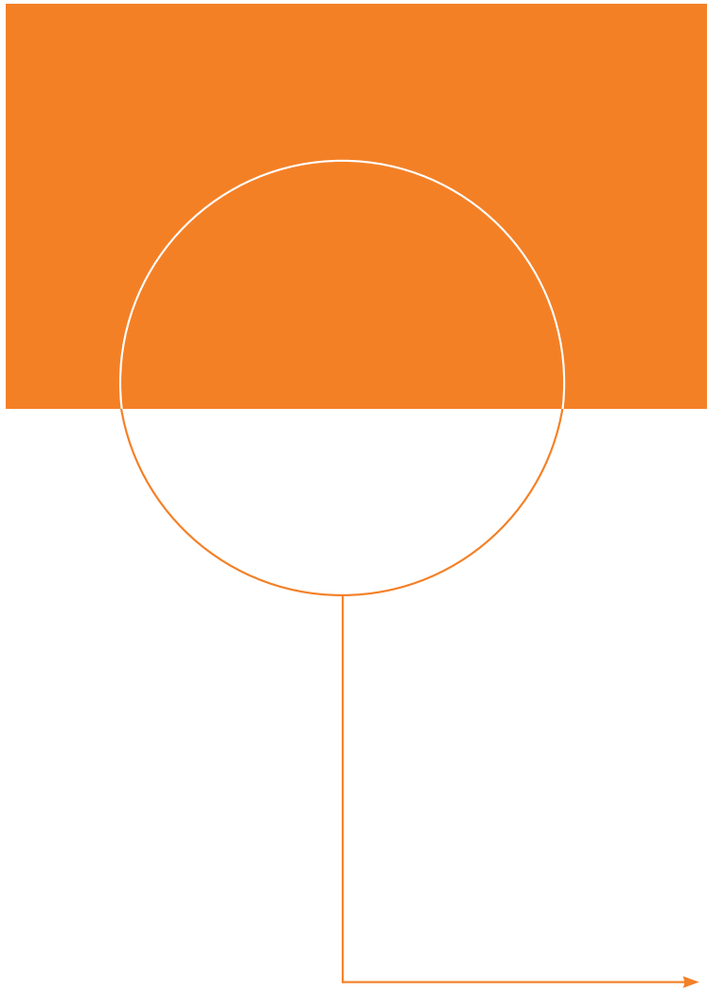

Master's thesis

2019

Elisabeth Børset

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

Elisabeth Børset

Investigating the Phototactic Response of Salmon Lice

Design and Analysis of Experiments

June 2019

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

NTNU

Norwegian University of
Science and Technology

Investigating the Phototactic Response of Salmon Lice

Design and Analysis of Experiments

Lektorutdanning i realfag
Submission date: June 2019
Supervisor: John Sølve Tyssedal

Norwegian University of Science and Technology
Department of Mathematical Sciences

Abstract

Salmon lice is one of the biggest challenges in the salmon industry today, causing great economic losses and threatening the animal welfare. None of the existing methods to reduce the infection rate caused by salmon lice is good enough, as they are not able to overcome the problem while preserving the animal welfare of the salmon.

This thesis is a part of the interdisciplinary project Profylax. Profylax aims to develop a method to reduce the infection rate of salmon lice by using light exposure. As a part of this project, there has been conducted two experiments, a pilot experiment and a main experiment, to investigate the phototactic response of salmon lice. A light source make up the independent factors in these experiments. The independent factors are colour, optical density and pulsation.

The main contribution of this thesis has been to illustrate how such experiments can be designed to provide informative and valid data, and various analysis methods to extract objective and valid results. Typically designs used are split-plot design and incomplete block design, the statistical models corresponding to these designs and evaluation of the regression parameters as well as the goodness of fit for the models.

Sammendrag

Lakselus er en av de største utfordringene i lakseindustrien i dag ved at de forårsaker store økonomiske tap og truer dyrevelferden. Ingen av de eksisterende metodene for å redusere infeksjonsraten forårsaket av lakselus er gode nok, da de ikke er klarer å bekjempe problemet samtidig som de bevarer dyrevelferden av laksen.

Denne masteroppgaven er en del av det tverrfaglige prosjektet Profylax. Profylax har som mål å utvikle en metode for å redusere infeksjonsraten av lakselus ved bruk av lyseksponering. Som en del av dette prosjektet har det blitt gjennomført to eksperiment, et pilotforsøk og et hovedforsøk, for å undersøke fototaktisk respons av lakselus. En lyskilde utgjør de uavhengige faktorene i disse eksperimentene, bestående av farge, lysintensitet og pulsering.

Hovedbidraget fra denne oppgaven har vært å illustrere hvordan slike eksperiment kan designes for å sikre informativ og gyldig data, og ulike analysemetoder for å trekke ut objektive og gyldige resultater. Eksperimentene baserer seg på typiske design som split plot design og ufullstendig blokkdesign, i tillegg til de statistiske modellene som korresponderer til disse designene og evaluering av regresjonsparametrene, samt hvor godt de ulike modellene beskriver dataene.

Preface

This master thesis concludes my masters degree at the Norwegian University of Science and Technology (NTNU). The work was conducted during the spring of 2019 at the Department of Mathematical Science.

In the summer 2018, I was asked to contribute with the statistical planning and analysis of experiments as a part of the interdisciplinary project, Profylax. Through the work on my master thesis I have learned a lot about statistics and I have, if possible, become even more fond of statistics as a discipline. This forms a good basis to become a mathematics teacher, as it is important to have solid knowledge about the subject to teach. In addition the thesis shows that mathematics can be used in research as well as for something that is relevant to society. As a mathematics teacher, this can be used as a motivating factor for the students to learn mathematics.

I would like to thank the people making it possible for me to complete this masters degree. First, I would thank my supervisor, John Sølve Tyssedal. Your expertise in statistics is admirable, but just as important is your ability to pass on knowledge and your desire to help. I have really learned a lot from you throughout this semester. I would also like to thank the people participating in Profylax: Jørgen Andreas Åm Vatn, Maria Arild Solstad, Live Forfang Bjørnstad, Anna Båtnes and Cecilie Miljeteig, for giving me the opportunity to be a part of this project.

To my friends at this study, these five years would never have been the same without you. Specially thanks to Tora Moe and Jenny Kvamme, for all the great laughs and support here at Matteland. To mum and dad, thank you for always believing in me and supporting me. To my sisters, thank you for all the professional and moral support.

Table of Contents

Abstract i

Sammendrag i

Preface ii

Table of Contents iv

List of Tables v

List of Figures viii

1 Introduction 1
- 1.1 Background and Motivation . . . 1
- 1.2 The Profylax Project . . . 2
- 1.3 Outline of the Thesis Work . . . 2

2 Experimental Designs and Statistical Models 3
- 2.1 Analysis of Variance . . . 3
  - 2.1.1 The Effects Model . . . 3
  - 2.1.2 Analysis of the Fixed Effects Model . . . 4
  - 2.1.3 Sum of Squares . . . 4
- 2.2 The Randomized Complete Block Design . . . 6
  - 2.2.1 Statistical Analysis of the RCBD . . . 6
  - 2.2.2 Comparing Block Means . . . 8
- 2.3 Incomplete Block Design . . . 9
  - 2.3.1 Notations and Symbols . . . 9
  - 2.3.2 Intrablock Analysis of Incomplete Block Design . . . 10
- 2.4 Split-Plot Design with two factors . . . 12
  - 2.4.1 Linear Model for the Split-Plot Design . . . 12
- 2.5 Single-Sample Repeated Measures ANOVA . . . 14

2.6 Regression Models . . . 15
2.6.1 Linear Regression Models . . . 15
2.6.2 Estimation of the Regression Coefficients . . . 16
2.6.3 Linear Mixed Models . . . 17
2.7 The Truncated Normal Distribution . . . 18
2.7.1 The Double Truncated Normal Distribution . . . 18
2.7.2 The Right Truncated Normal Distribution . . . 19
2.8 Checking Model Assumptions . . . 22

3 Experiments and Analysis **25**
3.1 Experimental Setup . . . 25
3.1.1 Independent Factors . . . 26
3.2 The Pilot Experiment . . . 27
3.2.1 Design of the Pilot Experiment . . . 27
3.2.2 Response Variable . . . 27
3.2.3 Execution of the Pilot Experiment . . . 28
3.2.4 Analysis of the Pilot Experiment . . . 28
3.2.5 Estimation of the Theoretical Response . . . 44
3.3 The Main Experiment . . . 46
3.3.1 Factors and Levels . . . 46
3.3.2 Response Variable . . . 47
3.3.3 Design of the Main Experiment . . . 47
3.3.4 Execution of the Main Experiment . . . 47
3.3.5 Analysis of the Main Experiment . . . 48

4 Summary and Conclusion **63**
4.1 The Pilot Experiment . . . 63
4.2 The Main Experiment . . . 64
4.3 Recommendations for Further Work . . . 65

Bibliography **67**

Appendix **69**
4.4 Data . . . 69
4.5 Results Pilot Experiment . . . 70
4.5.1 Linear Mixed Effects Model for data.pilot . . . 70
4.5.2 Implemented Algorithm in R - Finding Estimated Response . . . 71

List of Tables

2.1 The Analysis of Variance Table for the Single-Factor, Fixed Effects Model 6
2.2 The Analysis of Variance Table for a Randomized Complete Block Design 8
2.3 Intrablock Analysis of Variance Table for an Incomplete Block Design . . 12
2.4 The Analysis of Variance Table for a Split Plot Design 13

3.1 Independent Factors with Levels 27
3.2 Pilot Experiment: Analysis of Variance Table for a Split-Plot Design . . . 29
3.3 Pilot Experiment: Estimated Fixed Effects of the Linear Mixed Effects Model in (3.2) 30
3.4 Pilot Experiment: Analysis of Variance Table for the Full Model in (3.3) . 34
3.5 Pilot Experiment: Analysis of Variance Table for the Reduced Model in (3.4) 34
3.6 Pilot Experiment: Summary of the Linear Model in (3.6) 35
3.7 Estimated Response at Different Level Combinations 37
3.8 Pilot Experiment: Estimated- and Actual Distances 45
3.9 Main Experiment: Intrablock Analysis of Variance Table 49
3.10 Main Experiment: Summary of the Multiple Linear Model in (3.8) 51
3.11 Main Experiment: Omitted Estimated Parameters of the Two-Way Interactions 54
3.12 Main Experiment: Omitted Estimated Parameters of the Three-Way Interactions 55
3.13 Main Experiment: Estimated Contributions on the Response Variable . . 56

4.1 Summary of the Data from the Pilot Experiment 69
4.2 Summary of the Data from the Main Experiment 69
4.3 Summary of the Data including IR Light from the Main Experiment . . . 69

List of Figures

3.1 A schematic overview of the experimental setup used to detect the swimming behavior of salmon lice when exposed to different sources of light. (A) is an overview from above, (B) is an overview from the side and (C) is a picture of the setup from the side. The experimental setup consists of a camera (1) and an aquarium (2) with a raceway in the middle (shaded area) fitted to the width of the light source (3). The light source (3) was connected to a computer controlled filter wheel. The table legs where attached to two infrared lamps (4) (Miljeteig et al., 2014). 26
3.2 Pilot experiment: Box-plot of the response as a function of colour. 31
3.3 Pilot experiment: Box-plot of the response as a function of OD. 32
3.4 Pilot experiment: Box-plot of the response as a function of replicates. . . 33
3.5 Pilot experiment: Profile plot of the response as a function of OD in replicate II for the green level (green), the blue level (blue), the white level (white) and the red level (red). 38
3.6 Pilot experiment: Box-plot of the response as a function of colour for replicate II. 39
3.7 Pilot experiment: Box-plot of the response as a function of OD for replicate II. 40
3.8 Pilot experiment: Residual plot of the full model in (3.6) 41
3.9 Pilot experiment: Residual plot of the reduced model in (3.4) (red is taken out) 42
3.10 Pilot experiment: Q-Q plot of the full model in (3.3) 43
3.11 Pilot experiment: Q-Q plot of the reduced model in (3.4) (red is taken out) 44
3.12 Pilot experiment: Plot of the estimated distances (B.e, G.e and W.e) and the actual distances (B.a, G.a and W.a) for blue, green and white. 46
3.13 Main experiment: Box-plot of the response as a function of time 57
3.14 Main experiment: Box-plot of the response as a function of colour 58
3.15 Main experiment: Box-plot of the response as a function of OD 59
3.16 Main experiment: Box-plot of the response as a function of pulsation . . . 60
3.17 Main experiment: Residual plot of the model in (3.8) 61

3.18 Main experiment: Q-Q plot of the model in (3.8) 62

Chapter 1

Introduction

1.1 Background and Motivation

The Atlantic salmon industry has major challenges in facing the parasitic copepod *Lepeophtheirus salmonis*, referred to as salmon louse. Each year the salmon industry suffer major economic losses due to the salmon lice, and the salmon lice are also linked to a decrease in wild salmon populations (Glover et al., 2011).

The salmon lice feeds on the salmons skin components such as mucus and blood. As a consequence of this, the immune system of the salmon is decreased, which makes the salmon vulnerable to other infections. It is also shown that the infections of salmon lice leads to reduced growth rate of the salmon. Several methods have been developed and adapted to reduce the number of infections caused by the salmon lice. Among them, chemical treatments have been applied in the salmon industry. The use of chemical treatment have lead to adverse side effects such as stress on the salmon and the resistance to chemical treatments of the salmon lice. Due to this, the need for new methods to control the infections of salmon lice are required (Flamarique et al., 2009).

Researchers have investigated the phototactic response of salmon lice, and it has been shown that salmon lice are attracted towards light stimuli. A study done by Bron et al. (1993) have shown significant differences in the response of the salmon lice to different wavelengths of visible light. The study revealed that salmon lice had the highest response at 550 nm and lowest response at 400 nm. Fields et al. (2017) conducted a study on the response of salmon lice to flickering light, to simulate the reflection of light that arises from the salmons skin. All levels of pulsation showed to have a significant effect on the response variable. The level with shortest ON:OFF cycle (1.8:0.9 s) attracted 24 percent of the salmon lice. The level with cycle (3.5:0.9 s) attracted 37 percent of the salmon lice and the level with cycle (3.5:5.5 s) attracted 80 percent of the salmon lice. The level with cycle (3.5:16.5) did not show a significant increase in the amount of lice that was attracted towards the light source. Fields et al. (2017) also investigated the phototactic response of salmon lice to different levels of light intensity. The levels of light intensity were bright light, medium light and dim light. Their study showed that bright and medium light had a

significant effect on the response, where bright light had the highest effect. Dim light did not have a significant effect on the response.

These findings motivates to further investigate how different light sources affect the salmon lice, and to maybe find the light source which gives the best response from the salmon lice. By finding a light source with a significant effect on the salmon lice, it may be possible to develop an effective method to attract the salmon lice away from the salmon. This is one of the research ideas behind the Profylax project in which my thesis work is a part of.

1.2 The Profylax Project

Profylax is a collaborative project between different disciplines at the Norwegian University of Science and Technology (NTNU) and the Norwegian University of Life Science (NMBU). The project is initiated by Jørgen Andreas Åm Vatn, master student at NMBU. Other participants in the project are Anna Båtnes (researcher at NTNU), Cecilie Miljeteig (researcher at NTNU), Live Forfang Bjørnstad (Engineering and ICT, NTNU) and Maria Arild Solstad (Marine Technology, NTNU).

The main goal of this project is to develop a method to reduce the infection rate of salmon lice in the salmon industry. The aim of the experiments conducted so far is to investigate which light source gives the best response of the salmon lice, and it is of interest to look at the effect of interaction between different colours, optical densities and pulsations.

The experimental setup and execution of the experiments are provided by J.Å. Vatn, A. Båtnes and C. Miljeteig. Analysis of the videos and extraction of the centroids of the detected salmon lice are done by L.F Bjørnstad and M.A. Solstad. I was responsible for the experimental designs and the analysis of the results.

1.3 Outline of the Thesis Work

The aim of the thesis is to investigate the phototactic response of salmon lice and design and analyze the experiments in order to obtain valid and objective results. Two experiments have been executed to investigate the phototactic response of the salmon lice, one pilot experiment and one main experiment. In the planning of these experiments, it has been important to get an overview of the available equipment, resources and limitations which set the guidelines for conducting the experiments. These are factors that affect the choice of experimental design. The first step was to select a response variable which provided useful information about the phototactic response of salmon lice. After obtaining data from the experiments, a statistical analysis according to the experimental design was conducted.

Chapter 2 provides theory on experimental designs and methods to analyze the data obtained from these experiments. Chapter 3 presents the conduction of the experiments, the data obtained and the analysis of these. In Chapter 4, a summary of the statistical results and recommendations for further work are presented.

Chapter 2

Experimental Designs and Statistical Models

This chapter provides theory on experimental designs and methods to analyze the effect of independent factors on a response variable. The concepts of analysis of variance, different regression models and the truncated normal distribution are presented.

2.1 Analysis of Variance

Analysis of variance is a method for comparing the effect of treatments in an experiment (Montgomery, 2009). An important model used in this thesis is the fixed effects model.

2.1.1 The Effects Model

Imagine an experiment where we have a treatments and n observations within each treatment. Then the total number of observations becomes $an = N$.

We use a model to describe the data from the experiment. In this case we use the effects model, given as

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, .., a \\ j = 1, 2, .., n \end{cases} \tag{2.1}$$

where y_{ij} is the *ij*th observation, μ is the overall mean, τ_i is the effect of the *i*th treatment and ϵ_{ij} is a random error that contains variability from other sources in the experiment. In this model the response variable y_{ij} is a linear function of the model components. In the effects model the overall mean is constant and the *i*th treatment effects are deviations from the overall mean. That makes the effects model practical to use when testing the treatments means, which will be discussed later on.

The model errors are assumed to be normal and independently distributed with mean zero and variance σ^2. For all levels of the treatments, we assume the variance σ^2 of the

errors to be the same. Then we have that the observations y_{ij} are normal and independently distributed as

$$y_{ij} \overset{i.i.d}{\sim} N(\mu + \tau_i, \sigma^2)$$

In situations where testing only treatment means from treatments selected by the experimenter, the conclusion will lay only on the treatments in that particular experiment. Then we have a fixed effects model (Montgomery, 2009).

2.1.2 Analysis of the Fixed Effects Model

When analyzing the equality of the treatment means, it is convenient to separate the overall mean μ into $\mu_i = \mu + \tau_i$. As considering the effect of the *i*th treatment τ_i as deviation from the overall mean, we have by definition

$$\sum_{i=1}^{a} \tau_i = 0$$

The suitable hypothesis are

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = ... = \mu_a \\ &H_1 : \mu_i \neq \mu_j \text{ for at least one pair}(i, j) \end{aligned} \tag{2.2}$$

in the effects model. Or equally

$$\begin{aligned} &H_0 : \tau_1 = \tau_2 = ... = \tau_a = 0 \\ &H_1 : \tau_i \neq \tau_j \text{ for at least one pair}(i, j) \end{aligned} \tag{2.3}$$

(2.3) is another way of looking at the test, where we test if the treatment effects τ_i are 0 for all i.

The fundamental idea behind analysis of variance is that we compare the differences between the treatments with the differences within each treatment. If the differences between the treatments and the differences within each treatment are very similar, there is no reason to conclude that there are differences in the treatment means (Montgomery, 2009).

2.1.3 Sum of Squares

Let $y_{i.}$ denote the total of the observations of the *i*th treatment and $y_{..}$ denote the total of all observations in the experiment. We write the average of the observations under the *i*th treatment as $\bar{y}_{i.}$ and the average of all the observations as $\bar{y}_{..}$

$$\begin{aligned} y_{i.} &= \sum_{j=1}^{n} y_{ij} \\ \bar{y}_{i.} &= y_{i.}/n \\ y_{..} &= \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij} \\ \bar{y}_{..} &= y_{..}/N \end{aligned}$$

A measure of the variability over the whole data set is called the total corrected sum of squares, and is denoted as $SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$.

