issue, the industry is experimenting with various methods and operating models to gain control over the salmon lice. In addition, the Norwegian public authorities have implemented regulations related to lice management to minimize the negative impacts of salmon lice and promote ethical and efficient production of farmed salmon. One objective of this thesis is to investigate different regression models for count data using a baseline count variable. In addition to this, we are interested in applying the regression models to study the effect of the different non-medicinal treatment methods. Finally, we are interested in resuming the studies done in Mæland 2022 on re-infestation of salmon lice after delousing.

In this thesis, the development of salmon lice at 35 distinct locations in the Trøndelag region between 2018 and 2019 have been studied. Count data obtained from various salmon farms were analysed to examine the prevalence of salmon lice. The recorded number of salmon lice have been compared to different explanatory variables to determine their impact on the response variable.

To investigate the various delousing treatment methods and the re-infestation of salmon lice, generalized linear models including Poisson and negative binomial regression models, and multiple linear and random intercept models have been fitted to the observed count data. The results obtained in the thesis suggests that a multiple linear and random intercept model with a log-transformed response variable seemed to fit the salmon lice count data, while the Poisson and negative binomial models led to a poor model fit. According to the multiple linear model, it appeared that the *Optilicer* treatment method performed better than the *LiceFlusher* method, but when adding a random intercept to account for location based clustering, there were no indications that any of the treatment methods were better than the other. In the studies of re-infestation of mobile lice, the results suggested that the temperature of the sea, average weight, and the placement of salmons after delousing was associated with re-infestation. This result did not completely coincide with the results obtained in Mæland 2022, where also lice skirt was associated with re-infestation.

Sammendrag

Den norske lakseoppdrettsindustrien står overfor en betydelig utfordring med forekomst av lakselus (*Leopeohtheirus salmonis*). Lakselusen skader både villaks og oppdrettslaks ved å redusere vekstraten deres og til slutt forårsake deres død. For å takle denne utfordringen eksperimenterer industrien med ulike metoder og driftsmodeller for å få kontroll over utbredelsen av lakselus. I tillegg har norske myndigheter innført forskrifter knyttet til håndtering av lakselus for å minimere de negative konsekvensene av lakselus og fremme etisk og effektiv produksjon av oppdrettslaks.

Et mål med denne avhandlingen er å undersøke ulike regresjonsmodeller for telledata ved hjelp av en grunnlinjetellingsvariabel (baseline). I tillegg er vi interessert i å anvende regresjonsmodellene for å studere effekten av ulike ikke-medikamentelle behandlingsmetoder. Til slutt er vi interessert i å gjenoppta studiene gjort i Mæland 2022 om re-smitte av lakselus etter avlusning.

I dette prosjektet har utviklingen av lakselus ved 35 ulike lokasjoner i Trøndelag-regionen i perioden mellom 2018 og 2019 blitt studert. Telledataen som er innhentet fra ulike lakseoppdretterier har blitt analysert for å undersøke forekomsten av lakselus. Det registrerte antallet lus er blitt sammenlignet med ulike forklaringsvariabler for å bestemme deres innvirkning på responsvariabelen.

For å undersøke de ulike avlusningsmetodene og resmitte av lakselus har vi tilpasset generaliserte lineære modeller, inkludert Poisson- og negativ binomisk regresjonsmodeller, samt multiple lineære og random intercept modeller, til de observerte telledataene. Resultatene fra denne avhandlingen antyder at en multippel lineær og random intercept-modell med en log-transformert responsvariabel synes å være best egnet til å beskrive tellingsdataene for lakselus. Derimot ga Poisson- og negative binomiske modeller en dårlig tilpasning av dataen. Den multiple lineære modellen tydet på at *Optilicer* behandlingsmetoden presterte bedre enn *LiceFlusher*-metoden. Dette resultatet var ikke i samsvar med resultatene fra random intercept modellen, som ga ingen indikasjon på at noen av behandlingsmetodene var bedre enn de andre. I studien om re-smitte av lakselus antydet resultatene at sjøtemperatur, gjennomsnittlig vekt på laksen og plassering av laksen etter avlusning var assosiert med re-smitte. Dermed stemte ikke resultatene helt overens med de i Mæland 2022, hvor også lakseskjørt var koblet til re-smitte.

Table of Contents

List of Figures

List of Tables

1 Introduction

Please note that Sections 1 and 2 are substantially revised versions of Mæland 2022, except Section 2.4 which is new. Section 3 is also new and Section 4 is inspired by Mæland 2022, but concerns for the most part a new analysis.

1.1 Background

The Norwegian aquaculture industry has since the 1970s been facing a major issue with the salmon louse *Lepeophtherius salmonis* (Krøyer 1837), commonly referred to as salmon lice (Hamre et al. 2013, Thorvaldsen, Frank and Sunde 2019). In recent years, the presence of *Caligus elongatus* (Normann, 1832) has posed challenges for salmon farmers as well (Gaasø 2019, Hemmingsen et al. 2020). The parasites *Lepeophtherius salmonis* and *Caligus elongatus* will be collectively referred to as sea lice, while the term salmon lice will exclusively denote *Lepeophtherius salmonis*. In recent years, the growth of salmon farming has led to better conditions for the parasites to grow and spread, compared to their natural environment in seawater (Torrissen et al. 2013). The salmon lice attach to the skin of salmon (both wild and farmed), feeding on their blood, skin, and tissue. They cause skin lesions, tissue damage, and impaired movement. Infested salmon may experience reduced growth rates, delayed maturation, and increased vulnerability to other diseases, which can eventually lead to death (Finstad et al. 2011, Forseth et al. 2017). The life cycle of the salmon lice consists of eight stages, and these are classified into three developmental categories: sessile, mobile, and adult female lice (Hamre et al. 2013).

To control the population of salmon lice in aquaculture facilities worldwide, strict lice control regimes have been put in place. These regimes require all salmon farms to count and report the average number of salmon lice per salmon in the facility every week. In Norway, the salmon lice must be counted on at least ten random salmons in each cage, and the average count is referred to as the lice number. The counts must be reported for the three categories of developmental stages.

According to the regulations set by *The Ministry of Trade, Industry and Fisheries*, there should be no more than 0.5 adult female lice on average per salmon in the facility at all times. The restrictions are specifically imposed on adult female lice because they are the most prolific egg producers and play a significant role in the reproduction and population growth of lice. To ensure that the limit is not exceeded, measures such as delousing treatments and preventive measures must be implemented. The preventative measures aim to protect the salmon farms against salmon lice, and the two most commonly preventative measures include lice skirts and the use of cleaner fish. Lice skirts are typically made of a fine mesh material and are attached to the top of the salmon cages, creating a barrier that prevents sea lice from accessing the salmon. Cleaner fish are used in the salmon industry as a natural method of controlling sea lice. The cleaner fish, which are typically species such as wrasse (*Labridae*) or lumpfish (*Cyclopterus lumpus*), eat the salmon lice off the salmon helping to keep the parasite under control. The aim of the delousing treatments is to reduce the pressure of lice in cages where the lice pressure is high. The treatments can be split into five categories: Bath treatment, oral treatment, lice flusher, freshwater treatment and thermic treatment. The bath and oral treatments are referred to as medicinal treatment and has been extensively used to fight the problem of sea lice. Bath treatments are performed in two ways. The first method is an in-cage treatment where the salmon cage is lined with a tarpaulin and the volume of the water within the cage is reduced. The other method is a well-boat treatment and includes crowding and then pumping the salmon into a well boat. Then for both methods, the recommended treatment concentration for the chemotherapeutant is added and the salmon is held in the bath for the treatment period. After treatment the tarpaulin is removed, or the salmon are pumped out and the chemotherapeutant is released into the water. The oral treatment includes all treatments where chemotherapeutants are delivered through fish feed. The extensive use of medicinal treatments against salmon lice has led to the salmon lice developing a resistance towards the delousing chemicals. Thus, the non-medicinal methods lice flusher, freshwater treatment and thermic treatment are more used among salmon farmers worldwide today and make up the treatments considered in this thesis.

The lice flusher method is a delousing treatment that involves the use of a specialized vessel or barge equipped with high-pressure water jets to remove lice from the salmon. The freshwater treatment exploits the fact that salmon lice are sensitive to fresh water and generally cannot survive when water salinity is very low. By temporarily exposing salmon to fresh water, the salmon lice detach and can then be removed. Salmon lice release their hold at high water temperatures. This is exploited in the thermic treatment method, where the salmon is transferred to a treatment tank with heated salt water between 28° and 34° for about 30 seconds so that the salmon lice die and fall of the salmon.

1.2 The *RegLus*-project

The project *Taskforce Salmon Lice* aims to establish knowledge on how salmon lice spread within and between salmon farms. Project *RegLus* is deployed as a part of *Taskforce Salmon Lice* with the focus on studying salmon cages that are being treated for salmon lice with non-medicinal methods. The *RegLus*-project wants study and map the salmon lice throughout the delousing process and identify the stages in the delousing process where variations in the level of salmon lice can be observed. To do this, they have collected data from salmon lice counting's from delousing units using non-medicinal methods from 2018 to 2019. The counts of salmon lice were registered at four distinct time points during delousing treatments. We refer to the counting's done at the first time point as the first count, counting's done at the second time point as the second count, and so on.

The first count, henceforth referred to as the "0-sample", should ideally have been performed shortly before the delousing treatment started on a random sample of 20 fish from the cage. The second count, from now on referred to as the "1-sample", was done during crowding; this is the process when the salmon is gathered in a crowding net before delousing treatment. Figure 1 illustrates a crowding net in a delousing unit. One aim of the *RegLus*-project is to study the claim that some salmon lice fall off the salmon during the crowding process. A decline in the count of salmon lice between the second and the first count, i.e., between the 1-sample and the 0-sample, would support this claim. The third count, henceforth referred to as the "2-sample" was made on the delousing unit after treatment and before the salmon were placed back into a cage. Another objective of the *RegLus*-project is to investigate the effect of the various non-medicinal delousing treatment methods; lice flusher-, freshwater- and thermic treatment. This can be studied by investigating the prevalence of salmon lice in the 2-sample. Our main focus in this thesis will be on the 2-sample as compared to the 1-sample. Salmon lice can survive in the sea for some time without a host, and potentially re-attach to a new host. The fourth count, from now on referred to as the "3-sample", should ideally have been taken on a random sample of 20 fish from the cage within 40 hours after treatment. The *RegLus*-project aim to quantify re-infestation of salmon lice. An increase in *mobile* lice between the 2-sample and the 3-sample could be taken as evidence of re-infestation. This was studied in Mæland 2022 and will also be studied further in this thesis. We specify that for each of the four samples, the three salmon lice stages, mobile-, sessile- and adult female lice, and *Caligus elongatus*, were counted separately. This means that each sample contains four different counts.

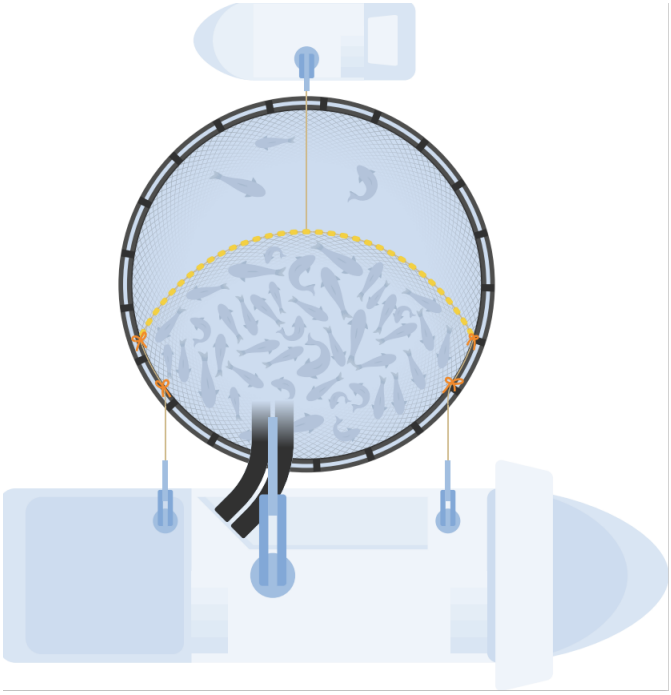


Figure 1: *Visualization of a crowding net in a salmon cage attached to a delousing unit.*

Source: Marit Nersten, 2021

1.3 Outline and aim of this thesis

Section 2 provides some details regarding statistical models for count data. Specifically, we will focus on Poisson regression, negative binomial regression and linear regression on log-transformed count data, including random intercept models. In this section, information is also provided regarding hypothesis testing and model validation. In Section 3, we introduce the concept of a baseline count and investigate how to best include a baseline in a regression analysis with relevant theory and simulation studies. Details regarding the dataset, pre-processing process and descriptive statistics is given in Section 4. Section 4 also provides the data analysis and results. Finally, a discussion with recommendations for further work is presented in Section 5.

2 Regression models for count data

The theory in this section is based on Fahrmeir et al. 2013 unless otherwise specified. In this section we will introduce three commonly used regression models for count data; Poisson, negative binomial and linear regression on log-transformed counts. These models are part of the wider framework of generalized linear models (GLM). In GLM, one assumes that the distribution of the response variable Y belongs to the exponential family, and that the mean, μ, is related to a linear predictor $\eta = \boldsymbol{x}^T\beta$ via a link function $g()$, such that $\mu = g^{-1}(\eta)$. The linear predictor is a linear combination of p covariates including the intercept. We will also introduce the random intercept model, which incorporates group-level variability and within-group dependencies by estimating unique intercepts for each group.

In this thesis, we consider response variables Y that are counts, that is, our response variables are non-negative integers $Y \in \{0, 1, 2, 3, ...\}$. In the application, the response variable is the number of salmon lice count on a sample of n salmon (typically $n = 20$). Therefore, we consider models where each count variable has a so-called exposure unit attached to it. This exposure unit depends on the context the data is collected. An exposure unit can refer to the amount of time it takes to measure a unit or to the sample size of the observational units. In the case of counting salmon lice, the exposure unit therefore refers to the sample size of salmon that is used to estimate the salmon lice number in each cage.

2.1 Poisson Regression

A Poisson regression model is typically used to model count variables. In the context of generalized linear models, this leads to Poisson regression models. In the following, we assume that we have count responses Y_i which are based on n_i exposure units. This information is available, therefore the values of $n_i > 0$ are known. For counts Y_i collected on a sample size $(0, n_i]$, we use n_i as an exposure unit for observation Y_i. In addition, we have p covariates $\boldsymbol{x}_i$ including the intercept for $i = 1, ..., n$ available. This allows us to formulate the following Poisson regression model.

$$Y_i \sim \text{Poisson}(n_i\lambda(\boldsymbol{x}_i, \boldsymbol{\beta}))\ i = 1, ..., n \text{ independent with} \tag{1}$$

$$\text{P}(Y_i = y_i) = \exp(-n_i\lambda(\boldsymbol{x}_i, \boldsymbol{\beta}))\frac{(n_i\lambda(\boldsymbol{x}_i, \boldsymbol{\beta}))^{y_i}}{y_i!}, \tag{2}$$

where n_i is known. Further we assume that the unit Poisson rate $\lambda(\boldsymbol{x}_i, \boldsymbol{\beta}) > 0$ satisfies

$$\lambda(\boldsymbol{x}_i, \boldsymbol{\beta}) := \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})) \geq 0 \tag{3}$$

for p unknown regression parameters $\boldsymbol{\beta}$ and known covariates $\boldsymbol{x}_i$. In Poisson regression we also know that the expected response is equal to the variance of the response. Thus we have

$$\text{E}(Y_i) = \mu_i = n_i\lambda(\boldsymbol{x}_i, \boldsymbol{\beta}) = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta} + \ln(n_i)) = \text{Var}(Y_i). \tag{4}$$

In the following, λ_i denotes $\lambda(\boldsymbol{x}_i, \boldsymbol{\beta})$. The log-likelihood for $\boldsymbol{\beta}$ is given by

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log f(y_i|\boldsymbol{\beta}) = \sum_{i=1}^{n} \log\left(\frac{(n_i\lambda_i)^{y_i}\exp(-n_i\lambda_i)}{y_i!}\right) = \sum_{i=1}^{n}[y_i\log(n_i\lambda_i) - n_i\lambda_i - \log(y_i!)] \\ &\propto \sum_{i=1}^{n}[y_i(\boldsymbol{x_i}^T\boldsymbol{\beta} + \log(n_i)) - \exp(\boldsymbol{x_i}^T\boldsymbol{\beta} + \log(n_i))]. \end{aligned} \tag{5}$$

The parameter estimates $\hat{\boldsymbol{\beta}}$ which maximize $l(\boldsymbol{\beta})$ are found via the so-called Fisher scoring algorithm, which involves the score vector

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{s_i}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x_i}(y_i - \exp(\boldsymbol{x_i}^T \boldsymbol{\beta} + \log(n_i))). \tag{6}$$

and the Fisher information matrix

$$\boldsymbol{F}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathrm{Cov}[\boldsymbol{s_i}(\boldsymbol{\beta})]. \tag{7}$$

It can be shown that the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is asymptotically unbiased and multivariate normal. The variance of each estimator is taken as the corresponding diagonal entries of $\boldsymbol{F}(\boldsymbol{\beta})$. Note that the exposure unit n_i is present in $l(\boldsymbol{\beta})$ and $s_i(\boldsymbol{\beta})$ only as an offset to the linear predictor. When using GLM in `R` to fit a Poisson regression model to count data with varying exposures, this is specified using the offset function, i.e. $\texttt{glm}(\texttt{y}_\texttt{i} \sim \texttt{x}_\texttt{i} + \texttt{offset}(\texttt{n}_\texttt{i}))$.

In the event that the counted response variables within the data exhibit a greater degree of variability than we assumed by the Poisson regression model, the model is considered to be overdispersed. This means that $\mathrm{Var}(Y_i) > \mathrm{E}(Y_i) = \lambda_i$, where $\lambda_i = n_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$. To address this, a dispersion parameter ϕ can be introduced by supposing $\mathrm{Var}(Y_i) = \phi \lambda_i$. The estimation of the dispersion parameter can be carried out via the average deviance or the average Pearson statistic of the model:

$$\hat{\phi}_P = \frac{P}{n-p} \quad \text{or} \quad \hat{\phi}_D = \frac{D}{n-p}, \tag{8}$$

where n is the number of observations, p is the number of parameters in the model, D is the deviance and P is the Pearson statistic. Overdispersion is indicated when the dispersion parameter ϕ exceeds 1, whereas under-dispersion is suggested if it is less than 1. The deviance D statistic is given as

$$D = 2\sum_{i=1}^{n} \{y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\} = \sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right) - \sum_{i=1}^{n} (y_i - \hat{\lambda}_i), \tag{9}$$

and the Pearson statistics is defined as

$$P = \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i}. \tag{10}$$

The deviance residual is defined as

$$d_{i,P} = \mathrm{sign}(y_i - \hat{\lambda}_i) \sqrt{2\left[y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right]}, \tag{11}$$

where $\mathrm{sign}(y_i - \hat{\lambda}_i) = 1$ if $y_i - \hat{\lambda}_i > 0$ and $\mathrm{sign}(y_i - \hat{\lambda}_i) = -1$ if $y_i - \hat{\lambda}_i < 0$. The Pearson residuals are defined as

$$r_{i,P} = \frac{(y_i - \hat{\lambda}_i)}{\sqrt{\mathrm{Var}(y_i)}} = \frac{(y_i - \hat{\lambda}_i)}{\sqrt{\hat{\lambda}_i}}. \tag{12}$$

The deviance and Pearson statistics adhere to an approximately χ^2-distribution with $n-p$ degrees of freedom and can be employed to assess the goodness of fit of the model. If D is less than $\chi^2_{\alpha,n-p}$, there is no proof to suggest that the model is not a good fit to the data. Typically, the Pearson statistic defined in Equation (10) serves as a test for overdispersion.

2.2 Negative Binomial Regression

The issues of model overdispersion is typical for count data. We can model overdispersion using a mixing approach. Consider a conditional Poisson regression model given random means and an independent mixing distribution for the random means. In particular, suppose that the random count variable Y_i is Poisson distributed, conditional on the parameter λ_i so that $f(Y_i = y_i|\lambda_i) = \frac{\exp(-n_i\lambda_i)(n_i\lambda_i)^{y_i}}{y_i!}$. Assume that the parameter λ_i is a random variable rather than being a completely deterministic function of $\boldsymbol{x}_i$. In particular, let $\lambda_i = \mu_i\nu_i$, where μ_i is a deterministic function of $\boldsymbol{x}_i$, typically $\mu_i = n_i\exp(\boldsymbol{x}_i\boldsymbol{\beta})$ and $\nu_i > 0$, often referred to as a random subject effect, is i.i.d. with density $g(\nu_i)$. Following Cameron and Trivedi 2005, the marginal density of y_i can be expressed as

$$P(Y_i = y_i) = \int P(Y_i = y_i|\nu_i)g(\nu_i)d\nu_i, \tag{13}$$

where $g(\nu_i)$ is the mixing distribution. Furthermore, let ν_i be gamma-distributed with mean $\mathrm{E}(\nu_i) = 1$ and variation $\mathrm{Var}(\nu_i) = \frac{1}{r}$ so that

$$g(\nu) = \frac{\nu_i^{r-1}\exp(-\nu_i r)r^r}{\Gamma(r)}, \text{ for } r > 0. \tag{14}$$

It then follows that

$$P(Y_i = y_i) = \frac{\Gamma(r + y_i)}{\Gamma(r)\Gamma(y_i + 1)}\left(\frac{r}{r + \mu_i}\right)^r\left(\frac{\mu_i}{\mu_i + r}\right)^{y_i}, y_i = 0, 1, 2, ..., \tag{15}$$

which we can recognize as a negative binomial distribution with,

$$\mathrm{E}(Y_i) = \mu_i \text{ and } \mathrm{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{r},$$

where, as before, $\mu_i = n_i\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$. Note that as r increases, $\mathrm{Var}(Y_i) \to \mathrm{E}(Y_i)$ and the distribution of Y_i is Poisson.

The parameters of interest are $\boldsymbol{\beta}$ and r. Assuming the response variables $Y_i, i = 1, 2, ..., n$ are i.i.d. negative binomial distributed, the log-likelihood function is derived as

$$l(\boldsymbol{\beta}, r) = \sum_{i=1}^{n}\left(\sum_{j=0}^{y_i-1}\log(j+r)\right) - \sum_{i=1}^{n}[\log\Gamma(y_i+1)+r\log r-r\log(\mu_i+r)+y_i\log\mu_i)-y_i\log(\mu_i+r)]. \tag{16}$$

Substituting $\mu_i = n_i\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$ in the log-likelihood function in Equation (16) and taking the derivatives with respect to $\boldsymbol{\beta}$ and r, we obtain the score functions

$$\boldsymbol{s}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, r) = \sum_{i=1}^{n}\left(r\boldsymbol{x}_i\frac{y_i - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta} + \log(n_i))}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta} + \log(n_i)) + r}\right), \tag{17}$$

and

$$\boldsymbol{s}_r(\boldsymbol{\beta}, r) = \sum_{i=1}^{n}\left(\sum_{j=0}^{y_i-1}\frac{1}{j+r}\right)+\log r-\log(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}+\log(n_i))+r)+\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta} + \log(n_i)) - y_i}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta} + \log(n_i)) + r}\right). \tag{18}$$

Following Nakashima 1997, we have that $\boldsymbol{F}_{12}(\boldsymbol{\beta}, r) = \boldsymbol{F}_{21}(\boldsymbol{\beta}, r) = 0$, and

$$\boldsymbol{F}_{11}(\boldsymbol{\beta}, r) = \sum_{i=1}^{n} \frac{r\mu_i \boldsymbol{x}_i \boldsymbol{x}_i^T}{\mu_i + r}, \tag{19}$$

and

$$\boldsymbol{F}_{22}(\boldsymbol{\beta}, r) = \sum_{i=1}^{n} \left(\mathrm{E}\left(\sum_{j=0}^{y_i - 1} \frac{1}{(j+r)^2} \right) - \frac{\mu_i}{r(\mu_i + r)} \right). \tag{20}$$

In order to estimate $\hat{\boldsymbol{\beta}}$, the Fisher scoring algorithm can be used

$$\hat{\boldsymbol{\beta}}^{t+1} = \hat{\boldsymbol{\beta}}^t + \boldsymbol{F}_{11}^{-1}(\hat{\boldsymbol{\beta}}) \boldsymbol{s}(\hat{\boldsymbol{\beta}}^t). \tag{21}$$

Similarly, for $\hat{r}$ one may use

$$\hat{r}^{t+1} = \hat{r}^t + \boldsymbol{F}_{22}^{-1}(\hat{r}^t) \boldsymbol{s}(\hat{r}^t). \tag{22}$$

In `R`, using `glm.nb` to fit the negative binomial regression model, $\boldsymbol{\beta}$ and r are estimated iteratively. An initial value of $\hat{r}$ is set, $\hat{\boldsymbol{\beta}}$ is estimated, then $\hat{\boldsymbol{\beta}}$ is used to update $\hat{r}$, etc until convergence of both.

When $t \to \infty$, the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{r}$ follows the asymptotic distribution

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{r} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\beta} \\ r \end{pmatrix}, \begin{pmatrix} \boldsymbol{F}_{11}^{-1}(\hat{\boldsymbol{\beta}}, \hat{r}) & 0 \\ 0 & \boldsymbol{F}_{22}^{-1}(\hat{\boldsymbol{\beta}}, \hat{r}) \end{pmatrix} \right), \tag{23}$$

where $\boldsymbol{F}_{11}(\boldsymbol{\beta})$ and $\boldsymbol{F}_{22}(r)$ are defined in Equation (19) and (20), respectively.

