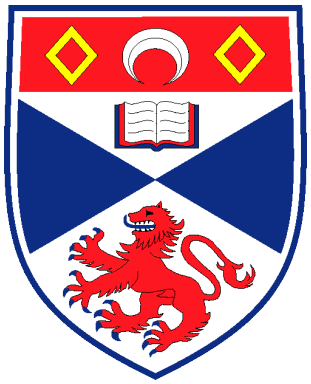
Estimating Anglerfish Abundance from Trawl Surveys, and Related Problems

Yuan Yuan



Thesis submitted for the degree of
DOCTOR OF PHILOSOPHY
in the School of Mathematics and Statistics
UNIVERSITY OF ST ANDREWS

April 2012

Declaration

I, Yuan Yuan, hereby certify that this thesis, which is approximately 60,000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in October 2007 and as a candidate for the degree of Doctor of Philosophy in Statistics in October 2007; the higher study for which this is a record was carried out in the University of St Andrews between 2007 and 2012.

date: $8\ February,\ 2012$ signature of candidate: _____

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Doctor of Philosophy in Statistics in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

date: $8\ February,\ 2012$ signature of supervisor: _____

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. We have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the electronic publication of this thesis:

Embargo on both all or part of printed copy and electronic copy for the same fixed period of two years on the following ground: publication would preclude future publication.

date: $8\ February, 2012$ signature of candidate: _____

date: $8\ February, 2012$ signature of supervisor: _____

Abstract

The content of this thesis was motivated by the need to estimate anglerfish abundance from stratified random trawl surveys of the anglerfish stock which occupies the northern European shelf (Fernandes *et al.*, 2007). The survey was conducted annually from 2005 to 2010 in order to obtain age-structured estimates of absolute abundance for this stock. An estimation method is considered to incorporate statistical models for herding, length-based net retention probability and missing age data and uncertainty from all of these sources in variance estimation.

A key component of abundance estimation is the estimation of capture probability. Capture probability is estimated from the experimental survey data using various logistic regression models with haul as a random effect. Conditional on the estimated capture probability, a number of abundance estimators are developed and applied to the anglerfish data. The abundance estimators differ in the way that the haul effect is incorporated. The performance of these estimators is investigated by simulation. An estimator with form similar to that conventionally used to estimate abundance from distance sampling surveys is found to perform best.

The estimators developed for the anglerfish survey data which incorporate random effects in capture probability have wider application than trawl surveys. We examine the analytic properties of these estimators when the capture/detection probability is known. We apply these estimators to three different types of survey data in addition to the anglerfish data, with different forms of random effects and investigate their performance by simulation. We find that a generalization of the form of estimator typically used on line transect surveys performs best overall. It has low bias, and also the lowest bias and mean squared error among all the estimators we considered.

Acknowledgements

A number of people deserve my sincere thanks for making it possible for me to pursue my PhD at the University of St Andrews and present the work in this thesis:

- I owe a huge debt of thanks to my supervisor Dr. David Borchers, who agreed to take me on as a research student at the very beginning and provide me with intellectual guidance and all the support he has given to me, an international student. He has been such a good teacher to me, helped me to understand so many statistical concepts, provoked much thought and shaped the idea of my future career. I also want to thank Dr. Paul Fernandes, who provided me with all the anglerfish survey data and answered all my questions about them. I want to given my special thanks to Prof. Steve Buckland, for all the discussions with him when I was working on the woodrats data and for all the support during the last three months. I gratefully acknowledge the financial support from the Fisheries Research Fund (now Marine Scotland Science).

- The people at CREEM deserve a large part of my gratitude. The people here make CREEM such a nice working environment that is so friendly and welcoming. Individual recognition needs to go to Rhona Rodger who provided all the administrative support, Phil Le Feuvre who provided all technical support that kept the work going. Also I want to thank Dr. Joanne Potts for giving me access to the woodrats data and helping me while I was working on it, Lindesay Scott-Hayward, Glenna Evans and Cornelia Oedekoven for all the talks in the coffee room, Vanessa Cave and Leslie New for offering to proof-read my draft. Special thanks to Leslie New for taking me as a flatmate when I was homeless in St Andrews. I want to thank Danielle Harris for putting up with me as an officemate, for all your encouragement and help when things were not going well both inside and outside the office. Finally, I want to thank my friends in China, especially 常鹤 for being such a good friend and a 'sister' for more than 10 years.