The total variability in the data set can be split into a sum of squares of the differences between the treatment averages and the overall average of all observations, plus a sum of squares of the differences of observations within treatments and the treatment average

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 \tag{2.4}$$

We can also write (2.4) as $SS_T = SS_{Tr} + SS_E$, where

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$$

$$SS_{Tr} = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2$$

SS_{Tr} is the sum of squares of the differences between treatments and SS_E is the sum of squares of the differences within treatments.
A pooled estimation of the variance within each treatment (σ^2) is

$$MS_E = \frac{SS_E}{N - a}$$

and if the treatment effects are zero, we have that

$$MS_{Tr} = \frac{SS_{Tr}}{a - 1}$$

is also an estimate of σ^2. MS_E and MS_{Tr} is called mean squares. The mean squares are as shown calculated by taking the SS divided by its associated degrees of freedom. The expected values of the mean squares are

$$E[MS_E] = \sigma^2$$
$$E[MS_{Tr}] = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a - 1}$$

By comparing MS_{Tr} and MS_E we can test if the treatment means are equal.

The test-statistics

$$F_0 = \frac{MS_{Tr}}{MS_E} \tag{2.5}$$

is F distributed with $a - 1$ and $N - a$ degrees of freedom. If the null-hypothesis is false, we have that MS_{Tr} is bigger than MS_E, and we reject H_0 for values of F_0 bigger than $F_{\alpha,a-1,N-a}$ (Montgomery, 2009).

The analysis of variance table for the Single-Factor, Fixed Effects Model is given in Table 2.1 (Montgomery, 2009, p. 70).

Table 2.1: The Analysis of Variance Table for the Single-Factor, Fixed Effects Model

Source of variation	Sum of Squares	df	Mean Square	F_0
Between treatments	$SS_{Tr} = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2$	a-1	MS_{Tr}	$F_0 = \frac{MS_{Tr}}{MS_E}$
Error within treatments	$SS_E = SS_T - SS_{Tr}$	N-a	MS_E	
Total	$SS_T = \sum_{i=1}^{a}\sum_{i=1}^{n}(y_{ij} - \bar{y}_{..})^2$	N-1		

2.2 The Randomized Complete Block Design

Complete randomization of the runs in an experiment is the best tool we have to avoid that nuisance factors are influencing the analysis of the experiment too much. Nuisance factors are factors that might have a large effect on the response variable, but which we are not interested in and perhaps not able to control. When we know that nuisance factors are present and may influence the response variable, a design technique called blocking can be used to eliminate the effect of the nuisance factors.

In a general randomized complete block design (RCBD), we may assume we have a treatments that we want to compare the effect of and b blocks. Within each block, we apply each of the a treatments at random. Since we only have randomization within each block, we have a restriction on randomization. In this case we would have variability between each block and variability within each block (Montgomery, 2009).

2.2.1 Statistical Analysis of the RCBD

The analysis of variance from Section 2.1 can easily be applied to RCBD. In this case the effects model would look like:

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \begin{cases} i = 1, 2, .., a \\ j = 1, 2, .., b \end{cases} \tag{2.6}$$

We have that y_{ij} is the (ij)th observation, μ is the overall mean, τ_i is the effect of the *i*th treatment, β_j is the effect of the *j*th block and ϵ_{ij} is the random error term. ϵ_{ij} is normal distributed with mean zero and variance σ^2. The treatment and block effects are deviations from the overall mean, so we have

$$\sum_{i=1}^{a} \tau_i = 0$$
$$\sum_{j=1}^{b} \beta_j = 0$$

As in Section 2.1 we are testing the equality of the treatment means. The hypothesis then becomes

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = ... = \mu_a \\ &H_1 : \mu_i \neq \mu_j \text{ for at least one pair}(i, j) \end{aligned} \tag{2.7}$$

in the effects model. Or equivalent

$$\begin{aligned} &H_0 : \tau_1 = \tau_2 = ... = \tau_a = 0 \\ &H_1 : \tau_i \neq \tau_j \text{ for at least one pair}(i, j) \end{aligned} \tag{2.8}$$

(2.8) comes from the fact that $\mu_i = \frac{\sum_{j=1}^{b}(\mu+\tau_i+\beta_j)}{b} = \mu + \tau_i$
Modifying the single-factor analysis of variance to RCBD we have that $y_{i.}$ is the total of the observations of the *i*th treatment, y_j. is the total of observations under block j, $y_{..}$ is the total of all observations in the experiment and $N = ab$ is the total number of observations. We write the average of the observations under the *i*th treatment as $\bar{y}_{i.}$, the average of the observations in block j as $\bar{y}_{.j}$, and the average over all the observations as $\bar{y}_{..}$

$$\begin{aligned} y_{i.} &= \sum_{j=1}^{b} y_{ij} \\ \bar{y}_{i.} &= y_{i.}/b \\ y_{.j} &= \sum_{i=1}^{a} y_{ij} \\ \bar{y}_{.j} &= y_{j.}/a \\ y_{..} &= \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij} \\ \bar{y}_{..} &= y_{..}/N \end{aligned} \tag{2.9}$$

The total corrected sum of squares (SS_T) can then be expressed as

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{..})^2 = b\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2 + a\sum_{j=1}^{b}(\bar{y}_{.j}-\bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{.j}-\bar{y}_{i.}+\bar{y}_{..})^2 \tag{2.10}$$

where we have that

$$\begin{aligned} SS_E &= \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{.j} - \bar{y}_{i.} + \bar{y}_{..})^2 \\ SS_{Tr} &= b\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 \\ SS_{Blocks} &= a\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2 \end{aligned}$$

In this case we have that SS_{Tr} is the sum of squares of differences between treatments, SS_{Blocks} is the sum of squares of differences between blocks and SS_E is the sum of squares between cells minus the sum of squares for treatments and blocks. The mean squares are then

$$E[MS_E] = \sigma^2$$
$$E[MS_{Tr}] = \sigma^2 + \frac{b\sum_{i=1}^{a}\tau_i^2}{a-1}$$
$$E[MS_{Blocks}] = \sigma^2 + \frac{a\sum_{j=1}^{b}\beta_j^2}{b-1}$$

An analysis to check the equality of treatment means are then performed by using the test-statistics

$$F_0 = \frac{MS_{Tr}}{MS_E}$$

If there is no differences between the treatment means (i.e. if the null hypothesis is true), the test statistics is $F_{a-1,(a-1)(b-1)}$ distributed. We reject the null hypothesis if $F_0 > F_{a-1,(a-1)(b-1)}$ (Montgomery, 2009).

The analysis of variance table for a Randomized Complete Block Design is given in Table 2.2 (Montgomery, 2009, p. 126)

Table 2.2: The Analysis of Variance Table for a Randomized Complete Block Design

Source of variation	Sum of Squares	df	Mean Square	F_0
Treatments	SS_{Tr}	$a-1$	MS_{Tr}	$F_0 = \frac{MS_{Tr}}{MS_E}$
Blocks	SS_{Blocks}	$b-1$	MS_{Blocks}	
Error	SS_E	$(a-1)(b-1)$	MS_E	
Total	SS_T	N-1		

2.2.2 Comparing Block Means

In a randomized complete blocks design there is only randomization of treatments within blocks. This means that there is a restriction of the randomization, since the blocks are not randomized. Montgomery (2009) states that considering $F_0 = \frac{MS_{Blocks}}{MS_E}$ as an exact F test on the equality of block means, is not a good method. This is due to that the normality assumption is questionable. To get an idea of the block effects, it could be reasonable to look at the ratio $\frac{MS_{Blocks}}{MS_E}$. If this ratio is big, it implies that the blocking factor has an effect.

2.3 Incomplete Block Design

The theory in this section is taken from Toutenburg and Shalabh (2009), rewritten to fit the interest of this thesis.

In some cases, when the number of treatments in an experiment is large, it is not always possible to perform a complete randomized block design because it may increase the time, cost etc. to execute all treatments within one block. In these situations an incomplete block design is suitable. In an incomplete block design only some of the treatments are applied within each block, and the block sizes are then smaller than the number of treatments.

To analyze experiments performed as incomplete block designs, two methods are used - intrablock analysis and interblock analysis. Intrablock analysis is performed by eliminating the block effects, which are assumed fixed, and then estimating the treatment effects and testing the significance of these. However, since the design is incomplete, one may expect that the block totals also provide some information on the treatments. This is taken care of by an interblock analysis, where the block effects are considered as random effects. When both analysis are carried out, two estimates of the treatment effects are available from each of them. It is possible to pool these estimates together to obtain a better estimator of the treatment effects. The interblock analysis demands that the number of blocks are larger than the number of treatments, and is therefore not presented here.

2.3.1 Notations and Symbols

In this thesis, it is of interest to look at an incomplete block design where each treatment occurs one or zero times in each block. In addition we assume that the number of treatments in each block is the same for all blocks. Then let
v be the number of treatments to be compared
b be the number of blocks in the design
k be the number of treatments applied in each block
r be the number of times each treatment occurs in the design
n= vr = bk be the total number of observations in the design
$y_{i.} = \sum_j y_{ij}$ be the block total of the *i*th block
$B = (y_{1.}, y_{2.}.., y_{b.})^T$
$y_{.j} = \sum_i y_{ij}$ be the treatment total due to the *j*th treatment
$V = (y_{.1}, y_{.2}, ., y_{.v})^T$
$y_{..} = \sum_i \sum_j y_{ij}$ be the total of all observations in the design
$n_{ij} = 1$ if treatment j occurs in block i, and $n_{ij} = 0$ otherwise
and the incidence matrix denoted by

$$\text{N} = \begin{bmatrix} n_{11} & n_{12} & \dots & n_{1v} \\ n_{21} & n_{22} & \dots & n_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ n_{b1} & \dots & \dots & n_{bv} \end{bmatrix}$$

2.3.2 Intrablock Analysis of Incomplete Block Design

Assume y_{ij} is the *j*th response in the *i*th block in the model

$$y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij} \begin{cases} i = 1, 2, .., b \\ j = 1, 2, .., v \end{cases} \tag{2.11}$$

where μ is the general mean effect, β_i is the effect of the *i*th block, τ_j is the effect of the *j*th treatment and ϵ_{ij} is the i.i.d. random error with $\epsilon_{ij} \sim N(0, \sigma^2)$.

We find the intrablock estimators $\hat{\mu}$ and $\hat{\tau_j}$ by minimizing the least square function

$$L = \sum_i \sum_j (y_{ij} - \mu - \beta_i - \tau_j)^2 \tag{2.12}$$

The corresponding normal equations are then

$$\begin{aligned} \mu &: y_{..} = n\hat{\mu} + r \sum_j \hat{\tau_j} + k \sum_i \hat{\beta}_i \\ \beta_i &: y_{i.} = k\hat{\mu} + k\hat{\beta}_i + \sum_j n_{ij} \hat{\tau_j} \\ \tau_j &: y_{.j} = r\hat{\mu} + \sum_i n_{ij} \hat{\beta}_i + r\hat{\tau_j} \end{aligned} \tag{2.13}$$

Having $\sum_j \hat{\tau_j} = \sum_i \beta_j = 0$, the estimator of μ is found to be

$$\hat{\mu} = \bar{y}_{..} \tag{2.14}$$

The normal equations in (2.13) can be written in matrix form as

$$\begin{pmatrix} n & 1_b^T K & 1_v^T R \\ K1_b & K & N \\ R1_v & N^T & R \end{pmatrix} \quad \begin{pmatrix} \hat{\mu} \\ \hat{\beta} \\ \hat{\tau} \end{pmatrix} \quad = \begin{pmatrix} y_{..} \\ B \\ V \end{pmatrix} \tag{2.15}$$

where $B = (y_{1.}, y_{2.}, .., y_{b.})^T$ with $B_i = y_{i.}$ as the block total of the *i*th block and $V = (y_{.1}, y_{.2}, .., y_{.v})^T$ with $V_j = y_{.j}$ as the treatment total due to the *j*th treatment.

Next we multiply both sides of (2.15) by

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & I_b & -NR^{-1} \\ 0 & -N^T K^{-1} & I_v \end{pmatrix} \tag{2.16}$$

to remove the block effect from the normal equations.

The reduced normal equations are then

$$n\hat{\mu} + 1_b^T K \hat{\beta} + 1_v R \hat{\tau} = y_{..} \tag{2.17}$$

$$\left(K - NR^{-1}N^T\right) \hat{\beta} = B - NR^{-1}V \tag{2.18}$$

$$\left(R - N^T K^{-1} N\right) \hat{\tau} = V - N^T K^{-1} B \tag{2.19}$$

where K^{-1} =diag$\left(\frac{1}{k}, \frac{1}{k}, .., \frac{1}{k}\right)$ and R^{-1} =diag$\left(\frac{1}{r}, \frac{1}{r}, .., \frac{1}{r}\right)$. (2.19) can be written as

$$Q = C\hat{\tau} \tag{2.20}$$

where $Q = (Q_1, .., Q_v)^T = V - N^T K^{-1} B$ and $C = R - N^T K^{-1} N$. The $(v \times 1)$ vector Q is the vector of adjusted treatment totals, where Q_j is the adjusted total for the *j*th treatment computed as

$$Q_j = y_{.j} - \sum_{i=1}^{b} \frac{n_{ij} y_{i.}}{k}, \;\; j = 1, 2, .., v$$

(2.21)

So Q_j is found by subtracting the sum of the average contributions of b blocks from the *j*th treatment. The intrablock estimate of τ is then

$$\hat{\tau} = C^{-1} Q \tag{2.22}$$

Sum of Squares and Analysis of Variance

The total variability in the data is expressed by the total sum of squares as

$$SS_T = \sum_i \sum_j y_{ij}^2 - \frac{y_{..}^2}{n} \tag{2.23}$$

We can divide the total variability into

$$SS_T = SS_{Tr(adj)} + SS_{blocks(unadj)} + SS_E$$

The sum of squares for treatments is adjusted to dissociate the treatment and the block effects. The unadjusted block sum of squares is

$$SS_{blocks(unadj)} = \sum_{i=1}^{b} \frac{y_{i.}^2}{k} - \frac{y_{..}^2}{n} \tag{2.24}$$

with b-1 degrees of freedom. The adjusted treatment sum of squares is

$$SS_{Tr(adj)} = \sum_{j=1}^{v} Q_j \hat{\tau}_j \tag{2.25}$$

where $\hat{\tau_j}$ is the least square estimator of τ_j. $SS_{Tr(adj)}$ has v-1 degrees of freedom.

The error sum of squares is found by

$$SS_E = SS_T - SS_{Tr(adj)} - SS_{blocks(unadj)}$$

with n-b-v+1 degrees of freedom.

The null-hypothesis for testing equality in the treatment effect is then based on the statistics

$$F_0 = \frac{SS_{Tr(adj)}/(v-1)}{SS_E/(n-b-v+1)} \tag{2.26}$$

Table 2.3 shows the intrablock analysis of variance (Toutenburg and Shalabh, 2009, p. 191)

Table 2.3: Intrablock Analysis of Variance Table for an Incomplete Block Design

Source of variation	Sum of Squares	df	Mean Square	F_0
Between treatments(adj)	$SS_{Tr(adj)}$	$v-1$	$MS_{Tr(adj)}$	$F_0 = \frac{MS_{Tr(adj)}}{MS_E}$
Between blocks(unadj)	$SS_{Blocks(unadj)}$	$b-1$	$MS_{Blocks(unadj)}$	
Intrablock error	SS_E	$n-b-v+1$	MS_E	
Total	SS_T	n-1		

2.4 Split-Plot Design with two factors

Split-plot design is often used when one or more factors in an experiment are hard to change (Montgomery, 2009). As an example, when we are interested in how four different colours and six different optical densities affects the movement of salmon lice, we may conduct the experiment as a split plot design. Due to the experimental setup, it takes a lot of work and time to change between the different colours. It is then convenient to run all six levels of optical density in random order within one colour, before changing to another colour. The four colours are randomized within each replicate.

In a split-plot design with two factors we have one or more replicates which we call blocks. Each replicate is then divided into whole plots. The hard to change factor is called the whole plot factor, and in the example described above this is colour. Each whole plot consists of several parts called subplots. We call the factors that are easy to change the subplot treatment. All levels of the subplot treatment are then applied to each whole plot. In the example, optical density is the subplot treatment.

If there are other nuisance factors present, their effect on the data will confound with the whole plot factors. So the subplot error is usually smaller than the whole plot error in split-plot designs. Due to that, it is best to have the factors we are most interested in testing as the subplots (Montgomery, 2009).

2.4.1 Linear Model for the Split-Plot Design

The linear model for the split-plot design is

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \gamma_k + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + (\tau\beta\gamma)_{ijk} + \epsilon_{ijk} \begin{cases} i = 1, 2, .., r \\ j = 1, 2, .., a \\ k = 1, 2, .., b \end{cases} \tag{2.27}$$

In (2.27) μ is the overall average of all observations, and ϵ_{ijk} is the random error variable.

To simplify, let the whole plot factor be denoted as A and the subplot factor be denoted as B. y_{ijk} is the response of the *i*th block, *j*th factor A and *k*th factor B. $\tau_i, i = 1, 2, .., r$ are the block effects, β_j is the effect of the *j*th level of factor A and $(\tau\beta)_{ij}$ is the whole plot error. These three terms represent the whole plot.

γ_k is the effect of the *k*th level of factor B, $(\tau\gamma)_{ik}$ is the effect of the *i*th block times

the *k*th level of factor B, $(\beta\gamma)_{jk}$ is the interaction effect between the *j*th level of factor A and the *k*th level of factor B and $(\tau\beta\gamma)_{ijk}$ is the subplot error.

In a split-plot design the whole plot factor (A) is tested against the whole plot error and the subplot treatment (B) is tested against the interaction between blocks and subplot treatment (B). The interaction between whole plot factor (A) and subplot treatment (B) is tested against the subplot error.

Assuming that the interaction between blocks and B and the interaction between blocks and AB are very small, it is practical to pool these errors with ϵ_{ijk} to make up the subplot error. Let σ_ϵ^2 be the variance of ϵ_{ijk}. Then the linear model can be written as

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \gamma_k + (\beta\gamma)_{jk} + \epsilon_{ijk} \begin{cases} i = 1, 2, .., r \\ j = 1, 2, .., a \\ k = 1, 2, .., b \end{cases} \tag{2.28}$$

Having a levels of the whole plot factor (A), b levels of subplot factor (B) and r blocks, the expected mean squares are

$$\begin{aligned}
E(MS_{Blocks}) &= \sigma_\epsilon^2 + ab\sigma_\tau^2 \\
E(MS_A) &= \sigma_\epsilon^2 + b\sigma_{\tau\beta}^2 + \frac{rb\sum_{j=1}^{a}\beta_j^2}{a-1} \\
E(MS_B) &= \sigma_\epsilon^2 + \frac{ra\sum_{j=1}^{a}\gamma_k^2}{ab-1} \\
E(MS_W) &= \sigma_\epsilon^2 + b\sigma_{\tau\beta}^2 \ \textit{(whole plot error)} \\
E(MS_E) &= \sigma_\epsilon^2 \ \textit{(subplot error)} \\
E(MS_{AB}) &= \sigma_\epsilon^2 + \frac{r\sum_{j=1}^{a}\sum_{k=1}^{b}(\beta\gamma)_{jk}^2}{(a-1)(b-1)}
\end{aligned}$$

The analysis of variance table for a Split Plot Design with two factors is given in Table 2.4.