The Pearson statistic of the negative binomial regression model is defined as

$$\mathrm{P} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i + r^{-1}\hat{\mu}_i^2}, \tag{24}$$

and the formula for the Pearson residuals is given as

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i + r^{-1}\hat{\mu}_i^2}} \tag{25}$$

The deviance statistic in the negative binomial is defined as

$$\mathrm{D} = 2 \sum_{i=1}^{n} \left(y_i \log\left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i + r) \log\left(\frac{r + y_i}{r + \hat{\mu}_i} \right) \right). \tag{26}$$

The associated deviance residuals are expressed as

$$d_i = \mathrm{sign}(y_i - \hat{\mu}_i) \sqrt{2 \left(y_i \log\left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i + r) \log\left(\frac{r + y_i}{r + \hat{\mu}_i} \right) \right)}. \tag{27}$$

2.3 Multiple Linear Regression

In a multiple linear regression model, we assume that there is a linear relationship between a response variable and several explanatory variables. The response variable Y_i is a count variable,

which have a naturally skewed or kurtotic distribution. This type of variable tends to generate non-normal distributions. In order to model this in a linear regression model, a data transformation is needed for the counted response variable. A log-transformation of the count variables both improves the normality and the homoscedasticity of the model residuals and the transformation is specified by $\ln(\frac{Y_i+1}{n_i})$, where n_i is as before the exposure for count number i. The extra +1-term is added to avoid problems with the logarithm functions in case of a zero-count. The model, having a log-transformed response variable and not unlogged explanatory variables, is referred to as a log-level regression model. The model is written as

$$\log\left(\frac{Y_i+1}{n_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ip-1} + \epsilon_i, \tag{28}$$

for $i = 1, ..., n$. One can also write this in terms of an offset, i.e.,

$$\log(Y_i + 1) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ip-1} + \log(n_i) + \epsilon_i, \tag{29}$$

where $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. The multiple linear regression model can also be written in a matrix form. Let $\tilde{\boldsymbol{Y}}$ be the vector of means $\frac{Y_i+1}{n_i}, i = 1, ..., n$. Assuming we have n sampling units $(x_{i1}, ..., x_{ik}, y_i), 1 \leq i \leq n$, such that each sampling unit represents an instance of Equation (29), we get

$$\log(\tilde{\boldsymbol{Y}}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{30}$$

The error terms ϵ_i in the multiple linear regression model is assumed to be Gaussian

$$\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}). \tag{31}$$

The parameter vector of interest, $\boldsymbol{\beta}$, is estimated with either the maximum likelihood function or the least squares method. Both these methods give the same estimators when we assume a normal linear regression model, that is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\log(\tilde{\boldsymbol{Y}})$. The distribution of $\hat{\boldsymbol{\beta}}$ is given by $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.

2.4 Random Intercept Models

The theory in this section is based on Fahrmeir et al. 2013 unless otherwise specified. The random intercept model is the simplest model in the family of linear mixed models. The term mixed refers to the use of a mix of fixed and random effects as covariates to model the dependent variable. In general, fixed effects are quantitative covariates which represents the whole population being studied, while random effects are quantitative variables which measures the individual deviation from the population fixed effect.

For simplicity, we first only look at the case of just one covariate x (in addition to the intercept). Let

$$(x_{ij}, y_{ij}), i = 1, ..., m, j = 1, ..., n_i$$

denote the values of the covariate x and response variable y for subjects $j = 1, ..., n_i$ in clusters $i = 1, ..., m$. For modelling the relationship between x and y, we start with the classical linear model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij} \text{ , where } \epsilon_{ij} \text{ i.i.d. } N(0, \sigma^2). \tag{32}$$

In this model, we assume that all observations are independent.

If there is reason for assuming cluster-specific heterogeneity, e.g. that y_{ij} and y_{il} observed for the same cluster should not be independent, we can introduce cluster-specific parameters γ_{0i} and obtain

$$y_{ij} = \overbrace{\beta_0 + \beta_1 x_{ij}}^{\text{fixed part}} + \overbrace{\gamma_{0i} + \epsilon_{ij}}^{\text{random part}}, \tag{33}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ are the standard i.i.d. errors for the classical linear model. Further, β_0 is the fixed intercept, γ_{0i} is the cluster-specific deviation from the fixed intercept β_0. This is a random variable and not a model parameter. Thus, $\beta_0 + \gamma_{0i}$ is the random intercept for cluster i. β_1 is the fixed slope common to all clusters. We assume for the cluster deviation intercept

$$\gamma_{0i} \sim N(0, \tau_0^2)$$

and that γ_{0i} and ϵ_{ij} are independent. The random intercept model therefore comes across as a linear regression model with two error terms, where γ_{i0} is a cluster-level error that is shared between measurements on the same cluster i and ϵ_{ij} is the observation error of the measurement j in cluster i. The presence of a random intercept in the model creates a particular correlation or dependency structure among the responses, y_{ij}. Given the random intercepts γ_{i0}, the y_{ij} are still conditionally independent with

$$y_{ij}|\gamma_{i0} \sim N(\beta_0 + \beta_1 x_{ij} + \gamma_{0i}, \sigma^2).$$

The motivation for including a new random intercept is to ensure that we consider that observations within a cluster are correlated, while those between clusters are independent. We look at the joint marginal distribution of the responses. Following Fahrmeir et al. 2013, measurements y_{ij} for cluster i are correlated with within-subject correlation coefficient, often referred to as the intraclass correlation (ICC)

$$\text{Corr}(y_{ij}, y_{il}) = \frac{\tau_0^2}{\tau_0^2 + \sigma^2}, j \neq l. \tag{34}$$

Let $\boldsymbol{y}_i$ represent the vector of responses in cluster i, and $\boldsymbol{X}_i$ the $n_i \times 2$-matrix of covariates (including the intercept). Then

$$\boldsymbol{y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_{n_i} + \tau_0^2\boldsymbol{J}_{n_i}), \tag{35}$$

where $\boldsymbol{J}_{n_i}$ denotes and $(n_i \times n_i)$-matrix of ones and $\boldsymbol{I}_{n_i}$ is the identity matrix.

Let $\boldsymbol{V}_i$ be the marginal covariance matrix for $\boldsymbol{y}_i$;

$$\boldsymbol{V}_i = \sigma^2\boldsymbol{I}_{n_i} + \tau_0^2\boldsymbol{J}_{n_i}. \tag{36}$$

The fixed effects $\boldsymbol{\beta}$ are estimated using maximum likelihood (ML), while the random effect parameters σ^2 and τ_0^2 are estimated using restricted maximum likelihood (REML). According to Langaas and Hem 2018, REML is used to get a better estimator for the random effects than using regular ML, because it is less downwards biased. However, even though REML provides estimates that are closer, on average, to the true value of the parameters being estimated, linear mixed models does in general not give unbiased estimates for the parameters in $\boldsymbol{V}_i$.

The inverse matrix of $\boldsymbol{V}_i$ is used as the weighting matrix for the estimation of the fixed effect as

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{m} \boldsymbol{X}_i^T \hat{\boldsymbol{V}_i}^{-1} \boldsymbol{X}_i)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \hat{\boldsymbol{V}_i}^{-1} \boldsymbol{Y}_i,$$

and we get

$$\hat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, (\sum_{i=1}^{m} \boldsymbol{X}_i^T \hat{\boldsymbol{V}_i}^{-1} \boldsymbol{X}_i)^{-1}).$$

The random effect parameters are as mentioned estimated with restricted maximum likelihood (REML). The transformation method and the integration method are two different approaches used to obtain the estimates of the model parameters with REML. In short terms, the transformation method transforms the response variable and estimates the variance components indirectly, while the integration method integrates the likelihood function over the random effects and estimates the random effects and variance components together. In the following, we define $\vartheta = (\sigma_0^2, \tau_0^2)$. The integration method can be given as

$$l_{REML}(\vartheta) = \log \int L(\boldsymbol{\beta}, \vartheta) d\boldsymbol{\beta},$$

and one can demonstrate that the REML log-likelihood is

$$l_{REML}(\vartheta) = l_P(\vartheta) - \frac{1}{2}\log|\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}(\vartheta)_i^{-1} \boldsymbol{X}_i|.$$

$l_P(\vartheta)$ is the profile log-likelihood given by $l_P(\vartheta) = -\frac{1}{2}\log|\boldsymbol{V}(\vartheta)| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\vartheta))^T \boldsymbol{V}(\vartheta)^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\vartheta))$. The REML estimator for ϑ is found by maximizing $l_{REML}(\vartheta)$.

2.5 Hypothesis testing

Hypothesis testing is used to check the significance of the covariates in the different regression models. Due to the nature of the response variables and the variation of the model's characteristics, the hypothesis testing differs slightly across the multiple linear, Poisson, negative binomial and random intercept regression models. However, common for all the models, to test the significance of a particular regression coefficient, β_j, the hypothesis statement is typically given by

$$H_0 : \beta_j = 0,$$

vs.

$$H_1 : \beta_j \neq 0.$$

In a multiple linear regression model, the t-test is used to check the significance of the individual covariance in the model. Each covariate's coefficient is tested against the null hypothesis of no association. The test statistic for the t-test is based on the t-distribution:

$$T_j = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)} \sim t_{n-p-1},$$

where

$$\mathrm{SE}(\hat{\beta}_j)^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_{ji} - \bar{x})^2}.$$

The null hypothesis is not rejected if the test statistic, T_j, lies in the acceptance region:

$$-t_{\alpha/2,n-2} < T_j < t_{\alpha/2,n-2}.$$

For GLM regression models such as the Poisson and the negative binomial model, the hypothesis testing for individual covariates is typically done using Wald tests. Wald tests assesses the significance of the coefficient estimates by comparing them to a standard normal distribution. Following Fahrmeir et al. 2013, the Wald statistic is given as

$$w = t_j^2 = \left(\frac{\hat{\beta}_j}{a_{jj}}\right)^2, \tag{37}$$

where a_{jj} is the j-th diagonal element of the asymptotic covariance matrix $\boldsymbol{A} = \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$. The test is typically based on the statistic t_j, which asymptotically follows a standard normal distribution $N(0,1)$. The null hypothesis is rejected if the absolute value of t_j, denoted as $|t_j|$, is greater than the critical value $z_{1-\alpha/2}$. Here, $z_{1-\alpha/2}$ represents the $(1-\alpha/2)$-quantile of the standard normal distribution.

The likelihood ratio test compares the fit of a full model, which includes the covariate of interest, with the fit of a reduced model that excludes the covariate. This test can also be employed to compare different models. By comparing the likelihoods of nested models, the test determines whether the more complex model significantly improves the fit compared to the simpler, reduced model. For notation simplicity, let A refer to the full model and B refer to the reduced model that is nested within the larger model. That is, B is a sub-model of A. The null hypothesis (H_0) assumes that the reduced model is correct, while the alternative hypothesis (H_1), suggests that the more full provides a significantly better fit to the data. The test statistic, denoted as $-2\ln\lambda$, is calculated as twice the difference in log-likelihoods between the two models:

$$-2\ln\lambda = -2(\ln L(\hat{\beta_B}) - \ln L(\hat{\beta_A})), \tag{38}$$

which is asymptotically χ^2-distributed under the null hypothesis. The degrees of freedom are determined by subtracting the number of parameters in the reduced model from the number of parameters in the full model. The p-values are calculated in the upper tail of the χ^2-distribution.

2.6 Model Validation

The goodness of fit of a model can be calculated using the residual deviance and the null deviance:

$$\frac{\text{null deviance - residual deviance}}{\text{null deviance}} \cdot 100\%. \tag{39}$$

The residual deviance is twice the difference between the log-likelihood of the saturated model and the log-likelihood of the proposed model, where the saturated model consists of the observed values y_i. The expected mean from the model fit is defined as $\hat{\lambda}_i = \exp(\boldsymbol{x_i}^T\hat{\boldsymbol{\beta}})$, giving the residual deviance $D = 2(L(y) - L(\hat{\lambda}))$. Finally, the null deviance is the residual deviance of the model that only contains an intercept.

Another goodness of fit measure is the Pearson statistic. The Pearson statistic is Chi-squared distributed with $n-p$ degrees of freedom and is calculated by squaring and summing all the Pearson residuals. A Pearson residual is given as

$$r_i = \frac{y_i - \mathrm{E}[\hat{Y}_i]}{\sqrt{\mathrm{Var}[\hat{Y}_i]}}. \tag{40}$$

If the Pearson statistic is larger than $\chi^2_{\alpha,n-p}$ for a significance level α, the null hypothesis is rejected, indicating that the model does not fit with the distribution that have been observed.

When hypothesis testing is difficult, and when models are non-nested, R^2 or information criteria such as the AIC can be useful.

R^2 gives the fraction of variance explained.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{TSS}{RSS}, \tag{41}$$

where TSS is the total sum of squares, calculated as the sum of the squared differences between each observed value of the dependent variable and the mean of the dependent variable. RSS is the residual sum of squares, calculated as the sum of the squared differences between each observed value of the dependent variable and its corresponding predicted value

Here we aim to have a large value, with the aim of explaining as much of the variance of the data as possible. R^2 does not penalize the number of parameters in the model. That is, adding more variables always increase the value of R^2. Thus, the adjusted R^2, given as

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{42}$$

is a better choice to measure model fit. The AIC is defined as

$$\text{AIC} = -2l(\hat{\boldsymbol{\beta}}) + 2p. \tag{43}$$

Here, l is the log-likelihood, n the number of observations and p is the number of regression parameters (not including the intercept). Among a range of competing models, the model with the smallest AIC is chosen. The penalty term for the number of parameters in the AIC prevents overfitting of the models.

3 Regression Models With a Baseline Count

The theory and method in this section is based on *A comparison of different ways of including baseline counts in negative binomial models for data from falls prevention trials* by Zheng et al. 2018, with some adjustments made according to our motivating dataset. In this section, we consider baseline and follow-up counts. The purpose of this section is to investigate how to best include a baseline count when modelling a follow-up count. The baseline count is an initial measurement and is used for comparison over time to look for changes. The follow-up count is the response variable in the model and used to model the actual change from the baseline count. In our analysis, both the baseline count and the follow-up count are counted numbers of salmon lice. For example, the baseline count may refer to the counted numbers of salmon lice in the 1-sample and the follow-up count may refer to the number of lice in the 2-sample. We assume that both the baseline count and the follow-up count have a known exposure unit attached to it.

In Mæland 2022, we included baseline counts as log-transformed covariates in the regression models based on the the simulation results by Zheng et al. 2018. Zheng et al. 2018 conducted a simulation mimicking a "falls prevention trial" where individuals were observed for some time prior to and after treatment. Then y_{i0} was the number of falls for person i before treatment (over a time t_{i0}), and y_{i1} was the number of falls for person i after treatment (counted over a time t_{i1}).

In our data, the same salmon are not counted twice, but a random sample is taken from a cage before "treatment" and thereafter a new random sample is taken from the same cage after treatment. Our goal is therefore to reproduce and extend the work of Zheng et al. 2018 in this setting and use our results to inform our choice of model in the application. In the essence, the simulation experiment of Zheng et al. 2018 is based on creating correlated count data.

Specifically, let ν_i be gamma distributed with $\mathrm{E}(\nu_i) = 1$ and $\mathrm{Var}(\nu_i) = \frac{1}{r}$. For the simulations, we redefine the variance as $\alpha := \frac{1}{r}$. Let λ_{i0} be the rate of occurrences before treatment and let λ_{i1} be the rate of occurrences after treatment where λ_{i1} includes information on treatment or no treatment. Similarly to what we saw in Section 2.2, let $\lambda_{i0} = \nu_i\mu_{i0}$ and let $\lambda_{i1} = \nu_i\mu_{i1}$, where ν_i is the same at both time points. Further, $\mu_{i0} = n_{i0}\mu_0$ and $\mu_{i1} = n_{i1}\exp(\beta_0 + \beta_1 x_i)$, where x_i represents treatment ($x_i = 1$) or no treatment ($x_i = 0$). The two counts y_{i0} and y_{i1} are, conditional on λ_{i0} and λ_{i1}, assumed Poisson-distributed;

$$Y_{i0}|\lambda_{i0} \sim \text{Poisson}(\lambda_{i0}), \tag{44}$$

$$Y_{i1}|\lambda_{i1} \sim \text{Poisson}(\lambda_{i1}), \tag{45}$$

and the relationship between them is determined by the gamma-distributed variable ν_i.

Zheng et al. 2018 simulates from this model and considers negative binomial regression models for the follow-up count y_{i1} using four different linear predictors. Based on this, the performance of the negative binomial model was investigated with the following four linear predictors: (i) ignoring y_{i0}, (ii) including y_{i0} as a covariate, (iii) including $\log(y_{i0})$ as a covariate and (iv) including $\log(y_{i0})$ as an log-transformed offset. The simplest model is the one excluding the baseline count in the linear predictor. This model is from now on referred to as the NB_{null} -model. The linear predictor is given as

$$\eta_{i1} = \beta_0 + \beta_1 x_i + \log(n_{i1}), \tag{46}$$

where $\log(n_{i1})$ is the offset of the model. The next model includes the unlogged baseline count in the linear predictor. This model is henceforth referred to as the $NB_{unlogged}$ model. The linear predictor is given as

$$\eta_{i1} = \beta_0 + \beta_1 x_i + \kappa\frac{y_{i0}}{n_{i0}} + \log(n_{i1}), \tag{47}$$

where κ is the coefficient associated with the unlogged baseline count. The third model includes the logarithm of the baseline count. This model is from now on referred to as the NB_{logged} model. The linear predictor in this model is given as

$$\eta_{i1} = \beta_0 + \beta_1 x_i + \zeta \log\left(\frac{y_{i0}}{n_{i0}}\right) + \log(n_{i1}), \tag{48}$$

with ζ being the coefficient for the logged baseline count. In practice, an extra +1 is added to all the baseline counts to allow the log-transformation also when y_{i0} is zero. According to Zheng et al. 2018, the choice of value to add do not substantially affect the estimation of β_1. The last model includes the baseline count as a log-transformed offset in the linear predictor. Here, the linear predictor is written as

$$\eta_{i1} = \beta_0 + \beta_1 x_i + \log\left(\frac{y_{i0}}{n_{i0}}\right) + \log(n_{i1}). \tag{49}$$

This model is from now on referred to as NB_{offset}. Again, +1 is added to all the baseline counts before the log-transformation in the offset.

Having introduced the linear predictors we use in the simulations, we also want to investigate how to theoretically best set up a negative binomial regression model to include the correlation between y_{i0} and y_{i1} with the rates λ_{i0} and λ_{i1} for the baseline and follow-up count. In order to do this, we are interested in finding an expression for the expectations of the follow-up count y_{i1} that incorporates the expectations of the baseline count y_{i0}. The expectations of y_{i0} and y_{i1} given by the subject effect (ν_i) in equation (44) and (45) are

$$\mathrm{E}(y_{i0}|\nu_i) = \lambda_{i0} = \nu_i \mu_{i0}, \tag{50}$$

$$\mathrm{E}(y_{i1}|\nu_i) = \lambda_{i1} n_{i1} \exp(\beta_0 \ \beta_1 x_i) = \nu_i \mu_{i1}. \tag{51}$$

ν_i is the same in Equation (50) and (51), and by combining these two equations one obtains,

$$\begin{aligned} \mathrm{E}(y_{i1}|\nu_i) &= \frac{\mathrm{E}(y_{i0})}{n_{i0}\mu_0}\mu_{i1} \\ &= \frac{\mathrm{E}(y_{i0})}{n_{i0}\mu_0} n_{i1} \exp(\beta_0 + \beta_1 x_i). \end{aligned} \tag{52}$$

Taking the logarithms of both sides yields

$$\log(\mathrm{E}(y_{i1}|\nu_i)) = \beta_0 + \beta_1 x_i + \log\left(\frac{1}{\mu_0}\right) + \log\left(\frac{\mathrm{E}(y_{i0})}{n_{i0}}\right) + \log(n_{i1}). \tag{53}$$

Further, assuming $\mathrm{E}(y_{i0}) \approx y_{i0}$ in (53), and defining the constant $\beta_0^* := \beta_0 + \log\left(\frac{1}{\mu_0}\right)$ the expression further simplifies to

$$\log(\mathrm{E}(y_{i1}|\nu_i)) = g(\mu_i) = \beta_0^* + \beta_1 x_i + \log\left(\frac{y_{i0}}{n_{i0}}\right) + \log(n_{i1}). \tag{54}$$

The expression given in Equation (54) may suggest that it is most appropriate to incorporate the logarithmic transformed baseline count as an offset or a covariate in the model, when our aim is

inference on β_1. Taking the exponential of this expression will further show how the follow-up count is explained by the baseline count. This gives

$$\begin{aligned} \mathrm{E}(y_{i1}|\nu_i) = \hat{y}_{i1} &= \exp(\beta_0^x + \beta x_i + \log\left(\frac{y_{i0}}{n_{i0}}\right) + \log(n_{i1})) \\ &= \frac{y_{i0}}{n_{i0}} \cdot \exp(\beta_0^* + \beta x_i) \cdot n_{i1}. \end{aligned} \tag{55}$$

One can see that the expected follow-up count $\hat{y}_{i1}$ is given by the constant term $\frac{y_{i0}}{n_{i0}}$ times the exponential term $\exp(\beta_0 + \beta_1 x_i)n_{i1}$. This suggest that there is a linear relationship between the expected follow-up count and the baseline count. One can also investigate the relationship between the expected follow-up count and the baseline count if the logarithm of the baseline count is included as a covariate in the model. In this case, the expression in Equation (54) would look like

$$\log(\mathrm{E}(y_{i1}|\nu_i)) = g(\mu_i) = \beta_0^* + \beta x_i + \gamma \log\left(\frac{y_{i0}}{n_{i0}}\right) + \log(n_{i1}), \tag{56}$$

where γ is the coefficient associated with the log-transformed baseline count. This expression is identical to the one given in (54), given that the coefficient γ associated with the logged baseline count is set to 1. Taking the exponential, the expression is given as

$$\begin{aligned} \mathrm{E}(y_{i1}|\nu_i) = \hat{y}_{i1} &= \exp(\beta_0^* + \beta x_i + \gamma \log\left(\frac{y_{i0}}{n_{i0}}\right) + \log(n_{i1})) \\ &= \left(\frac{y_{i0}}{n_{i0}}\right)^{\gamma} \cdot \exp(\beta_0^* + \beta x_i) \cdot n_{i1}. \end{aligned} \tag{57}$$

If the estimated value of γ is close to one, one can expect the behaviour of this model and the offset model to be quite similar. Otherwise, the value of γ will decide how much the baseline count is affecting the follow-up count. Having $\gamma > 1$, the baseline count can pay a great impact on the follow-up count and having $\gamma < 1$ results in the baseline count having a lower impact on the follow-up count.

The two models presented above, including the baseline count as a log-transformed offset and including the logarithmic baseline count as a covariate in the models, are clearly the most appealing given the theory presented so far in this section. However, for a simpler model, one could think that including the unlogged baseline count as a covariate in the model also would explain the relationship between the baseline and follow-up count sufficiently. In this case, the expression given in Equation (54) would be

$$\log(\mathrm{E}(y_{i1}|\nu_i)) = g(\mu_i) = \beta_0^* + \beta x_i + \kappa \frac{y_{i0}}{n_{i0}} + \log(n_{i1}), \tag{58}$$

where κ is the coefficient associated with the unlogged baseline count. Taking the expectation, one obtains

$$\begin{aligned} \mathrm{E}(Y_{i1}|\nu_i) = \hat{y}_{i1} &= \exp(\beta_0^* + \beta x_i + \kappa \frac{y_{i0}}{n_{i0}} + \log(n_{i1})) \\ &= \exp(\kappa \frac{y_{i0}}{n_{i0}}) \cdot \exp(\beta_0^* + \beta x_i) \cdot n_{i1}. \end{aligned} \tag{59}$$

Here, it is not so easy to see how the follow-up count is directly affected by the baseline count since the baseline count is incorporated in an exponential function. One model that is even simpler is the one not including the baseline count at all. In this model, the follow-up count is only affected by the given covariates in the model.

Following Mæland 2022, we know that the log-transformed multiple linear regression model performed well on the count data of mobile lice for the 3-sample. Thus, we are also interested in investigating the performance of the different simulations and linear predictors using the log-transformed linear multiple regression model. This model is discussed in detail in Section 2.1.3. In this case the follow-up count y_{i1} will be log-transformed as

$$\log(y_{i1} + 1),$$

while the baseline count is unlogged. The four linear predictors described in this section will also be used for the log-transformed model. The log-transformed model using the linear predictor described in Equation (46) is henceforth referred to as $LM(logged)_{null}$. The log-transformed model using the linear predictor described in Equation (47) and Equation (48) is henceforth referred to as $LM(logged)_{unlogged}$ and $LM(logged)_{logged}$, respectively. Finally, the offset model using the linear predictor given in Equation (49) is henceforth referred to as $LM(logged)_{offset}$. In total we therefore investigate eight unique models based on two different regression models and four linear predictors.

3.1 Simulation of data sets

We are interested in studying the eight different models with the motivation of determining the best linear predictor among the four givens above. In order to do this properly, we want to simulate data in order to test the different model's effectiveness before applying them to our motivating dataset. By generating data with known properties based on our observed data, we can assess whether the different models accurately can capture the patterns and relationships within the data. We can then use model selection methods, e.g., AIC, type-I error rates and the models estimated values with standard deviation to evaluate the performance of models with the different linear predictors.

There were two different types of simulations done. The first simulation, henceforth referred to as simulation 1, were a simplification of the events of counting salmon lice in the salmon cages. The second simulation, henceforth referred to as simulation 2, tries to incorporate the real-life events of the salmon lice sampling from the cages. Both simulations will be further explained in detail.