- I want to express my special thanks to Prof. Peter Jupp, whom I first met when I arrived in the School of Mathematics and Statistics in 2006, who told me

"Don't worry" on the first day we met and then for the next five years, who proof-read my thesis and gave me encouragement and support all the time. He has always been there for answering all my questions, either statistical or non-statistical, and sometime even non-sense questions. I have learned so much from him about everything. St Andrews feels like a second home because of you.

- I want to thank Angelika Studeny for being my friend, for listening to me, for sharing both happiness and sorrow in life, and for being such good company during all the time when we were inside the dark tunnel of getting a PhD. Thank you so much for talking me through those difficult months in 2011, for keeping me calm when I was completely lost. Without you I cannot imagine where I would be now. Thank you so much for everything that I can express here and for things that I cannot.

- Finally, I, 袁媛, owe my deepest and lifelong gratitude to my parents 丁香 and 袁振奋, who made me a person driven by curiosity and believing in 有志者事竟成. It was they who sent me to UK and encouraged me to stay to pursue my interest in Statistics. They have always been there to support me unconditionally and encouraged me to keep going. They have never lost their confidence in me no matter what has happened. I also want to thank my uncle, 丁建淮, for his unconditional support and for being there for my parents when I am not there.

Naturally, despite all the valuable assistance of people listed above, this thesis is not perfect and I bear the sole responsibilities for all the mistakes.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Motivating problem and structure of the thesis

Key Idea: introduce the anglerfish surveys and the structure of the thesis

FIGURE 1.1. An image of an anglerfish

1.1 Description of anglerfish species

The motivating application concerns a demersal trawl survey of the anglerfish stock in Scottish waters. Anglerfish (*Lophius piscatorius* and *Lophius budegassa*), is shown in Figure 1.1. There are more than 200 species of anglerfish, many of which are fished commercially throughout the world, notably in north-western Europe, eastern North America, Africa and the Far East. The anglerfish is also known as the goosefish in North America and the monkfish in Europe. Its tail meat is low-fat and has firm texture with a mildly sweet flavour. It is quite often compared to lobster tail in taste and texture. Anglerfish is widely used in cooking and is becoming more and more popular in the market. In some Asian countries, especially in Japan, the anglerfish's tail meat is considered to be a delicacy and often fetches a premium price in the fish markets.

Anglerfish, being deep sea creatures, are considered to be one of the strangest animals on earth. They have an intimidating and ugly appearance. with their giant mouth which can be as long as their own body, they can swallow a prey twice of their own size. More distinctively, female anglerfish have a built-in illuminated fishing pole and they use it to bait prey. In terms of their unusual method of reproduction, the male, which is much smaller than the female, is a permanent parasite mate on the female. A female can carry up to six male parasites (Thomas, 1976; Burton, 1976).

As anglerfish becomes more and more popular in fish markets, the fisheries management plays an increasingly important role to maintain the fishery. For the fisheries management, in order to answer important questions such as the age distribution of the population and the distribution of the stock, abundance estimation is a vital component. It is feasible to obtain anglerfish abundance estimates from trawl surveys because of anglerfish's sedentary behaviour.

1.2 The anglerfish surveys

Anglerfish are considered as one of the most commercially important species to the Scottish fleet fishing (Anon, 2005). The anglerfish stock around the British Isles is of major fishery interest. In 2005, Fisheries Research Services (now Marine Scotland Science) initiated the anglerfish project for the anglerfish stock which occupies the

European Northern shelf (see Fernandes *et al.*, 2007, for details). This new project is unique for two reasons: 1) it aims at providing reliable absolute abundance estimates of anglerfish on the Northern shelf; and 2) the fishing industry is involved throughout this project, from planning to the execution of the surveys, such as allowing observers from industry on the vessels during the survey and participation in discussion about the survey designs and results presentation.

There are three components of the anglerfish project: the annual abundance survey, the experimental survey and the video camera survey. The annual abundance survey is a bottom trawl survey with the haul as the trawling gear illustrated in Figure 1.2. In order to estimate the absolute abundance from the annual abundance survey data, studies of the selectivity (i.e. catchability in Fernandes *et al.* (2007)) of the survey trawl were also essential, and these were carried out in a separate sister project including the experimental and video camera survey, both of which were designed to collect data for estimating the capture probability of the gear in Figure 1.2. Experimental surveys with specially designed gear to collect the fish that escaped under the footrope of the main net were conducted in 2006 and 2007. These surveys provide information required to estimate detection probability. The data comprise fish length l, the haul number i in which the fish was caught, and whether or not the fish was caught in the main net (see Reid *et al.*, 2007b, for details). The data collected by the video camera mounted on the sweeps were used to derive a model for estimating the proportion of anglerfish herded by the trawl doors and sweeps (see Reid *et al.*, 2007a, for full details).