Table 2.4: The Analysis of Variance Table for a Split Plot Design

Source of variation	Sum of Squares	df	Mean Square	F_0
Blocks	SS_{Blocks}	$r-1$	MS_{Blocks}	$F_0 = \frac{MS_{Blocks}}{MS_E}$
Whole plot factor (A)	SS_A	$a-1$	MS_A	$F_0 = \frac{MS_A}{MS_W}$
Whole plot error	SS_W	$(r-1)(a-1)$	MS_W	$F_0 = \frac{MS_W}{MS_E}$
Subplot treatment (B)	SS_B	$b-1$	MS_B	$F_0 = \frac{MS_B}{MS_E}$
AB	SS_{AB}	(a-1)(b-1)	MS_{AB}	$F_0 = \frac{MS_{AB}}{MS_E}$
Subplot error	SS_E	$(r-1)(b-1)[1+(a-1)]$	MS_E	
Total	SS_T	$rab-1$		

Now both the subplot treatment (B) and the AB interaction are tested against the subplot error mean square. In order to estimate the individual effects, one may assume the usual restrictions $\sum_{i=1}^{r}\tau_i = 0$, $\sum_{j=1}^{a}\beta_j = 0$, $\sum_{k=1}^{b}\gamma_k = 0$, $\sum_{i=1}^{r}(\tau\beta)_{ij} = 0, j =$

$1, .., a$, $\sum_{j=1}^{a}(\tau\beta)_{ij} = 0, i = 1, .., r$, $\sum_{j=1}^{a}(\beta\gamma)_{jk} = 0, k = 1, .., b$, and $\sum_{k=1}^{b}(\beta\gamma)_{jk} = 0, j = 1, .., a$.

2.5 Single-Sample Repeated Measures ANOVA

Hedeker and Gibbons (2006) describes the single-sample repeated measures ANOVA as a special case of a split plot design where there is only one replication. There is then no blocking effect, but the model is used to describe rates of change over time. With N subjects and n measurements occasions, we have the linear model

$$y_{ij} = \mu + \pi_i + \tau_j + \epsilon_{ij} \begin{cases} i = 1, 2, .., N \\ j = 1, 2, .., n \end{cases} \tag{2.29}$$

where y_{ij} is the observation for subject i at occasion j, μ is the overall mean, π_i is the individual difference component for subject i, τ_j is the effect of time, assumed to be the same for all subjects, and ϵ_{ij} is the error for subject i at occasion j. In addition we assume the random components distributed as $\pi_i \sim N(0, \sigma_\pi^2)$, having σ_π^2 as the between-subject variance, and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, where σ_ϵ^2 is the within-subject variance. Referring to the example described in Section 2.4, the different subjects are analogue to the different levels of colour, and the different measurement occasions are analogue to the different levels of optical density. In this case the linear model is extended to

$$y_{ij} = \mu + \beta_i + \pi_i + \tau_j + \epsilon_{ij} \begin{cases} i = 1, 2, .., N \\ j = 1, 2, .., n \end{cases} \tag{2.30}$$

where β_i is the effect of colour i. β_i and π_i are completely confounded, so we have that $(\beta_i + \pi_i) \sim N(\beta_i, \sigma_\pi^2)$. As β_i and τ_i are deviations from the overall mean, we have that $\sum_i \beta_i = \sum_j \tau_j = 0$.

Let $\bar{y}_{..}$ be the grand mean and $\bar{y}_{i.}$ the colour mean $(i = 1, .., N)$. $\bar{\epsilon}_{..}$ is the grand mean of errors and $\bar{\epsilon}_{i.}$ is the mean of errors of colour i.

$$\begin{aligned} \bar{y}_{..} &= \frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n} y_{ij} = \mu + \bar{\pi}_{.} + \bar{\epsilon}_{..} \\ \bar{y}_{i.} &= \frac{1}{n}\sum_{j=1}^{n} y_{ij} = \mu + \beta_i + \pi_i + \bar{\epsilon}_{i.} \end{aligned} \tag{2.31}$$

The sum of squares for colour is thus

$$\begin{aligned} & SS_c \\ & \quad = \sum_{i=1}^{N}\sum_{j=1}^{n}(\bar{y}_{i.} - \bar{y}_{..})^2 \\ & \quad = \sum_{i=1}^{N}\sum_{j=1}^{n}(\pi_i - \bar{\pi}_{.} + \beta_i + \bar{\epsilon}_{i.} - \bar{\epsilon}_{..})^2 \\ & \quad = \sum_{i=1}^{N}\sum_{j=1}^{n}(\pi_i - \bar{\pi}_{.} + \beta_i)^2 + \sum_{i=1}^{N}\sum_{j=1}^{n}(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})^2 + R \end{aligned} \tag{2.32}$$

where R consists of cross terms with expectation equal zero. The expectation of SS_c is then

$$\begin{aligned} & E[SS_c] \\ & \quad = E\left[\sum_{i=1}^{N}\sum_{j=1}^{n}(\bar{y}_{i.} - \bar{y}_{..})^2\right] \\ & \quad = n\sum_{i=1}^{N}E\left[(\pi_i - \bar{\pi}_{.})^2\right] + n\sum_{i=1}^{N}\beta_i^2 + n\sum_{i=1}^{N}E\left[(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})^2\right] \end{aligned} \tag{2.33}$$

So the expectation of the mean square of colour becomes

$$\begin{aligned} & E[MS_c] \\ & \quad = \frac{n\sum_{i=1}^{N}E\left[(\pi_i - \bar{\pi}_{.})^2\right] + n\sum_{i=1}^{N}\beta_i^2 + n\sum_{i=1}^{N}E\left[(\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})^2\right]}{N-1} \\ & \quad = n\sigma_{\pi}^2 + \frac{n}{N-1}\sum_{i=1}^{N}\beta_i^2 + \sigma_{\epsilon}^2 \end{aligned} \tag{2.34}$$

2.6 Regression Models

2.6.1 Linear Regression Models

In a completely randomized design, linear regression models can be used to calculate the effects of the predictors on the response variable. Suppose in an experiment having the data $(y_i, x_{i1}, .., x_{ik})$, $i = 1, .., n$ with n observations of the response y and the predictors $(x_1, .., x_k)$. The aim is to understand the effect of the predictors on the response variable y. In this case, where we have more than one predictor, we a use multiple regression to model the relationship between the response variable and the predictors.

The response variable y is random and its distribution relay on the predictors. When the response variable y is continuous and shows an approximately normal distribution conditional on the predictors, we use the classical linear regression model given by

$$y_i = \beta_0 + \beta_1 x_{i1} + .. + \beta_k x_{ik} + \epsilon_i,\ i = 1, .., n \tag{2.35}$$

and the conditional mean of y is

$$\eta_i = E[y|x_{i1}, .., x_{ik}] = \beta_0 + \beta_1 x_{i1} + .. + \beta_k x_{ik} = \mathbf{x}_i^T \boldsymbol{\beta} \tag{2.36}$$

i.e. the conditional mean of y is a linear combination of the k predictors. $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ is also called the linear predictor of the random variable y_i. Here $\mathbf{x}_i^T = (1, x_{i1}, .., x_{ik})$, where x_{ij} is the value of the *j*th covariate, $j = 1, ..k$ for the *i*th observation. $\boldsymbol{\beta} = (\beta_0, .., \beta_k)^T$ are the unknown parameters and ϵ_i is the random deviation from the expected value of observation y_i, also called random error. We assume the errors to be independent and normally distributed with mean zero and variance σ^2. This means that the errors are independent of the predictors (Fahrmeir et al., 2013).

2.6.2 Estimation of the Regression Coefficients

The goal of estimating the unknown coefficients in (2.36) is to minimize the sum of the squared deviations

$$LS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - x_i^T \boldsymbol{\beta})^2 \tag{2.37}$$

with respect to $\boldsymbol{\beta} \in R^p$.

Finding the estimators that minimizes (2.37) is the same as setting the vector of the first derivatives to zero, solving for $\boldsymbol{\beta}$ and show that the matrix of the second derivatives is positive definite.

$$\frac{\partial LS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \tag{2.38}$$

Then the equation

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} \tag{2.39}$$

have a unique solution given by the least square estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.40}$$

assuming the matrix $\mathbf{X}^T\mathbf{X}$ is invertible. The second derivatives of (2.37) are

$$\frac{\partial^2 LS(\boldsymbol{\beta})}{\partial^2 \boldsymbol{\beta}} = 2\mathbf{X}^T\mathbf{X} \tag{2.41}$$

Assuming normally distributed errors like in (2.35), we can find the maximum likelihood estimators. Assuming equal variances for all observations we have:

$$\boldsymbol{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

The likelihood is then given by

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) \tag{2.42}$$

and the log-likelihood given by

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{2.43}$$

We maximize the log-likelihood by finding the vector of first derivatives of (2.43) and setting it to zero. The first two terms are then zero. Maximizing $-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ is the same as minimizing $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$, and we have that the maximum likelihood estimator is the same as the least square estimator in (2.40) (Fahrmeir et al., 2013).

2.6.3 Linear Mixed Models

In experiments where we have repeated measurements on the same subject, the assumption of independent observations does not hold. The observations within each subject may be correlated. Linear mixed models is a method for taking into account the correlations caused by this when estimating the model parameters. By expanding the linear predictor in (2.35) with random effects in addition to the fixed effects $\boldsymbol{\beta}$, we obtain a mixed effects model.

The random intercept model is used when the estimated regression lines for each subject reveals different intercepts, but the slopes are the same across subjects. Imagine having m subjects, n_i observations within each subject and one predictor. Then the random intercept model becomes

$$y_{ij} = \beta_0 + \beta_1 xij + \gamma_{0i} + \epsilon_{ij} \begin{cases} i = 1, .., m \\ j = 1, .., n_i \end{cases} \tag{2.44}$$

In this model the fixed effects are β_0, which is the common fixed intercept for all subjects, and β_1, which is the fixed slope parameter of predictor x and the same across all subjects. The random effects are ϵ_{ij}, which are the independent normally distributed errors with mean zero and variance σ^2, and γ_{0i}, which is the random deviation for each subject from the common fixed intercept.

Each subject is a random sample from a larger data set, so the parameters γ_{0i} are assumed to be independent and random with

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

and the τ_{0i}s and ϵ_{ij}s are assumed to be mutually independent. We then have that

$$y_{ij} \sim N(\beta_0 + \beta_1 x_{ij}, \tau_{0i}^2 + \sigma^2) \tag{2.45}$$

Repeated measurements y_{ij} for subject i are correlated within each subject with covariance

$$\begin{aligned} &Cov(y_{ij}, y_{il}) \\ &= E[(y_{ij} - \mu_{ij})(y_{il} - \mu_{il})] \\ &= E[(\beta_0 + \beta_1 x_{ij} + \gamma_{0i} + \epsilon_{ij} \\ &- (\beta_0 + \beta_1 x_{ij}))(\beta_0 + \beta_1 x_{il} + \gamma_{0i} + \epsilon_{il} - (\beta_0 + \beta_1 x_{il}))] \\ &= E(\gamma_{0i}^2) + E(\epsilon_{ij})E(\gamma_{0i}) + E(\gamma_{0i})E(\epsilon_{il}) + E(\epsilon_{ij})E(\epsilon_{il}) \\ &= \tau_0^2, \ \ (j \neq l) \end{aligned} \tag{2.46}$$

which gives the correlation coefficient

$$Corr(y_{ij}, y_{il}) = \frac{Cov(y_{ij}, y_{il})}{\sqrt{var(y_{ij})}\sqrt{var(y_{il})}} = \frac{\tau_0^2}{\tau_0^2 + \sigma^2},\ (j \neq l)$$

(Fahrmeir et al., 2013).

2.7 The Truncated Normal Distribution

In cases when a random variable X is normally distributed, but there are lower and/or upper bounds for the values that X can take, we have a truncated normal distribution. If only a lower bound exists, the distribution is left truncated, and when there is only a upper bound, the distribution is right truncated. In cases of both bounds, the distribution is double truncated (Ryan, 2011).

A truncated distribution is often used in experiments when the underlying variate x cannot be observed in its whole range, and we would like to predict the behavior of the random variable for the whole range. An example of this is the experiment where we investigate the relation between movement of salmon lice and different sources of light. The salmon lice can only move within the aquarium that is used in the experiment, but without this limitation, some of the salmon lice in the experiment would most likely move beyond the range of the aquarium.

2.7.1 The Double Truncated Normal Distribution

Wiik (2013) presents the cumulative density function of a random variable X which is truncated by $X \in (a, b]$ as

$$P\left(X \leq x | a < X \leq b\right) = F(x|a < X \leq b) = \begin{cases} 0 & \text{for } x \leq a \\ \frac{F(x)-F(a)}{F(b)-F(a)} & \text{for } a < x \leq b \\ 1 & \text{for } x > b \end{cases} \quad (2.47)$$

By differentiating (2.47) we get the corresponding probability function for values $a < X \leq b$

$$f(x|a < X \leq b) = \frac{g(x)}{F(b) - F(a)} \quad (2.48)$$

where g(x)=F'(x) for values within the interval $(a, b]$ and g(x) = 0 otherwise.

Assume $X \sim N(\mu, \sigma^2)$ truncated by $X \in (a, b)$. Then the distribution of X is given by

$$\begin{aligned} & f(x|a < X < b) \\ & = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\left[\frac{1}{\sqrt{2\pi\sigma^2}}\int_a^b exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)dx\right]^{-1} \\ & = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\left[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right]^{-1} \end{aligned} \quad (2.49)$$

where $\Phi(.)$ is the cumulative distribution function of a standardized normally distributed random variable and $\phi(.)$ is the corresponding probability density function.

2.7.2 The Right Truncated Normal Distribution

In this thesis it is of interest to investigate the distribution of a random normally distributed variable X_w with mean μ and variance σ^2 which is truncated by $(X_w \geq b)$. The cumulative density function is then given by

$$F(x_w|X_w \geq b) = \frac{F(x_w) - F(b)}{F(\infty) - F(b)} = \frac{F(x_w) - F(b)}{1 - F(b)} \text{ for } x_w \geq b \tag{2.50}$$

and the corresponding distribution function is obtained by differentiating (2.50)

$$f(x_w|X_w \geq b) = \frac{f(x_w)}{1 - F(b)} = \frac{1}{\sigma}\phi\left(\frac{x_w - \mu}{\sigma}\right)\left[1 - \Phi\left(\frac{b - \mu}{\sigma}\right)\right]^{-1} \text{ for } x_w \geq b \tag{2.51}$$

where $\phi(.)$ and $\Phi(.)$ is defined as in Section 2.7.1.

The mean of this truncated value is then

$$\begin{aligned}
&E[X_w|X_w \geq b] \\
&= \int_b^\infty x_w f(x_w|X_w \geq b) dx_w \\
&= \left[1 - \Phi\left(\frac{b - \mu}{\sigma}\right)\right]^{-1} \int_b^\infty x_w \frac{1}{\sigma}\phi\left(\frac{x_w - \mu}{\sigma}\right) dx_w \\
&= \left[1 - \Phi\left(\frac{b - \mu}{\sigma}\right)\right]^{-1} \left[\int_{\frac{b-\mu}{\sigma}}^\infty \sigma z exp\left(-\frac{z^2}{2}\right) dz + \mu \int_{\frac{b-\mu}{\sigma}}^\infty exp\left(-\frac{z^2}{2}\right) dz\right] \\
&= \frac{\sigma\phi\left(\frac{b-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{b-\mu}{\sigma}\right)} + \mu \\
&= \mu + \sigma\lambda(\alpha)
\end{aligned} \tag{2.52}$$

where $\alpha = \frac{b-\mu}{\sigma}$ and $\lambda(\alpha) = \frac{\phi\left(\frac{b-\mu}{\sigma}\right)}{1-\Phi\left(\frac{b-\mu}{\sigma}\right)}$.

To derive the corresponding variance, we first calculate $E[X_w^2|X_w \geq b]$. Let $k =$

$\left[1-\Phi\left(\frac{b-\mu}{\sigma}\right)\right]^{-1}$. Then

$$\begin{aligned}
&\frac{1}{k}E[X_w^2|X_w \geq b] \\
&= \int_b^\infty x_w^2 \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma} exp\left(-\frac{1}{2}\left(\frac{x_w-\mu}{\sigma}\right)^2\right) dx_w \\
&= \sigma \int_b^\infty \left(\frac{x_w^2}{\sigma^2} - \frac{2\mu x_w}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right) \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x_w-\mu}{\sigma}\right)^2\right) dx_w \\
&+ \sigma \int_b^\infty \frac{2\mu x_w - \mu^2}{\sigma^2} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x_w-\mu}{\sigma}\right)^2\right) dx_w \\
&= \sigma \int_b^\infty \left(\frac{x_w-\mu}{\sigma}\right)^2 \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x_w-\mu}{\sigma}\right)^2\right) dx_w \\
&+ 2\mu E\left[X_w|X_w \geq b\right]\left[1-\Phi\left(\frac{b-\mu}{\sigma}\right)\right] - \mu^2\left[1-\Phi\left(\frac{b-\mu}{\sigma}\right)\right]
\end{aligned}$$

Let $z = \frac{x_w-\mu}{\sigma}$. Then $dx_w = \sigma dz$ and hence

$$\begin{aligned}
&\frac{1}{k}E\left[X_w^2|X_w \geq b\right] \\
&= 2\mu E\left[X_w|X_w \geq b\right]\left[1-\Phi\left(\frac{b-\mu}{\sigma}\right)\right] - \mu^2\left[1-\Phi\left(\frac{b-\mu}{\sigma}\right)\right] \\
&+ \sigma \int_{\frac{b-\mu}{\sigma}}^\infty z^2 \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}z^2\right) \sigma dz
\end{aligned}$$

Next we integrate by parts with $u = z\sigma^2\frac{1}{\sqrt{2\pi}}$ and $dv = zexp\left(-\frac{1}{2}z^2\right)dz$. Then

$$\begin{aligned}
&\frac{1}{k}E\left[X_w^2|X_w \geq b\right] \\
&= 2\mu E\left[X_w|X_w \geq b\right]\left[1-\Phi\left(\frac{b-\mu}{\sigma}\right)\right] - \mu^2\left[1-\Phi\left(\frac{b-\mu}{\sigma}\right)\right] \\
&+ \sigma^2\left(\frac{b-\mu}{\sigma}\right)\phi\left(\frac{b-\mu}{\sigma}\right) + \sigma^2\left[1-\Phi\left(\frac{b-\mu}{\sigma}\right)\right] \\
&\Rightarrow E\left[X_w^2|X_w \geq b\right] \\
&= 2\mu E\left[X_w|X_w \geq b\right] - \mu^2 \\
&+ \sigma^2\left(\frac{b-\mu}{\sigma}\right)\frac{\phi\left(\frac{b-\mu}{\sigma}\right)}{1-\Phi\left(\frac{b-\mu}{\sigma}\right)} + \sigma^2
\end{aligned}$$

Using $var[Y] = E[Y^2] - (E[Y])^2$ gives

$$\begin{aligned} & var\left[X_w | X_w \geq b\right] \\ & = \sigma^2 \left[1 + \frac{\left(\frac{b-\mu}{\sigma}\right)\phi\left(\frac{b-\mu}{\sigma}\right)}{1-\Phi\left(\frac{b-\mu}{\sigma}\right)} - \left(\frac{\phi\left(\frac{b-\mu}{\sigma}\right)}{1-\Phi\left(\frac{b-\mu}{\sigma}\right)}\right)^2\right] \\ & = \sigma^2\left[1 + \alpha\lambda(\alpha) - (\lambda(\alpha))^2\right] \end{aligned} \tag{2.53}$$

Assume now a random normally distributed variable X_{nw} with mean μ and variance σ^2, truncated by $X_{nw} < b$. The distribution function of this random variable is then

$$f(x_{nw}|X_{nw} < b) = \frac{f(x_{nw})}{F(b)} = \frac{1}{\sigma}\phi\left(\frac{x_{nw}-\mu}{\sigma}\right)\left[\Phi\left(\frac{b-\mu}{\sigma}\right)\right]^{-1} \quad \text{for } x_{nw} < b \tag{2.54}$$

By similar calculations as in (2.52), the corresponding mean of this variable is found to be

$$E[X_{nw}|X_{nw} < b] = \mu - \sigma\frac{\phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)} \tag{2.55}$$

and the corresponding variance, found by similar calculations as in (2.53), is

$$var[X_{nw}|X_{nw} < b] = \sigma^2\left[1 - \left(\frac{b-\mu}{\sigma}\right)\frac{\phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)} - \left(\frac{\phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)}\right)^2\right] \tag{2.56}$$

In the salmon lice experiment described in Section 2.7, imagine the wall of the aquarium where the light source is positioned to be the limit b. Let us consider an observed value $\bar{x}_{nw}$ for the mean distance the salmon lice which have responded towards the light source, but not into the wall, have moved. Let $\bar{X}_{nw}$ be the random variable in (2.54). The observed value $\bar{x}_{nw}$ is then an estimate of $E[\bar{X}_{nw}|\bar{x}_{nw} < b]$.