For both simulations, 2000 sets of data were simulated in `R`, using the mixed Poisson distribution described above with the number of cages m ($i = 1, ..., m$) set to 100. For simplicity, we assumed that we only had one covariate in the model. The covariate, x_i, either took the value 0 or the value 1. We assumed that the first $k = m/2$ cages took the value 0 and the second k cages took the value 1. The rate of the follow-up count was adjusted according to $\boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_i$. Setting $\beta_0 = 1$ for each simulation, only the value of β_1 adjusted the rate of the follow-up count. The mean baseline was set to $\mu_{i0} = 0.5$, close to the observed average baseline count from our motivating dataset of 0.41. Following Zheng et al. 2018, three levels of intervention effect were considered: $\beta_1 = -0.2$, $\beta_1 = -0.1$ and $\beta_1 = 0$ for checking empirical type-I error rate. The variance of the underlying mixing distribution, α, was set at two levels: $\alpha = 3$ to give a large level of overdispersion and $\alpha = 0.5$ to give a lower level of overdispersion. In the simulations, we wanted to capture two properties of the salmons in the salmon cages, based on the real-life scenario. First, we wanted every salmon to be unique, but there being small differences between each salmon. This is referred to as the salmon-level subject effect. Secondly, we wanted all salmons coming from the same cage to be more similar than salmons coming from different cages. We refer to this as the cage-level subject effect. In order to incorporate these properties in the simulations, α was created as an $m \times 1$ Gaussian vector with mean 3 (or 0.5) and a small standard deviation, instead of α being a constant. With the small standard deviation of α, we simulated that each cage had its own α-value, α_i, attached to it. The α_i value was then used as the variance in a gamma distribution, simulating subject effects for all the salmons in the cage.

In both simulations, the subject effects was then used to create the random Poisson parameters in the baseline and follow-up count, namely $\nu_i\mu_{i0}$ and $\nu_i\mu_{i1}$, where ν_i is the subject effect and μ_{i0} and μ_{i1} were known and based on the average baseline and follow-up rate, respectively. These two parameters were then used to create the conditional Poisson distributions described in Equation (44) and (45). Having these distributions, we wanted to extract the total baseline and follow-

up count. This was done by summing up all the generated values from the conditional Poisson distribution for the baseline and follow-up.

The two simulations were quite similar, but differed in the number of salmon in the cages and how the salmon was sampled in the counting process. In simulation 1, the number of salmon in each cage was set to 20. This was a huge simplification compared to the real-life salmon cages, but mimics the simulation of Zheng et al. 2018. We included this simple simulation in order to compare it with simulation 2. Consequently, we may be able to use the simulation results to say something about whether the model is affected by the actual sampling process of salmon lice. In simulation 1, using the salmon-level subject effect described above, a baseline and follow-up count of mobile lice on the 20 salmons were created using the Poisson mixture model for each of the 100 cages. Following each simulation, the generated baseline and follow-up count were used to create four different negative binomial models and four different linear regression models with the four linear predictors. In simulation 2, the number of salmon in each cage was set to 100000 in order to create a more realistic sampling scenario. This was motivated by the real dataset having 103615 salmons in each cage on average. In order to simulate the actual counting process of mobile lice, a random selection of 20 salmons from the 100000 salmons were used to create the baseline count of mobile lice. Then, another random selection of 20 salmons were used to create the follow-up count of salmon lice in the cage. Both the baseline count and the follow-up count were created using the Poisson mixture model. Following each simulation, the baseline and follow-up count were used to create the negative binomial and log-transformed regression models for each scenario.

From the simulated datasets, $\hat{\beta}_1$ and their standard errors (SE($\hat{\beta}_1$)) were recorded. In addition, the AIC was recorded, and the type-I error rate was calculated from the model fits to each simulated dataset. The type-I error rate was calculated as the proportion of significant results from the Wald test of β_1 among replicates when $\beta_1 = 0$.

As an example of how the data has been simulated, the `R`-code for the simulation of the baseline and follow-up count is presented in Appendix C.

3.2 Simulation results

Our main focus was comparing the different ways of including a baseline count in a negative binomial regression model and a log-transformed linear regression model. But we were also interested in comparing the results from simulation 1 and simulation 2. In this section, we present the results from the simulations in terms of $\hat{\beta}_1$ with standard errors, the power of the statistical models and the AIC. In all scenarios the eight different models behaved quite similarly.

In Figure 26, the simulation results from simulation 1 and simulation 2 is presented using $\alpha = 3$ with three different levels of β_1. A similar figure showing the values of the estimated $\hat{\beta}_1$ and their standard deviation using $\alpha = 0.5$ is presented in Appendix A. The figure shows the mean value of the estimated $\hat{\beta}_1$'s and the standard deviation for the four negative binomial and four log-transformed linear models. In all the scenarios, $\hat{\beta}_1$ is close to the underlying value and the standard deviation is quite similar for all the models. We note that the standard deviation is largest for the *null*-models, i.e. the models using the linear predictor described in Equation (46), where the baseline count excluded from the linear predictor.

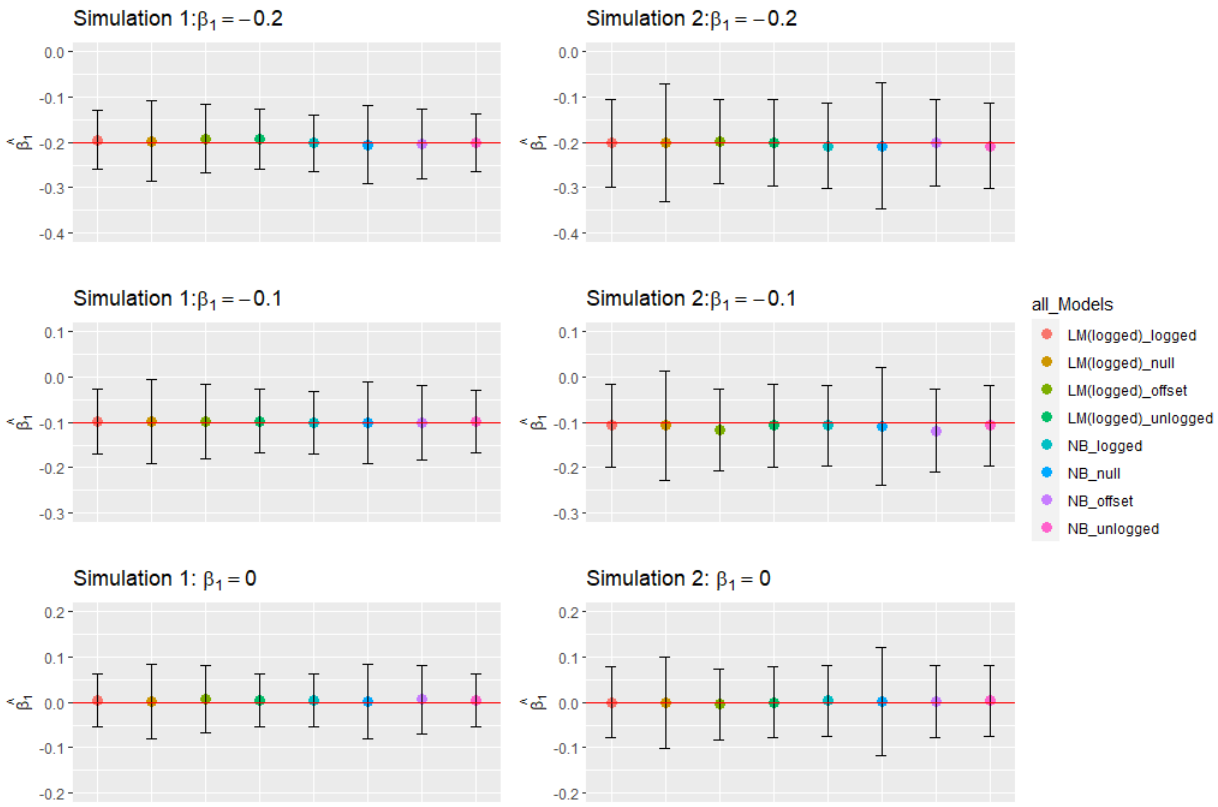


Figure 2: *Estimated* $\hat{\beta}_1$ *from simulation 1 and simulation 2 using* $\alpha = 3$ *with* $\beta_1 = -0.2$, $\beta_1 = -0.1$ *and* $\beta_1 = 0.0$.

Figure 3 shows the type-I error rates for simulation 1 and simulation 2 using $\alpha = 3$. with Clopper-Pearson confidence intervals as described in Clopper and Pearson 1934. The similar figure using $\alpha = 0.5$ is presented in Appendix A. The type-I error rates are in general higher for simulation 1 than for simulation 2. Type-I error rates refers to the incorrect rejection of a true null hypothesis, meaning that the models are falsely detecting a relationship or effect when it does not exist. For simulation 1, only the models including the baseline count as an offset and the models excluding the baseline count, have a type-I error rate close to the nominal level of 5%. The models using the other linear predictors, that is, including the baseline count as a covariate and as a log-transformed covariate have a much higher type-I error rate on average. The type-I error rates of these models are ranging from 10% to 13%, which is over twice as high as the nominal level. However, for simulation 1, the Clopper-Pearson confidence intervals are wide, resulting in the nominal level being encompassed by the confidence intervals. In simulation 2, the type-I error rates are in general closer to the nominal level of 5%. The models including the baseline count as a log-transformed offset, i.e. NB_{offset} and $LM(logged)_{offset}$, seem to have the most appropriate type-I error rate in both simulation 1 and simulation 2.

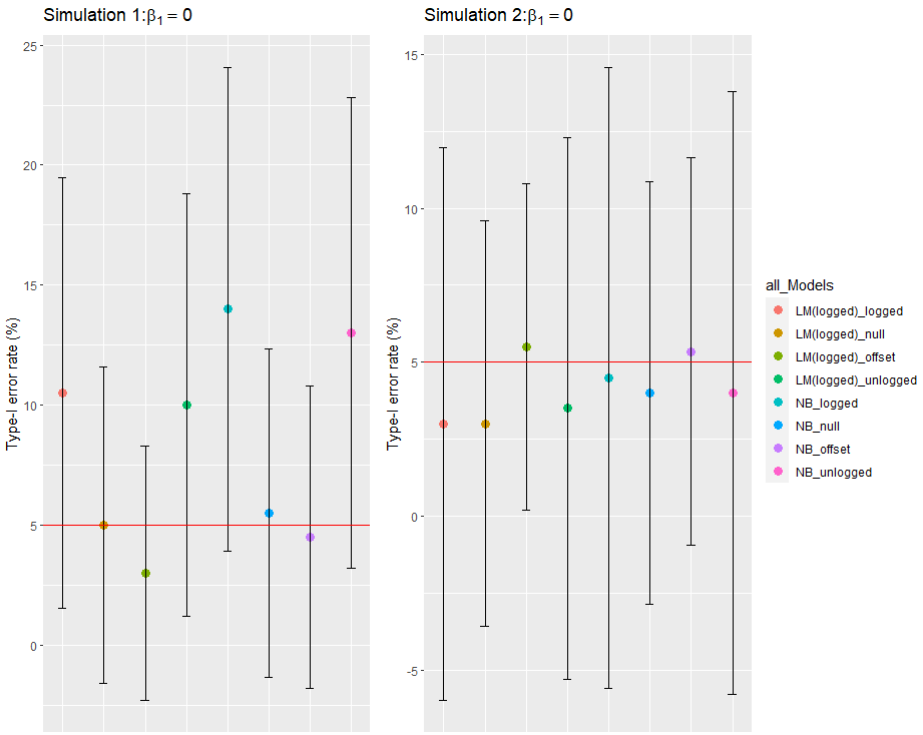


Figure 3: *Type-I error rates displayed as the empirical power with* $\beta_1 = 0.0$ *from simulation 1 and simulation 2 having* $\alpha = 3$ *and* $\beta_1 = 0$*. The values on the y-axis are presented in percentages (%).*

Table 1 shows the AIC of the negative binomial regression models fitted to the simulated data in simulation 1 and simulation 2 in parenthesis. In general, it appears that the AIC is higher for simulation 2 than it is for simulation 1, with some few exceptions. The models including the baseline count as an offset appears to have the largest AIC compared to the rest of the models in both simulation 1 and simulation 2. That is, the NB_{offset}-model has the largest AIC, followed by the model excluding the baseline count, i.e. NB_{null}. The trend is similar for almost all scenarios. One of the reasons for the offset-models having a larger AIC may be due to the models lack of flexibility. An offset model is less flexible compared to a model using a covariate because it provides a fixed adjustment or constraint on the relationship between the predictors and the response variable. A covariate model on the other hand, allows for more flexibility in capturing the relationship between the predictors and the response variable as the model parameters associated with the covariate can be estimated.

The log-transformed linear regression models use a transformation of the response variable, and is therefore not similar to the response variable used in the negative binomial regression model. Therefore, the AIC from the log-transformed linear regression model cannot be directly compared with the AIC from the negative binomial regression model. A table of the AIC from the log-transformed models is presented in Appendix B to compare the results from the different linear predictors from simulation 1 and simulation 2.

	$\alpha = 3$			$\alpha = 0.5$		
	$\beta_1 = -0.2$	$\beta_1 = -0.1$	$\beta_1 = 0.0$	$\beta_1 = -0.2$	$\beta_1 = -0.1$	$\beta_1 = 0.0$
$NB_{null}(sim2)$	653 (751)	760 (760)	769 (769)	653 (723)	659 (714)	666 (798)
$NB_{unlogged}(sim2)$	646 (752)	703 (761)	710 (770)	646 (724)	652 (715)	658 (799)
$NB_{logged}(sim2)$	645 (752)	701 (761)	709 (770)	645 (724)	652 (715)	659 (602)
$NB_{offset}(sim2)$	711 (839)	716 (849)	724 (870)	712 (818)	720 (809)	727 (603)

Table 1: *AIC from* NB_{null}*,* $NB_{unlogged}$*,* NB_{logged} *and* NB_{offset} *from simulation 1 and simulation 2 using* $\alpha = 3$ *and* $\alpha = 0.5$ *with* $\beta_1 = -0.2$*,* $\beta_1 = -0.1$ *and* $\beta_1 = 0.0$*.*

3.3 Remarks on the baseline models

All models performed reasonably well in both simulations. Hence, one can argue that including the baseline count in any of the four different ways is a sound decision. The extension of the simulation of Zheng et al. 2018 to the current simulation did not seem to alter their results, also when modelling the log-transformed follow-up counts with a linear model. We also note that the results from simulation 1 and simulation 2 were quite similar.

Seeing that we are interested in capturing all underlying factors that determine the level of salmon lice in the 2-sample, we want to include the baseline count in the regression models. We will later see that there also is a positive correlation between the salmon lice abundance in the 2-sample and in the 1-sample, what we have referred to as the baseline count in this section. We therefore omit all the models excluding the baseline count. Going forward, we will include the baseline count as a log-transformed offset in all the regression models. This is also compatible with the results in Zheng et al. 2018, where the models including the baseline count as a log-transformed covariate and as a log-transformed offset performed best among the four models. The reason why we prefer the offset-version is the somewhat unrealiable results for type-I error rate found in Figure 3 for the other models.

More specifics regarding how the baseline count is included in the regression analysis will be further discussed in Section 4.1 Including the baseline count as an offset will also be a continuation of the work done in Mæland 2022, where the baseline count was included as a log-transformed covariate in the regression models.

4 Application of methods to salmon lice data

In lines with the *RegLus*-project, we are interested in combining the theory of the respective regression models for count data with the observed count data of salmon lice, by applying the data to the models. In this section, we first present a description of the data with various summary statistics and visualizations. Then we present the results of the four different regression models applied to the observed count data.

The data considered used in this thesis was collected from various locations in mid-Norway in co-operation with the *RegLus* project. The salmon farms were ones undergoing delousing treatments. Figure 4 displays production area 1-13 along the Norwegian coast. In this project, we only use observations from study production area 6 and 7, marked with * in the figure. In total, there were 35 unique salmon farms registered in the dataset.

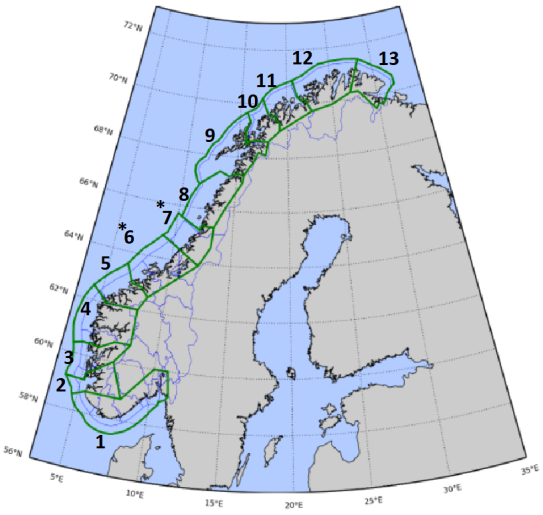


Figure 4: Production areas 1-13 along the Norweagian coast

Source: *Forskrift om produksjonsområder for akvakultur av matfisk i sjø av laks, ørret og regnbueørret, 2017*

The process of counting the salmon lice and sampling the data was performed in the period 2018 to 2019. The lice counting was executed by operating technicians on the facility as a part of the ordinary salmon lice registrations. The different parameters registered for every delousing treatment included company, location, start-time and end-time of delousing, number of salmons in the cage, average weight of fish in the cage, cage location of the salmon after treatment, presence of lice skirt before and after treatment, sample size of salmons used to count the salmon lice in a cage, mortality, treatment method and sea temperature.

The salmon lice number were the average number of lice per salmon and were calculated on a random sample of a small number of salmons in the cage. At least 20 salmons in the cage were inspected to calculate the lice number. The number of adult female lice, mobile lice, sessile lice and *Caligus elongatus* were recorded, and the lice numbers were calculated as the sample mean for each of the four stage groups.

The delousing treatments used in this data set include freshwater treatment, lice flusher and thermic treatment, and are explained in detail in Section 1.1. In the dataset, there was a large variation in the use of the different methods. In addition, there were three different lice flusher methods registered in the dataset: *Hydrolicer*, *FLS-delousing* and *Skamik*. Seeing that we were interested in studying lice flusher as a delousing treatment method, and not the different types of lice flusher methods, it was reasonable to merge them together to one treatment variable. The new

treatment variable, consisting of *Hydrolicer*, *FLS-delousing* and *Skamik*, was called *LiceFlusher* in the regression analysis.

4.1 Response Variable

Figure 5 presents the sum over all cages of the average number of mobile, sessile and adult female lice in the 0-sample, 1-sample, 2-sample and 3-sample. The figure illustrates how the lice number varies throughout the delousing process. From the figure, we see that the total number of salmon lice is dominated by mobile lice count in the 0-,1-, and 3-sample. In the 2-sample, the lice number of mobile, sessile and adult female lice are more evenly distributed, but the sum of mobile lice is still the largest among the three. For our main analysis, the response variable is the total count of sessile, mobile and adult female lice in the 2-sample per cage. We use the total count of *all* salmon lice as a response variable because our primary objective is studying the delousing treatment methods, which adversely is affecting all three stages of salmon lice. We will also extent the work done in Mæland 2022 by considering *mobile* lice in the 3-sample. Here, only mobile lice as that is the only reliable measure of re-infestation.

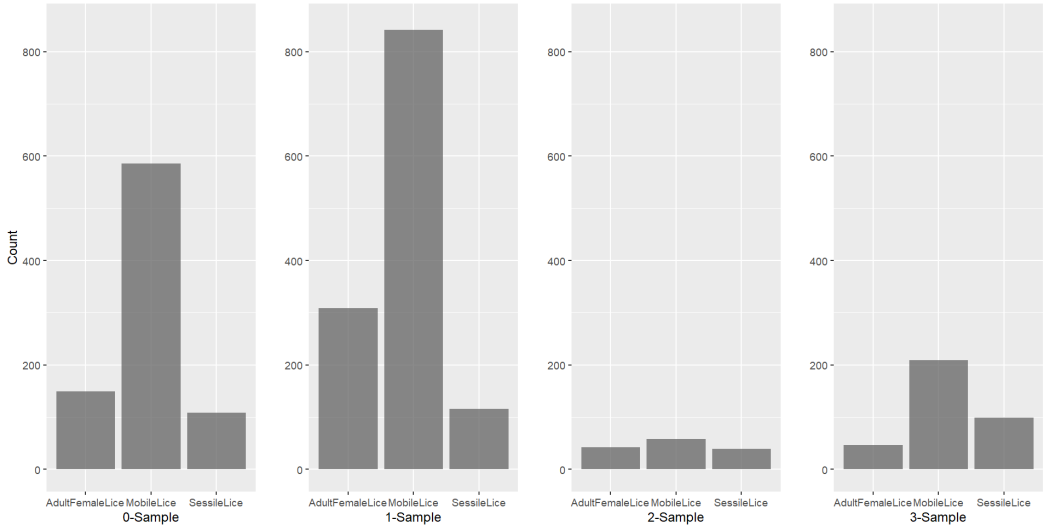


Figure 5: *The sum of the reported salmon lice number of mobile, sessile and adult female lice in the 0-sample, 1-sample, 2-sample and 3-sample.*

As previously stated, our primary focus in this project is to examine the impact of various non-medicinal treatment techniques. Therefore, we choose to study the prevalence of salmon lice in the 2-sample, where the count of salmon lice was done straight after the delousing treatment was done. Consequently, *SalmonLiceCount2* = *MobileLiceCount2* + *SessileLiceCount2* + *AdultFemaleLiceCount2* is used as the response variable in the regression analysis. The total counts are calculated using the sample count, e.g., *MobileLiceCount2* is calculated as $MobileLiceCount2 = MobileLice2 \cdot SampleCount2$, where *SampleCount2* refers to the amount of salmon the lice was counted on. The salmon lice *Caligus elongatus* (*C. elongatus*) is not used in the analysis even though the lice numbers are registered in the dataset. The main reason is that this is a different species than *Lepeophtheirus salmonis*. In addition, the lice numbers for *Caligus elongatus* are more uncertain and there is a lack of lice registrations of *C.elongatus* in the dataset. As a baseline for the 2-sample counts, we could have used either the 0-sample or the 1-sample. The 0-sample was deemed too unreliable as a baseline because counts were taken as much as 30 days prior to delousing treatment. As we were only interested in what happens immediately prior and immediately following to the delousing treatment, we chose to use the reported average of salmon lice in the 1-sample as a baseline count in the regression models, namely *SalmonLiceCount1* = *MobileLiceCount1* + *SessileLiceCount1* + *AdultFemaleLiceCount1*. The reported lice numbers of salmon lice in the 1-sample plotted against the reported lice numbers of salmon lice in the 2-sample, and the

log-transformed salmon lice number in the 1-sample against the log-transformed lice number in the 2-sample is presented in Figure 6. The plot indicates a linear relationship between the log-transformed lice number in the 1-sample and the 2-sample. For both the unlogged and the logged case, an increase in the salmon lice number in the 1-sample indicates an increase in the salmon lice number in the 2-sample. As discussed in Section 3, we decided that the baseline count is best included in a negative binomial and log-linear multiple regression model as a log-transformed offset. For simplicity, this offset will not be written out in the model specifications in this section.

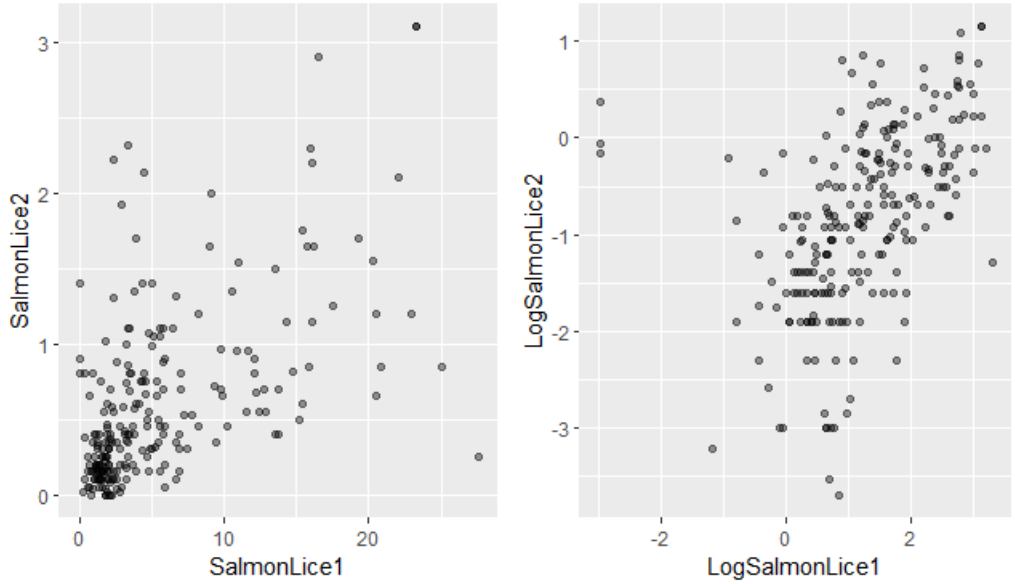


Figure 6: *The reported salmon lice number in the 1-sample vs the reported salmon lice number in the 2-sample(left) and the log-transformed salmon lice number in the 1-sample (right) plotted against the log-transformed salmon lice number in the 2-sample. The data points with the lightest black colour indicate a single point, while a darcer black colour indicates overlapping points.*

4.2 Pre-processing

The first step of the pre-processing of the data was to remove the duplicates. There were some data that was gathered from the same delousing treatment. This gave multiple values of the lice numbers in the four different samples. The values from the same treatment were for most part the same, and thus it was reasonable to only keep one observation for each sample from each delousing treatment. We assumed that the first data point in a duplicate set gave sufficient information. The next step of the pre-processing was to remove cages that were missing registered lice numbers (NAs) or other important values like for instance the sample size.

The original dataset consisted of 299 observations from salmon cages at different salmon farms in the Trøndelag-area. After removing the duplicates in the dataset, i.e., data gathered from the same delousing unit, the data set is reduced to 266 unique observations. Removing NAs in the data, further reduced the dataset to 241 unique observations. From a statistical perspective, the dataset provided is considered small. Dealing with small datasets can pose several challenges. One such challenge is that outlier values can disproportionately affect the model's accuracy. Small datasets are also more susceptible to overfitting, where the model becomes too complex and fits the data too closely. Therefore, simpler models tend to be more suitable for small datasets.