The annual abundance survey used bottom trawls to collect data which are independent of the fish industry. This type of data is considered as one of the most important and effective methods for the stock management of commercial species. The abundance survey has been conducted annually from 2005 to 2010 to obtain age-structured estimates of absolute abundance for this stock. In 2006, 2007 and 2009 this survey was extended south into Irish waters with the participation of the Irish Marine Institute in association with Bord Iascaugh Mhara. The abundance survey from 2005 to to 2007 was conducted in November, while from 2008 to 2010 the survey was conducted in April for more survey time given better weather conditions. This change of survey time is also indicated by the fishermen's advice that the better weather in April allows high catch rates of anglerfish. Almost all fish are aged as well as measured for length in the annual abundance survey, and the age-structured

estimates of absolute abundance can then be obtained given the information of the capture probability of the trawl gear.

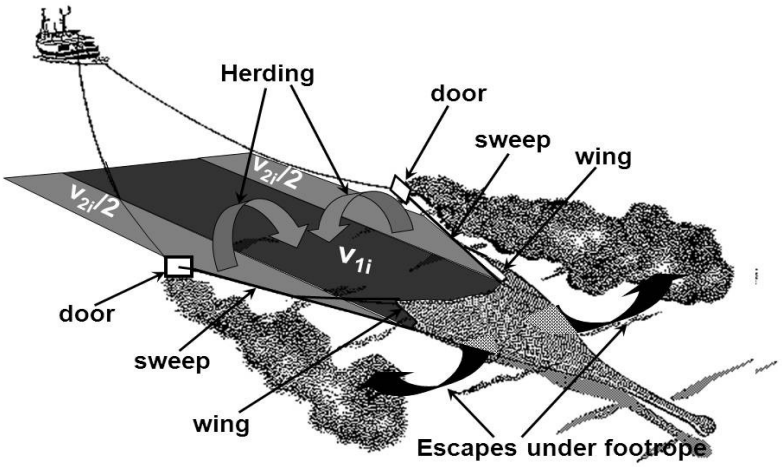


FIGURE 1.2. A schematic diagram of a haul, showing the area swept by the whole gear, i.e. the doors, and the wings. Movement of fish herded by the doors into the path of the net is indicated by the grey arrows and escapement of fish under the footrope is indicated by the dark black arrows. In addition, two swept areas for haul i are defined: v_{1i} is the area swept by the wings for haul i (the black area), and v_{2i} is the area swept by the doors minus that by the wings for haul i (the grey area).

The anglerfish abundance survey was designed with a stratified-random sampling protocol. This protocol partition the survey area into strata and then samples are selected randomly within each stratum. Figure 1.3 plots the survey strata, which encompasses the northern shelf of the British Isles, north of latitude 56° to a northerly limit of $62^\circ 30'$ north, with depth limited to 1000 metres. Given such a large geographic study area in the anglerfish abundance survey, stratification into strata is helpful to ensure the sampling effort is spread out over the whole study area. The principal basis for stratification is depth, with strata set at 0-140 m, 140-200 m, 200-500 m and 500-1000 m. The density in the shallow water 0-140 m and very deep water 500-1000 m is expected to be low, and the density in the water 140-200 m and 200-500 m strata is expected to be high. Four regions were considered as distinct areas to be surveyed: Rockall, west of Scotland, north of Scotland and east of Scotland.