In this example, $\Phi\left(\frac{b-\mu}{\sigma}\right) = r$ corresponds to that a portion $(1-r)$ of the total amount of salmon lice in the experiment have moved into the wall. In general, we have $\Phi\left(\frac{b-\mu}{\sigma}\right) = r \leftrightarrow \frac{b-\mu}{\sigma} = `\Phi^{-1}(r) = \Phi_r$. We can now calculate $\phi\left(\frac{b-\mu}{\sigma}\right) = \phi\left(\Phi_r\right)$.

This gives the following equations

$$\hat{\mu} - \frac{\hat{\sigma}\phi\left(\Phi_r\right)}{r} = \bar{x}_{nw} \tag{2.57}$$

$$\frac{b-\hat{\mu}}{\hat{\sigma}} = \Phi_r \tag{2.58}$$

We are now able to find an estimate for μ and σ, and thus calculate the estimated mean and variance of the random normally variable distributed variable $\bar{X}_w$ truncated by $\bar{X}_w \geq b$. This random variable corresponds to the mean distance the salmon lice which have moved into the wall possibly could have moved.

A second alternative is to estimate a value for the mean distance the salmon lice which have moved into the wall, could have moved without the restriction of the wall, with $\bar{x}_w$, by assuming the random variables $\bar{X}_{nw}$ and $\bar{X}_w$ to be independent. We then have

$$Z = (\bar{X}_w - \bar{X}_{nw}) \sim N\left(0, \sigma^2\left(\frac{1}{m_w} + \frac{1}{m_{nw}}\right)\right) \tag{2.59}$$

where m_w is the number of salmon lice which have moved into the wall and m_{nw} is the number of salmon lice which have not. We get the truncated distribution of Z by condition on the values $\bar{x}_{nw}$ and $\bar{x}_w \geq b$ such that

$$\begin{aligned} &P(Z \leq z | X_w \geq b \cap \bar{X}_{nw} = \bar{x}_{nw}) \\ &= P(Z \leq z | \bar{X}_w - \bar{X}_{nw} \geq b - \bar{x}_{nw}) \end{aligned} \tag{2.60}$$

The expected value of Z is then obtained by inserting the mean and variance into (2.52), and we get

$$E[Z|Z \geq b - \bar{x}_{nw}] = \frac{\sqrt{\frac{1}{m_w} + \frac{1}{m_{nw}}}\sigma\phi\left(\frac{b-\bar{x}_{nw}}{\sqrt{\frac{1}{m_w} + \frac{1}{m_{nw}}}\sigma}\right)}{1 - \Phi\left(\frac{b-\bar{x}_{nw}}{\sqrt{\frac{1}{m_w} + \frac{1}{m_{nw}}}\sigma}\right)} \tag{2.61}$$

Let $E[Z|Z \geq b - \bar{x}_{nw}] = \gamma$. We then have $\bar{x}_w = \bar{x}_{nw} + \gamma$. This equation requires that we have an estimate for σ^2. The estimate can be found by calculating the variance of the distance the salmon lice which have not moved into the wall have moved from the data, which then will be an estimate of the variance in (2.56). Further we can use (2.56) to calculate an estimate for σ^2. The estimated mean for the distance of movement for the total amount of salmon lice can now be calculated as

$$E[\bar{x}_t] = \bar{x}_{nw} \times r + \bar{x}_w \times (1 - r) \tag{2.62}$$

2.8 Checking Model Assumptions

Before doing analysis on the data in an experiment, we have to investigate the data. This includes checking the relationship of the response variable and the predictors, the distribution of the errors and other assumptions the model we use rely on (Fahrmeir et al., 2013).

One of the assumptions of the linear model, is that the errors are normally distributed and independent of the predictors. As for the model in (2.1), we have that

$$\epsilon_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$$

The estimated errors are called residuals and are defined as

$$\hat{\epsilon}_{ij} = y_{ij} - \hat{y}_{ij} \tag{2.63}$$

where $\hat{y}_{ij}$ is an estimation of the observation y_{ij}. As the $\hat{\epsilon}_{ij}$ are predictions of the errors, they are not identical to the errors.

When the variance of ϵ_{ij} do not systematically vary across different predictors, that is the variance do not increase or decrease by one or more predictors, we have homoscedastic error variances. In the opposite case, we talk about heteroscedastic variances. A method for detecting heteroscedastic variances, i.e. errors that are dependent, is to plot the residuals against the predicted values of y_{ij}. By plotting the fitted values of the model on the x-axis and the residuals on the y-axis, we are able to see if the residuals are independent of the predictors. If the residuals are spread around a horizontal line without any distinct structure, it is an indication of independence (Fahrmeir et al., 2013).

Q-Q plots (quantile-quantile plots) are graphical tools used for evaluating the distribution of a sample, and can thus be used to check if the errors are normal distributed. A Q-Q plot is constructed by plotting the quantiles of two samples, **X** and **Y**, against each other. If the two samples are identically distributed, the plot will be a straight line with slope 1 and intercept 0. In cases where the sample **Y** is a linear function of **X**, the Q-Q plot will be a straight line, but there may be changes of the slope and intercept (Wilk and Gnanadesikan, 1968). Loy et al. (2016) describes this with an example of plotting a normal distributed sample $(x_1, x_2, .., x_n) \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, against the quantiles of a standard normal distributed sample $(z_1, z_2, .., z_n) \overset{i.i.d}{\sim} N(0, 1)$. The corresponding Q-Q plot will then be a line where the slope is an estimate of σ^{-1}, and the intercept an estimate of $-\frac{\mu}{\sigma}$.

For the linear model, we assume the response to be a linear function of the predictors. A scatter plot is a simple graphical device for checking the relationship between the response variable and the predictors (Ryan, 2011). In cases of categorical predictors it is useful to compare the mean and standard deviation of the response variable against each level of the predictors (Fahrmeir et al., 2013)

Chapter 3

Experiments and Analysis

The purpose of this chapter is to describe the design of the experiments conducted to investigate the phototactic response of salmon lice, present the data and the analysis of these using the models and methods described in Chapter 2.

First, the experimental setup is described and the different sources of light which make up the independent factors in the experiments. Second, the conduction of the pilot experiment and analysis of the results are presented. Then the design of the main experiment is presented and the data obtained from the main experiment are analyzed. Model fitting and analysis of the experiments are done using the statistical programming software R.

3.1 Experimental Setup

The experimental setup is provided by Cecilie Miljeteig, Anna Båtnes and Jørgen Vatn. To track the swimming behavior of the salmon lice, we recorded the salmon lice in a tank exposed to different sources of light. The tank was placed in a dark room with about 8-12 degrees Celcius. In these experiments, we have used the same setup as in Miljeteig et al. (2014), but with a different tank. A schematic overview of the setup is shown in Figure 3.1.

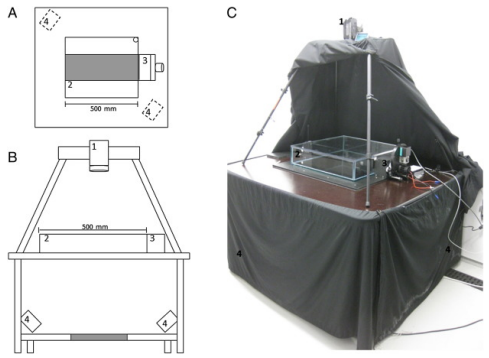

Figure 3.1: A schematic overview of the experimental setup used to detect the swimming behavior of salmon lice when exposed to different sources of light. (A) is an overview from above, (B) is an overview from the side and (C) is a picture of the setup from the side. The experimental setup consists of a camera (1) and an aquarium (2) with a raceway in the middle (shaded area) fitted to the width of the light source (3). The light source (3) was connected to a computer controlled filter wheel. The table legs where attached to two infrared lamps (4) (Miljeteig et al., 2014).

The raceway has measures equal 40 cm $\times$ 12.8 cm $\times$ 10 cm. The tank was placed on the top of a table. To make the objects detectable in the recorded videos, near-infrared (IR) light was directed towards the tank through a hole in the table. The water depth was 4 cm, and the camera was placed 60 cm above the bottom of the tank. The light source was placed on the short side of the raceway. A light source was connected to a computer, so the switch between different optical densities and pulsations could be done without interruption of the environment. Switching between different colours on the other hand, had to be done manually. The experimental setup was covered in black fabric during the experiments to avoid possible external light and to minimize the effect of air currents on the aquarium. A detailed description of the experimental setup can be found in (Vatn, 2019).

3.1.1 Independent Factors

In this project, light is defined as irradiance at 400–700 nm. The different sources of light make up the independent factors in the experiments. To adjust the irradiance, a light emitting diode was attached to a filter wheel. The filter wheel contained neutral optical density (OD) filters with increasing OD. OD is logarithmic, decreasing the irradiance to 10% of the previous level for each increasing OD number. OD from 1 to 7 were used in the experiments, called OD1 up to OD7, respectively. OD1 corresponds to letting 10% of all the light through. So OD2 is then letting 1% of all of the light through and so on.

In addition to OD, the factors colour (wavelength) and pulsation make up the independent factors in the experiments. Colour consists of the four levels white, blue with an emission peak at 455 nm, green with peak at 525 nm and red with peak at 640 nm. The first level of pulsation, pulsation1, corresponds to light being 0.1 seconds on and 0.1 seconds off, pulsation2 corresponds to light being 2 seconds on and 3 seconds off and pulsation3 corresponds to light being 5 seconds on and 5 seconds off. Pulsation0 corresponds to no

pulsation, that is the light is constantly on.

The independent factors with levels are summed up in Table 3.1.

Table 3.1: Independent Factors with Levels

Colour	OD	Pulsation
blue	OD7	pulsation0
green	OD6	pulsation1
white	OD5	pulsation2
red	OD4	pulsation3
	OD3	
	OD2	
	OD1	

3.2 The Pilot Experiment

3.2.1 Design of the Pilot Experiment

The purpose of the pilot experiment was to get an understanding of the effects of colour and OD on the response, and thus possibly reduce the number of levels before the main experiment if any shows to be non-significant. All levels of colour and OD7 to OD2 were used in the pilot experiment.

Due to that the switching between levels of colour had to be done manually, the pilot experiment was conducted as a split-plot design with colour as the whole plot factor and OD as the subplot factor as described in Section 2.4. This leads to a restriction on the randomization, and the influence of possible nuisance factors will confound with colour. To be able to estimate the whole plot error, the experiment was carried out with three replicates, corresponding to blocks in the split-plot design. The whole plot factor, colour, was randomized within each replicate. To reduce the experimental error, each level of OD was applied within one run of colour. Each level of colour was run for 63 minutes. The first 15 minutes was conducted in darkness, before switching to OD7, which was run for 8 minutes. The OD was then increased every 8. minute until OD2.

Since we have multiple measurements of OD within the same colour, with increasing levels over a time period, this is in fact a repeated measurement experiment. If we assume the correlation between each level of OD to be the same within each colour, analyzing the data as a split plot experiment will still apply when using analysis of variance.

3.2.2 Response Variable

Detection and extraction of the centroids of the detected salmon lice are provided by Maria Arild Solstad and Live Forfang Bjørnstad. In Bjørnstad and Solstad (2018) they describe how the detection and extraction was executed. The responses obtained with different types of light was recorded with a camera of 1920 $\times$ 1080 pixels, obtaining 25 frames per second. The number of salmon lice detected was reduced when a phototactic response

towards the light source occurred. We assume that the salmon lice became invisible in the video when moving into light reflections close to the aquarium wall.

The mean position of the salmon lice before the light was turned on is used as a baseline. The response variable used in this analysis is the mean position of the salmon lice (distance from baseline in cm), where positive values corresponds to salmon lice moving towards the light source starting from the baseline. For each of the OD-levels, only the single last minute of recording was used to calculate the mean position. Due to the decreasing number of salmon lice detected as the response of salmon lice towards the light occurred, and the assumption that these "disappeared" into the light, these salmon lice were assumed to have full response. The number of salmon lice which obtained full response influenced the result in a large extent. To be able to compare the differences in the treatment means, the mean position of the salmon lice were calculated by excluding the 5 percent highest and 5 percent lowest positions of the salmon lice. A summary of the data from the pilot experiment is shown in Table 4.1 in the Appendix.

3.2.3 Execution of the Pilot Experiment

Each replicate was carried out in three separate days, so all levels of colour were executed each day. The day before each experiment, the aquarium was washed and filled up with filtrated seawater to prevent bubbles on the glass of the aquarium. Then about 200 salmon lice were transferred into the aquarium and left undisturbed in the dark over night. The next day the salmon lice were stirred carefully to make sure that the salmon lice were evenly spread out. This was done in the dark to make sure of dark acclimation of the salmon lice. The tank was then left undisturbed for 15 minutes to ensure further dark acclimation and for the currents to die out. After 15 minutes the camera was turned on and the experiment started. Between each whole-plot run, a new colour was set and the salmon lice were again stirred. Then another 15 minutes with dark followed before the next run started. After one replicate, the aquarium was emptied, washed and filled up with new filtrated seawater. New salmon lice were counted and put into the tank over night. Then the same procedure was repeated the next day. In replicate I, salmon lice from earlier experiments were used. In replicate II fresh salmon lice were used, and for replicate III salmon lice from replicate II were used again.

3.2.4 Analysis of the Pilot Experiment

Analysis of all three Replicates

Assuming the correlation between each level of OD to be the same within each level of colour, we apply the split-plot analysis in Section 2.4. Using the function aov in the package stats, we test if there are any differences in the treatment means. Colour was tested against the whole plot error, while OD and the interaction between OD and colour were tested against the subplot error. The corresponding model is

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \gamma_k + (\beta\gamma)_{jk} + \epsilon_{ijk} \begin{cases} i = 1, 2, 3 \\ j = 1, .., 4 \\ k = 1, .., 6 \end{cases} \tag{3.1}$$

where y_{ijk} is the response from the *i*th replicate, *j*th colour and *k*th level of OD. μ is the average response of all observations, τ_i is the main effect of the *i*th replicate, β_j is the main effect of the *j*th colour, $(\tau\beta)_{ij}$ is the whole plot error, γ_k is the main effect of the *k*th level of OD and $(\beta\gamma)_{jk}$ is the effect of the interaction between the *j*th colour and the *k*th level of OD. ϵ_{ijk} is the subplot error. Descriptive statistics of the data used for analysis is shown in Table 4.1 in the Appendix.

The analysis of variance table is given in Table 3.2.

Table 3.2: Pilot Experiment: Analysis of Variance Table for a Split-Plot Design

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Replicates	2	434.16	217.08		
Colour	3	146.49	48.83	1.15	0.4024
Whole Plot Error	6	254.63	42.44		
OD	5	82.61	16.52	4.37	0.0029
colour:OD	15	83.56	5.57	1.47	0.1623
Subplot Error	40	151.27	3.78		

The analysis shows that only OD is statistically significant at a 5% significance level with a p-value equal 0.0029, while colour and the interaction between OD and colour are not statistically significant at a 5% significance level.

By continuing the analysis with a linear mixed effects model as in Section 2.6.3 we were able to compare the main effects of each level of OD. Assuming dependence between the observations both within groups of colours and within groups of replicates, we consider colours and replicates as random effects in our model. The random intercept model is then

$$y_{ijk} = \beta_0 + \beta_k + \gamma_{0i} + \gamma_{0j} + \epsilon_{ijk} \begin{cases} i = 1, 2, 3 \\ j = 1, .., 4 \\ k = 1, .., 6 \end{cases} \tag{3.2}$$

where β_0 is the common fixed intercept for both replicates and colours, β_k is the fixed main effect of the *k*th level of OD and is the same across all colours and replicates. γ_{0i} is the random deviation for each replicate from the fixed intercept, and γ_{0j} is the random deviation from each colour from the fixed intercept. ϵ_{ijk} is the random error variable.

The estimated main fixed effects are found by the method of maximum likelihood as described in Section 2.6.2 and the summary of the mixed effects analysis is given in Section 4.5.1 in the Appendix.

An extract from the summary in Section 4.5.1, showing only the fixed effects, is given in Table 3.3.

Table 3.3: Pilot Experiment: Estimated Fixed Effects of the Linear Mixed Effects Model in (3.2)

	Estimate	Std. Error	t value
(Intercept)	0.56	1.82	0.31
OD2	2.65	0.84	3.14
OD3	2.20	0.84	2.61
OD4	2.55	0.84	3.03
OD5	1.64	0.84	1.94
OD6	0.20	0.84	0.23

The model in (3.2) is fitted using dummy coding with the function lmer in the package stats. OD7 is used as the reference level. The intercept, β_0, is thus the estimated mean of the response variable at OD7. The estimated main effect of each of the other levels of OD are the deviation from the mean of the response variable at OD7.

Comparing each level of OD against the estimated mean effect at OD7 is performed by a t-test. The critical value in the t-distribution at 5 degrees of freedom and at a 5% significance level is 2.571. The estimated main effects of OD4 up to OD2 have t-values greater than the critical value, and are thus statistically significant at a 5% significance level. OD2 has the largest estimated main effect equal 2.65. OD4 has the second largest estimated main effect equal 2.55. Due to the design of the experiment being a repeated measurement experiment, time will be a factor influencing the result. At OD2 the salmon lice have been exposed to light for 48 minutes, while at OD7 the salmon lice have only been exposed to light for 8 minutes. In addition, the distance of which the salmon lice can move will decrease as they approach the light source.

By looking at the summary of the fixed effects model in Section 4.5.1 in the Appendix, we see that colour has an estimated variance equal 6.72, which is relatively high. This indicates that the effect of colour do vary between the levels of colour. Visualization of the data with the box-plot in Figure 3.2 reveals that blue, green and white seems to affect the response variable more than red. The median of the response variable for all four levels of colour are close to zero, but the distribution shows that blue, green and white have a higher percent of salmon lice that have moved towards the light source. The box-plots in Figure 3.3 shows that the median of the response variable of all levels of OD are close to zero, but we can still observe an increase in percentage of salmon lice that have moved towards the light source from OD7 up to OD2.

The summary of the fixed effects model in Section 4.5.1 in the Appendix shows that replicate has an estimated variance equal 7.19, which is relatively high. This indicates that the effect of the different replicates varies. The plot in Figure 3.4 shows that replicate I has low impact on the response variable. Replicate III has low impact with some outliers in both direction, which may indicate experimental error. Replicate II differs from the two other replicates with a higher influence on the response variable. As described in Section 3.2.3, the salmon lice used in replicate I and III were used in earlier experiments, while for replicate II fresh salmon lice were used. This might be a factor influencing the results. Based on these results, we continue the analysis by considering only replicate II.

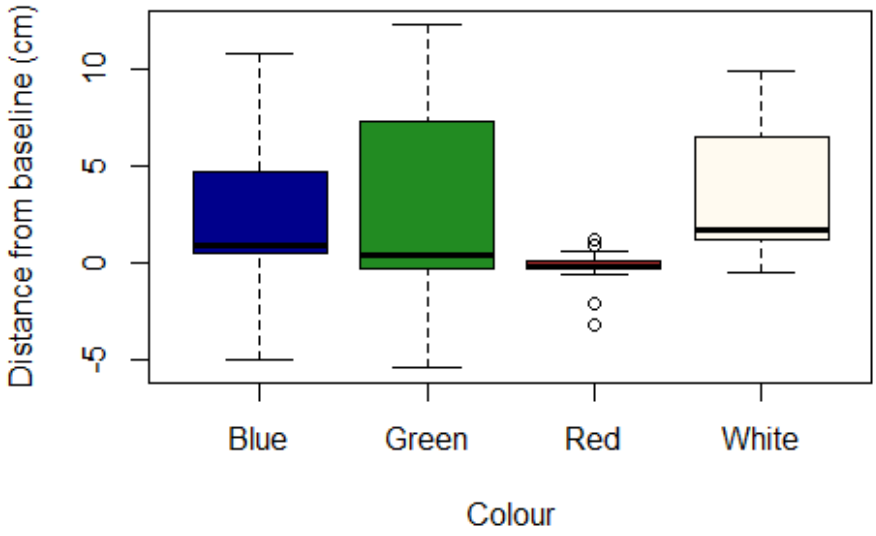

Figure 3.2: Pilot experiment: Box-plot of the response as a function of colour.