The dataset included information for every operational cage following a delousing treatment at the examined sites from 4th of April 2018, to 26th of October 26 2019. Table 2 provides an overview of the various variables utilized in the analysis, along with their explanations.

Variable Name	Explanation of Variable
SalmonLice1	Lice number of salmon in the 1-sample, given as a reported average calculated from a sample of at least 20 salmon. The lice number refers to the sum of the lice number of mobile lice, adult female lice and pre-adult lice.
SalmonLice2	Lice number of salmon in the 2-sample, given as a reported average calculated from a sample of at least 20 salmon. The lice number refers to the sum of the lice number of mobile lice, adult female lice and pre-adult lice.
SampleCount1	The number of salmons used to count the salmon lice in the 1-sample. SampleCount1 is usually 20.
SampleCount2	The number of salmons used to count the salmon lice in the 2-sample. SampleCount2 is usually 20.
SalmonLiceCount1	The count of salmon lice in the 1-sample. Calculated as SalmonLice1 · SampleCount1.
SalmonLiceCount2	The count of mobile lice in the 2-sample. Calculated as SalmonLice2 · SampleCount2.
LogSalmonLice2	The log-transformed lice number of salmon in the 2-sample. Calculated as $\log(\frac{SalmonLiceCount2+1}{SampleCount2})$.
Date	Date of the observation
SeaTemperature	Temperature of the sea, ° C
AverageWeight	Average weight of the salmon in the salmon cage, g
NumberOfFish	Number of fish in the cage
Method	Delousing method used, divided into freshwater treatment, thermic treatment and lice flusher
Location	Location of the salmon cage.

Table 2: *Variables used in the analysis with explanation and units.*

In both the Poisson and negative binomial regression model, all the elements in the response variable *SalmonLiceCount2* and baseline *SalmonLiceCount1* were transformed to integers by rounding down to the nearest integers in order to fit the model criterions. The presence of non-integer counts, calculated using the sample size of salmon and the registered salmon lice numbers from the dataset, may suggest that there were inaccuracies in the data. In the multiple linear regression model, the response variable was log-transformed using the natural log. That is, the transformation $\log(SalmonLiceCount2+1)$ was used to improve the normality and homoscedasticity of the model residuals. The details and justification of the log-transformation is presented in Section 4.6.1.

4.3 Data visualization

This section presents the data in the form of summary statistics and various visualizations. Table 3 displays the summary statistics for the response variable and the continuous explanatory variables used in the data visualizations and regression analysis. The percentage of zeros in each group of lice numbers is provided in parentheses.

Variable	Mean	Sd.	Median	Min.	Max.
SalmonLice1 (1.24%)	7.66	8.05	4.90	0.00	38.20
SalmonLice2 (2.49%)	0.59	0.57	0.40	0.00	3.10
SampleCount1	20.70	3.59	20.00	20.00	40.00
SampleCount2	29.80	18.45	20.00	20.00	120.00
SalmonLiceCount1 (1.24%)	155.00	160.60	99.00	0.00	764.00
SalmonLiceCount2 (2.49%)	18.40	23.23	9.00	0.00	171.00
NumberOfFish	115000.00	54950.00	134221.00	12274.00	196000.00
AverageWeight (g)	2710.00	1147.00	2813.00	700.00	5424.00
SeaTemperature (°)	11.20	1.91	11.30	4.00	14.80

Table 3: *Summary statistics for the various variables used in the regression analysis.*

The lice number and the log-transformed lice number of salmon lice in the 2-sample, i.e. *SalmonLice2* and *LogSalmonLice2*, have been graphed alongside the explanatory variables employed in the analysis. The latter is included to check the model assumptions of there being a linear relationship between the log-transformed response and the various explanatory variables. The reason being that the multiple linear regression model and the random intercept model both use a log-transformed response variable. In addition, both the Poisson and negative binomial regression model use a logarithmic function to link the response variable and the linear predictor.

In Figure 7, the salmon lice number and log-transformed salmon lice number in the 2-sample are plotted against the explanatory variable *NumberOfFish*. Common for both *SalmonLice2* and *LogSalmonLice2* is that there is appears to be a correlation with *NumberOfFish*. For *SalmonLice2*, the number of salmon lice appears to increase with increasing number of salmon in the cage. The log-transformed count, *LogSalmonLice2*, on the other hand, tends to decrease with an increasing number of salmons in the cage. Seeing that there is a correlation between the numbers of salmon in the cages and the salmon lice numbers, *NumberOfFish* is included in the regression analysis as an explanatory variable.

Figure 7: *The reported salmon lice number in the 2-sample (left) and the log-transformed reported salmon lice number in the 2-sample (right) plotted against the number of fish (NumberOfFish) in the salmon cage. The data points with the lightest black colour indicate a single point, while a darker black colour indicates overlapping points.*

Figure 8 illustrates the relationship between the explanatory variable *AverageWeight* and the salmon lice numbers and log-transformed salmon lice number in the 2-sample. One can see from the figure that there is a positive correlation between the average weight of the salmons and the unlogged and logged salmon lice numbers in the 2-sample. That is, an increase in *AverageWeight* also gives an increase in *SalmonLice2* and *LogSalmonLice2*. For *SalmonLice2*, the linear trend becomes more pronounced when the average weight of salmon reaches 2500 grams or higher, while for lower weights, the lice abundance seems to be more randomly scattered. For *LogSalmonLice2*, the linear trend appears to range all values of *AverageWeight*. This suggests that the average weight of salmon serves as a reliable explanatory variable for the regression analysis.

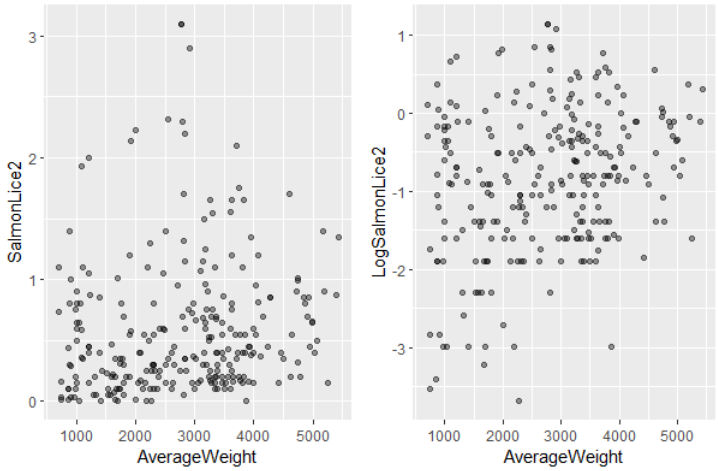


Figure 8: *The reported salmon lice number in the 2-sample (left) and the log-transformed reported salmon lice number in the 2-sample (right) plotted against the average weight (g) of salmon lice in a salmon cage (AverageWeight). The data points with the lightest black colour indicate a single point, while a darker black colour indicates overlapping points.*

In Figure 9, the reported salmon lice number and log-transformed salmon lice number in the 2-sample are plotted against the explanatory variable *SeaTemperature*. There appears to be a linear relationship between the temperature of the sea and the lice abundance in the 2-sample, both on a regular and on a log-transformed scale. The linear trend is positive and thus the salmon lice number increase with an increasing temperature. The observed correlation suggests that *SeaTemperature* is a suitable explanatory variable to include in the regression analysis.

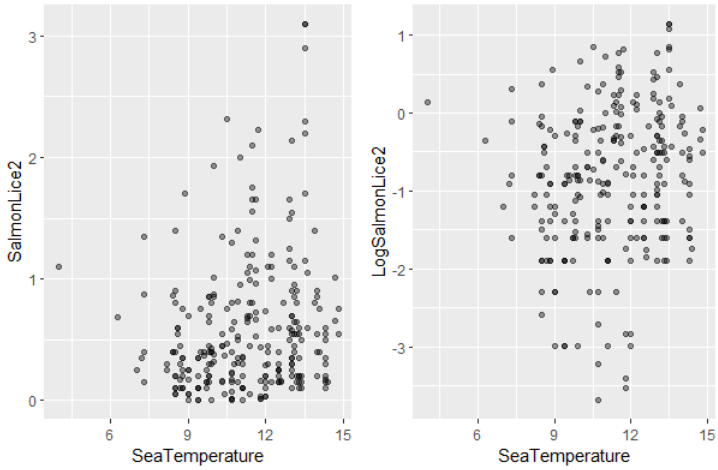


Figure 9: *The reported salmon lice number in the 2-sample (left) and the log-transformed reported salmon lice number in the 2-sample (right) plotted against the temperature in the sea (°C) in the salmon cage (SeaTemperature). The data points with the lightest black colour indicate a single point, while a darker black colour indicates overlapping points.*

In Figure 10, a box plot of the covariate *Method* and the reported lice numbers of salmon lice in the 2-sample and the log-transformed number of salmon lice number in the 2-sample is presented. There is a modest difference in the registered lice numbers and log-transformed lice numbers following the three different methods. The median of the registered salmon lice numbers in the 2-sample (both unlogged and logged) is lowest for the *Freshwater* treatment. The quartiles for the *Freshwater* treatment are considerably smaller compared to those for *LiceFlusher* and *Optilicer* treatments. In the dataset, the *Freshwater* treatment was only applied 11 times, whereas *LiceFlusher* and *Optilicer* treatments were used 118 and 112 times, respectively. Consequently, comparing the

effectiveness of *LiceFlusher* and *Optilicer* treatments to the Freshwater treatment may not be entirely fair. With the aim of investigating the effect of the three treatment methods, we include *Method* as an explanatory variable in the regression analysis. *Method* is coded as a factor variable, considering the three distinct methods.

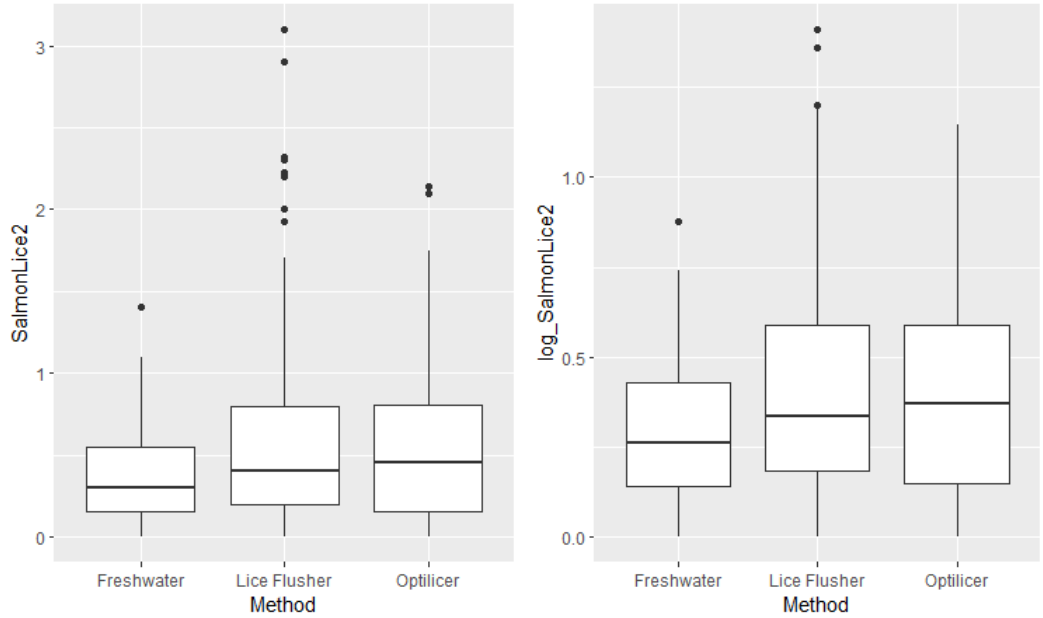


Figure 10: *The reported salmon lice number in the 2-sample (left) plotted against the three different delousing treatment methods and the log-transformed salmon lice number in the 2-sample plotted against the treatment methods (left). The data points with the lightest black colour indicate a single point, while a darker black colour indicates overlapping points.*

Figure 11 shows the correlation plot between each pair of the numerical variables used in the regression analysis. The plot presents the Pearson correlation coefficient between each pair of variables on the upper diagonal, the scatter plots of each pair in the lower diagonal, and finally, the distribution of each variable on the main diagonal. By the Pearson correlation coefficient, it is observed high correlation between every pair of variables except *SeaTemperature* and *AverageWeight*. For instance, *LogSalmonLice2* and *LogSalmonLice1* are highly correlated with correlation coefficient 0.485.

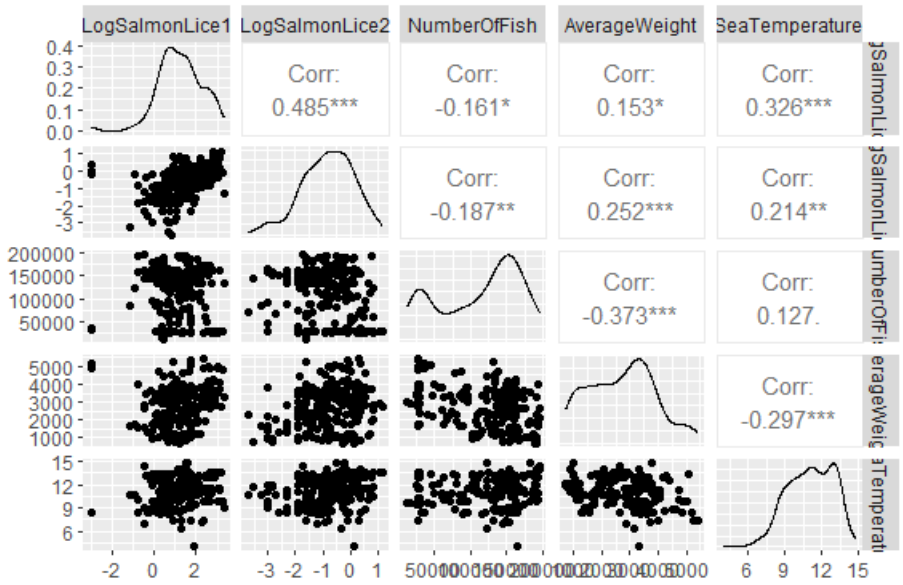


Figure 11: *The Pearson correlation coefficients calculated for every pair of numeric variables, scatter plots for each pair, and the generated distribution plots for each individual variable*

The 35 different salmon farms were registered in the dataset as a variable determining the location of the salmon farm. This variable, *Location*, was not included in the regression analysis. The reason being that this categorical variable had too many levels for one-hot encoding or including it as a factor variable. However, it was included as a random intercept term in the random intercept model to assess if the salmon farms could be grouped based on the location.

4.4 Poisson regression

The 2-sample counts of salmon lice were first analysed using a Poisson regression model with a logarithmic link function and treatment method (as a factor variable), number of fish, average weight and sea temperature as covariates, and sample sizes of the 1-sample and 2-sample and the baseline count as offsets in the model. A regression model was fitted in `R` to analyse the count of salmon lice in the 2-sample. The summary output of the model is presented in Table 4, which includes the estimated regression coefficients along with their standard errors, t-values and the p-values.

Table 4: *Regression coefficients with associated estimate, standard error, t-value and p-value from the Poisson regression for the count model for salmon lice in the 2-sample.*

Coefficient	Estimate	Std.Error	t-value	p-value
Intercept	2.18	0.061	35.64	$< 2.00 \cdot 10^{-16}$
LiceFlusher	0.42	0.042	9.89	$< 2.00 \cdot 10^{-16}$
Optilicer	0.69	0.042	16.33	$< 2.00 \cdot 10^{-16}$
NumberOfFish	$-2.52 \cdot 10^{-6}$	$1.22 \cdot 10^{-7}$	-20.75	$< 2.00 \cdot 10^{-16}$
AverageWeight	$-1.71 \cdot 10^{-5}$	$6.36 \cdot 10^{-6}$	-2.69	0.0072
SeaTemperature	0.058	0.0037	15.39	$< 2.00 \cdot 10^{-16}$

AIC: 2122, Null deviance: 25723 on 237 degrees of freedom, Residual deviance: 24533 on 232 degrees of freedom.

A goodness-of-fit test for the model was performed by using the residual deviance of the fitted model. The residual deviance D is 24533 for the Poisson model. The corresponding quantile of the $\chi^2_{\alpha,n-p}$ distribution is $\chi^2_{0.05,226} = 262$ for the model. Since $24533 > 262$ the model was rejected

at significance level $\alpha = 0.05$. The explained deviance was calculated as 4.63% based on the null deviance and the residual deviance provided in Table 4, using Equation (39).

To test for overdispersion, a hypothesis test was performed using the Pearson statistic. By substituting the estimated mean $\hat{\lambda}$ in Equation (10), the Pearson statistic for the Poisson model was calculated as $P = 32765$. The observed Pearson statistic yielded an estimated overdispersion parameter of $\hat{\phi}_P = \frac{P}{n-p} = 145$ for the model. The null hypothesis (H_0), stated as H_0 : No overdispersion, written as $H_0 : \phi \leq 1$, is tested against the alternative hypothesis, $H_1 : \phi > 1$, indicating the presence of overdispersion. Under the null hypothesis, the Pearson statistic, P, follows a chi-squared distribution with $n - p = 226$ degrees of freedom. Since the calculated p-value exceeded the critical value $\chi^2_{0.05,226}$, the null hypothesis was rejected, leading to the conclusion that overdispersion existed.

In Figure 12, we present plots depicting the relationship between the fitted values and the Pearson and deviance residuals for the Poisson model. The scatter of both the Pearson and deviance residuals around zero was evident, with residuals exhibiting considerable magnitude. These findings indicated that the Poisson model was not well-suited for accurately representing the data. Furthermore, it was apparent that the residuals' magnitude is greater for value between 15 and 25 compared to the rest of the values. This observation suggests a violation of the assumption of constant error variance, indicating the presence of heteroscedasticity in the data.

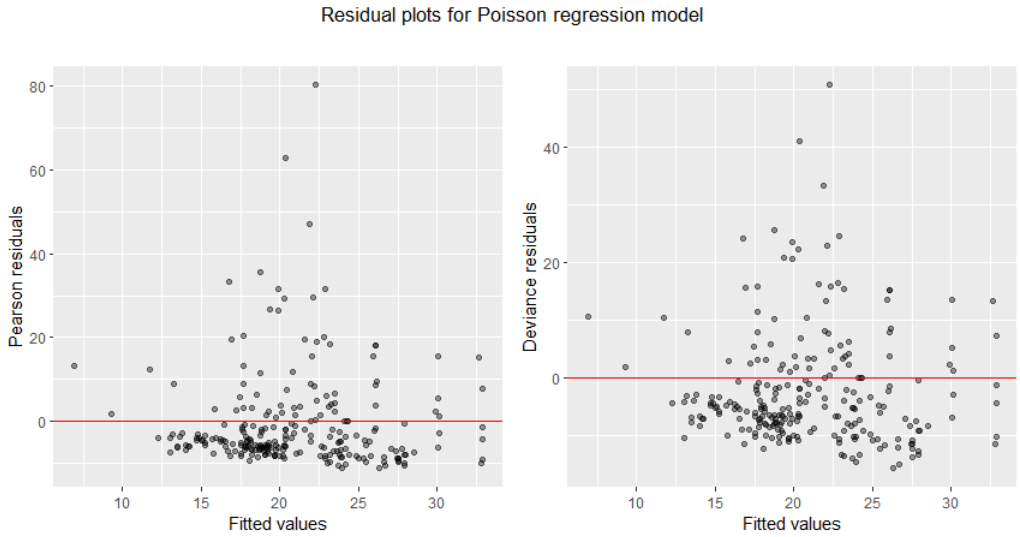


Figure 12: *Plot of Pearson and deviance residuals against fitted values from the Poisson regression model. To visualize overlapping, the data points are partially transparent. The data points with the lightest black colour indicate a single point, while a darker black colour indicates overlapping points.*

Figure 13 displays a frequency plot comparing the observed and fitted values obtained from the Poisson regression model. The observed values exhibit left skewness, revealing a significant number of zeros and small values. Conversely, the plot of the fitted values illustrated that the Poisson model was unable to adequately capture the abundance of zeros and small values, resulting in a more right-skewed distribution. This plot therefore further supports the notion that the Poisson regression model may not was a suitable fit for the observed count data of salmon lice in the 2-sample.

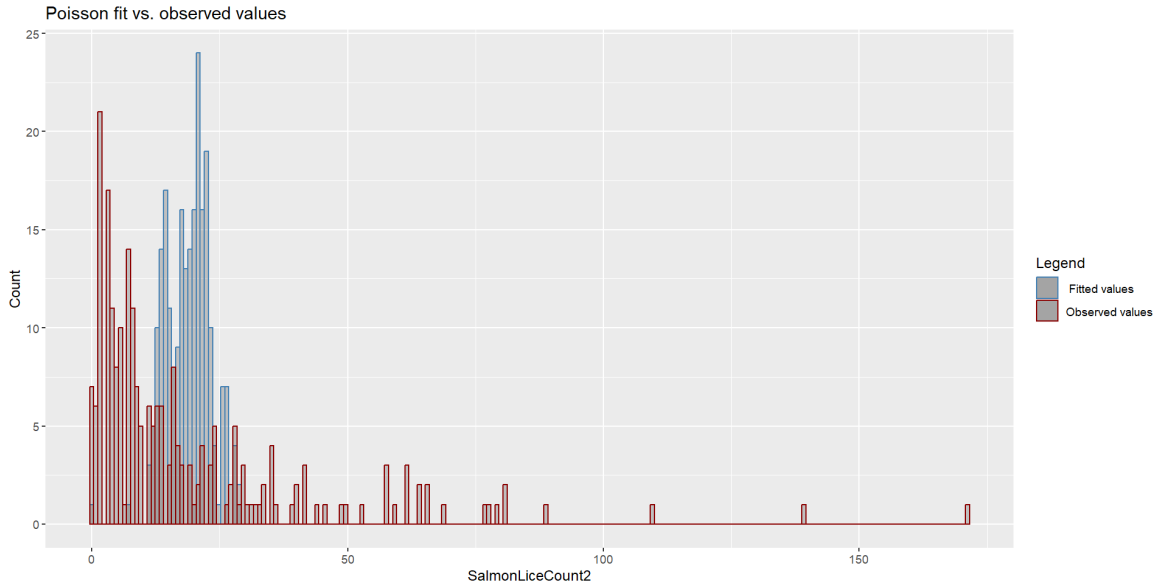


Figure 13: *The frequency of observed (red) and fitted (blue) values from the Poisson model for the count of salmon lice in the 2-sample.*

4.5 Negative Binomial Regression

The negative binomial regression model extends the Poisson regression model, by relaxing the assumption of equality between the mean and the variance with the use of a dispersion parameter which accounts for extra variability in the data. Consequently, we employed a negative binomial regression model to analyse the count of salmon lice in the 2-sample. Table 5 presents the summary output obtained from the negative binomial regression model fitted in `R`. According to the p-values from the Wald test, all covariates, except for the treatment method *LiceFlusher* (against the reference *Freshwater*) as well as the *AverageWeight* covariate, demonstrated statistical significance at a significance level of 0.05.

Table 5: *Regression coefficients with associated estimate, standard error, z-value and p-value from the negative binomial regression for the count model for salmon lice in the 2-sample.*

Coefficient	Estimate	Std.Error	t-value	p-value
Intercept	2.54	0.26	9.82	$< 2.00 \cdot 10^{-16}$
LiceFlusher	0.24	0.16	1.51	0.13
Optilicer	0.60	0.16	3.81	0.00014
NumberOfFish	$-3.07 \cdot 10^{-6}$	$5.93 \cdot 10^{-7}$	-5.17	$< 2.30 \cdot 10^{-7}$
AverageWeight	$-5.11 \cdot 10^{-5}$	$3.00 \cdot 10^{-5}$	-1.70	0.089
SeaTemperature	0.051	0.017	2.94	< 0.0033

AIC: 1735, Null deviance: 1368.4 on 237 degrees of freedom, Residual deviance: 1209.6 on 232 degrees of freedom.

Figure 14 displays the Pearson and deviance residuals plotted against the fitted values for the negative binomial regression model. Both plots exhibit scattered residuals around zero, similar to the residual plots observed in the Poisson regression model. These residual plots also indicate the presence of heteroscedasticity, where the error variance is largest for values between 15 and 25. It is therefore evident that the assumption of constant variance is violated. This pattern is more clearly visualized in the deviance plot on the right side compared to the Pearson plot on the left side. When comparing these residual plots to the similar plots for the Poisson model depicted in Figure 12, the magnitude of the residuals for the negative binomial model is smaller. The Pearson and deviance residuals of the Poisson regression extend up to 80.1 and 50.8, respectively, whereas the corresponding residuals for the negative binomial model reach up 16.3 and 7.5, respectively.

This indicates that the negative binomial regression model provides a better fit to the data than the Poisson regression model.