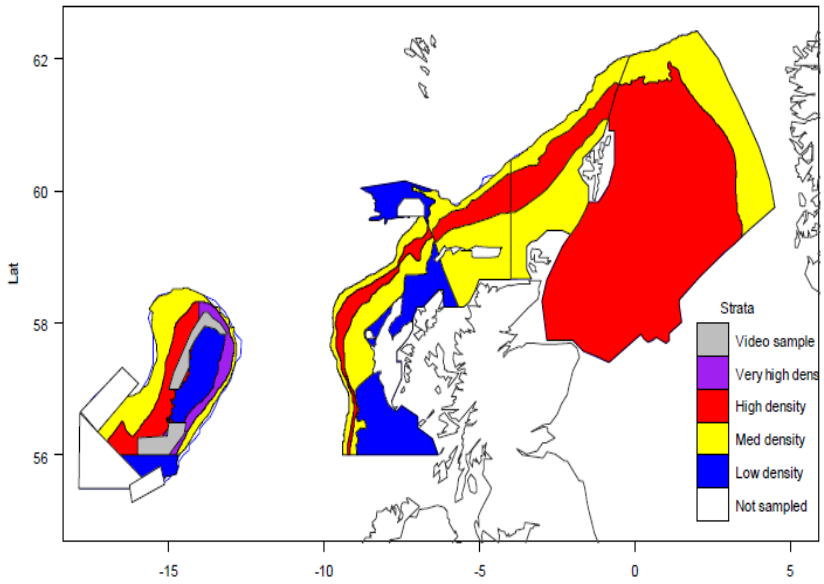


FIGURE 1.3. Map of the northern continental shelf around the British Isles showing the areas surveyed by the anglerfish abundance surveys from 2006 to 2010. Areas are shaded according to the scale given in the legend on the right corner. The colour in this legend indicates the sampling intensity. Those areas that were not surveyed are unshaded and also unlabelled.

It can be seen from Figure 1.3 that the survey area is so large that there were multiple vessels in the annual abundance survey from 2005 to 2010. Therefore, to avoid the problem caused by inconsistent gear specification in a multi-vessel survey, all the vessels in the annual abundance survey used the same sampling tool (i.e., trawl gear) purchased by Marine Scotland Science. The trawl gear was rigged in a consistent manner without modification during the survey. Furthermore, this type of trawl used in the abundance survey was considered as the most effective tool to catch anglerfish by the fishing industry; see Fernandes *et al.* (2007) for further details of the survey methods. In addition, the abundance surveys in 2006, 2007, and 2009 were extended into Irish waters, however, the catch data collected in Irish waters have no age information for each captured fish. The problem caused by missing age data in obtaining abundance estimates by age is addressed in Part III.

For both the abundance survey and the experimental survey, the trawling gear, haul, is illustrated in Figure 1.2. There is some responsive movement of anglerfish from the area within the doors to the area within the wings, and this needs to be taken into account when estimating the capture probability of the gear in Part II. As shown in Figure 1.2, the area swept by the whole gear is $v_{1i} + v_{2i}$. In addition to these swept

areas for each trawl, the primary anglerfish data gathered on the survey and used in estimation are the length and age of each fish. The age classes of anglerfish are the yearly intervals, from age 0 up to age 15.

In this case detection probability depends on fish length (shorter fish tend to evade the net to a greater extent) and haul (modelled as a random effect). The detection probability cannot be estimated reliably from the abundance survey itself, so separate experimental surveys were conducted to estimate it. Both random effects (haul in this case) and an observed explanatory variable (fish length class in this case) affect detection probability.

1.3 Outline of the thesis

The major component of this thesis is the abundance estimation of the Northern Shelf anglerfish stock. Accurate estimation of the capture probability of the gear is essential in estimating absolute abundance. For the gear used in the abundance survey, there are two factors causing the imperfect capture probability: herding defined in Figure 1.2 and incomplete net retention caused by the escapement beneath the footrope. The statistical analysis for the capture probability concentrates on the net retention probability, which is estimated from the experimental survey data. Motivated by the anglerfish abundance estimation, another important component of this thesis is a study of abundance estimation with random effects in different forms.

Part II describes all the components of the capture probability and then presents all the models that are applied to the experimental survey data for the estimation of the net retention probability. This includes both fixed-effects and random-effects models. Then conditional on the chosen model for the net retention probability, Part III develops abundance estimation methods based on the Horvitz-Thompson (Horvitz & Thompson, 1952) estimator for the anglerfish annual abundance survey, and then studies the properties of these estimators by simulation. In the process of applying the Horvitz-Thompson estimator conditional on a mixed-effects model for the net retention probability with haul as the random effect, it becomes apparent that there are different ways to include the random effects in the Horvitz-Thompson-like estimator. Therefore, Part IV starts with developing a Horvitz-Thompson-like estimator

with random effects in various forms, and then describes the further application of the estimator to different types of wildlife survey data.

Part II

Estimation of capture probability

Chapter 2

Capture probability in anglerfish survey

Key Idea: describe the capture probability for the anglerfish survey and the experimental survey which was designed for its estimation

As introduced in Chapter 1, this study aims to estimate the absolute abundance of anglerfish stock of northwest Scotland. If there is knowledge of the whole-gear capture probability, then absolute abundance can be estimated given the catch data. Note that the capture probability is also referred to as the *whole-gear selectivity* or *catchability* in fisheries research. For its estimation, Somerton *et al.* (1999) suggested focusing on the three components of the trawl catching process – vertical herding, horizontal herding and escapement from the net. The presence of vertical herding leads to an increase in the effective fishing height of a trawl. However, for dermersal species such as anglerfish, according to Godø & Totland (1996), vertical herding is unlikely to occur as the anglerfish tend to stay on the seabed and unlikely to dive from the seabed into the trawl path in the trawl catching process. The presence of horizontal herding results in an increase in the effective fishing width of a trawl, and this herding occurs when fish avoid the trawl doors, mud clouds, and bridles by swimming into the path of the trawl (Somerton *et al.*, 1999).