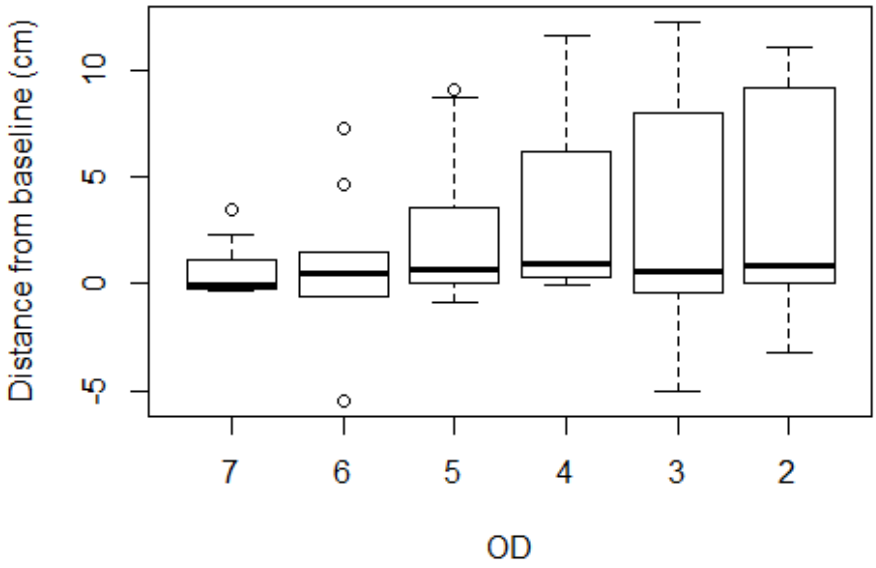

Figure 3.3: Pilot experiment: Box-plot of the response as a function of OD.

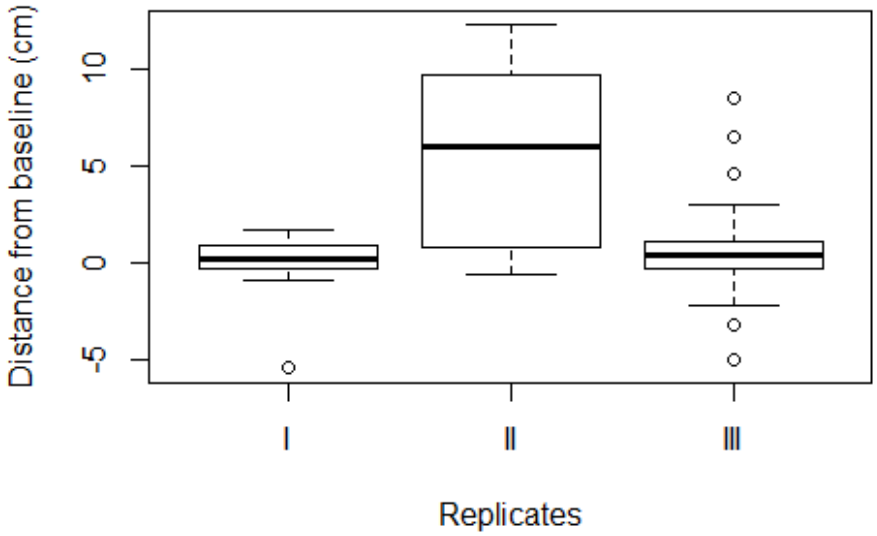

Figure 3.4: Pilot experiment: Box-plot of the response as a function of replicates.

Analysis of Replicate II

By analyzing only replicate II we consider the experiment as a special case of the split-plot design with only one replicate. With colours as subjects and the levels of OD applied within each subject, we have a single-sample repeated measurement as in Section 2.5. The linear model describing the relationship between the response variable and the factors is then given by

$$y_{ij} = \mu + \beta_i + \pi_i + \tau_j + \epsilon_{ij} \begin{cases} i = 1, .., 4 \\ j = 1, .., 6 \end{cases} \tag{3.3}$$

where in this case β_i is the effect of colour i, π_i is the individual difference component for colour i, τ_j is the effect of OD level j and ϵ_{ij} is the error for colour i at OD level j. In addition we assume the random components distributed as $\pi_i \sim N(0, \sigma_\pi^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, where σ_ϵ^2 is the subplot error. The effect of colour is then confounded with the differences between colours, so we have $(\beta_i + \pi_i) \sim N(\beta_i, \sigma_\pi^2)$ with σ_π^2 as the whole plot error.

By using the function aov in the package stats we performed an analysis of variance on the data set, where OD was tested against the subplot error, σ_ϵ^2. The analysis of variance table is given in Table 3.4.

Table 3.4: Pilot Experiment: Analysis of Variance Table for the Full Model in (3.3)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Colour	3	295.96	98.65		
OD	5	140.32	28.06	7.91	0.0008
Subplot error	15	53.22	3.55		

We see that OD has a significant effect on the response variable with p-value = 0.0008. Due to the restriction in randomization of colours, we meet some challenges in the analysis of the treatment effects of colour. Montgomery (2009) claims that by comparing the mean square of colour against the subplot error we can still get an idea of the effects of colour. $\frac{MS_Colour}{MS_E} = \frac{98.65}{3.55} \approx 27.79$ is a high value which indicates that colour do have an effect on the response variable.

Looking at the plot in Figure 3.6 we see that red differs from the three other colours with low effect on the response variable. This leads to a high variability between the colours, which then make the whole plot error big due to that the effect of colours confound with the error between the colours. By performing an analysis of variance on the data without red included, we get a much lower mean square of colour which is reduced from 98.65 to 16.73. By excluding red, we have the reduced model given by

$$y_{ij} = \mu + \beta_i + \pi_i + \tau_j + \epsilon_{ij} \begin{cases} i = 1, .., 3 \\ j = 1, .., 6 \end{cases} \tag{3.4}$$

where there is only three levels of colour. The analysis of variance table on the data set without red included, is shown in Table 3.5.

Table 3.5: Pilot Experiment: Analysis of Variance Table for the Reduced Model in (3.4)

	Df	Sum Sq	Mean Sq
Colour	2	33.45	16.73
OD	5	178.81	35.76
Subplot error	10	14.16	1.42

An upper limit of the whole plot error, σ_π^2, can now be found using (2.34). We get

$$\begin{aligned} 16.73 &= 6\sigma_\pi^2 + \frac{6}{3}\sum_{i=1}^{4}\beta_i^2 + 1.42 \\ \frac{16.73 - 1.42}{6} &= \sigma_\pi^2 + \frac{1}{3}\sum_{i=1}^{4}\beta_i^2 \\ \Leftrightarrow \sigma_\pi^2 &= 2.55 - \frac{1}{3}\sum_{i=1}^{4}\beta_i^2 \leq 2.55 \end{aligned} \tag{3.5}$$

showing that the whole plot error is relatively small. We can now perform an approximate analysis of the factor effects using a multiple linear regression model, assuming independence between all observations.

Using the classical linear regression model in (2.35), we formulate the following linear model

$$y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij} \begin{cases} i = 1, .., 4 \\ j = 1, .., 6 \end{cases} \tag{3.6}$$

where μ is the grand mean effect of all levels, β_i is the main effect of colour i, τ_j is the main effect of OD level j and ϵ_{ij} is the random deviation from the expected value of observation y_{ij}. In addition we have the constraints $\sum_{i=1}^{4} \beta_i = 0$ and $\sum_{j=1}^{6} \tau_i = 0$. The estimated main effects are found by the method of maximum likelihood as described in Section 2.6.2. Summary of the linear model is given in Table 3.6.

Table 3.6: Pilot Experiment: Summary of the Linear Model in (3.6)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.5580	0.3845	14.45	0.0000
OD7	-3.9079	0.8598	-4.55	0.0004
OD6	-2.3541	0.8598	-2.74	0.0153
OD5	-0.1452	0.8598	-0.17	0.8681
OD4	1.5258	0.8598	1.77	0.0963
OD3	2.3993	0.8598	2.79	0.0137
colourR	-5.7283	0.6660	-8.60	0.0000
colourB	2.2447	0.6660	3.37	0.0042
colourG	3.3860	0.6660	5.08	0.0001

The model in (3.6) was fitted by using the function lm in the package stats with effect coding. The estimated main effects of the levels sum to zero as specified in (3.6). From the summary in Table 3.6, we see that the grand mean effect of all levels is equal 5.56. We find the estimated response at each level combination by using the model in (3.6). The estimated response at different level combinations is summed up in Table 3.7.

The summary in Table 3.6 reveals that red has a significant estimated main effect equal -5.73. That is, for a given level of OD, the estimated response decreases by -5.73 for red. The estimated main effect of both blue and green shows to be significant at a 5% significance level. The estimated main effect of green is highest with the value 3.39. The estimated main effect of blue is equal 2.24 and the estimated main effect of white is found to be 0.10 using the constraint $\sum_{i=1}^{4} \beta_i = 0$.

Analyzing the levels of OD, we see that OD3 has a significant estimated main effect, which is equal 2.40. That corresponds to the estimated response increasing by 2.40 for OD3 at a given level of colour. The estimated main effect of OD2 is found using the constraint $\sum_{j=1}^{6} \tau_i = 0$, and is equal 2.48. The estimated main effect of OD2 is thus the highest among the levels of OD. Since the response variable is the mean position of the salmon lice (distance from baseline in cm), the response variable at each level of OD is affected by the position of the salmon lice at the previous level of OD.

The fact that this is a repeated measurement experiment, as in Section 2.5, the effect of colours is accumulated with the effect of OD, where OD can be seen as rates of change over time. By looking at the profile plot in Figure 3.5, we see that OD6 has relatively small effect on the response variable for white compared to green. In contrast, OD4 for white

has high effect on the response variable compared to both green and blue. The Profile plot in Figure 3.5 also show that the responses for green, blue and white are increasing up to OD4, and then tends to flatten out. This matches the results we see in Table 3.6, where the difference in estimated main effects between OD4, OD3 and OD2 are relatively small. Since the range of which the salmon lice can move is limited by the length of the raceway, this is a factor which will influence the result. The salmon lice which have moved towards the light and into the wall, have obtained maximum response. These salmon lice are then not able to respond further, but it is possible that the salmon lice move away from the light, so increasing the OD-levels will not show a higher effect for these salmon lice.

The estimated responses in Table 3.7 shows that green at OD2 has the overall highest estimated response. Comparing these results with the profile plot in Figure 3.5 we see that green at OD3 gives the overall highest response, and that the response decreases from OD3 to OD2 for green. This strengthens our assumption of that the salmon lice might be limited by the wall, and that the response variable may be right truncated normal distributed.

As discussed in Section 3.2.2, we assume that the decreasing number of detected salmon lice is due to that they disappear into the light, and we assume full response for these salmon lice. This assumption reduces the validity of the results, since there may be other reasons that they can not be detected. It is possible that the response in reality is lower than these results show.

Table 3.7: Estimated Response at Different Level Combinations

Colour	OD	Estimated Response
red	OD7	-0.0782
red	OD6	-2.5244
red	OD5	-0.3155
red	OD4	1.3555
red	OD3	2.2290
red	OD2	2.3118
blue	OD7	3.8948
blue	OD6	5.4486
blue	OD5	7.6575
blue	OD4	9.3285
blue	OD3	10.2020
blue	OD2	10.2848
green	OD7	5.0361
green	OD6	6.5899
green	OD5	8.7988
green	OD4	10.4698
green	OD3	11.3433
green	OD2	11.4261
white	OD7	1.7477
white	OD6	3.3015
white	OD5	5.5104
white	OD4	7.1814
white	OD3	8.0549
white	OD2	8.1377

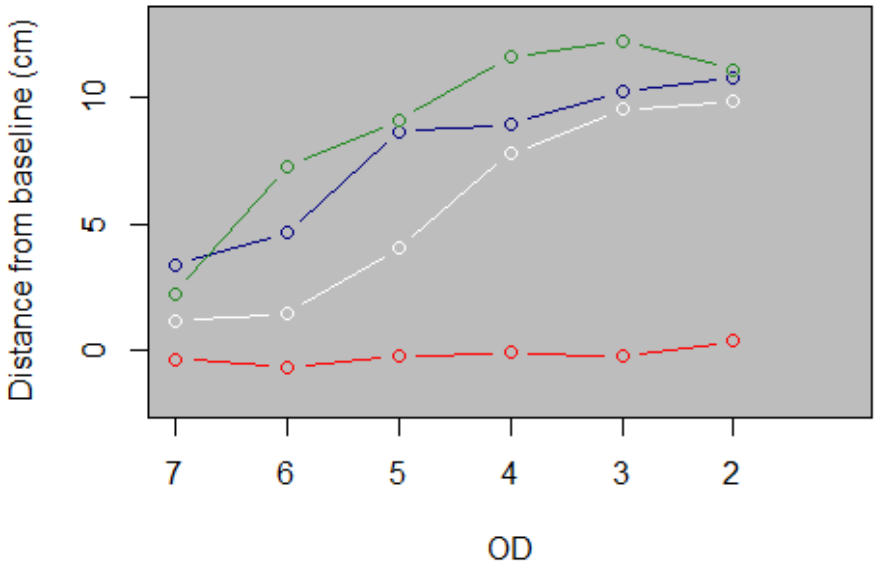

Figure 3.5: Pilot experiment: Profile plot of the response as a function of OD in replicate II for the green level (green), the blue level (blue), the white level (white) and the red level (red).

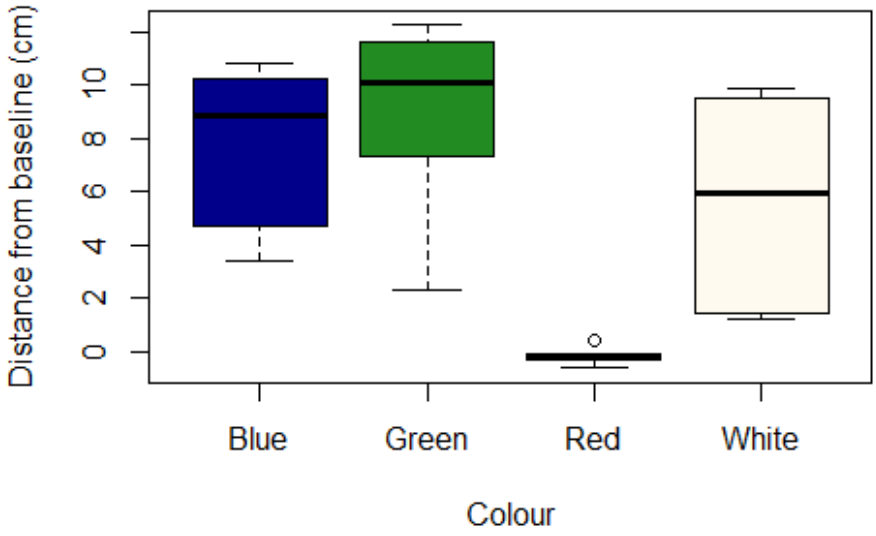

Figure 3.6: Pilot experiment: Box-plot of the response as a function of colour for replicate II.

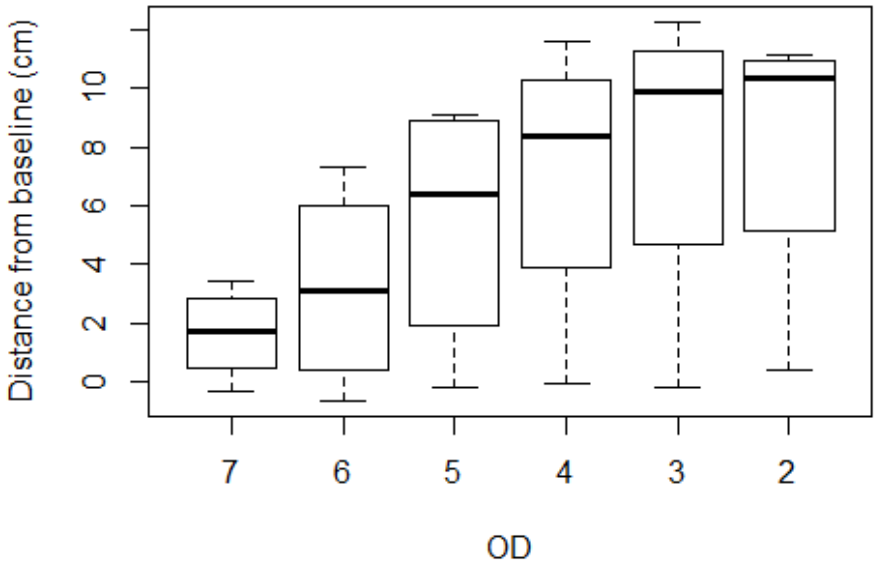

Figure 3.7: Pilot experiment: Box-plot of the response as a function of OD for replicate II.

Checking model assumptions

To investigate if the full model in (3.6) is a good fit to the data, we check if the model assumptions hold. By making a residual plot as described in Section 2.8, we are able to investigate if the residuals behave as independent variables with constant variance. The residual plot in Figure 3.8 indicates some correlations in the residuals. For fitted values between minus five to zero and five to ten, the residuals are positive. For fitted values between zero and five, the residuals are negative. This is an indication of that the model is not a good fit to the data. In multiple linear regression we assume all observations to be independent, so we have a violation of the model assumptions. In addition, the response variable in this experiment is in fact right truncated normal distributed. By using a multiple linear regression model, we have assumed the response variable to be normally distributed. The residual plot in Figure 3.9 of the reduced model in (3.4) where we exclude red, we see that the residuals are more evenly spread around a horizontal line. This indicates that a multiple linear regression model without red included is a better fit to the data than the full model. This may be due to that the effect of red deviates largely from the other levels of colour, which leads to a big variability between the colours.

The Q-Q plot for the full model in (3.6) in Figure 3.10 shows an approximately straight line, which is consistent with the assumption of normally distributed errors in (3.6). The

Q-Q plot in Figure 3.11 of the reduced model where red is excluded shows some deviations from the straight line. This indicates that the errors might not be normal distributed for the reduced model in (3.4). The reason for this may be that more lice move into the light reflections for blue, green and white making the respective distribution truncated.