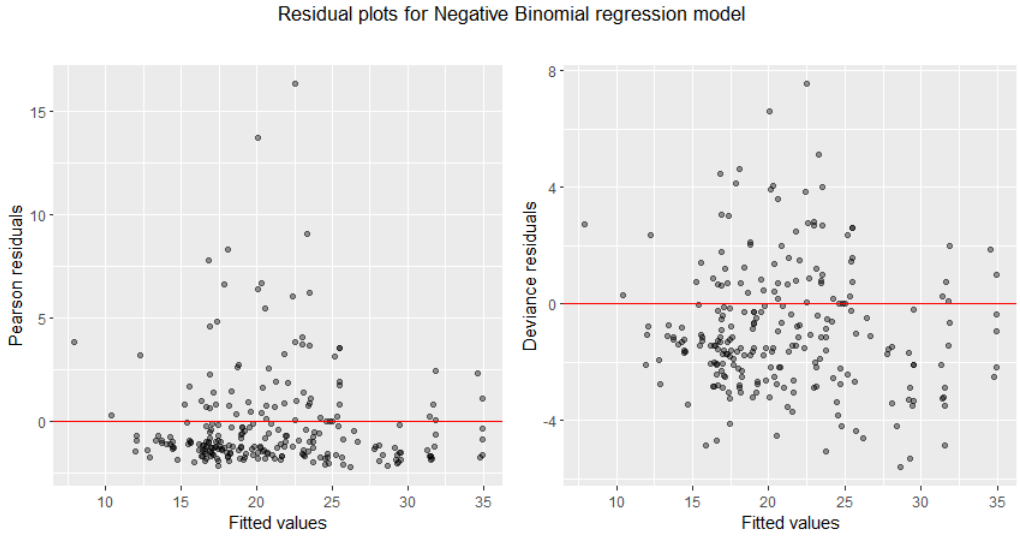


Figure 14: *Plot of Pearson and deviance residuals against fitted values from the negative binomial regression model. To visualize overlapping, the data points are partially transparent. The data points with the lightest black colour indicate a single point, while a darker black colour indicates overlapping points.*

Figure 15 portrays a frequency plot comparing the fitted and observed values of the count of salmon lice in the 2-sample. The plot reveals that the observed values display left skewness, including some zeros and small values. The fitted values obtained from the negative binomial model struggle to accurately capture these values. In general, the fitted values exhibit a greater right skewness compared to the observed values. This indicates that the negative binomial regression model does not effectively fit the data.

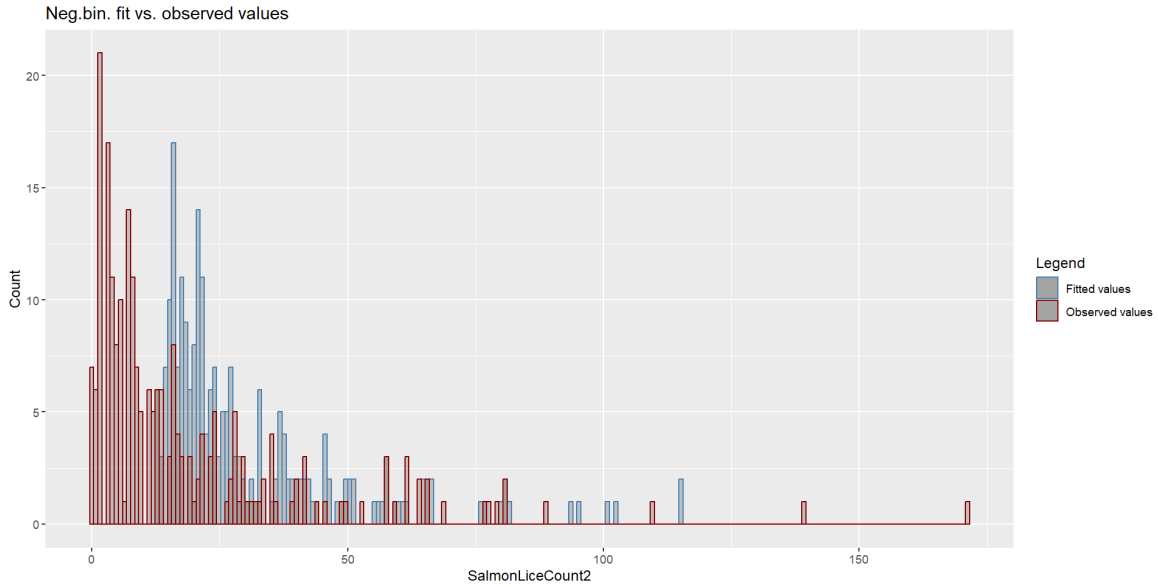


Figure 15: *The frequency of fitted (blue) and observed (red) values from the negative binomial regression model for the count of salmon lice in the 2-sample.*

The AIC value obtained from the negative binomial model is lower than the AIC of the Poisson

model, providing additional evidence that the negative binomial model was a better fit. However, a more direct comparison between the two models can be done using a likelihood ratio test. The parameter vector for the Poisson regression model is denoted as $\boldsymbol{\theta_0} = \boldsymbol{\beta} = (\beta_0, ..., \beta_5)$ and the parameter vector for the negative binomial regression model is $\boldsymbol{\theta} = (\beta_0, ..., \beta_5, \phi)$. This implies that $\boldsymbol{\theta_0}$ is a subset of $\boldsymbol{\theta}$, indicating that the Poisson model is nested within the negative binomial model. Hence, a likelihood ratio test can be employed to compare the models for the 2-sample data, considering the Poisson and negative binomial distributions. The likelihood ratio test aims to evaluate the null hypothesis $H_0 : \phi = 0$ against the alternative $H_1 : \phi > 0$. The test can be computed using the `R`-function `lrtest` available in the `lmtest` package. The results of this test are presented in Table 6. The low p-value, less than $2 \cdot 10^{-16}$ indicates that the negative binomial model is a more appropriate choice for the data compared to the model.

Number of df.	log-likelihood	df	chisq.	p-value
6	-14721			
7	-4676	1	20089	$< 2 \cdot 10^{-16}$

Table 6: *Summary of the likelihood ratio test between the Poisson regression model and the negative binomial model for salmon lice in the 2-sample.*

4.6 Multiple Linear Regression

Even though the negative binomial regression model fits the data better than the Poisson regression model according to the residual plots and AIC, the negative binomial model was still not a good fit to the observed count data, as seen in the frequency plot given in Figure 15. We therefore try to fit a simpler model, namely the log-transformed multiple linear regression model. First, we investigate how a log-transformation affects the response variable, *SalmonLiceCount2*.

4.6.1 Effect of log-transformed response

Figure 16 showcases histograms of the counts and log-transformed counts of salmon lice in the 2-sample with the associated normal distribution displayed as a blue line. The histogram on the left side shows the unlogged counts, i.e. *SalmonLiceCount2*. This histogram illustrates a left-skewed distribution, indicating a high density of counts with small values and a small density of counts with large values. On the right side, the histogram of the log-transformed counts, i.e. *LogSalmonCount2*, displays a more symmetric and normal distribution. This is attributed to the log-transformation, which helps in achieving a more reasonable fit to the assumption of normality in the multiple linear regression model. Consequently, using the log-transformed count response variable is considered a more appropriate choice than the unlogged count for this model.

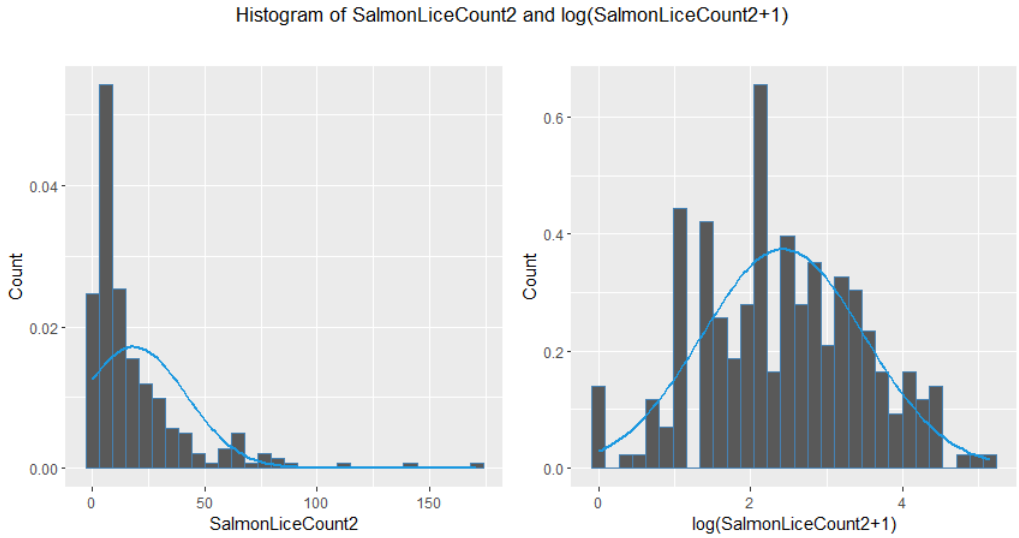


Figure 16: *Histogram of the estimated counted salmon lice numbers (left) and the log-transformed estimated counted salmon lice numbers (right) in the 2-sample. The blue line shows the belonging estimated density function.*

4.6.2 The fitted log-linear model

Table 21 presents the output summary of the multiple linear regression model with *LogSalmonLiceCount2* as response variable. The table includes the estimated regression coefficients along with their standard error, t-value and p-value. The p-values from the Wald test indicate that none of the terms except *SeaTemperature* are significant up to a significance level 0.05. From the estimated coefficient of *SeaTemperature*, the model suggests that the delousing is less effective when the temperature is high. In this model, we used *Freshwater* as a reference level in the factor variable *Method*. The summary output therefore indicates that neither *LiceFlusher* nor *Optilicer* performs better or worse than *Freshwater*. A model using *LiceFlusher* as a reference level was fitted to see if *Optilicer* performed better or worse than *LiceFlusher*. The summary output from this model is presented in Appendix B. The result indicated significance of *Optilicer*, indicating that *Optilicer* performed better than *LiceFlusher* ($\hat{\beta}_{LiceFlusher} = -0.26$, p-value : 0.022).

Table 7: *Regression coefficients with associated estimate, standard error, t-value and p-value from the log-transformed multiple linear regression for the count model for salmon lice in the 2-sample.*

Coefficient	Estimate	Std.Error	t-value	p-value
Intercept	-1.77	0.55	-3.22	0.0015
LiceFlusher	0.26	0.34	0.78	0.44
Optilicer	0.11	0.34	0.32	0.75
NumberOfFish	$7.88 \cdot 10^{-7}$	$1.32 \cdot 10^{-6}$	0.59	0.55
AverageWeight	$5.41 \cdot 10^{-5}$	$6.52 \cdot 10^{-5}$	0.83	0.41
SeaTemperature	-0.090	0.037	-2.41	0.0033

AIC: 581. Residual standard error : 1.04 on 235 degrees of freedom. R^2 : 0.56 and R^2_{adj} : 0.55. F-statistic: 59 on 5 and 235 degrees of freedom. p-value: $< 2 \cdot 10^{-16}$.

In Figure 17, the residual plot of the log-transformed multiple linear regression is presented, where the studentized residuals are plotted against the fitted values. The plot illustrates that the residuals are scattered around zero, with a relatively low magnitude, ranging from -2.2 to 5.5. In contrast to the residual plots for the Poisson and negative binomial regression models, this plot does not exhibit clear signs of heteroscedasticity, and the error variance of the studentized residuals seems relatively constant. This could be attributed to the log-transformation applied to the response variable, which has the potential to enhance the homoscedasticity of the model residuals. However,

comparing this residual plot directly to the ones from the Poisson and negative binomial models is not straightforward due to the different types of residuals used in each model.

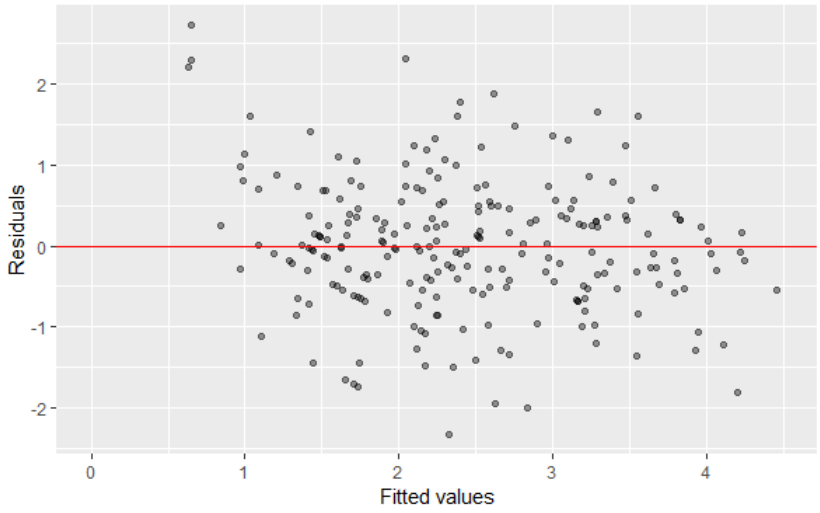


Figure 17: *Plot of studentized residuals against the fitted values from the log-transformed multiple linear regression for the model of salmon lice in the 2-sample. To visualize overlapping, the data points are partially transparent. The data points with the lightest black colour indicate a single point, while a darker black colour indicate overlapping points.*

Figure 18 displays a frequency plot of the observed values and fitted values obtained from the log-transformed multiple linear regression model. The multiple linear regression model appears to fit the data well. It also looks like the model also succeeds in fitting the small values, compared to the Poisson and negative binomial model.

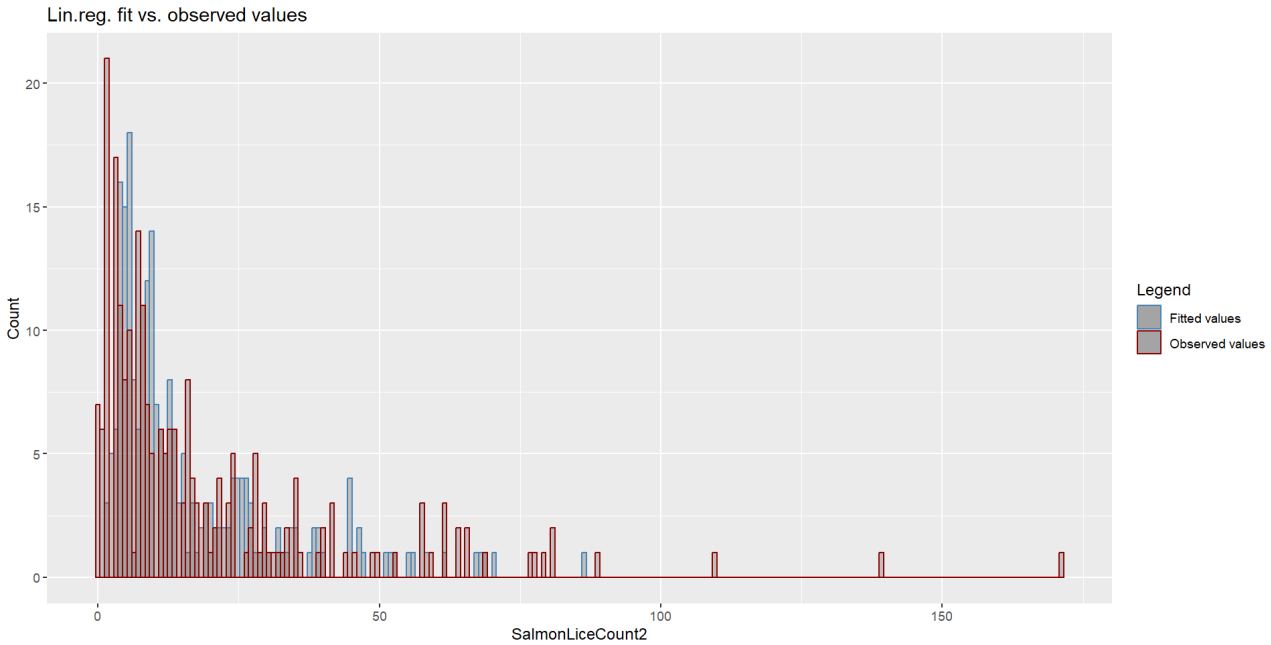


Figure 18: *Frequency plot of and fitted (blue) and observed (red) values from the log-transformed multiple linear regression model of counted salmon lice in the 2-sample.*

4.7 Random Intercept Model

4.7.1 Specifications of the model

Based on the results from the Poisson, negative binomial and log-transformed multiple linear regression model, the log-transformed multiple linear regression model clearly fits the observed count data best. However, none of the models considers that samples from the same location might be correlated. Determined by the positive results of the log-transformed multiple linear regression model, we use the log-transformed count of salmon lice in the 2-sample, i.e., *LogSalmonLiceCount2*, as a response variable in the random intercept model.

4.7.2 The fitted random intercept model

The random intercept model was fitted in `R` with the function `lmer` from the `lme4` package. The random intercept was based on clustering the data based on their location, i.e., the *Location* variable. The summary output from the model fit is presented in Table 8 and 9. Table 8 presents the estimations of the random effects and Table 9 presents the estimations of the fixed effects.

From Table 8, we can see that $\hat{\tau_0}^2 = 0.3546$ and $\hat{\sigma_0}^2 = 0.4913$. We can therefore calculate the ICC of the model using Equation (34) as ICC $= \frac{0.3546}{0.3546+0.4913} = 0.4192$. This ICC value suggests a moderate level of clustering or between-group variation. Specifically, 42% of the variability in the outcome can be attributed to differences between the locations, while the remaining 58% of the variation is due to variations within each location.

Table 8: *Random effects with associated variance and standard deviation from the random intercept model for the count model for salmon lice in the 2-sample.*

Groups	Name	Variance	Std.Dev
Location	(Intercept)	0.3546	0.5955
Residual		0.4913	0.7009

Number of obs: 241, groups: Location,35.

From Table 9, none of the parameters in the models were significant at a 5%-level according to the random intercept model. We note that *SeaTemperature* is no longer significant, as it was in the multiple linear regression model. Also in this model, *Freshwater* was used as a reference level in the factor variable *Method*. The summary output therefore indicates that neither *LiceFlusher* nor *Optilicer* outperforms *Freshwater*. A model using *LiceFlusher* as a reference level was fitted to see if *Optilicer* performed better or worse than *LiceFlusher*. The summary output from this model is presented in Appendix B and indicated no significant variables ($\hat{\beta_{LiceFlusher}} = -5.45, p - value$: 0.69). Therefore, the random intercept model did not suggest that *Optilicer* performed better (or worse) than *LiceFlusher*.

Table 9: *Regression coefficients for the fixed effects with associated estimate, standard error, t-value and p-value from the random intercept regression for the count model for salmon lice in the 2-sample.*

Coefficient	Estimate	Std.Error	t-value	p-value
Intercept	-1.47	0.70	-2.11	0.034
LiceFlusher	0.33	0.45	0.73	0.50
Optilicer	0.43	0.44	0.97	0.35
NumberOfFish	$-7.64 \cdot 10^{-7}$	$2.20 \cdot 10^{-6}$	-0.35	0.72
AverageWeight	$-6.58 \cdot 10^{-5}$	$8.01 \cdot 10^{-5}$	-0.82	0.42
SeaTemperature	-0.071	0.040	-1.77	0.076

The individual factor level random effects are not included in the output summaries. However, the location-specific intercept coefficients can be found with the `coef`-function in `R`. The output from

this function, alongside the relative frequency of the location (in %) is presented in Table 10. Due to privacy reasons, the names of the locations are not included in the table. It is clear that each location has its own intercept. The slope coefficient for the fixed effects is not presented, but they are the same as the ones given in Table 9 and stay fixed for all the locations.

Location	(Intercept)	Relative frequency of Location (%)
Location 1	-2.816	5.603
Location 2	-2.877	0.431
Location 3	-3.134	2.586
Location 4	-3.544	0.431
Location 5	-2.173	4.741
Location 6	-2.998	1.293
Location 7	-3.065	3.017
Location 8	-3.266	2.155
Location 9	-3.770	1.724
Location 10	-3.781	9.483
Location 11	-2.559	0.862
Location 12	-2.498	0.431
Location 13	-2.370	0.431
Location 14	-2.551	16.379
Location 15	-2.940	1.293
Location 16	-3.692	0.431
Location 17	-2.484	1.724
Location 18	-2.745	1.724
Location 19	-2.969	4.741
Location 20	-2.574	4.310
Location 21	-4.024	3.017
Location 22	-3.392	13.362
Location 23	-2.661	0.862
Location 24	-3.491	2.586
Location 25	-2.854	1.724
Location 26	-3.980	0.431
Location 27	-3.571	0.862
Location 28	-2.596	1.724
Location 29	-3.216	1.293
Location 30	-2.711	1.724
Location 31	-2.941	1.724
Location 32	-3.307	2.586
Location 33	-3.661	0.862
Location 34	-3.060	1.724

Table 10: *Location-specific intercept coefficients from the random intercept model for the count model of salmon lice in the 2-sample.*

Figure 19 presents the residuals plot of the residuals against the fitted values from the random intercept model. The plot shows that the residuals are scattered around zero. The magnitude of the residuals stretches from -2.6 to 2.6 and is quite similar to the one from the log-transformed linear regression model. The error variance appears to be quite constant, indicating presence of homoscedasticity.

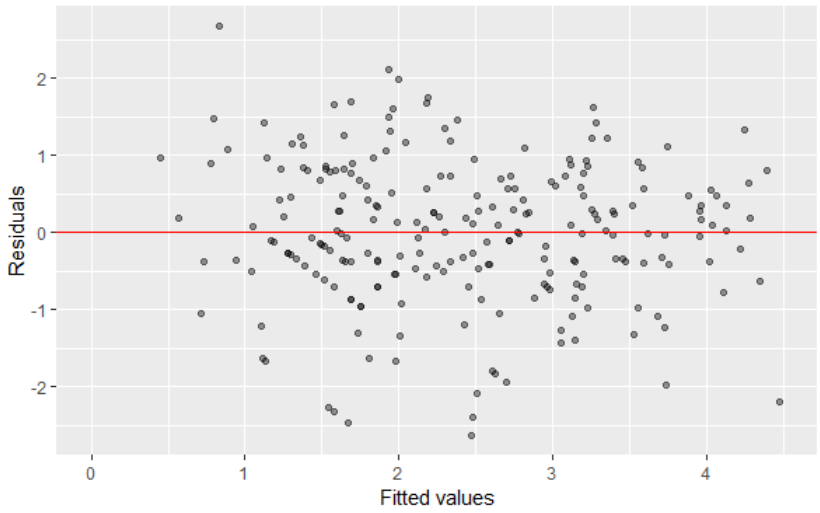


Figure 19: *Plot of residuals against the fitted values from the random intercept regression for the model of salmon lice in the 2-sample. To visualize overlapping, the data points are partially transparent. The data points with the lightest black colour indicate a single point, while a darker black colour indicate overlapping points.*

A frequency plot of fitted and observed values of the count of salmon lice in the 2-sample is presented in Figure 20. The plot shows that the random intercept model fits the data well. The fitted values from the models manages to fit the small observed values and the shape of the fitted frequency plot coincides with the shape of the observed frequency plot.

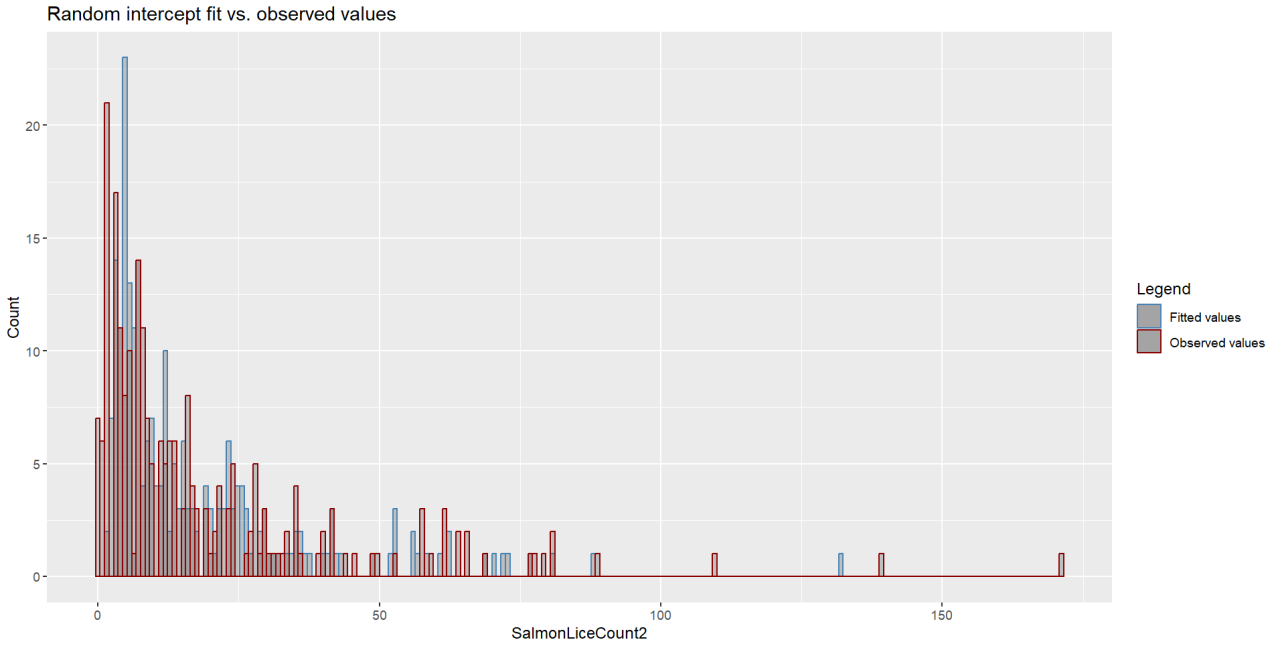


Figure 20: *The frequency of observed (blue) and fitted (red) values from the random intercept model of counted salmon lice in the 2-sample.*

4.8 A continuation of investigating re-infestation of salmon lice after delousing treatment

In Mæland 2022, we studied regression models for count data with the application of investigating re-infestation of salmon lice after delousing treatment. We did this by looking at the count of mobile lice in the 3-sample using the count of mobile lice in the 2-sample as a baseline count. The data was modelled using the Poisson, negative binomial and log-transformed multiple linear regression models. We used only the count of mobile lice, and not the total amount of salmon lice, because mobile lice are the only lice that can move freely in the water and therefore re-attach to the salmon after delousing treatment. By looking at residual plots and frequency plots of the observed and fitted values, the log-transformed multiple linear regression model fitted the data best among the three models, just as seen for the current study of salmon lice in the 2-sample. Without first investigating how to best include the baseline count in the regression models, it was included as a log-transformed covariate based on the results from Zheng et al. 2018. In Figure 5 we see evidence of re-infestation in that the number of mobile lice has substantially increased shortly after delousing treatment. With the presence of significant covariates in our models, we can say that these factors could have protected against or worsened re-infestation of mobile lice. Table 11 presents the summary output of the multiple linear model of mobile lice in the 3-sample from Mæland 2022, including the baseline count of mobile lice in the 2-sample as a log-transformed covariate. The data showed indications of re-infestation and the model suggested that *AverageWeight* and *SeaTemperature* were factors that could worsen the re-infestation and that *Placement* and *LiceSkirt* could protect it. That is, an increase in the average weight and sea temperature increased the amount of mobile lice in the 3-sample, and that the presence of lice skirt and placing the salmons in a new cage (instead of the old cage) after delousing decreased the amount of mobile lice in the 3-sample.