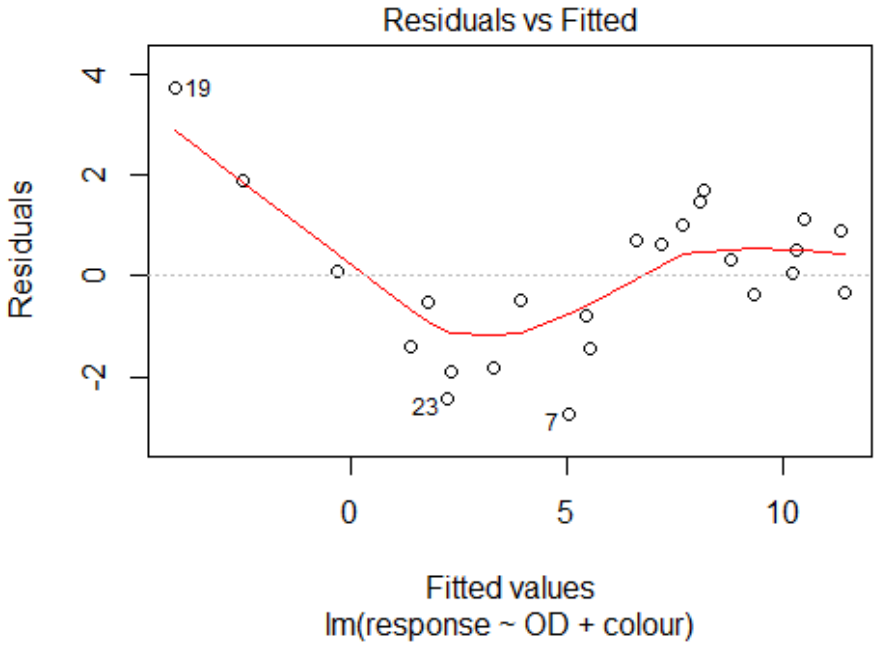

Figure 3.8: Pilot experiment: Residual plot of the full model in (3.6)

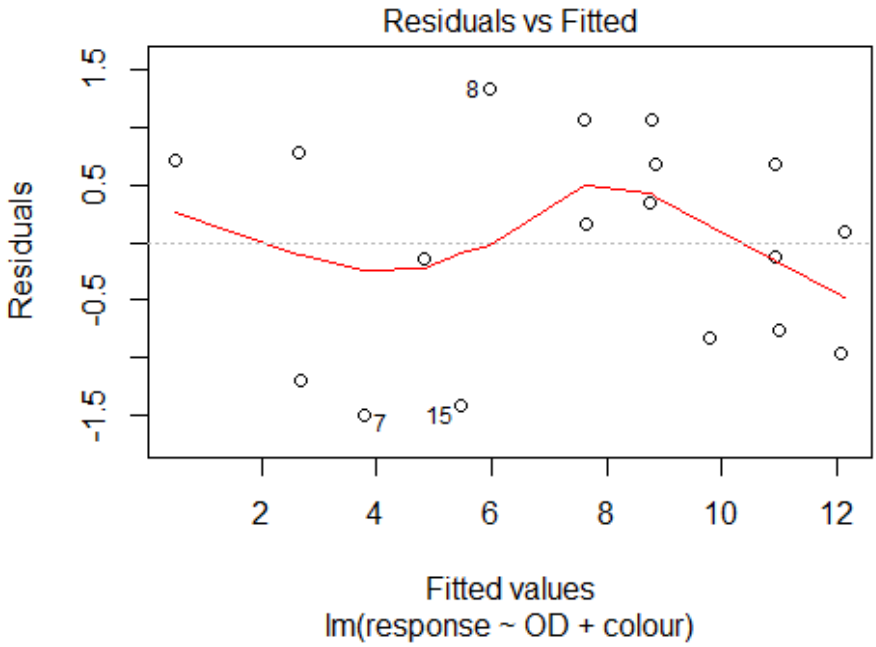

Figure 3.9: Pilot experiment: Residual plot of the reduced model in (3.4) (red is taken out)

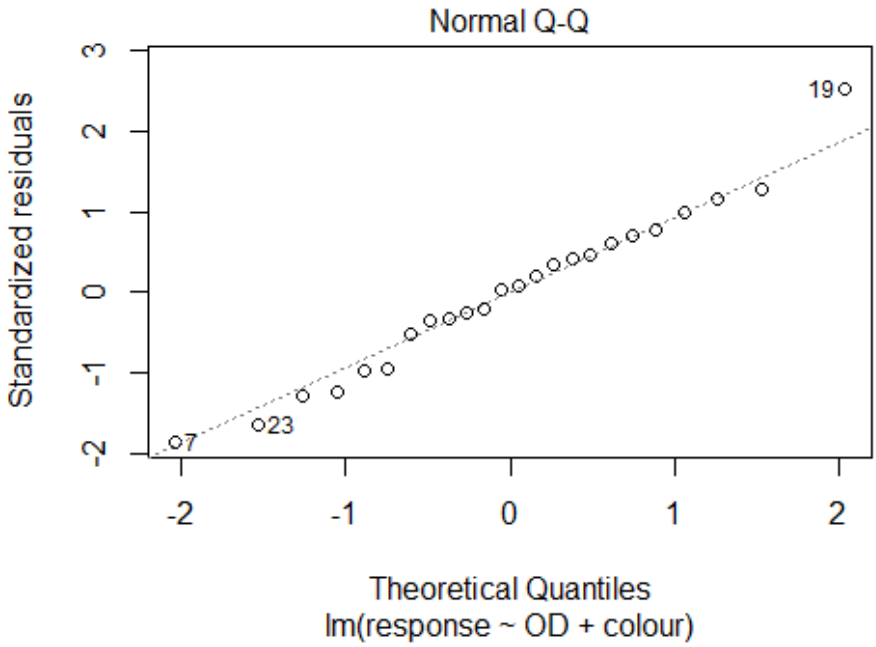

Figure 3.10: Pilot experiment: Q-Q plot of the full model in (3.3)

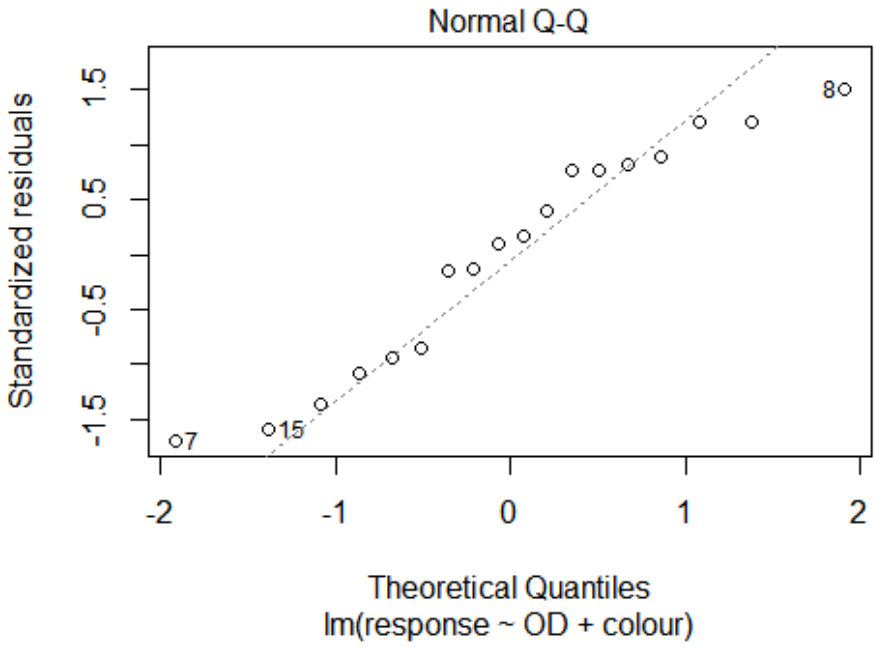

Figure 3.11: Pilot experiment: Q-Q plot of the reduced model in (3.4) (red is taken out)

3.2.5 Estimation of the Theoretical Response

The plot in Figure 3.5 indicates that the salmon lice might be limited by the wall of the aquarium for blue, green and white, and hence we assume that the response variable is right truncated normal distributed.

By using the method described in Section 2.7.2, we are able to estimate the distance the salmon lice could have moved in 48 minutes without this limitation. The estimation is done for blue, green and white at OD2. The limit b in this case, will be the distance between the baseline and the wall where the light source is positioned. Let $\bar{X}_{nw}$ denote the response for these salmon lice which have not moved into the wall after 48 minutes. We have the random variable $\bar{X}_{nw}$ distributed as the random variable in (2.54). In this case, we have found the variance of $\bar{X}_{nw}$ from the data, and then used (2.56) to find an estimate of σ^2. From this we obtained the expected value of Z from (2.61). Finally, the total estimated distance for all salmon lice, $E[\bar{x}_t]$, was found using (2.62). All calculations are done using an implemented algorithm in R. The algorithm is shown in Section 4.5.2 in the Appendix.

Since we in the pilot experiment have assumed that the decrease in detected amount of salmon lice were due to that they disappeared into the light, and thus assumed full response for these, we obtain a large amount of salmon lice which moved into the wall. Comparing

the response of the salmon lice which had not moved into the wall, against the limit b, we see that this distance are quite big. As an example, for blue the limit b is equal 24.70 cm, the distance of salmon lice which have not moved into the wall, $\bar{x}_{nw}$, is equal 9.12 cm and the proportion of salmon lice which have moved into the wall is equal 0.41. Since the difference between b and $\bar{x}_{nw}$ is this big, it is unlikely that the proportion of salmon lice which have moved into the wall is that high. Not only does this assumption weaken the results from the pilot experiment, it also makes it difficult to estimate a realistic distance of the salmon lice without the limitation of the wall. In order to demonstrate the method described in Section 3.12, we subtract 9 cm from the limit b.

Table 3.12 shows the estimated distances the salmon lice could have moved without the limitation of the wall and the actual distances the salmon lice moved within the aquarium. White has the biggest difference between the estimated- and actual distance. This might be due to that a greater amount of salmon lice had moved into the wall for white compared to blue and green.

Table 3.8: Pilot Experiment: Estimated- and Actual Distances

Colour	Estimated Distance	Actual Distance
blue	11.89	10.80
green	13.16	11.11
white	13.12	9.85

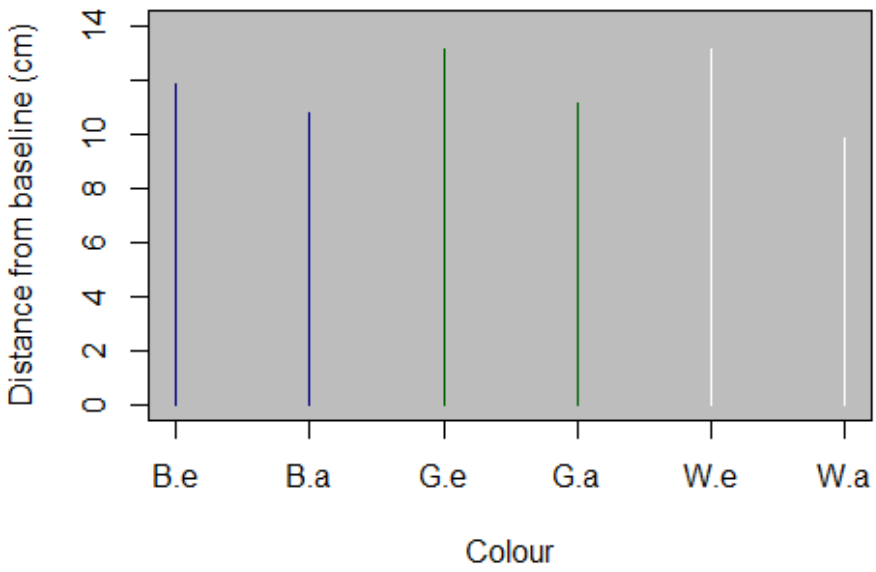

Figure 3.12: Pilot experiment: Plot of the estimated distances (B.e, G.e and W.e) and the actual distances (B.a, G.a and W.a) for blue, green and white.

3.3 The Main Experiment

3.3.1 Factors and Levels

The aim of the main experiment was to investigate the phototactic response of the salmon lice, and compare the effect of the different sources of light on the response variable. Based on the results from the pilot experiment, we were able to reduce the number of levels in our main experiment. The analysis in Section 3.2.4 shows that red has low effect on the response variable, and is thus excluded from the main experiment. From the analysis in Section 3.2.4, we see that OD3 and OD2 have significant estimated main effects. To investigate if there is a significant difference between distinct levels of OD without time being a confounding factor, we decided to include OD5, OD3 and OD1 in the main experiment. OD5 corresponds to the level at which some salmon lice responded in the pilot experiment, but not all. OD3 is included in the main experiment because it showed to have a significant main effect on the response variable in the pilot experiment and OD1 is included to maybe provoke an even stronger response. In addition, all four levels of pulsation were included in the main experiment.

As a supplement to the different sources of light, we decided to look at how time influ-

ence the response variable. Time has levels time1 up to time7, where time1 corresponds to the first minute with light exposure, time2 corresponds to the second minute of light exposure and so on.

3.3.2 Response Variable

Detection and extraction of the centroids of the detected salmon lice were provided by Maria Arild Solstad and Live Forfang Bjørnstad as described in Bjørnstad and Solstad (Pre-print). The salmon lice were recorded with a camera of $815\times$ 2448 pixles, obtaining 16 frames per second. In the main experiment, there was no problem with salmon lice disappearing into the light when responding towards the light. The mean position of the salmon lice (distance from light source in cm) before the light was turned on is used as baseline in the experiment. The response variable used in the main experiment is the mean position of the salmon lice (distance from baseline in cm) for each minute. The response variable was calculated by taking the mean position for each minute, and then subtracting each position from the baseline. Positive values corresponds to salmon lice moving towards the light source. A summary of the data from the main experiment is shown in Table 4.2 in the Appendix.

3.3.3 Design of the Main Experiment

To eliminate the effect of nuisance factors on the response variable, we designed the main experiment with three replicates. Within each replicate, each combination of the levels of factors were run in a randomized order. Each combination of the levels was run for 10 minutes. The first three minutes was conducted with no light at all, and then the seven last minutes were conducted with light.

Three levels of colour, three levels of OD and four levels of pulsation make up a total of 36 runs within each replicate. Due to practical reasons, each replicate was conducted over two days, with eighteen runs each day. This is an incomplete block design, where the days corresponds to blocks in our experiment.

3.3.4 Execution of the Main Experiment

The day before each experiment, the aquarium was washed and filled up with filtrated seawater to prevent bubbles on the glass of the aquarium, and 150-200 salmon lice were counted and transferred to the raceway. The salmon lice were then left undisturbed in the dark over night. The morning after the water and air temperature was measured, the near-infrared light turned on and the light source set to the right combination of treatments. After the experimental setup was ready, the salmon lice were stirred carefully to make sure the salmon lice were evenly spread out and then dark acclimatized for five minutes before the first run started. Between each run, the salmon lice were stirred and the light source set to a new combination of treatments. Then the salmon lice were dark acclimatized for five minutes, before the next run started.

For each day with experiments, different salmon lice were used. The first day we used salmon lice at 8 days past moulting, the second day salmon lice at 9 days past moulting,

the third day salmon lice at 12 days past moulting, the fourth and fifth days salmon lice at 1 day past moulting and the sixth day salmon lice at 2 days past moulting.

As described in Section 3.1, near-infrared (IR) light was used in the experiment to make the salmon lice detectable in the videos. To investigate if the IR light had any effect on the response variable, three runs with IR light were executed in the same way as described above. The light source was replaced with one of the IR lamps used in the experiment. The first run was conducted the first day of experiments, the second run was conducted the second day of experiments and the third run was conducted the fifth day of experiments. Summary of the data including IR light is shown in Table 4.3 in the Appendix.

3.3.5 Analysis of the Main Experiment

Comparing Factor Means

The main experiment is analyzed with intrablock analysis as described in Section 2.3.2. To investigate if there is any difference in the factor means with intrablock analysis of variance, we consider the block effects as fixed. We analyze the effect of time, colour, OD, pulsation and the interactions of these. To analyze the effect of IR light on the response variable, we include IR light as a factor in our model. The corresponding model is

$$\begin{aligned} y_{rijkm} = \mu &+ \beta_r + \tau_i + \gamma_j + \alpha_k + \delta_m + \lambda \\ &+ (\tau\gamma)_{ij} + (\tau\alpha)_{ik} + (\tau\delta)_{im} \\ &+ (\gamma\alpha)_{jk} + (\gamma\delta)_{jm} + (\alpha\delta)_{km} \\ &+ (\tau\gamma\alpha)_{ijk} + (\tau\gamma\delta)_{ijm} + (\tau\alpha\delta)_{ikm} \\ &+ (\gamma\alpha\delta)_{jkm} + (\tau\gamma\alpha\delta)_{ijkm} + \epsilon_{rijkm} \end{aligned} \tag{3.7}$$

$$\begin{cases} r = 1, .., 6 \\ i = 1, .., 3 \\ j = 1, .., 3 \\ k = 1, .., 4 \\ m = 1, .., 7 \end{cases}$$

where μ is the mean effect of all observations, β_r is the main effect of the *r*th block, τ_i is the main effect of the *i*th colour, γ_j is the main effect of the *j*th level of OD, α_k is the main effect of the *k*th pulsation and δ_m is the main effect of the *m*th time. λ is the main effect of the IR light. ϵ_{rijkm} is the i.i.d random error with $\epsilon_{rijkm} \sim N(0, \sigma^2)$. In addition we have the two-way interactions, where $(\tau\gamma)_{ij}$ is the effect of the interaction between the *i*th colour and the *j*th level of OD. The rest of the two-way coupled coefficients are interpreted in the same way.

For the three-way interactions we have $(\tau\gamma\alpha)_{ijk}$ which is the effect of the interaction between the *i*th colour, the *j*th level of OD and the *k*th pulsation. The rest of the three-way coupled coefficients are interpreted in the same way. The four-way interaction, $(\tau\gamma\alpha\delta)_{ijkm}$, is the effect of the interaction between the *i*th colour, the *j*th level of OD, the *k*th pulsation and the *m*th time.

The model was fitted using the function aov in the package stats. Summary of the intrablock analysis of variance of the model in (3.7) is shown in Table 3.9.

Table 3.9: Main Experiment: Intrablock Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
IR	1	1.44	1.44	0.64	0.4230
time	6	113.54	18.92	8.46	0.0000
colour	2	13.16	6.58	2.94	0.0536
OD	2	45.75	22.87	10.23	0.0000
pulsation	3	20.32	6.77	3.03	0.0291
block	5	337.79	67.56	30.22	0.0000
time:colour	18	17.58	0.98	0.44	0.9797
time:OD	12	9.42	0.79	0.35	0.9787
colour:OD	4	22.44	5.61	2.51	0.0411
time:pulsation	18	16.98	0.94	0.42	0.9833
colour:pulsation	6	31.41	5.24	2.34	0.0306
OD:pulsation	6	95.33	15.89	7.11	0.0000
time:colour:OD	24	9.66	0.40	0.18	1.0000
time:colour:pulsation	36	20.72	0.58	0.26	1.0000
time:OD:pulsation	36	22.87	0.64	0.28	1.0000
colour:OD:pulsation	12	117.86	9.82	4.39	0.0000
time:colour:OD:pulsation	72	36.08	0.50	0.22	1.0000
Residuals	513	1146.90	2.24		

The intrablock analysis of variance shows that time, OD and pulsation have a significant effect on the response variable at a 5% significance level. Colour has a p-value equal 0.054, and is thus significant at a 10% significance level. We see that colour and pulsation contributes less to the variation in the data, compared to time and OD. Block has a very high mean square compared to the other main effects, which indicates that there is difference in the response due to block. Evaluating the interactions, the summary shows that all interactions are significant, except those who include time. The summary also shows that IR light do not have a significant effect on the response variable, with a p-value equal 0.42.

Comparing Level Means

Due to complete randomization of the treatments within each replicate, we use the multiple linear regression models in Section 2.6.1 to compare the effects of each level of the factors. We still consider the block effects as fixed when estimating the effects of the levels. Since the interactions including time showed to be non-significant, we exclude these from our

model. The reduced model is then

$$y_{rijkm} = \mu + \beta_r + \tau_i + \gamma_j + \alpha_k + \delta_m$$
$$+ (\tau\gamma)_{ij} + (\tau\alpha)_{ik} + (\gamma\alpha)_{jk} + (\tau\gamma\alpha)_{ijk} + \epsilon_{rijkm} \begin{cases} r = 1, .., 6 \\ i = 1, .., 3 \\ j = 1, .., 3 \\ k = 1, .., 4 \\ m = 1, .., 7 \end{cases} \tag{3.8}$$

where we have μ as the grand mean effect of all levels. β_r is the main effect of the *r*th block, τ_i is the main effect of the *i*th colour, γ_j is the main effect of the *j*th level of OD, α_k is the main effect of the *k*th pulsation and δ_m is the main effect of the *m*th time. ϵ_{rijkm} is the i.i.d random error with $\epsilon_{rijkm} \sim N(0, \sigma^2)$. In addition we have the two-way interactions, where $(\tau\gamma)_{ij}$ is the effect of the interaction between the *i*th colour and the *j*th level of OD, $(\tau\alpha)_{ik}$ is the effect of the interaction between the *i*th colour and the *k*th pulsation and $(\gamma\alpha)_{jk}$ is the effect of the interaction between the *j*th level of OD and the *k*th pulsation. For the three-way interaction we have $(\tau\gamma\alpha)_{ijk}$ which is the effect of the interaction between the *i*th colour, the *j*th level of OD and the *k*th pulsation.
In addition we have the constraints
$\sum_{r=1}^{6} \beta_r = 0$
$\sum_{i=1}^{3} \tau_i = 0$
$\sum_{j=1}^{3} \gamma_j = 0$
$\sum_{k=1}^{4} \alpha_k = 0$
$\sum_{m=1}^{7} \delta_m = 0$
For the two-way interactions we have the constraints
$\sum_{j=1}^{3} (\tau\gamma)_{ij} = 0, i = 1, .., 3$
$\sum_{i=1}^{3} (\tau\gamma)_{ij} = 0, j = 1, .., 3$
$\sum_{k=1}^{4} (\tau\alpha)_{ik} = 0, i = 1, .., 3$
$\sum_{i=1}^{3} (\tau\alpha)_{ik} = 0, k = 1, .., 4$
$\sum_{k=1}^{4} (\gamma\alpha)_{jk} = 0, j = 1, .., 3$
$\sum_{j=1}^{3} (\gamma\alpha)_{jk} = 0, k = 1, .., 4$
For the three-way interactions we have the constraints
$\sum_{i=1}^{3} (\tau\gamma\alpha)_{ijk} = 0, j = 1, .., 3, k = 1, .., 4$
$\sum_{j=1}^{3} (\tau\gamma\alpha)_{ijk} = 0, i = 1, .., 3, k = 1, .., 4$
$\sum_{k=1}^{4} (\tau\gamma\alpha)_{ijk} = 0, i = 1, .., 3, j = 1, .., 3$
The model was fitted using the function lm in the package stats. A summary of the model in (3.8) is shown in Table 3.10.

Table 3.10: Main Experiment: Summary of the Multiple Linear Model in (3.8)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5400	0.0482	11.20	0.0000
time1	-0.5685	0.1181	-4.81	0.0000
time2	-0.3119	0.1181	-2.64	0.0085
time3	-0.1868	0.1181	-1.58	0.1142
time4	0.0076	0.1181	0.06	0.9489
time5	0.0453	0.1181	0.38	0.7017
time6	0.2790	0.1181	2.36	0.0184
colourB	-0.2278	0.0704	-3.23	0.0013
colourG	0.1727	0.0690	2.50	0.0125
OD1	0.2504	0.0684	3.66	0.0003
OD3	0.0996	0.0686	1.45	0.1473
pulsation0	-0.0840	0.0852	-0.99	0.3244
pulsation1	-0.0578	0.0853	-0.68	0.4979
pulsation2	0.2194	0.0904	2.43	0.0155
block1	-0.1565	0.1231	-1.27	0.2040
block2	1.0533	0.1243	8.47	0.0000
block3	0.4427	0.1243	3.56	0.0004
block4	-0.9955	0.1242	-8.02	0.0000
block5	-0.5446	0.1242	-4.39	0.0000
colourB:OD1	0.1803	0.0976	1.85	0.0651
colourG:OD1	-0.1008	0.0982	-1.03	0.3049
colourB:OD3	-0.3117	0.0981	-3.18	0.0015
colourG:OD3	0.3074	0.0986	3.12	0.0019
colourB:pulsation0	-0.1500	0.1205	-1.24	0.2136
colourG:pulsation0	0.2612	0.1247	2.09	0.0366
colourB:pulsation1	0.1844	0.1205	1.53	0.1264
colourG:pulsation1	-0.0738	0.1184	-0.62	0.5333
colourB:pulsation2	0.2674	0.1207	2.22	0.0270
colourG:pulsation2	-0.2261	0.1277	-1.77	0.0771
OD1:pulsation0	0.0266	0.1194	0.22	0.8237
OD3:pulsation0	0.3840	0.1209	3.18	0.0016
OD1:pulsation1	-0.7090	0.1181	-6.00	0.0000
OD3:pulsation1	0.2798	0.1216	2.30	0.0217
OD1:pulsation2	0.2732	0.1190	2.30	0.0220
OD3:pulsation2	-0.3744	0.1206	-3.10	0.0020
colourB:OD1:pulsation0	0.5011	0.1738	2.88	0.0041
colourG:OD1:pulsation0	-0.2102	0.1717	-1.22	0.2214
colourB:OD3:pulsation0	0.2551	0.1679	1.52	0.1290
colourG:OD3:pulsation0	-0.2992	0.1698	-1.76	0.0784
colourB:OD1:pulsation1	-0.5061	0.1716	-2.95	0.0033
colourG:OD1:pulsation1	0.1431	0.1823	0.79	0.4327
colourB:OD3:pulsation1	0.3152	0.1695	1.86	0.0634
colourG:OD3:pulsation1	-0.0711	0.1681	-0.42	0.6723
colourB:OD1:pulsation2	-0.4386	0.1683	-2.61	0.0093
colourG:OD1:pulsation2	0.2967	0.1675	1.77	0.0769
colourB:OD3:pulsation2	0.4925	0.1729	2.85	0.0045
colourG:OD3:pulsation2	0.0043	0.1707	0.03	0.9799

To analyze the effect of the levels, we have used effect coding in R. The estimated main effects of the levels sum to zero as specified in the model in (3.8). The estimated main effects are thus the deviation from the overall mean effect of all levels. Looking at the summary in Table 3.10 we see that the grand mean effect of all levels is equal 0.54.

Evaluating the main effect of blocks in our model, we see that the estimated main effect of block1 is not significant. Block1 corresponds to the first day of experiments, with salmon lice at 8 days past moulting. Block2, which corresponds to the second day with salmon lice at 9 days past moulting, has a significant estimated main effect equal 1.05. Block3 with salmon lice at 12 days moulting, has a significant estimated main effect equal 0.44. Block4 and block5, both with salmon lice at 1 one day past moulting have significant negative estimated effects. Using $\sum_{r=1}^{6} \beta_r = 0$, we find the estimated main effect of block6 to be 0.20. Block6 corresponds to the sixth day of experiments, with salmon lice at 2 days past moulting. Since blocks are confounded with the days past moulting of the salmon lice used that particular day (block), it is hard to tell if the estimated main effects are due to blocks or the days past moulting of the salmon lice.

The main effect of time is increasing as the levels of time increases. The summary shows that the first minute has a significant negative estimated main effect. The sixth minute has a significant main effect equal 0.28. The estimated main effect of the seventh minute is found to be 0.74. This corresponds to the plot in Figure 3.13, where we see that the means of the response variable are close to zero for all levels of time, but there is a slight increase as the levels of time increases, showing an approximately linear relationship between the levels of time and the response variable. We also observe that as the time increases, spread in the data increases. This indicates that the salmon lice swim both towards and away from the light source when exposed to light.

Analyzing the levels of colour, the summary shows that blue has a significant estimated main effect equal -0.23 and green has a significant estimated main effect equal 0.17. White has an estimated main effect equal 0.06. The plot in Figure 3.14 shows that the medians of the response for each of the levels are small, so there is no strong relationship between colours and the response variable. This indicates that the colours do not have a high influence on the response variable.

OD1 has a significant estimated main effect equal 0.25 and the estimated main effect of OD5 is found to be -0.35. The estimated main effect of OD3 is equal 0.10 and is not significant. The plot in Figure 3.15 shows that the medians of the response variable are close to zero for all levels of OD, but there are outliers in both direction for OD1. This might be an indication of that the salmon lice moves towards and away from the light source for OD1.

Evaluating the effects of pulsation, we see that pulsation2 has a significant estimated main effect equal 0.22. The estimated main effects of Pulsation0 and pulsation1 are not significant, equal -0.08 and -0.06 respectively. The estimated main effect of pulsation3 is found to be -0.08.

Analyzing the two-way interactions, we first find the estimated parameters of the omitted interactions. These are found by using the constraints in (3.8). As an example, we have the constraint

$$colourB : OD1 + colourB : OD3 + colourB : OD5 = 0 \tag{3.9}$$

From (3.9) we find the estimated parameter of the interaction, colourB:OD5, to be 0.13. Summary of the omitted estimated parameters of the two-way interactions are shown in Table 3.11.

The interaction between green and OD3 has a significant estimated parameter equal 0.3074, which is the highest of the interactions between colour and OD. Evaluating the estimated parameters between colour and pulsation, we see that the interactions between green and pulsation0, blue and pulsation2 and white and pulsation3 have the highest estimated parameters equal 0.26, 0.27 and 0.26 respectively. Among the interactions between OD and pulsation there are several significant estimated parameters. The interaction between OD5 and pulsation1 has the highest estimated parameter equal 0.43, followed by the interaction between OD1 and pulsation3 with estimated parameter equal 0.41.

The estimated parameters of the omitted three-way interactions are found by using the constraints in (3.8). Finding the estimated parameter of the interaction colourB:OD1:pulsation3, we have the constraint

$$\begin{aligned} &colourB:OD1:pulsation0 + colourB:OD1:pulsation1 \\ &+ colourB:OD1:pulsation2 + colourB:OD1:pulsation3 \\ &= 0 \end{aligned} \tag{3.10}$$

Using (3.10), we find the estimated parameter of colourB:OD1:pulsation3 to be 0.44. Summary of the omitted estimated parameters for the three-way interactions is found in Table 3.12.

The estimated contributions on the response variable can be found by summing estimated main effects and estimated interaction parameters. As an example, consider the level combination blue, OD1 and pulsation0. The estimated contribution on the response variable is then found as

$$\begin{aligned} &\hat{\tau}_{colourB} + \hat{\gamma}_{OD1} + \hat{\alpha}_{pulsation0} + \hat{(\tau\gamma)}_{colourB:OD1} \\ &+ \hat{(\tau\alpha)}_{colourB:pulsation0} + \hat{(\gamma\alpha)}_{OD1:pulsation0} \\ &+ \hat{(\tau\gamma\alpha)}_{colourB:OD1:pulsation0} \\ &= -0.2278 + 0.2504 - 0.0840 + 0.1803 - 0.1500 + 0.0266 + 0.5011 \\ &= 0.4966 \end{aligned} \tag{3.11}$$

Summary of the estimated contributions from all level combinations are shown in Table 3.13.

The summary in Table 3.13 shows that the level combination green, OD1 and pulsation2 has the highest estimated contribution on the response variable equal 0.89, followed by the level combination green, OD3 and pulsation0 and white, OD1 and pulsation2 with estimated contributions equal 0.84 and 0.82 respectively. To find the total estimated response at a given level combination, we need to sum the estimated contribution from that level combination and intercept at a given level of time and block. We find the highest estimated response with the level combination green, OD1 and pulsation2 at time7 and block2 to be 3.22.

Table 3.11: Main Experiment: Omitted Estimated Parameters of the Two-Way Interactions

	Estimate
colourB:OD5	0.1314
colourG:OD5	-0.4082
colourW:OD1	-0.0795
colourW:OD3	0.0043
colourW:OD5	0.0752
colourB:pulsation3	-0.3018
colourG:pulsation3	0.0387
colourW:pulsation0	-0.1112
colourW:pulsation1	-0.1106
colourW:pulsation2	-0.0413
colourW:pulsation3	0.2631
OD1:pulsation3	0.4092
OD3:pulsation3	-0.2894
OD5:pulsation0	-0.4106
OD5:pulsation1	0.4292
OD5:pulsation2	0.1012
OD5:pulsation3	-0.1198

Table 3.12: Main Experiment: Omitted Estimated Parameters of the Three-Way Interactions

	Estimate
colourB:OD1:pulsation3	0.4436
colourB:OD3:pulsation3	-1.0628
colourG:OD1:pulsation3	-0.2296
colourG:OD3:pulsation3	0.3660
colourB:OD5:pulsation0	-0.7562
colourB:OD5:pulsation1	0.1909
colourB:OD5:pulsation2	-0.0536
colourB:OD5:pulsation3	0.6192
colourG:OD5:pulsation0	0.5094
colourG:OD5:pulsation1	-0.0720
colourG:OD5:pulsation2	-0.3010
colourG:OD5:pulsation3	-0.1364
colourW:OD1:pulsation0	-0.2909
colourW:OD1:pulsation1	0.3630
colourW:OD1:pulsation2	0.1419
colourW:OD1:pulsation3	-0.2140
colourW:OD3:pulsation0	0.0441
colourW:OD3:pulsation1	-0.2441
colourW:OD3:pulsation2	-0.4968
colourW:OD3:pulsation3	0.6968
colourW:OD5:pulsation0	0.2468
colourW:OD5:pulsation1	-0.1189
colourW:OD5:pulsation2	0.3546
colourW:OD5:pulsation3	-0.4825

Table 3.13: Main Experiment: Estimated Contributions on the Response Variable

	Estimate
colourB:OD1:pulsation0	0.4966
colourG:OD1:pulsation0	0.3159
colourB:OD3:pulsation0	-0.0348
colourG:OD3:pulsation0	0.8417
colourB:OD1:pulsation1	-0.8856
colourG:OD1:pulsation1	-0.3752
colourB:OD3:pulsation1	0.2817
colourG:OD3:pulsation1	0.6568
colourB:OD1:pulsation2	0.5243
colourG:OD1:pulsation2	0.8855
colourB:OD3:pulsation2	0.1650
colourG:OD3:pulsation2	0.2029
colourB:OD1:pulsation3	0.6763
colourB:OD3:pulsation3	-2.1715
colourG:OD1:pulsation3	0.4630
colourG:OD3:pulsation3	0.6174
colourB:OD5:pulsation0	-1.8472
colourB:OD5:pulsation1	0.3003
colourB:OD5:pulsation2	0.0880
colourB:OD5:pulsation3	-0.3264
colourG:OD5:pulsation0	-0.3095
colourG:OD5:pulsation1	-0.3599
colourG:OD5:pulsation2	-0.7920
colourG:OD5:pulsation3	-0.8806
colourW:OD1:pulsation0	-0.2335
colourW:OD1:pulsation1	-0.2884
colourW:OD1:pulsation2	0.8192
colourW:OD1:pulsation3	0.6067
colourW:OD3:pulsation0	0.3919
colourW:OD3:pulsation1	0.0263
colourW:OD3:pulsation2	-0.5341
colourW:OD3:pulsation3	0.7519
colourW:OD5:pulsation0	-0.5787
colourW:OD5:pulsation1	-0.0778
colourW:OD5:pulsation2	0.4142
colourW:OD5:pulsation3	-0.6365

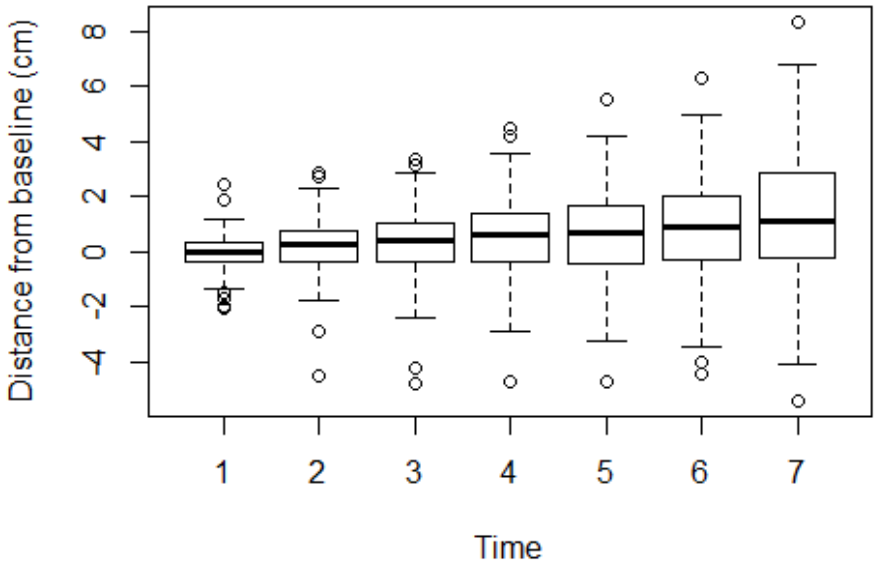

Figure 3.13: Main experiment: Box-plot of the response as a function of time

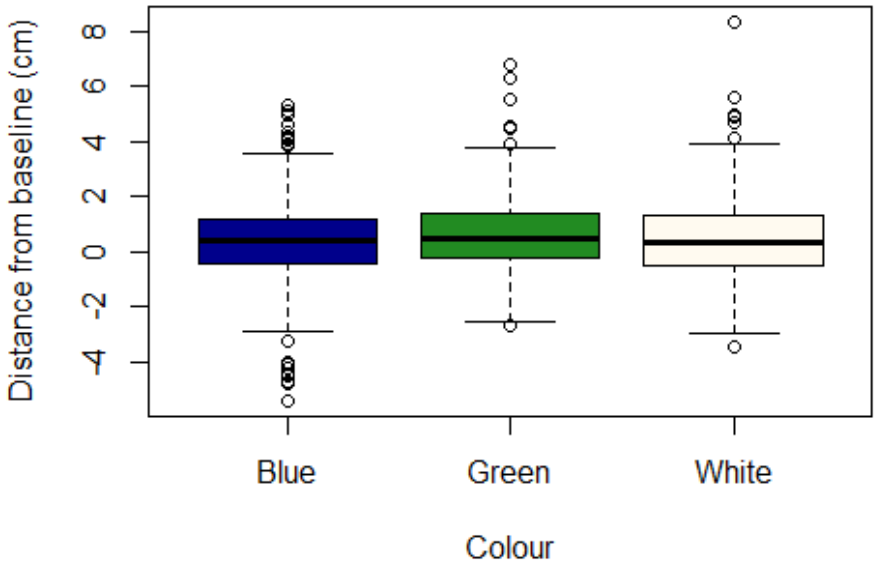

Figure 3.14: Main experiment: Box-plot of the response as a function of colour

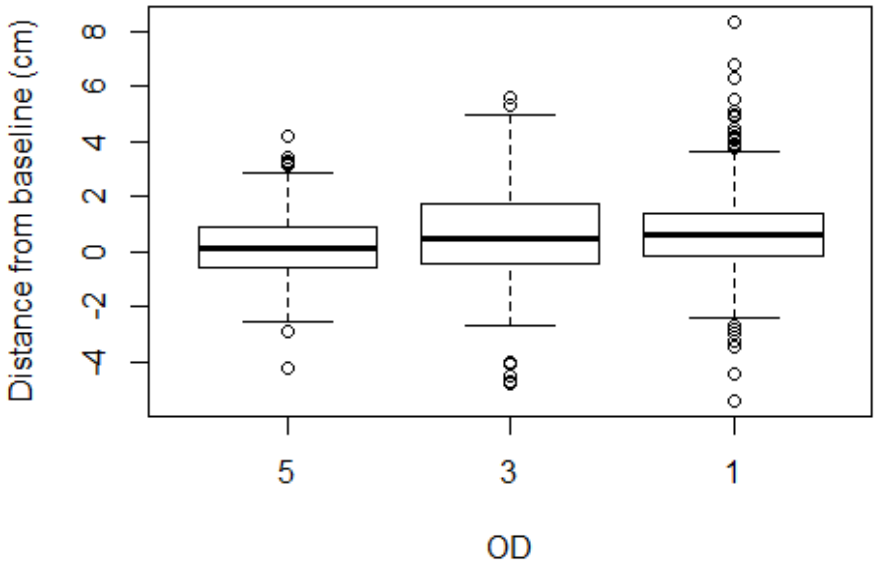

Figure 3.15: Main experiment: Box-plot of the response as a function of OD

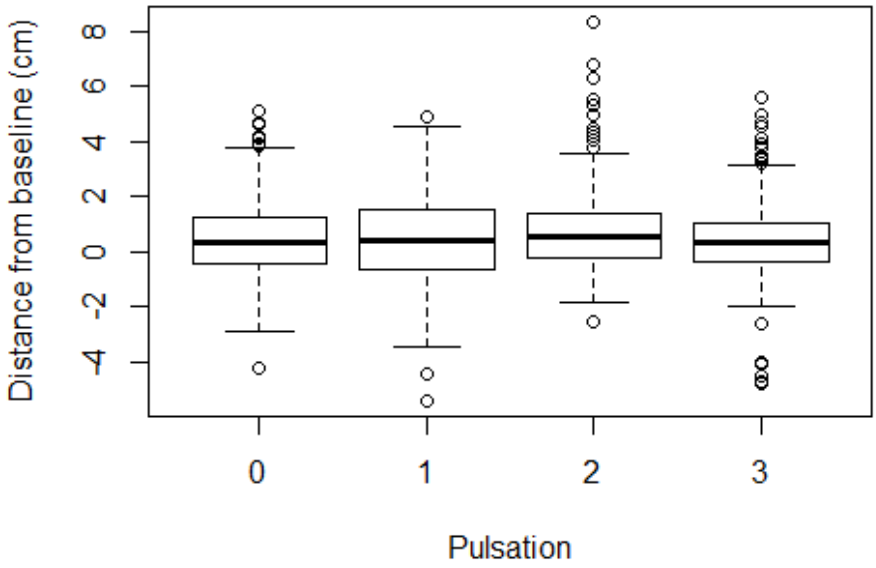

Figure 3.16: Main experiment: Box-plot of the response as a function of pulsation

Checking Model Assumptions

For the linear regression model in (3.8), we assume the residuals to be normal distributed and independent of the factors. The residual plot in Figure 3.17 of the model shows that for fitted values between -2 and 2, there is no distinct pattern of the residuals. For fitted values smaller than -2 and greater than 2, the residuals shows a tendency to be positive. The Q-Q plot in Figure 3.18 shows that the residuals are approximately normal distributed. For values lower than -2 and greater than 2, we see that the Q-Q plot deviates some from the straight line. From these two plots, it seems like the model in (3.8) fits data between -2 and 2 best.