Coefficient	Estimate	Std.Error	t-value	p-value
Intercept	-5.29	0.46	-11.59	$2.00 \cdot 10^{-16}$
Placement	-0.58	0.16	-3.52	0.00055
LiceFlusher	0.28	0.32	0.88	0.38
Optilicer	-0.20	0.28	-0.72	0.47
LiceSkirt	-0.31	0.15	-2.10	0.037
NumberOfFish	$1.13 \cdot 10^{-6}$	$1.41 \cdot 10^{-6}$	0.80	0.42
AverageWeight	$2.33 \cdot 10^{-4}$	$5.60 \cdot 10^{-5}$	4.17	$4.89 \cdot 10^{-5}$
SeaTemperature	0.054	0.032	1.69	0.046
log(MobileLice2 + 1)	1.11	0.23	4.90	$2.23 \cdot 10^{-6}$

Residual standard error: 0.71 on 169 degrees of freedom. R^2 : 0.42 and R^2_{adj} : 0.39. F-statistic: 15.06 on 8 and 169 degrees of freedom. P-value: $2.2 \cdot 10^{-16}$.

Table 11: *Regression coefficients with associated estimate, standard error, t-value and p-value from the multiple normal regression for the log-transformed count of mobile lice in the 3-sample.*

We are now interested in studying the re-infestation of salmon lice with modified regression models. The main difference will be how the baseline count of mobile lice in the 2-sample is included in the regression models. Instead of including the baseline count as a log-transformed covariate, we now want to include the baseline count as a log-transformed offset. In addition, we also want to use the random intercept model to see if clustering based on the location of the salmon cages further improves the fit of the regression models and alters the results of the 3-sample count models.

Seeing that we are interested in studying the 3-sample, we use a different set of response variables and covariates than what have used so far in this thesis. The response variable is now based on the count of mobile lice in the 3-sample. All the covariates used so far are included, but we also include *Placement* and *LiceSkirt* as covariates in the regression model. Both of these explanatory variables are only relevant for what happens when the salmon is brought back to the salmon cages *after* delousing treatment and are therefore not relevant for the studies of the 2-sample. The different variables used in the analysis of investigating re-infestation of salmon lice is presented in Table 12 with explanation.

Variable Name	Explanation of Variable
MobileLice2	Lice number of mobile lice in the 2-sample. Reported average of mobile lice in the 2-sample calculated from a sample of at least 20 salmon
MobileLice3	Lice number of mobile lice in the 3-sample. Reported average of mobile lice in the 3-sample calculated on a sample of at least 20 salmon
SampleCount2	The number of salmons used to count the salmon lice in the 2-sample. SampleCount2 is usually 20.
SampleCount3	The number of salmons used to count the salmon lice in the 3-sample. SampleCount3 is usually 20.
MobileLiceCount2	The count of mobile lice in the 2-sample. Calculated as MobileLice2 · SampleCount2.
MobileLiceCount3	The count of mobile lice in the 3-sample. Calculated as MobileLice3 · SampleCount3.
LogMobileLiceCount3	The log-transformed lice number of mobile lice in the 3-sample. Calculated as $\log(\frac{MobileLiceCount3+1}{SampleCount3})$.
Date	Date of the observation
SeaTemperature	Temperature of the sea, ° C
AverageWeight	Average weight of the salmon in the cage, g
NumberOfFish	Number of fish in the cage
Method	Delousing method used, divided into freshwater treatment, thermic treatment and lice flusher
Placement	Placement indicator variable, coded as: 0 - salmon placed back in same cage after treatment, 1 - salmon placed in new cage after treatment
LiceSkirt	Skirt indicator variable, coded as: 0 - Lice skirt around the salmon cage, 1 - No lice skirt around the salmon cage

Table 12: *The variables used in the analysis with explanation and units.*

4.8.1 Poisson Regression Model

Following Mæland 2022, we start with the most common regression model for count data, namely the Poisson regression model. Table 13 presents the summary output from the model fit with *MobileLiceCount3* as response variable. The summary output includes estimated regression coefficients along with their corresponding standard errors, t-values, and p-values.

Table 13: *Regression coefficients with associated estimate, standard error, t-value and p-value from the Poisson regression for the model of mobile lice in the 3-sample.*

Coefficient	Estimate	Std.Error	t-value	p-value
Intercept	-0.37	0.21	-1.72	0.085
Placement	-0.28	0.066	-4.20	$2.71 \cdot 10^{-5}$
LiceFlusher	0.67	0.16	4.08	$4.56 \cdot 10^{-5}$
Optilicer	0.55	0.15	3.61	0.00031
LiceSkirt	-0.42	0.059	-7.20	$6.09 \cdot 10^{-13}$
NumberOfFish	$2.47 \cdot 10^{-6}$	$5.63 \cdot 10^{-7}$	4.39	$1.16 \cdot 10^{-5}$
AverageWeight	$1.93 \cdot 10^{-4}$	$2.15 \cdot 10^{-5}$	8.97	$< 2.00 \cdot 10^{-16}$
SeaTemperature	0.14	0.012	11.65	$< 2.00 \cdot 10^{-16}$

AIC: 2392.4, Null deviance: 1947.3 on 181 degrees of freedom, Residual deviance: 1663.4 on 174 degrees of freedom.

The p-values from the Wald test indicated that all the terms were significant. A goodness-of-fit analysis of the model was conducted using the residual deviance. The critical value in the $\chi^2_{\alpha,n-p}$ distribution was determined as $\chi^2_{0.05,166} = 197$ for a significance level $\alpha = 0.05$. Since the residual deviance $D = 1663 > 197 = \chi^2_{0.05,166}$, it can be concluded that the Poisson model is not a good fit to the data.

A hypothesis test was then performed to assess overdispersion using the Pearson statistic. The null hypothesis, H_0 suggested the absence of overdispersion was formulated as $H_0 : \phi \leq 1$. The null hypothesis was tested against the alternative hypothesis, $H_1 : \phi > 1$, which posited the presence of overdispersion. Under the null hypothesis, the Pearson statistic, P, adhered to a $\chi^2_{\alpha,n-p}$ distribution. As $P = 2380$ exceeds $\chi^2_{\alpha,n-p} = 197$, the null hypothesis was reached and a conclusion of overdispersion was reached.

A figure of the deviance and Pearson residuals plotted against the fitted values from the Poisson regression model is presented in Appendix A.

4.8.2 Negative Binomial Regression Model

To accommodate a greater variance in the count data than the Poisson model, the negative binomial regression model was applied to the model of mobile lice in the 3-sample. Table 14 presents the summary output from the model fit in `R`. The p-values from the Wald test indicated that all the terms, excluding *LiceFlusher* and *Optilicer*, exhibited significance at a level of 0.05. Using Equation (24), the Pearson statistic was calculated as $P = 263$, less than the Pearson statistic obtained from the Poisson model. Despite this, the corresponding quantile $\chi^2_{0.05,166} = 197$ was still lower than the Pearson statistic, implying that the fitted model failed to align with the actual distribution. The residual plot is presented in Figure 28 in Appendix A.

Table 14: *Regression coefficients with associated estimate, standard error, t-value and p-value from the negative binomial regression for the model of mobile lice in the 3-sample.*

Coefficient	Estimate	Std.Error	t-value	p-value
Intercept	-0.053	0.57	-0.093	0.93
Placement	-0.48	0.20	-2.44	0.015
LiceFlusher	0.41	0.40	1.03	0.30
Optilicer	0.36	0.35	1.02	0.31
LiceSkirt	-0.32	0.18	-1.77	0.076
NumberOfFish	$3.89 \cdot 10^{-6}$	$1.69 \cdot 10^{-6}$	2.20	0.022
AverageWeight	$1.87 \cdot 10^{-4}$	$6.78 \cdot 10^{-5}$	2.75	0.0059
SeaTemperature	0.14	0.039	3.50	0.00047

AIC: 1295.5, Null deviance: 230.9 on 181 degrees of freedom, Residual deviance: 200.6 on 174 degrees of freedom.

4.8.3 Log-transformed Multiple Linear Regression Model

Both the Poisson regression model and the negative binomial regression model failed to fit the observed count data of mobile lice. We fitted a log-transformed multiple linear regression to the model of mobile lice in the 3-sample in `R`. The summary output from the model fit is given in Table 15. The p-values from the Wald test indicate that only *Placement*, *AverageWeight* and *SeaTemperature* are significant up to a significance level 0.05. With positive estimated coefficient for *SeaTemperature* and *AverageWeight*, the prevalence of lice in the 3-sample increase with increasing sea temperature and increased average weight of the salmon. With negative estimated coefficient for *Placement* (coded as a factor variable; 0 for the same cage after delousing and 1 for a new cage after delousing), the prevalence of salmon lice in the 3-sample decreased when the salmon was placed in a new cage after delousing treatment. This result is not fully consistent with the results in Mæland 2022, where also *LiceSkirt* showed significance.

Table 15: *Regression coefficients with associated estimate, standard error, t-value and p-value from the log-transformed multiple linear regression for the model of mobile lice in the 3-sample.*

Coefficient	Estimate	Std.Error	t-value	p-value
Intercept	0.47	0.54	0.87	0.39
Placement	-0.57	0.19	-3.01	0.0030
LiceFlusher	0.21	0.37	0.56	0.57
Optilicer	0.076	0.33	0.23	0.82
LiceSkirt	-0.26	0.18	-1.48	0.14
NumberOfFish	$3.13 \cdot 10^{-6}$	$1.64 \cdot 10^{-6}$	1.90	0.059
AverageWeight	$1.84 \cdot 10^{-4}$	$6.58 \cdot 10^{-5}$	2.80	0.0057
SeaTemperature	0.096	0.037	2.57	0.011

Residual standard error: 0.85 on 174 degrees of freedom. R^2 : 0.44 and R^2_{adj} : 0.41. F-statistic: 19.22 on 7 and 174 degrees of freedom. P-value: $< 2.2 \cdot 10^{-16}$.

Figure 21 presents a plot of the residuals versus the fitted values in the log-transformed multiple linear regression model. The residual appears to be randomly shattered around zero. In addition, the magnitudes of the residuals are relatively low, further indicating that the model was a good fit to the data.

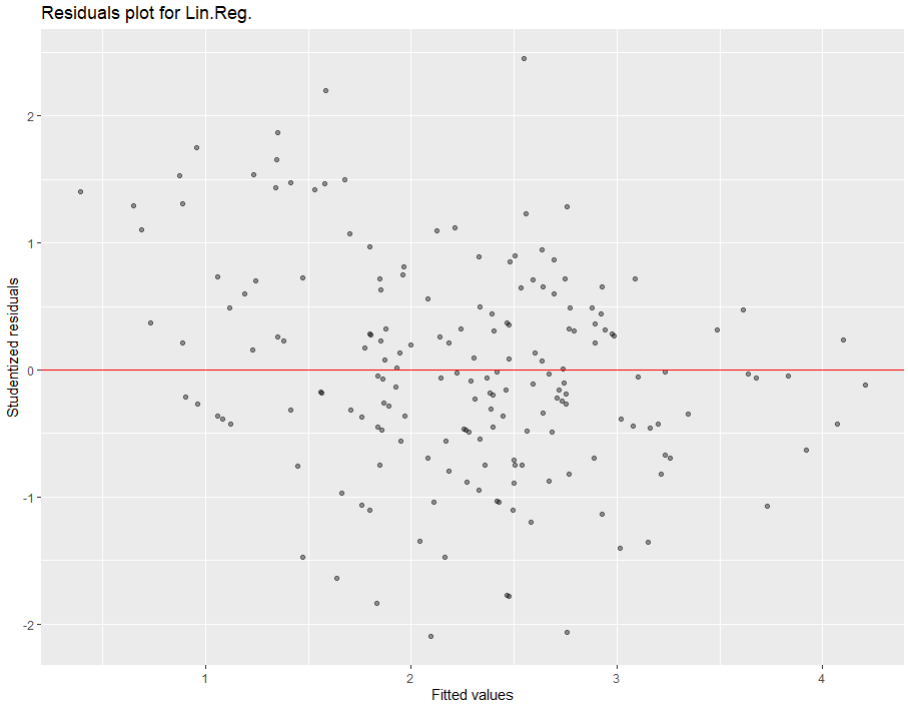


Figure 21: *Plot of residuals against the fitted values from the negative binomial regression for the model of mobile lice in the 3-sample. To visualize overlapping, the data points are partially transparent. The data points with the lightest black colour indicate a single point, while a darker black colour indicate overlapping points.*

Figure 22 presents a frequency plot of the fitted values from the log-transformed multiple linear regression model against the observed count data of mobile lice in the 3-sample. The fitted values from the model fits the observed count data reasonably well.

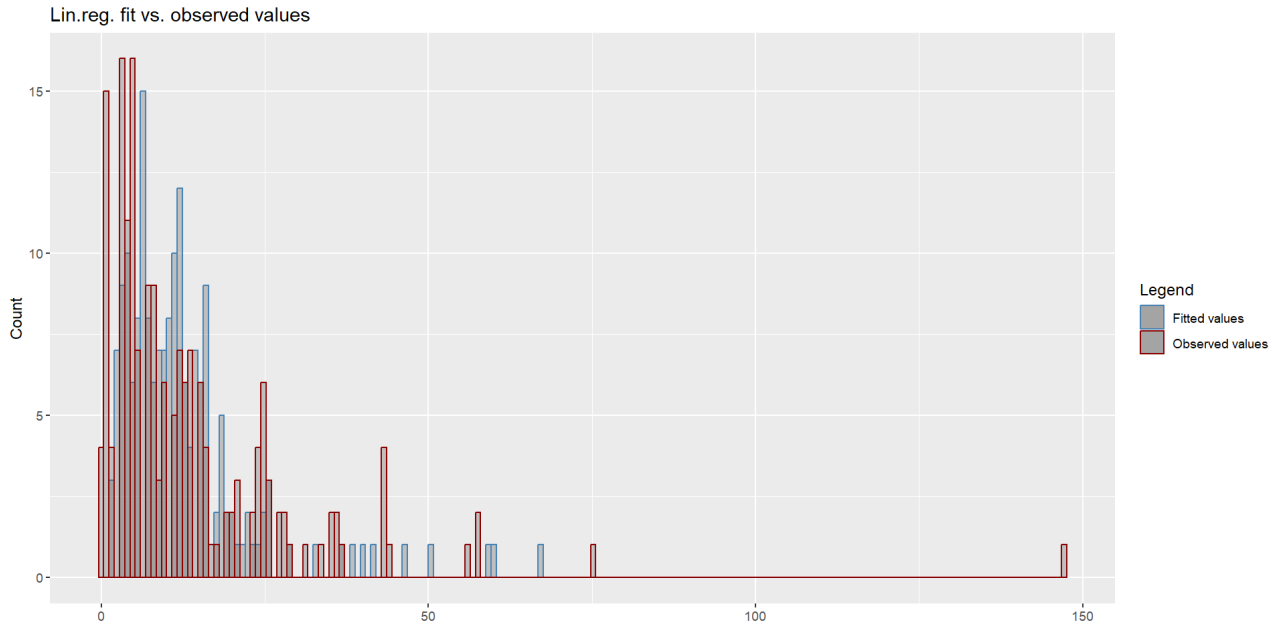


Figure 22: *The frequency of observed (blue) and fitted (red) values from the log-transformed multiple linear regression model of counted mobile lice in the 3-sample.*

4.8.4 Random Intercept Model

Knowing that the data can be clustered based on location, we fit a random intercept regression to the model of mobile lice in the 3-sample in `R`. The summary output from the model is presented in Table 16. Also for this model, the p-values obtained from the Wald test suggests that *Placement*, *AverageWeight* and *SeaTemperature* are significant up to a level of 0.05. This result is consistent with the results obtained in the log-transformed multiple linear model above, and thus not consistent with the results obtained in Mæland 2022.

Table 16: *Regression coefficients with associated estimate, standard error, t-value and p-value from the random intercept regression for the model of mobile lice in the 3-sample.*

Coefficient	Estimate	Std.Error	t-value	p-value
Intercept	0.37	0.66	0.57	0.53
Placement	-0.63	0.22	-2.91	0.008
LiceFlusher	0.13	0.45	0.29	0.76
Optilicer	0.21	0.42	0.51	0.61
LiceSkirt	-0.23	0.23	-0.99	0.32
NumberOfFish	$1.27 \cdot 10^{-6}$	$2.10 \cdot 10^{-6}$	0.63	0.51
AverageWeight	$2.00 \cdot 10^{-4}$	$7.30 \cdot 10^{-5}$	2.75	0.008
SeaTemperature	0.12	0.042	2.78	0.006

In Figure 23 a plot of the residuals for the random intercept model is presented. We see that the residuals are randomly scattered around the horizontal axis of 0.

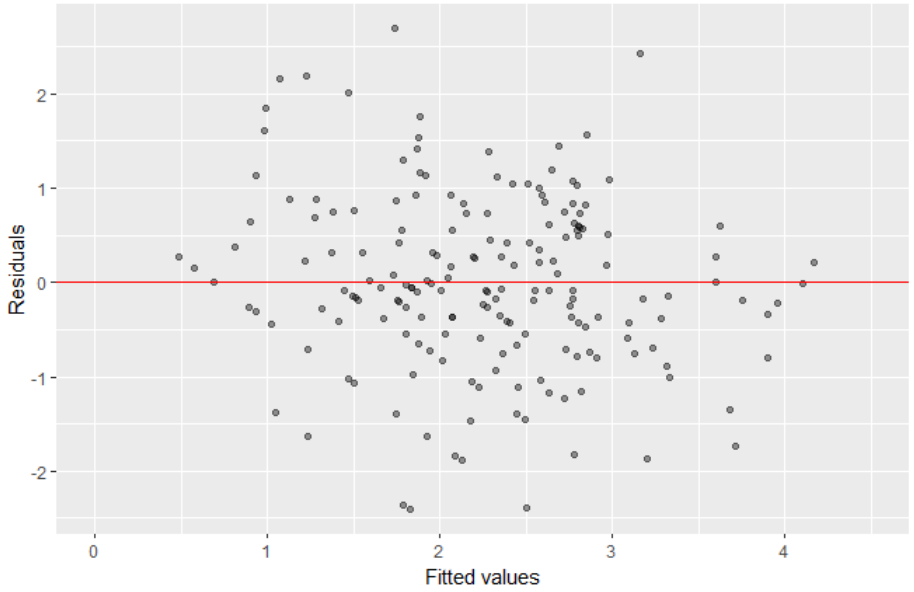


Figure 23: *Plot of residuals against the fitted values from the random intercept regression for the model of mobile lice in the 3-sample. To visualize overlapping, the data points are partially transparent. The data points with the lightest black colour indicate a single point, while a darker black colour indicate overlapping points.*

Figure 24 presents the frequency plot of the observed data and the fitted data from the random intercept model. From the figure it appears that the random intercept model fits the observed data well.

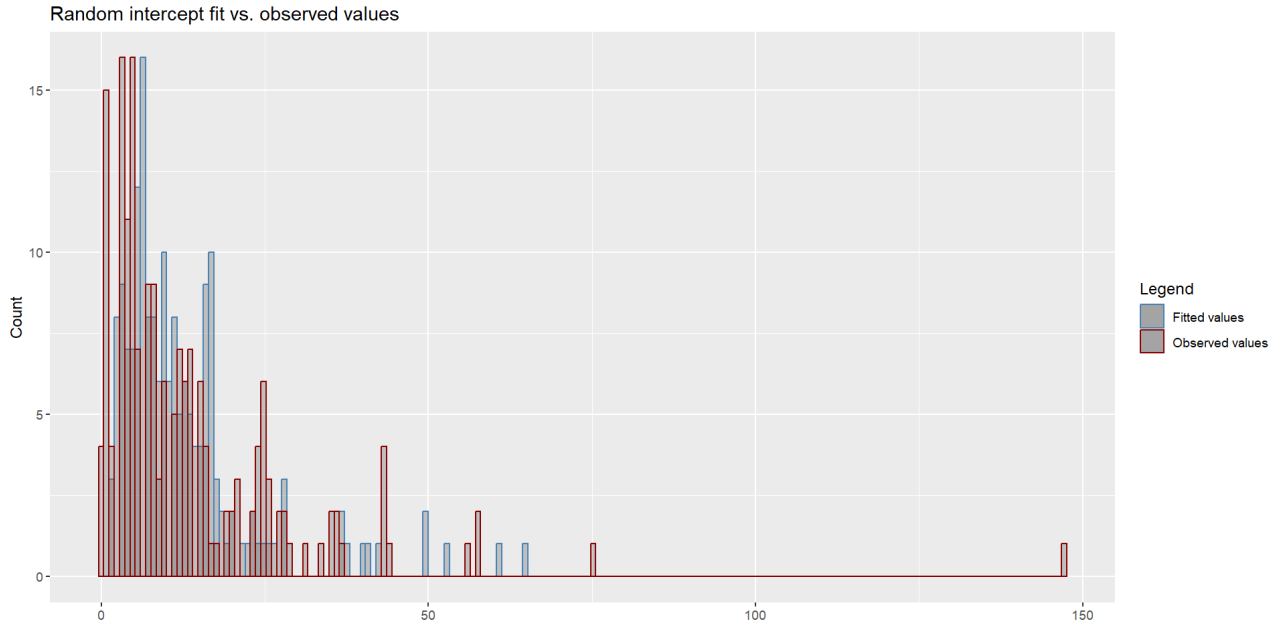


Figure 24: *The frequency of observed (blue) and fitted (red) values from the random intercept model of the counted mobile lice in the 3-sample.*

5 Discussion

5.1 Remarks on the regression models

The model for the counted salmon lice in the 2-sample, and the model for counted mobile lice in the 3-sample was fitted in R with a Poisson, negative binomial, log-transformed multiple linear and random intercept regression model. It is natural to think that the Poisson regression model and the negative regression model would perform well on the count data as these are the two most common models for count data. However, since the Poisson regression model assumes that the estimated value and the variance of the data is equal, it was not suitable for the given count data of salmon lice in the 2-sample or the mobile lice in the 3-sample. That was seen from the lack of fit in the frequency plots and the large residuals. The negative binomial model is supposed to be more suitable for models with larger variance than the mean as it handles overdispersion. Even though the negative binomial model outperformed the Poisson regression model, it did not fit the observed data well, as was observed in the frequency- and residual plots. In both the Poisson and negative binomial regression model, the residual plots showed evidence of heteroscedasticity, meaning that the error variance was not constant. This was a violation of the model assumptions of constant error variance, and we concluded that these models were not a good fit to the observed count data.

The log-transformed multiple linear regression fitted the observed count data better than the Poisson and negative binomial model by the look of the frequency plots and the model residuals. The frequency plot showed that the fitted data from the model was well suited for the observed count data, and the residual plots showed homoscedastic residuals with a relatively low magnitude. The reason for the log-transformed multiple linear models performing better than the Poisson and negative binomial models may be due to two things; the log-transformation of the response variable and the linear models simplicity. As mentioned in Section 2.3, a log-transformation of the response variable is known to improve normality and more importantly, the homoscedasticity of the model residuals. This effect was visualized in the residual plots. Furthermore, the dataset we worked with was very small with few observations and limited covariates. As we mentioned in Section 4.2, simpler models often tend to be more suitable for small datasets. The multiple linear model is considered simpler than the Poisson and negative binomial models due to its assumptions about the response variable, the straightforward interpretation of coefficients, the less complex modelling process and the preference for simplicity when the assumptions of multiple linear regression are met.

The random intercept model also used a log-transformed response variable. From the residual plots for the model of the salmon lice in the 2-sample and the mobile lice in the 3-sample, we saw that the residuals had a quite low magnitude with a constant error variance. Furthermore, the frequency plot of the fitted values from the random intercept model indicated that the model fitted the observed count data well. Both the frequency plot and the residual plot were quite similar to the ones obtained from the multiple linear model, and we concluded that the random intercept model with log-transformed counts was a good fit to the data.

In Table 17, the model statistics from the model fits of the count of salmon lice in the 2-sample is presented. A similar table of the model statistics from the model fits of the count of mobile lice in the 3-sample is presented in Appendix B. One has to take into account that the log-likelihood and the AIC of the Poisson and negative binomial cannot be directly compared with the log-likelihood and AIC of the multiple linear and random intercept due to the log-transformation of the response variable in the two latter models. However, the magnitude of the AIC and log-likelihood is significantly higher for the Poisson and negative binomial model compared to the multiple linear and random intercept model.

	Poisson	Neg.bin	Multiple linear	Random Intercept
Degrees of freedom	7	7	7	8
Log-likelihood	-1054	-861	-283	-286
AIC	2122	1735	581	589

Table 17: *A comparison of the fitted regression models for the count of salmon lice in the 2-sample.*

Because of the lack of fit, high residuals, high AIC and low log-likelihood of the Poisson and negative binomial model, we did not use the results obtained in these models to study the count of salmon lice in the 2-sample and the count of mobile lice in the 3-sample. The results from these models would not have explained enough and not been certain. As an example, we may look at the summary outputs from the models that indicated that some or all covariates in the regression model was significant. In the case of the study of salmon lice in the 2-sample, the multiple linear only indicated one significant covariate, and the random intercept model indicated zero significant covariates. Therefore, the significant covariates outputted from the Poisson and negative binomial regression models showed false positives due to model misspecifications. The multiple linear and random intercept regression models with a log-transformed response variable gave a better model fit and is therefore considered to be more accurate to the observed data. We therefore only used the results from the multiple linear and random intercept model in this thesis.