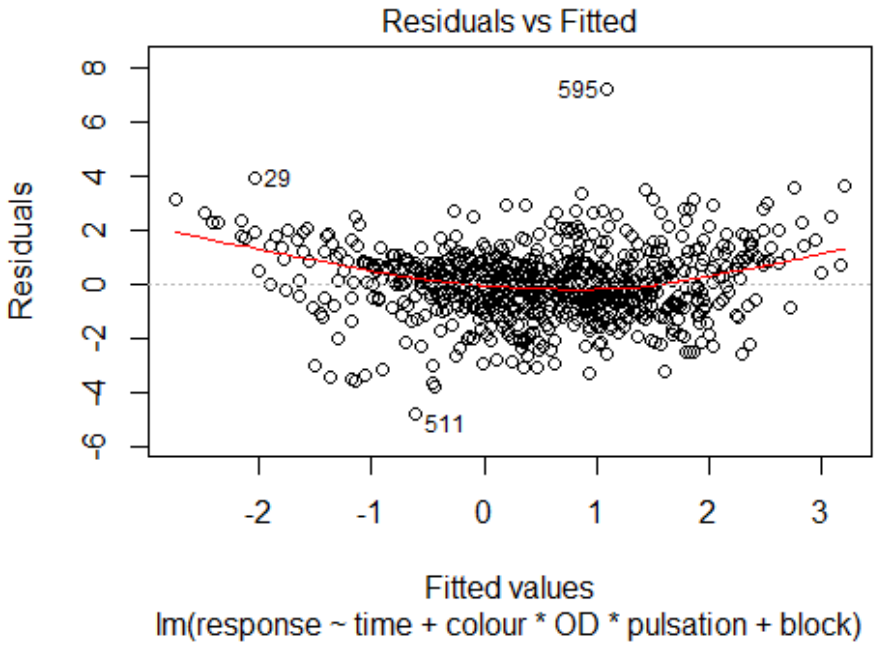

Figure 3.17: Main experiment: Residual plot of the model in (3.8)

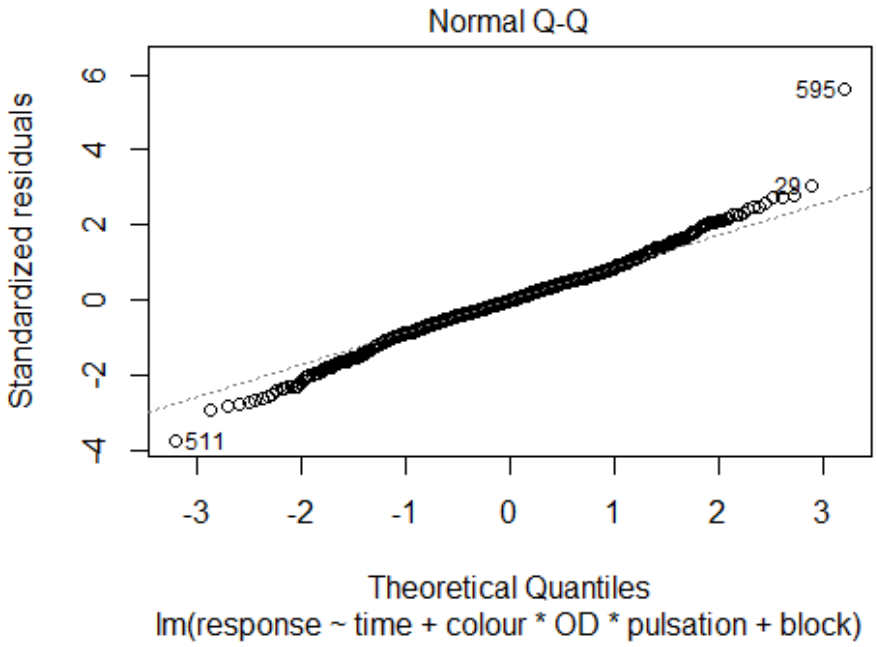

Figure 3.18: Main experiment: Q-Q plot of the model in (3.8)

Chapter 4

Summary and Conclusion

The aim of this thesis was to design and analyze experiments in order to obtain valid and objective findings about the phototactic response of salmon lice. The response variable in these experiments is defined as the distance at which the salmon lice have moved when exposed to the different treatments of light. The independent factors in these experiments are colour, optical density (OD) and pulsation. Colour consists of the levels white, blue with emission peak at 455 nm, green with emission peak at 525 nm and red with emission peak at 640 nm. Optical density consists of the levels OD1 to OD7, where OD1 corresponds to letting 0.1 of all the light through. Optical density is logaritmic, decreasing the irradiance by 10% of the previous level for each increasing OD number. Pulsation consists of the levels pulsation0 to pulsation3. Pulsation0 corresponds to the light being constantly on, pulsation1 corresponds to light being 0.1 seconds on and 0.1 seconds off, pulsation2 corresponds to light being 2 seconds on and 3 seconds off and pulsation3 corresponds to light being 5 seconds on and 5 seconds off.

First a pilot experiment was conducted to possibly reduce the number of levels of the factors colour and optical density, before the main experiment was conducted with the factors colour, optical density and pulsation. This chapter provides an overview of the experimental designs, the main results and recommendations for further work.

4.1 The Pilot Experiment

The pilot experiment was conducted as a split-plot experiment, with increasing levels of OD within each level of colour such that the effect of OD was confounded with time. Due to the increasing levels of OD within each level of colour, the pilot experiment was in fact conducted as a repeated measurement experiment.

The analysis shows that red (with emission peak at 640 nm) do not have a significant main effect on the response variable, and was thus excluded from the main experiment. OD2 to OD7 were included in the pilot experiment, and the analysis shows that OD3 and OD2 have a significant effect on the response variable. In the main experiment we decided to include OD5, OD3 and OD1. OD5 was included to have a level in the transition phase,

where only some of the salmon lice responded. OD3 was included in the main experiment based on the analysis showing a significant effect on the response variable at OD3, and OD1 was included to maybe provoke an even stronger response.

4.2 The Main Experiment

The main experiment was conducted as an incomplete block design, and the results were analyzed with intrablock analysis of variance by considering the block effects as fixed.

The analysis shows that there is difference in the response due to block, and block2 with salmon lice at 9 days past moulting shows to have the highest estimated main effect equal 1.05. Block4 with salmon lice at 1 day past moulting shows to have the lowest estimated main effect equal -1.00. It is not possible to tell if the block effects are due to blocks or days past moulting of the salmon lice, since they are confounded.

The analysis shows that phototactic response of the salmon lice increases as the intensity of light increases. OD1 shows to have a significant estimated main effect equal 0.25 at a 5% significance level. OD3 with estimated main effect equal 0.10 is not significant, while the analysis shows that OD5 has a negative estimated main effect on the response variable compared to the mean effect of all levels of the factors. This correspond to previous studies which showed that bright light had a higher effect on the response compared to medium and dim light (Fields et al., 2017).

Comparing the effect between blue, green and white, the analysis shows that green (with emission peak at 525 nm) has the highest significant estimated main effect equal 0.17. The estimated main effect of white is 0.06 and the estimated main effect of blue (with emission peak at 455 nm) is equal -0.23 compared to the mean effect of all levels of the factors. This results are consistent with previous research, which have shown that a wavelength at 550 nm obtained the best response and a wavelength at 400 nm obtained the lowest response (Bron et al., 1993).

The estimated effect of pulsation2 with an ON:OFF cycle (2:3 s) shows to have a significant estimated main effect equal 0.22, which is the highest among the levels of pulsation. Pulsation1 with cycle (0.1:0.1 s), pulsation0 (light constantly on) and pulsation3 with cycle (5:5 s) do not have significant main effects equal -0.06, -0.08 and -0.08 respectively. The study by Fields et al. (2017) showed that all levels of pulsation in their study had a significant estimated effect on the response variable. Pulsation with an ON:OFF cycle (1.8:0.9) had the lowest effect on the phototactic response, and attracted 24 percent of the salmon lice. Pulsation with cycle (3.5:5.5 s) attracted 80 percent of the salmon lice, and an increase of the OFF time to 16.5 s did not show a significant higher effect on the phototactic response.

Comparing the pilot experiment and the main experiment, we observe that the response variable in the pilot experiment obtained a greater value for each of the treatments. In the pilot experiment the salmon lice were exposed to light for 48 minutes, while in the main experiment the salmon lice were exposed to light for seven minutes. This could be an indication that time is a factor with high effect on the phototactic response. In the main experiment, the last and seventh minute has the highest estimated main effect on the response variable equal 0.74.

The analysis shows that contribution from the level combination green, OD1 and pulsation2 gives the highest estimated phototactic response of salmon lice. The estimated contribution to the response from this level combination is equal 0.89. The highest estimated response is obtained with this level combination at time7 and block2, and is equal 3.22. The level combination blue, OD3 and pulsation3 has an estimated contribution on the response variable equal -2.17, which is the lowest of all the level combinations. The estimated response at time7 and block2 with this level combination is equal 0.16.

4.3 Recommendations for Further Work

This study have showed that the level combination green, OD1 and pulsation2 make up the light source with the highest estimated contribution on the response variable. If light is to be used as a method to reduce the infection rate of salmon lice, it would be interesting to investigate if this light source have an effect on salmon lice which are already attached to a host salmon.

In these experiments, a raceway with length 40 cm was used, which leads to a limitation for the salmon lice. That is, salmon lice which moved into the wall where the light source was positioned, were not able to respond any further. By using an aquarium with a longer raceway, it would be possible to observe the range the salmon lice potential could have moved. The range of the phototactic response would be important findings in the research on whether light can used as a method to reduce the infection rate of salmon lice.

The analysis shows that time influenced the length of which the salmon lice moved. By carrying out experimental runs for a longer time period, one can investigate further the effect of time on the phototactic response of salmon lice. It could be interesting to look at when the effect of time decreases, and do not longer show a significant effect on the response variable.

Bibliography

Bjørnstad, L. F., Solstad, M. A., 2018. Investigation of light sensitivity of the copepod lepeophtheirus salmonis using computer vision. Unpublished project thesis.

Bjørnstad, L. F., Solstad, M. A., Pre-print. Investigation of light response and swimming behaviour of salmon lice lepeophtherius salmonis using feature detection and tracking.

Bron, J. E., Sommerville, C., Rae, G. H., 1993. Aspects of the behavior of copepodid larvae of the salmon louse lepeophtheirus salmonis (krøyer, 1837). New York : CRC Press, pp. 125–142.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Regression. Model, Methods and Applications. Springer-Verlag Berlin Heidelberg.

Fields, D., Skiftesvik, A., Browman, H., 2017. Behavioural responses of infective- stage copepodids of the salmon louse (lepeophtheirus salmonis, copepoda:caligidae) to host-related sensory cues. Journal of Fish Diseases 41 (6).

Flamarique, I. N., Gulbransen, C., Galbraith, M., Stucchi, D., 2009. Monitoring and potential control of sea lice using an led-based light trap.(report). Canadian Journal of Fisheries and Aquatic Sciences 66 (8).

Glover, K. A., Stølen, B., Messmer, A., Koop, B. F., Torrissen, O., Nilsen, F., 2011. Population genetic structure of the parasitic copepod lepeophtheirus salmonis throughout the atlantic. Marine Ecology Progress Series 427, 161–172.

Hedeker, D., Gibbons, R. D., 2006. Longitudinal Data Analysis. John Wiley and Sons.

Loy, A., Follett, L., Hofmann, H., 2016. Variations of q–q plots: The power of our eyes! The American Statistician 70 (2), 202–214.
URL https://doi.org/10.1080/00031305.2015.1077728

Miljeteig, C., Olsen, A. J., Båtnes, A. S., Altin, D., Nordtug, T., Alver, M. O., Speed, J. D. M., Jenssen, B. M., 2014. Sex and life stage dependent phototactic response of the marine copepod calanus finmarchicus (copepoda: Calanoida). Journal of Experimental

Marine Biology and Ecology 451, 16 – 24.
URL http://www.sciencedirect.com/science/article/pii/S0022098113003730

Montgomery, D. C., 2009. Design and Analysis of Experiments, 7th Edition. Arizona State University.

Ryan, T. P., 2011. Statistical Methods for Quality Improvement, 3rd Edition. John Wiley Sons.

Toutenburg, H., Shalabh, 2009. Incomplete Block Designs. Springer New York, New York, NY, pp. 181–244.
URL https://doi.org/10.1007/978-1-4419-1148-3_6

Vatn, J. A., 2019. Metode for kartlegging av den fototaktiske svømmeresponsen til lepeophtheirus salmonis.

Wiik, H. E., 2013. Non-regular design with censored data. Unpublished project thesis.

Wilk, M. B., Gnanadesikan, R., 1968. Probability plotting methods for the analysis of data. Biometrika 55 (1), 1–17.
URL http://www.jstor.org/stable/2334448

Appendix

4.4 Data

Table 4.1: Summary of the Data from the Pilot Experiment

	replicate	colour	OD	response3
1	I :24	B:18	2:12	Min. :-5.4727
2	II :24	G:18	3:12	1st Qu.:-0.2337
3	III:24	R:18	4:12	Median : 0.5656
4		W:18	5:12	Mean : 2.1042
5			6:12	3rd Qu.: 3.1193
6			7:12	Max. :12.2426

Table 4.2: Summary of the Data from the Main Experiment

	replicate	block	time	colour	OD	pulsation	response
1	I :252	1:126	1:108	B:252	1:252	0:189	Min. :-5.4045
2	II :252	2:126	2:108	G:252	3:252	1:189	1st Qu.:-0.3541
3	III:252	3:126	3:108	W:252	5:252	2:189	Median : 0.4009
4		4:126	4:108			3:189	Mean : 0.5400
5		5:126	5:108				3rd Qu.: 1.3042
6		6:126	6:108				Max. : 8.3361
7			7:108				

Table 4.3: Summary of the Data including IR Light from the Main Experiment

	replicate	block	time	colour	IR	OD	pulsation	response
1	I :259	1:133	1:111	0: 21	off:756	0: 21	0:210	Min. :-5.4045
2	II :259	2:133	2:111	B:252	on : 21	1:252	1:189	1st Qu.:-0.3532
3	III:259	3:126	3:111	G:252		3:252	2:189	Median : 0.4059
4		4:126	4:111	W:252		5:252	3:189	Mean : 0.5328
5		5:133	5:111					3rd Qu.: 1.2919
6		6:126	6:111					Max. : 8.3361
7			7:111					

4.5 Results Pilot Experiment

4.5.1 Linear Mixed Effects Model for data.pilot

```
Formula: response3 ~ OD + (1 | replicate/colour)
   Data: data.pilot

REML criterion at convergence: 327

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.38565 -0.46965  0.02598  0.56559  1.78009

Random effects:
 Groups           Name        Variance Std.Dev.
 colour:replicate (Intercept) 6.716    2.592
 replicate        (Intercept) 7.188    2.681
 Residual                     4.270    2.066
Number of obs: 72, groups:  colour:replicate, 12; replicate, 3

Fixed effects:
            Estimate Std. Error t value
(Intercept)   0.5640     1.8198   0.310
OD2           2.6506     0.8436   3.142
OD3           2.2036     0.8436   2.612
OD4           2.5537     0.8436   3.027
OD5           1.6358     0.8436   1.939
OD6           0.1976     0.8436   0.234

Correlation of Fixed Effects:
    (Intr) OD2    OD3    OD4    OD5
OD2 -0.232
OD3 -0.232  0.500
OD4 -0.232  0.500  0.500
OD5 -0.232  0.500  0.500  0.500
OD6 -0.232  0.500  0.500  0.500  0.500
```

4.5.2 Implemented Algorithm in R - Finding Estimated Response

```
#NA: Salmon lice that we assume have gone into the wall
truncated_pilot = function(coordinates){
  coor = coordinates
  #removing the first 14 minutes (of dark)
  v = coor[-c(1:21000),]
  v = as.matrix(v)
  v[v>1900]= NA
  # 1500 rows are equivalent to one minute
  m = 1500
  #converting from pixles to cm and finding the baseline
  ratio = 40/1920
  matrix.baseline = v[1:m,]
  matrix.baseline.cm = matrix(0,nrow(matrix.baseline),ncol(
      matrix.baseline))
  for (i in 1:nrow(matrix.baseline)){
    for (j in 1:ncol(matrix.baseline)){
      if (!is.na(matrix.baseline[i,j])){
        matrix.baseline.cm[i,j] = matrix.baseline[i,j]*
            ratio
      }
    }
  }

  baseline = mean(matrix.baseline.cm[matrix.baseline.cm>0])
  #Matrix that corresponds to the last minute of OD2
  matrix.od2 = v[72000:73500,]
  matrix.od2.cm = matrix(0,nrow(matrix.od2),ncol(matrix.od2
      ))

  for (i in 1:nrow(matrix.od2)){
    for (j in 1:ncol(matrix.od2)){
      if (!is.na(matrix.od2[i,j])){
        matrix.od2.cm[i,j] = matrix.od2[i,j]*ratio
      }
    }
  }

  #Distances between the baseline and the position at OD2
  distance = matrix(0,nrow(matrix.od2.cm),ncol(matrix.od2.
      cm))
  for (i in 1:nrow(matrix.od2.cm)){
    for (j in 1:ncol(matrix.od2.cm)){
      if (matrix.od2.cm[i,j]!=0){
        distance[i,j] = matrix.od2.cm[i,j]-baseline
```

```
        }
      }
    }

    #finding x.nw and variance of x.nw
    x.nw = mean(distance[distance>0])
    g = as.vector(distance)
    var.x.nw = var(g[distance>0])
    #finding m.w and m.nw
    l = v[73500,]
    m.w = sum(is.na(l))
    m.nw = length(l)-m.w
    #porpotion.w = 1-r and porpotion.nw = r
    porpotion.w = m.w/(m.nw+m.w)
    porpotion.nw = m.nw/(m.w+m.nw)
    #findning the limit b
    b =  31 - baseline
    #The inverse cumulative
    ic = qnorm(porpotion.nw)
    #The probability density
    pd = dnorm(ic)
    #The estimated variance s2
    s2 = var.x.nw/(1-ic*pd/porpotion.nw-(pd/porpotion.nw)^2)
    s= sqrt(s2)
    #The expected value of z
    o = (b-x.nw)/(sqrt((1/m.w)+(1/m.nw))*s)
    dp.z = dnorm(o)
    c.z = pnorm(o)
    gamma = sqrt((1/m.w)+(1/m.nw))*s*dp.z/(1-c.z)
    x.w = gamma + x.nw
    x.t = x.nw*porpotion.nw + x.w*porpotion.w
    return(c(estimated.distance = x.t))
}
```