5.2 Comparison of treatment methods

In this thesis we were interested in studying the non-medicinal treatment methods *Freshwater*, *LiceFlusher* and *Optilicer*. More specifically, we wanted to look at the effects of the treatment methods and see if any of them performed better than the others. In Section 4.4, we presented a box plot of the three treatment methods against the number of salmon lice in the 2-sample (with and without a log-transformation). This plot indicated only minor differences in the prevalence of salmon lice in the 2-sample following the three different treatment methods. We therefore get an early indication that none of the treatment methods outperforms the others. We also noted a significant imbalance in the dataset, with the *Freshwater* treatment being applied only 11 times, while the *LiceFlusher* and *Optilicer* treatments were used 118 and 112 times respectively, which suggests that a fair comparison of the effectiveness between the *LiceFlusher* and *Optilicer* treatments versus the *Freshwater* treatment might not be possible. We still had to analyse the methods more thoroughly. The treatment method variable, *Method*, was included in the regression analysis and we fitted four different regression models for count data. The treatment method variable, *Method*, was first coded as a factor variable in `R`, where *Freshwater* was used as the reference level in order to compare it with *LiceFlusher* and *Optilicer*. Then we fitted another model using *LiceFlusher* as the reference leve, in order to compare it with *Optilicer*.

As mentioned, we only used the results from the multiple linear and random intercept model to study the treatment methods. The results from the first analysis, i.e., using *Freshwater* as the reference level, showed that neither the log-transformed multiple linear nor the random intercept model indicated any significant result for *LiceFlusher* or *Optilicer* in the summary outputs. Based on this we can say that neither of these two treatment methods distinguishes from *Freshwater*. Using *LiceFlusher* as a reference level in *Method*, the results from the multiple linear model showed significance of *Optilicer*, indicating that this method performed better than *LiceFlusher*. The random intercept model on the other hand, did not show significance of any of the covariates. This model does therefore not suggest that the *Optilicer* method performed better than the *LiceFlusher* method.

The reason for *LiceFlusher* and *Optilicer* not showing any significance against *Freshwater* may be due to the significantly greater usage of these two methods compared to *Freshwater*. For the same reason, the comparison between *LiceFlusher* and *Optilicer* when using *LiceFlusher* as a reference appears more equitable. Seeing that the linear model showed significance of *Optilicer* and the random intercept model did not, it may have been that the treatment methods were not randomly distributed across locations. Therefore it impossible determine, based on the given information, whether it is the treatment methods itself that produce the different results, or whether it is the

locations (and implicitly also companies) that led to different results.

One can also stress that the nature of the dataset and more specifically - the design of the study - makes it difficult to give a fair evaluation of the treatment methods. The dataset is very small and there are multiple factors that differs between the observations besides the treatment methods. The study this dataset is based on is an observational study, not specifically designed to analyse the treatment methods. To say something certain about the treatment methods, we would recommend a more clinical design on the study, with the purpose of studying the effect of the treatment methods. In a clinical trial designed to investigate the effect of one specific factor, a controlled experimental design is typically used. Generally speaking, in such an experiment, participants are randomly assigned to different groups representing the levels of the factor under investigation, and outcome measures are collected and analysed to compare the outcomes between the groups, accounting for potential confounding variables. This design allows for isolating the impact of the specific factor, which in this case would have been the treatment methods, and obtaining a clearer understanding of its relationship with the outcome of interest.

However, as the intention of the *RegLus*-project was not only to study the effect of the different treatment methods, but also to e.g., study the re-infestation after delousing, an observational study works well. It is a useful study in cases like these, where there are several areas of interest in the dataset, e.g., to spot trends in the dataset. Even though the dataset is small and noisy, and it is difficult to say something certain about the treatment methods based on the design of the study, significant results on the treatment methods could have indicated that one treatment method was better than the other. Now, since the only significant result is given in the linear model comparing the *LiceFlusher* method to the *Optilicer* method, and not in the random intercept model, we can not distinguish these two treatment methods.

5.3 Continuation on re-infestation

In Mæland 2022 we studied regression models on count data, and the study of interest were re-infestation of salmon lice after delousing treatment. Therefore, we looked at the count of salmon lice, more specifically of mobile lice, in the 3-sample with the count of mobile lice in the 2-sample as a baseline count. We followed the results of the simulation on how to include the baseline count from Zheng et al. 2018, where the simulations done was based on negative binomial regression models. The results obtained in Zheng et al. 2018 showed that the best way to include a baseline count was as a log-transformed covariate in the linear predictor. In Mæland 2022 we found that the log-transformed multiple linear regression model fitted the data best. Therefore, it was not obvious for those simulated count data that the models including the baseline count as a log-transformed covariate would perform better than the other models using different linear predictors. We found in Section 3, with simulations and parameters based on the real sampling scenario and the observed data, that the models (*both* negative binomial and log-transformed linear) including the baseline count as a log-transformed offset performed best for our count data. Using this baseline, we wanted to once again study the re-infestation of salmon lice after delousing.

The results from the log-transformed multiple linear regression model indicated that the variables *SeaTemperature*, *Placement* and *AverageWeight* were significant. Further, we were interested in seeing if these results stayed the same when we incorporated clusters of the data based on location. That is, we used random intercept to model the number of mobile lice in the 3-sample. The random intercept model, with a log-transformed response variable, fitted the data well. The results from the model showed the same as the log-transformed multiple linear regression model, namely that the coviarates *SeaTemperature*, *Placement* and *AverageWeight* were significant. In Mæland 2022, in addition to *SeaTemperature*, *Placement* and *AverageWeight*, the variable *LiceSkirt* was also significant. The significance of *LiceSkirt* indicated a lower count of mobile lice in the 3-sample with the presence of a lice skirt on the salmon cage. The results obtained in this thesis, including the count of mobile lice in the 2-sample as a log-transformed offset, does therefore not coincide with the results in Mæland 2022. In the case of the multiple linear regression model, the only difference from the model used in this thesis and the one used in Mæland 2022, was how the baseline count was included. In Mæland 2022 it was included as a log-transformed covariate, and in this thesis it was included as a log-transformed offset. From the simulations done in Section 3, we saw that all

the linear predictors performed quite well and that there were no significant differences between the models. Therefore, we cannot say that the model including the baseline count as a log-transformed covariate is better than the model including the baseline count as a log-transformed offset. We therefore conclude that the reason for the different results obtained in this thesis in contrast to the results obtained in Mæland 2022, could be due to the different ways of including the baseline count in the model, and in particular the *LiceSkirt* finding seems unreliable.

5.4 Problems with the dataset

The analysis of the salmon lice data encountered several potential issues. Firstly, the dataset was notably small, and improving the model's accuracy would have been possible with a larger dataset incorporating more covariates and observations. For instance, it would have been valuable to include environmental factors such as salinity, wind, and current in the model. Wind direction and currents can potentially influence the spread of salmon lice in water, making them relevant variables for regression analysis. Although the salinity parameter was absent in this dataset, its inclusion would have been interesting since previous records indicate a correlation between salinity and salmon lice prevalence. According to Dalvin, Ø. Karlsen and Samuelsen 2020, salmon lice struggle to survive in low salinity conditions and eventually detach from the host.

Furthermore, investigating the impact of cleaner fish usage in the cages would have been informative. Cleaner fish, such as wrasse (*Labridae*) and lumpfish (*Cyclopterus lumpus*), are commonly employed in salmon farms as a biological delousing method. These cleaner fish consume salmon lice on the skin of the salmon. Jevne and Reitan 2019 highlight that using cleaner fish in salmon farms can delay the time it takes for adult female lice to reach the threshold value of 0.1 per salmon at the start of the production cycle. Thus, the number of cleaner fish deployed in a cage should have been included as an explanatory variable in the regression analysis. However, this dataset lacks information regarding the number of cleaner fish deployed, making it a worthwhile variable for future research.

According to Torrissen et al. 2013, the density of salmon farms significantly affects the prevalence of salmon lice at individual sites within an area. Therefore, including a distance parameter, such as the distance to the nearest site, in the regression analysis would have been beneficial. This parameter could have illustrated how neighbouring cages impact the prevalence of salmon lice in a specific cage. Additionally, previous studies, including C. Karlsen 2021, demonstrate that the distance from the site to the coastline has a clear effect on the prevalence of salmon lice and should have been recorded as an explanatory variable.

Uncertainty arises from the estimation of the number of salmon lice in the 2-sample and 3-sample. The count was derived from a small sample of salmon multiplied by the registered average mean of salmon lice. However, relying on a sample size of 20 salmon may not accurately represent the average lice abundance in a cage. Additionally, the inconsistency in sample size, e.g., there were cases where the sample sizes was as high as 120, led to instances of higher counts due to larger sample sizes.

Another problem with the dataset is its size. There are very few observations in the dataset, and the data was only collected from a one-year period (2018-2019). It is possible that a one-year production cycle is not completely representative for e.g. a ten year production cycle, and therefore more data should have been collected and preferably over more than one year.

5.5 Conclusion and further work

In this thesis we have studied regression models for count data with applications to salmon lice data. We have used the models to fit data for both salmon lice in the 2-sample in order to study the different treatment methods and continued the studies in Mæland 2022, namely looking at mobile lice in the 3-sample to study re-infestation of mobile lice after delousing treatment.

In both studies, the multiple linear and random intercept regression model with a log-transformed

response variable fitted the observed data best. Given that the Poisson and negative binomial regression models were significantly inferior to the two log-transformed regression models, the results of these models were not used in the further analysis of the salmon lice. In summary, addressing the limitations mentioned above, such as the size of the dataset, inclusion of relevant environmental factors, improving sampling methods, could enhance the analysis of the salmon lice data, providing a more comprehensive understanding of the factors influencing their prevalence. For further work it would therefore have been interesting doing the same analysis with a larger dataset with more observations and more covariates. For the studies of re-infestation of mobile lice in the 3-sample and the investigation of the different delousing methods, results from such a analysis could have given a better indication of which aspects and factors in the dataset to study further. From this, one could for instance have set up a new study that would have been designed to study specific factors or aspects of the prevalence of salmon lice in the four samples.

References

Cameron, A Colin and Pravin K Trivedi (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Clopper, Charles J and Egon S Pearson (1934). 'The use of confidence or fiducial limits illustrated in the case of the binomial'. In: *Biometrika* 26.4, pp. 404–413.

Dalvin, S, Ø Karlsen and O Samuelsen (2020). *Topic: Sea Lice*. URL: https://www.hi.no/en/hi/temasider/species/sea-lice.

Fahrmeir, Ludwig et al. (2013). *Regression models*. Springer.

Finstad, Bengt et al. (2011). 'The effect of sea lice on Atlantic salmon and other salmonid species'. In: *Atlantic salmon ecology*, pp. 253–276.

Forseth, Torbjørn et al. (2017). 'The major threats to Atlantic salmon in Norway'. In: *ICES Journal of Marine Science* 74.6, pp. 1496–1513.

Gaasø, Maria (2019). 'Sea lice (Lepeophtheirus salmonis and Caligus elongatus) during freshwater treatment (master thesis)'. In: *Norwegian University of Science and Technology*. URL: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2656633.

Hamre, Lars A et al. (2013). 'The salmon louse Lepeophtheirus salmonis (Copepoda: Caligidae) life cycle has only two chalimus stages'. In: *PloS one* 8.9, e73539.

Hemmingsen, Willy et al. (2020). 'Caligus elongatus and other sea lice of the genus Caligus as parasites of farmed salmonids: a review'. In: *Aquaculture* 522, p. 735160.

Jevne, Lone Sunniva and Kjell Inge Reitan (2019). 'How are the salmon lice (Lepeophtheirus salmonis Krøyer, 1837) in Atlantic salmon farming affected by different control efforts: A case study of an intensive production area with coordinated production cycles and changing delousing practices in 2013–2018'. In: *Journal of fish diseases* 42.11, pp. 1573–1586.

Karlsen, Camilla (2021). 'Investigation of a shortened sea phase effect on salmon lice (master thesis)'. In.

Langaas, Mette and Ingeborg Gullikstad Hem (2018). *Linear Mixed Models*.

Mæland, Ingrid Langevei (2022). 'Investigating re-infestation of salmon lice (Project assignment)'. In: *Norwegian University of Technology*.

Nakashima, Eiji (1997). 'Some methods for estimation in a Negative-Binomial model'. In: *Annals of the Institute of Statistical Mathematics* 49, pp. 101–115.

Thorvaldsen, Trine, Kevin Frank and Leif Magne Sunde (2019). 'Practices to obtain lice counts at Norwegian salmon farms: status and possible implications for representativity'. In: *Aquaculture Environment Interactions* 11, pp. 393–404.

Torrissen, Ole et al. (2013). 'Salmon lice–impact on wild salmonids and salmon aquaculture'. In: *Journal of fish diseases* 36.3, pp. 171–194.

Zheng, Han et al. (2018). 'A comparison of different ways of including baseline counts in negative binomial models for data from falls prevention trials'. In: *Biometrical journal* 60.1, pp. 66–78.

Appendix

A Additional Figures

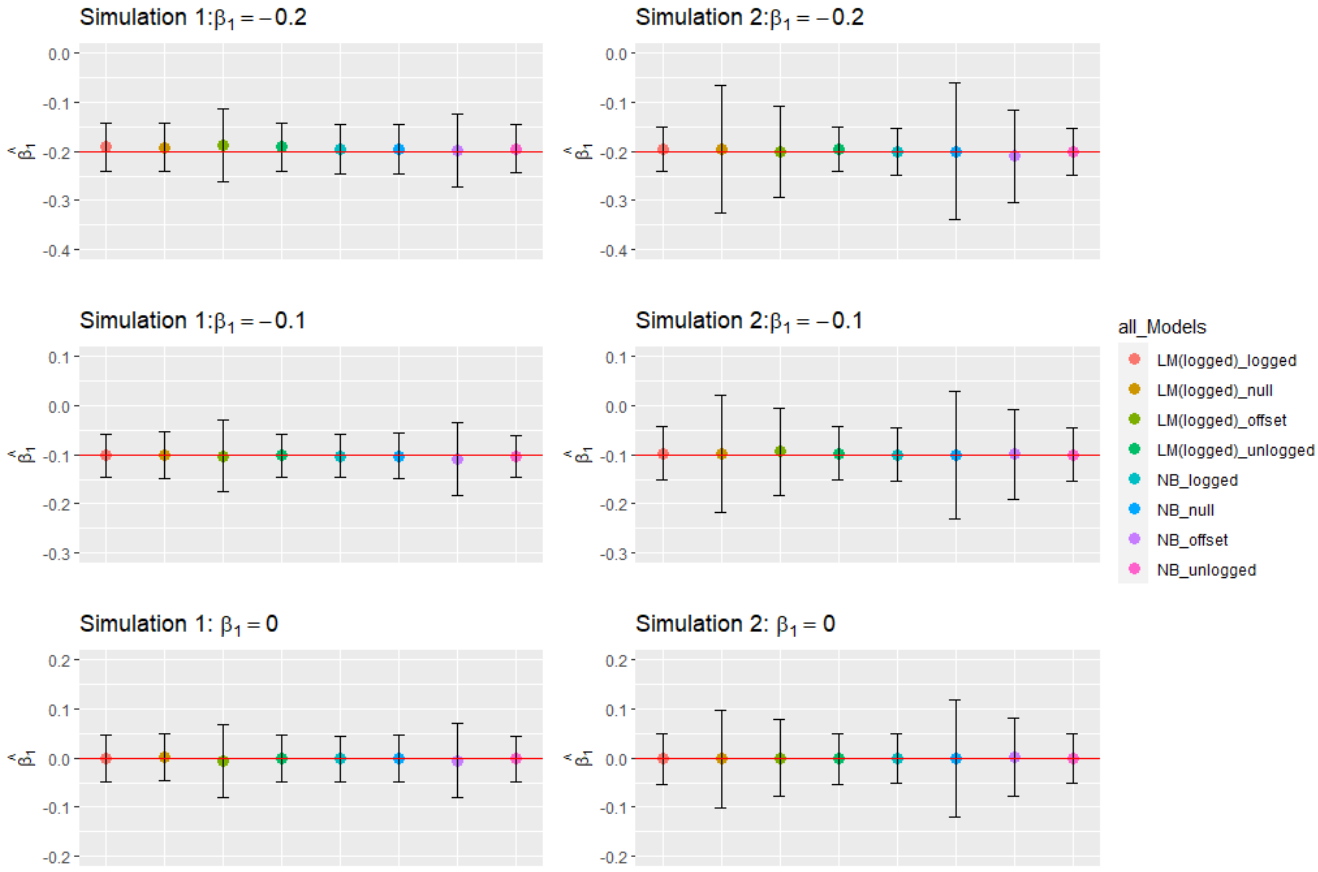


Figure 25: *Estimated* $\hat{\beta}_1$ *from simulation 1 and simulation 2 using* $\alpha = 0.5$ *and* $\beta_1 = -0.2$, $\beta_1 = -0.1$, $\beta_1 = 0.0$.

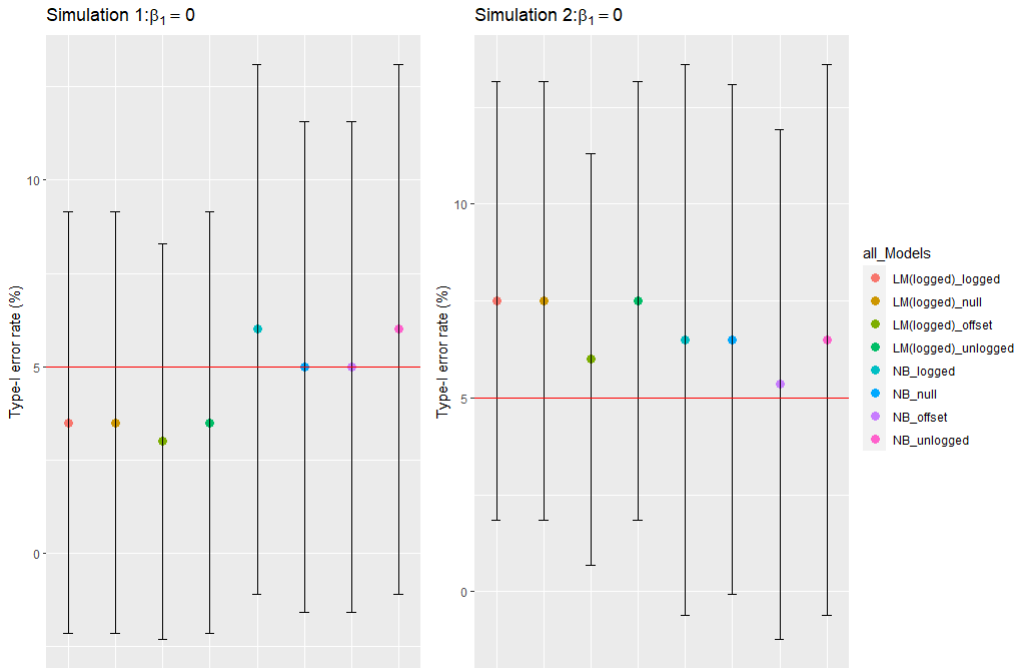


Figure 26: *Type-I error rate from simulation 1 and simulation 2 using* $\alpha = 0.5$ *and* $\beta_1 = -0.2$, $\beta_1 = -0.1$, $\beta_1 = 0.0$.

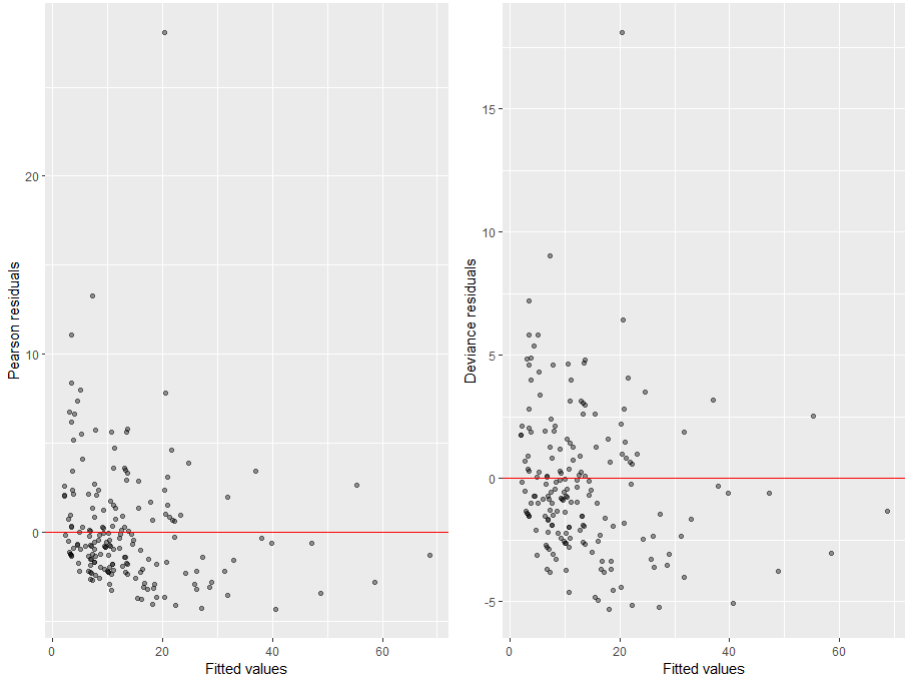


Figure 27: *Plot of residuals against the fitted values from the Poisson regression for the model of mobile lice in the 3-sample. To visualize overlapping, the data points are partially transparent. The data points with the lightest black colour indicate a single point, while a darker black colour indicate overlapping points.*

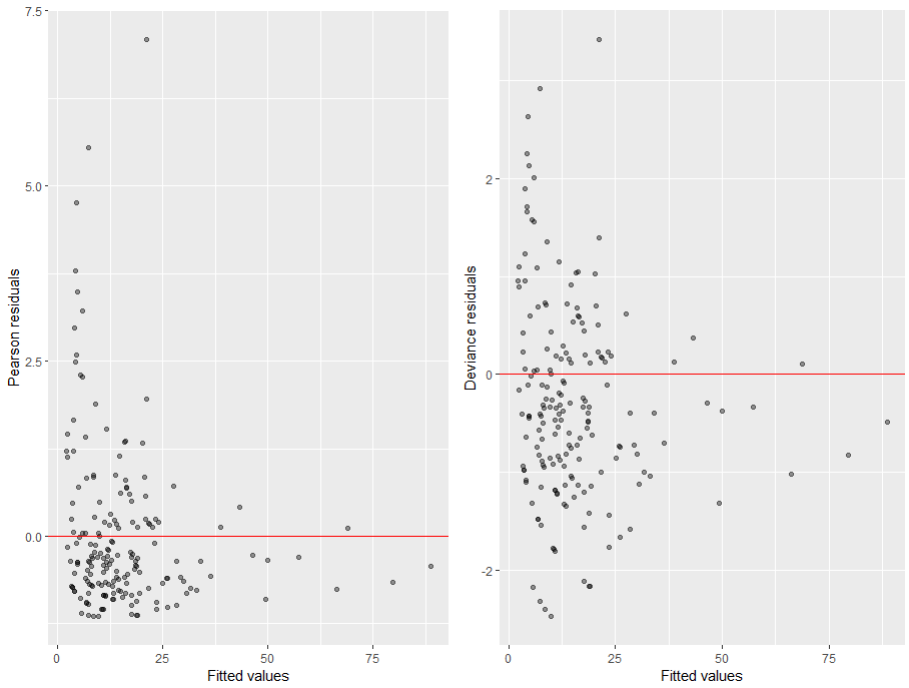


Figure 28: *Plot of residuals against the fitted values from the negative binomial regression for the model of salmon lice in the 3-sample. To visualize overlapping, the data points are partially transparent. The data points with the lightest black colour indicate a single point, while a darker black colour indicate overlapping points.*

B Additional Results

	$\alpha = 3$			$\alpha = 0.5$		
	$\beta_1 = -0.2$	$\beta_1 = -0.1$	$\beta_1 = 0.0$	$\beta_1 = -0.2$	$\beta_1 = -0.1$	$\beta_1 = 0.0$
$LM(logged)_{null}(sim2)$	14 (125)	124 (124)	124 (318)	14 (69)	10 (69)	7 (316)
$LM(logged)_{unlogged}(sim2)$	7 (126)	68 (125)	67 (320)	7 (70)	3(70)	-0.9 (318)
$LM(logged)_{logged}(sim2)$	7 (126)	67 (125)	65 (320)	8 (70)	3 (70)	-0.6 (318)
$LM(logged)_{offset}(sim2)$	88 (206)	97 205)	96 (398)	89 (160)	86 (161)	84 (352)

Table 18: *AIC from* $LM(logged)_{null}$, $LM(logged)_{unlogged}$, $LM(logged)_{logged}$ *and* $LM(logged)_{offset}$ *from simulation 1 and simulation 2 using* $\alpha = 3$ *and* $\alpha = 0.5$ *with* $\beta_1 = -0.2$, $\beta_1 = -0.1$ *and* $\beta_1 = 0.0$.

	Poisson	Neg.bin	Multiple linear	Random Intercept
Degrees of freedom	8	9	9	10
Log-likelihood	-1186	-636	-223	-247
AIC	2389	1290	463	513

Table 19: *A comparison of the fitted regression models for the count of mobile lice in the 3-sample.*

Table 20: *Regression coefficients with associated estimate, standard error, t-value and p-value from the multiple linear regression for the count model for salmon lice in the 2-sample.*

Coefficient	Estimate	Std.Error	t-value	p-value
Intercept	0.72	0.45	1.58	0.12
Optilicer	-0.26	0.12	-2.31	0.022
Freshwater	-0.15	0.27	-0.55	0.59
NumberOfFish	$8.05 \cdot 10^{-7}$	$1.09 \cdot 10^{-6}$	0.74	0.46
AverageWeight	$6.65 \cdot 10^{-5}$	$5.40 \cdot 10^{-5}$	1.23	0.22
SeaTemperature	-0.021	0.032	-0.66	0.51

AIC: 581, Residual standard error: 0.83 on 235 degrees of freedom.
$R^2 : 0.52 and R^2_{adj} : 0.51, F - statistic : 48.04 on 5 and 235 degrees of freedom, p - value :< 2.2 \cdot 10^{-16}$

Table 21: *Regression coefficients with associated estimate, standard error, t-value and p-value from the log-transformed random intercept regression for the count model for salmon lice in the 2-sample.*

Coefficient	Estimate	Std.Error	t-value	p-value
Intercept	1.12	0.54	2.09	0.0014
Optilicer	-5.45	0.13	-0.44	0.69
Freshwater	-0.26	0.36	-0.74	0.46
NumberOfFish	$6.14 \cdot 10^{-7}$	$1.73 \cdot 10^{-6}$	0.36	0.68
AverageWeight	$-2.18 \cdot 10^{-5}$	$6.51 \cdot 10^{-5}$	-0.34	0.77
SeaTemperature	-0.033	0.033	-0.99	0.29

C R-code examples

Packages

```
library("MASS")
library("PropCIs")
library("lmridge")
library("ggplot2")
library("latex2exp")
library("ggpubr")
library("tikzDevice")
library("readxl")
library("stats")
library("GGally")
library("car")
library("dplyr")
library("stargazer")
library("pscl")
library("reshape2")
library("summarytools")
library("AER")
library("lmtest")
library("lme4")
```

Simulation

```
#Simulation 2
```

```
obs = 100 #number of salmon cages
n_fisk = 100000 #number of salmons in each cage
#(n_fisk = 20 for simulation 1)
mu <- 0.5 #mu_0
alpha_vec <- rnorm(obs, mean = 3, sd = 0.001) #mean value varied between 3 and 0.5
sample_size = 20 #number of salmons to sample. Only used in simulation 2
beta_0 <- 1
beta_1 <- -0.2 #varied between -0.2, -0.1, 0.0
sim = 2000 #number of simulations

y_0 = c()
y_1 = c()

beta1_hat = c()
beta2_hat = c()
beta3_hat = c()
beta4_hat = c()

p1_values = c()
p2_values = c()
p3_values = c()
p4_values = c()

aic1 = c()
aic2 = c()
aic3 = c()
aic4 = c()

#Log-transformed models
beta1log_hat = c()
beta2log_hat = c()
beta3log_hat = c()
beta4log_hat = c()

p1log_values = c()
p2log_values = c()
p3log_values = c()
p4log_values = c()

aic1log = c()
aic2log = c()
aic3log = c()
aic4log = c()

x = c(rep(0,obs/2), rep(1,obs/2)) #group
n = rep(20,obs) #for the nb model

alpha_vec <- rnorm(obs, mean = 3, sd = 0.001)
sample_size = 20
beta_0 <- 1
beta_1 <- -0.1
sim = 2000

for (j in 1:sim){
  print(j)

for (i in 1:obs){
  s_fisk <- rgamma(n_fisk, shape = 1/alpha_vec[i], scale = alpha_vec[i])
```

```
  lambda_baseline = mu * s_fisk
  temp = rpois(n = n_fisk , lambda = lambda_baseline)
  rand = sample(1:n_fisk , sample_size)
  y_0[i] = sum(temp[rand])
  lambda_followup = exp(beta_0 + beta_1*x[i])*mu*s_fisk
  temp2 = rpois(n = n_fisk , lambda = lambda_followup)
  rand2 = sample(1:n_fisk , sample_size)
  y_1[i] = sum(temp2[rand2])

}
  model1 = glm.nb(y_1 ~ x + offset(log(n)))
  model2 = glm.nb(y_1 ~ x + log(y_0 + 1) + offset(log(n)))
  model3 = glm.nb(y_1 ~ x + y_0 + offset(log(n)))
  model4 = glm.nb(y_1 ~ x + offset(log(n) + log(y_0 + 1)))

  model1_log = lm(log(y_1 + 1) ~ x + offset(log(n)))
  model2_log = lm(log(y_1 + 1) ~ x + log(y_0 + 1) + offset(log(n)))
  model3_log = lm(log(y_1 + 1) ~ x + y_0 + offset(log(n)))
  model4_log = lm(log(y_1 + 1) ~ x + offset(log(n) + log(y_0 + 1)))

  beta1_hat[j] = summary(model1)$coefficient[2]
  beta2_hat[j] = summary(model2)$coefficient[2]
  beta3_hat[j] = summary(model3)$coefficient[2]
  beta4_hat[j] = summary(model4)$coefficient[2]

  p1_values[j] = anova(model1)$'Pr(>Chi)'[2]
  p2_values[j] = anova(model2)$'Pr(>Chi)'[2]
  p3_values[j] = anova(model3)$'Pr(>Chi)'[2]
  p4_values[j] = anova(model4)$'Pr(>Chi)'[2]

  aic1[j] = AIC(model1)
  aic2[j] = AIC(model2)
  aic3[j] = AIC(model3)
  aic4[j] = AIC(model4)

  beta1log_hat[j] = summary(model1_log)$coefficient[2]
  beta2log_hat[j] = summary(model2_log)$coefficient[2]
  beta3log_hat[j] = summary(model3_log)$coefficient[2]
  beta4log_hat[j] = summary(model4_log)$coefficient[2]

  p1log_values[j] = anova(model1_log)$'Pr(>F)'[1]
  p2log_values[j] = anova(model2_log)$'Pr(>F)'[1]
  p3log_values[j] = anova(model3_log)$'Pr(>F)'[1]
  p4log_values[j] = anova(model4_log)$'Pr(>F)'[1]

  aic1log[j] = AIC(model1_log)
  aic2log[j] = AIC(model2_log)
  aic3log[j] = AIC(model3_log)
  aic4log[j] = AIC(model4_log)
}

(beta1 = mean(beta1_hat))
(var(beta1_hat))
(beta2 = mean(beta2_hat))
(beta3 = mean(beta3_hat))
(beta4 = mean(beta4_hat))

(se1 = sd(beta1_hat))
```

```
(se2 = sd(beta2_hat))
(se3 = sd(beta3_hat))
(se4 = sd(beta4_hat))

(bias1 <- beta_1 - beta1)
(bias2 <- beta_1 - beta2)
(bias3 <- beta_1 - beta3)
(bias4 <- beta_1 - beta4)

(power1 = (length(p1_values[p1_values<0.05]))/sim * 100)
(power2 = (length(p2_values[p2_values<0.05]))/sim * 100)
(power3 = (length(p3_values[p3_values<0.05]))/sim * 100)
(power4 = (length(p4_values[p4_values<0.05]))/sim * 100)

cp1 <- exactci(power1/100*num_sim, num_sim, conf.level = 0.95)
cp2 <- exactci(power2/100*num_sim, num_sim, conf.level = 0.95)
cp3 <- exactci(power3/100*num_sim, num_sim, conf.level = 0.95)
cp4 <- exactci(power4/100*num_sim, num_sim, conf.level = 0.95)

cp1 <- cp1$conf.int[2]*100 - cp1$conf.int[1]*100
cp2 <- cp2$conf.int[2]*100 - cp2$conf.int[1]*100
cp3 <- cp3$conf.int[2]*100 - cp3$conf.int[1]*100
cp4 <- cp4$conf.int[2]*100 - cp4$conf.int[1]*100

(aic_1 = mean(aic1))
(aic_2 = mean(aic2))
(aic_3 = mean(aic3))
(aic_4 = mean(aic4))

(aic_1_log = mean(aic1log))
(aic_2_log = mean(aic2log))
(aic_3_log = mean(aic3log))
(aic_4_log = mean(aic4log))

(beta1_log = mean(beta1log_hat))
(beta2_log = mean(beta2log_hat))
(beta3_log = mean(beta3log_hat))
(beta4_log = mean(beta4log_hat))

(se1_log = sd(beta1log_hat))
(se2_log = sd(beta2log_hat))
(se3_log = sd(beta3log_hat))
(se4_log = sd(beta4log_hat))

(power1_log = (length(p1log_values[p1log_values<0.05]))/num_sim * 100)
(power2_log = (length(p2log_values[p2log_values<0.05]))/num_sim * 100)
(power3_log = (length(p3log_values[p3log_values<0.05]))/num_sim * 100)
(power4_log = (length(p4log_values[p4log_values<0.05]))/num_sim * 100)

cp1_log <- exactci(power1_log/100*num_sim, num_sim, conf.level = 0.95)
cp2_log <- exactci(power2_log/100*num_sim, num_sim, conf.level = 0.95)
cp3_log <- exactci(power3_log/100*num_sim, num_sim, conf.level = 0.95)
cp4_log <- exactci(power4_log/100*num_sim, num_sim, conf.level = 0.95)

cp1_log <- cp1_log$conf.int[2]*100 - cp1_log$conf.int[1]*100
cp2_log <- cp2_log$conf.int[2]*100 - cp2_log$conf.int[1]*100
cp3_log <- cp3_log$conf.int[2]*100 - cp3_log$conf.int[1]*100
cp4_log <- cp4_log$conf.int[2]*100 - cp4_log$conf.int[1]*100
```

```
(bias1_log <- beta_1 - beta1_log)
(bias2_log <- beta_1 - beta2_log)
(bias3_log <- beta_1 - beta3_log)
(bias4_log <- beta_1 - beta4_log)
```

Data Visualizations

```
data <- data.frame(SalmonLice1, SalmonLice2,
LogSalmonLice1,LogSalmonLice2, Method,
NumberOfFish, AverageWeight, Location,
SeaTemperature, Date,
SampleCount1, SampleCount2)

Sample_0 = c(rep("MobileLice",sum(Mobile0)),
rep("SessileLice",sum(Sessile0)),
rep("AdultFemaleLice",sum(AdultFemale0)))
Sample_1 = c(rep("MobileLice",sum(Mobile1)),
rep("SessileLice",sum(Sessile1)),
rep("AdultFemaleLice",sum(AdultFemale1)))
Sample_2 = c(rep("MobileLice",sum(Mobile2)),
rep("SessileLice",sum(Sessile2)),
rep("AdultFemaleLice",sum(AdultFemale2)))
Sample_3 =
c(rep("MobileLice",sum(Mobile3,na.rm=T)),
rep("SessileLice",sum(Sessile3,na.rm=T)),
rep("AdultFemaleLice",sum(AdultFemale3,
na.rm=T)))
ok <- c("Mobile_Lice", "Sessile_Lice", "Adult
Female_Lice")

dat <- data_frame(Sample_0)
dat1 <- data.frame(Sample_1)
dat2 <- data.frame(Sample_2)
dat3 <- data.frame(Sample_3)
```

```
p1 <- ggplot(dat) + geom_histogram(aes(x =
Sample_0), binwidth = 0.8, alpha = 0.7, stat =
"count") + labs(x = "0-Sample", y = "Count") +
ylim(0,850)
p2 <- ggplot(dat1) + geom_histogram(aes(x =
Sample_1), alpha = 0.7, binwidth = 0.8, stat =
"count") + labs(
  x = "1-Sample", y ="_") + ylim(0,850)
p3 <- ggplot(dat2) + geom_histogram(aes(x =
Sample_2), alpha = 0.7, binwidth = 0.8, stat =
"count") + labs(
x = "2-Sample", y = "") + ylim(0,850)
p4 <- ggplot(dat3) + geom_histogram(aes(x =
Sample_3), alpha = 0.7, binwidth = 0.8, stat =
"count") + labs(
  x = "3-Sample", y = "")+ ylim(0,850)
```

```
ggarrange(p1,p2,p3,p4,nrow=1,ncol = 4,
common.legend=T)

p1 <- ggplot(data, aes(x = NumberOfFish, y =
SalmonLice2)) + geom_point(alpha = 0.4)
p2 <- ggplot(data, aes(x = NumberOfFish, y =
LogSalmonLice2)) + geom_point(alpha = 0.4)

ggarrange(p1,p2, common.legend = T)

p3 <- ggplot(data, aes(x = AverageWeight, y =
SalmonLice2)) + geom_point(alpha = 0.4)
p4 <- ggplot(data, aes(x = AverageWeight, y =
LogSalmonLice2)) + geom_point(alpha = 0.4)

ggarrange(p3,p4, common.legend = T)

p5 <- ggplot(data, aes(x = SeaTemperature, y =
SalmonLice2)) + geom_point(alpha = 0.4)
p6 <- ggplot(data, aes(x = SeaTemperature, y =
LogSalmonLice2)) + geom_point(alpha = 0.4)

ggarrange(p5,p6, common.legend = T)


data <- data.frame(LogSalmonLice1,
LogSalmonLice2, NumberOfFish, AverageWeight,
SeaTemperature)
ggpairs(data)
```

Regression models

```
#Models for counts of salmon lice in the 2-sample using the count of
#salmon lice in the 1-sample as a baseline

ds <- read_excel("RegLusdatasett11okt.xlsx", sheet = 2)
behandling_nr = ds$`nr. merdbeh.`
ds = ds[-which(duplicated(behandling_nr)),]

bev1 = ds$bev_1
bev2 = ds$bev_2

fast1 = ds$fast_1
fast2 = ds$fast_2

kjm1 = ds$kjm_1
kjm2 = ds$kjm_2

Method = ds$met
Method[Method=="FLS-avluser"
| Method == "Hydrolicer"
| Method == "Skamik"] = "Mekanisk"
NumberOfFish = ds$ant
AverageWeight = ds$snt
```

```
Location = ds$lok
SeaTemperature = ds$sjotemperatur
Date = ds$dato_avlusing
Date = as.Date(Date, origin = "1899-12-30")
SampleCount1 = ds$`ant.fisk.telt-prove1`
SampleCount2 = ds$`ant.fisk.telt-prove2`

df = data.frame(bev1, bev2, fast1, fast2,
kjm1, kjm2, Method, NumberOfFish, AverageWeight, Location,
SeaTemperature,
Date, SampleCount1, SampleCount2)

df = df[-c(23,24,25,26,27,28,30,35,36,37,
38,39,112,113,114,115,116,117,118,119,120,
123,124,125,126,127,128,138,139,140,141,
230,233,251),] #removing NA's
bev1 = df$bev1
bev2 = df$bev2
fast1 = df$fast1
fast2 = df$fast2
kjm1 = df$kjm1
kjm2 = df$kjm2
Method = df$Method
NumberOfFish = df$NumberOfFish
AverageWeight = df$AverageWeight
Location = df$Location
SeaTemperature = df$SeaTemperature
Date = df$Date
SampleCount1 = df$SampleCount1
SampleCount2 = df$SampleCount2

SalmonLice1 = bev1 + as.numeric(fast1) + kjm1
SalmonLice2 = bev2 + fast2 + kjm2

SalmonLiceCount1 = SalmonLice1 * SampleCount1
SalmonLiceCount2 = SalmonLice2 * SampleCount2

LogSalmonLice1 = log(SalmonLice1 + 1/SampleCount1)
LogSalmonLice2 = log(SalmonLice2 + 1/SampleCount2)

LogSalmonLiceCount2 <- log(SalmonLiceCount2 + 1)

#Defining the regression models

rand_int <- lmer(LogSalmonLiceCount2 ~ as.factor(Method) + NumberOfFish +
AverageWeight + SeaTemperature + (1|Location) + offset(log(SampleCount2)
+ log(SalmonLice1 + 1) - log(SampleCount1)))
pois <- glm(SalmonLiceCount2 ~ as.factor(Method) + NumberOfFish +
AverageWeight + SeaTemperature + offset(log(SampleCount2)
+ log(SalmonLice1 + 1) - log(SampleCount1)))
lm <- lm(LogSalmonLiceCount2 ~ as.factor(Method) + NumberOfFish +
AverageWeight + SeaTemperature + offset(log(SampleCount2)
+ log(SalmonLice1 + 1) - log(SampleCount1)))
neg.bin <- glm.nb(SalmonLiceCount2 ~ as.factor(Method) + NumberOfFish +
AverageWeight + SeaTemperature + offset(log(SampleCount2)
+ log(SalmonLice1 + 1) - log(SampleCount1)))
```

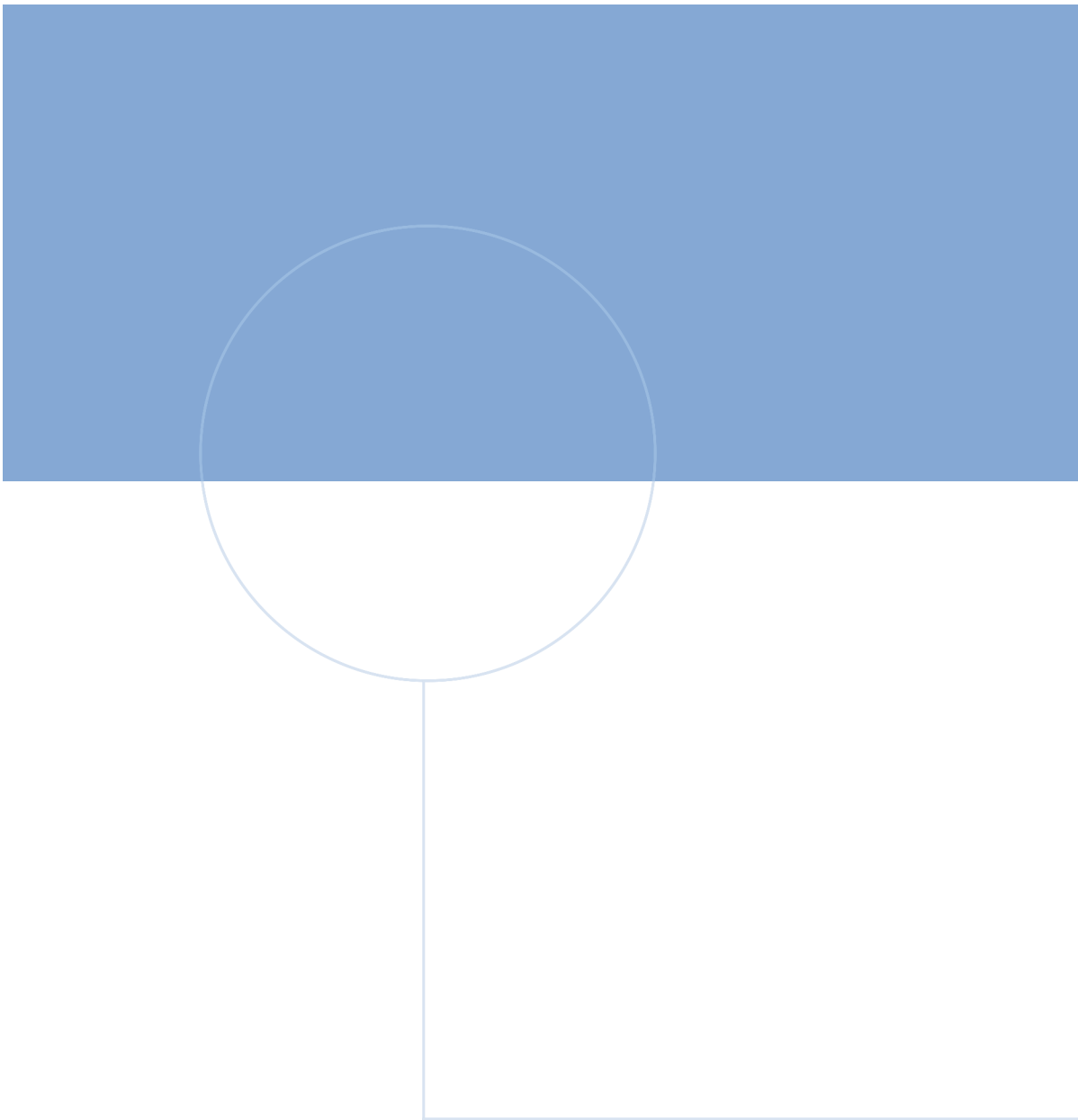
```
##Poisson regression model

dp = sum(residuals(pois,type ="pearson")^2)/pois$df.residual
P_3 = sum(residuals(pois, type = "pearson")^2)

dp #Dispersion parameter
P_3 #Pearson residual

summary(pois)

df <- summary(pois)$df
df
df[2] - df[1]

(chi_pois <- qchisq(0.95,166)) #Chi squared value

fit_pois <- pois$fitted.values
df_pois <- data.frame(fit_pois, SalmonLiceCount2)
cols = c("steelblue", "darkred")

#Plot of fitted values
ggplot(df_pois) + geom_histogram(aes(x = fit_pois, color = " Fitted values "),
alpha = 0.3, binwidth = 0.8) +
  geom_histogram(aes(x = SalmonLiceCount2, color = "Observed values"),
  alpha = 0.3, binwidth = 0.8) +
  labs(x = "SalmonLiceCount2", y = "Count",
  title = "Poisson fit vs. observed values",
  color = "Legend") +
  scale_color_manual(values = cols)

#Plot of residuals
pearson_pois <- residuals(pois, type = "pearson")
deviance_pois <- residuals(pois, type = "deviance")

residual_df <- data.frame(pearson_pois, deviance_pois, fit_pois)

ggplot(residual_df, aes(x = fit_pois, y =
pearson_pois)) + geom_point(alpha = 0.4)+
labs(title = "Residual plot for Poisson", x=
"Fitted values", y = "Pearson residuals") +
geom_hline(yintercept = 0, color = "red")

ggplot(residual_df, aes(x = fit_pois, y =
deviance_pois)) + geom_point(alpha = 0.4) +
  labs(title = "Residual plot for Poisson", x=
  "Fitted values", y = "Deviance residuals") +
  geom_hline(yintercept = 0, color = "red")

## Negative binomial regression model

dp = sum(residuals(neg.bin,type ="pearson")^2)/neg.bin$df.residual
P_3 = sum(residuals(neg.bin, type = "pearson")^2)

dp #Dispersion parameter
P_3 #Pearson residual

dp = sum(residuals(neg.bin,type ="deviance")^2)/neg.bin$df.residual
P_3 = sum(residuals(neg.bin, type = "deviance")^2)
```

P_3

```
df <- summary(neg.bin)$df
df
df[2] - df[1]

(chi_pois <- qchisq(0.95,166)) #Chi squared value

fit_nb <- neg.bin$fitted.values
df_nb <- data.frame(fit_nb, SalmonLiceCount2)

#Plot of fitted values

ggplot(df_nb) + geom_histogram(aes(x = fit_nb +
10, color = "Fitted values"), alpha = 0.3,
binwidth = 0.8) +
geom_histogram(aes(x = SalmonLiceCount2, color
= "Observed values"), alpha = 0.3, binwidth =
0.8) +
  labs(x = "SalmonLiceCount2", y = "Count",
  title = "Neg.bin. fit vs. observed values",
  color = "Legend") +
  scale_color_manual(values = cols)

#Plot of residuals
pearson_nb <- residuals(neg.bin, type = "pearson")
deviance_nb <- residuals(neg.bin, type = "deviance")

resid_df_nb <- data_frame(pearson_nb, deviance_nb, mobile3_count)

ggplot(resid_df_nb, aes(x = fit_nb, y =
pearson_nb)) + geom_point(alpha = 0.4) +
  labs(title = "Residual plot for Negative
  binomial", x= "Fitted values", y = "Pearson
  residuals") +
  geom_hline(yintercept = 0, color = "red")

ggplot(resid_df_nb, aes(x = fit_nb, y =
deviance_nb)) + geom_point(alpha = 0.4) +
  labs(title = "Residual plot for Negative
  binomial", x= "Fitted values", y = "Deviance
  residuals") +
  geom_hline(yintercept = 0, color = "red")

##Multiple linear regression model

fit_lm <- lm$fitted.values
fit_lm <- exp(fit_lm)  #inverse transformation
df_lm <- data.frame(fit_lm, SalmonLiceCount2)

#Plot of fitted values

ggplot(df_lm) + geom_histogram(aes(x = fit_lm,
color = "Fitted values"), alpha = 0.3, binwidth = 0.8) +
geom_histogram(aes(x = SalmonLiceCount2, color
= "Observed values"), alpha = 0.3, binwidth =
```

```
0.8) +
labs(x = "SalmonLiceCount2", y = "Count", title
= "Lin.reg. fit vs. observed values", color =
"Legend") +
  scale_color_manual(values = cols)

#Plot of residuals

residuals_lm <- lm$residuals

df_residuals_lm <- data.frame(residuals_lm, fit_lm)

ggplot(df_residuals_lm, aes(x = fit_lm, y =
residuals_lm)) + geom_point(alpha = 0.4) +
labs(title = "Residuals plot for Lin.Reg.", x =
"Fitted values", y = "Studentized residuals") +
geom_hline(yintercept = 0, color = "red")


#Random Intercept regression model

fit_rlm <- fitted(rand_int)
fit_rlm <- exp(fit_rlm) #inverse transformation

df_rlm <- data.frame(fit_rlm, SalmonLiceCount2)

#Plot of fitted values

ggplot(df_rlm) + geom_histogram(aes(x =
fit_rlm, color = "Fitted values"), alpha = 0.3,
binwidth = 0.8) +
geom_histogram(aes(x = SalmonLiceCount2, color
= "Observed values"), alpha = 0.3, binwidth =
0.8) +
labs(x = "SalmonLiceCount2", y = "Count", title
= "Random intercept fit vs. observed values",
color = "Legend") +
scale_color_manual(values = cols)

#Plot of residuals

residuals_rlm <- residuals(rand_int)

df_residuals_rlm <- data.frame(residuals_rlm, fit_lrm)

ggplot(df_residuals_rlm, aes(x = fit_rlm, y =
residuals_rlm)) + geom_point(alpha = 0.4) +
labs(title = "Residuals plot for Random
Intercept model", x = "Fitted values", y =
"Studentized residuals") +
geom_hline(yintercept = 0, color = "red")


#Model comparison

AIC(rand_int)
AIC(pois)
AIC(lm)
```

```
AIC(neg.bin)

get_df(rand_int)
get_df(pois)
get_df(lm)
get_df(neg.bin)


logLik(rand_int)
logLik(pois)
logLik(lm)
logLik(neg.bin)
```



NTNU

Norwegian University of
Science and Technology