Anglerfish are not entirely sedentary, there is some herding of fish within the area swept by the net doors (the light grey area in Figure 1.2) into the net's path (see Reid *et al.*, 2007a, for details). This is referred to as the (horizontal) herding factor in the anglefish survey, which means that a fraction h of anglerfish between the doors and the wings (refer to Figure 1.2) are herded into the net's path. This herding factor h was estimated to be 0.017 by Allen (2006), and this is the value for $\hat{h}$ used for the abundance estimation in Chapter 4.

The escapement from the net occurs in the anglerfish survey when fish dive under the footrope, and experimental survey was designed to collect data for estimating the proportion of escapement. This survey used auxiliary bags to collect the fish that escaped beneath the footrope; see Figure 2.1. The proportion of this escapement is referred to as the net retention probability, which gives the probability that a fish was retained in the main cod-end of a net given that it contacted the net. The cod-end is the narrow end of a trawl net given in Figure 1.2 – the part of the trawl net where the fish are retained. In the experimental survey, this means that the fish was not collected by the auxiliary bags but retained in the main cod-end. The estimation of this net retention probability is the main component of the capture probability for the anglerfish survey, and it is fully studied in Chapter 3.

This chapter starts with specifying the capture probability of the fishing gear used in the anglerfish abundance surveys (Section 2.1) with corresponding definition of swept areas of the gear. Then Section 2.2 describes the experimental survey data for estimating the net retention probability. Finally, Section 2.3 briefly reviews a general statistical methodology for the estimation of the capture probability as a function of fish size, which motivates the statistical consideration in Sections 3.1 and 3.3 of Chapter 3. In addition, a proposed probability model for the herding factor of the anglerfish survey is presented in Section 2.4. This is not applied here due to the lack of data.

2.1 Capture probability in anglerfish survey

In order to specify the capture probability for the anglerfish survey, two swept areas which represent separate components of the capture probability occurring in the trawl catching process are first defined. The definition of the swept areas for haul i

are given below, and Figure 1.2 provides a more explicit visual description. The two swept areas for haul i are defined as

- v_{1i} is the area swept by wings of a trawl net and referred to as the wing-swept area (depicted by the dark area in Figure 1.2); and
- v_{2i} is the area swept by doors minus that swept by wings of a trawl net, and is referred to as the door-swept area (depicted by the grey area in Figure 1.2).

As anglerfish are so sedentary, it is reasonable to assume that fish located between v_{1i} remain there until the footrope reaches them. They then enter the net with probability r (the "net retention probability") or go under it with probability $(1 - r)$. Fish that are between the doors and the wings ($v_{2i}/2$ on both sides of the trawl net in Figure 1.2) encounter the cable between the doors and the wings before they encounter the footrope. These fish move into the area between the wings with probability h (the herding factor illustrated by grey arrows in Figure 1.2) by the time they come abeam of the footrope.

Therefore, r and h are the two components that cause the imperfect capture probability in anglerfish surveys. In consideration of the random location of the hauls, the probability that a fish that is between the doors is also between the wings when the haul starts is $v_{1i}/(v_{1i}+v_{2i})$. It follows that the probability that a fish initially between the doors, ends up being between the wings when the footrope comes abeam of it is $(v_{1i}+hv_{2i})/(v_{1i}+v_{2i})$. Therefore, the capture probability for a fish initially between the doors of haul i is

$$p_i \quad = \quad r \left(\frac{v_{1i} + hv_{2i}}{v_{1i} + v_{2i}} \right). \tag{2.1}$$

To ensure that there is no other escape from the mesh of the main cod-end in anglerfish surveys, another experiment was conducted in a similar geographic area using the same gear as that used in the experimental survey, except that this time the mesh size of the main cod-end was smaller, with the same size as the bottom additional collecting bags. It is confirmed that the imperfect capture probability has nothing to do with the escape through the net mesh. Hence the main net does capture the bigger fish, i.e. the smaller fish are not available for the main net and the reason for this might be that it is easier for them to escape under the footrope of the main

net. Therefore, it is assumed that only h and r are the sources of imperfect capture probability for the anglerfish survey.

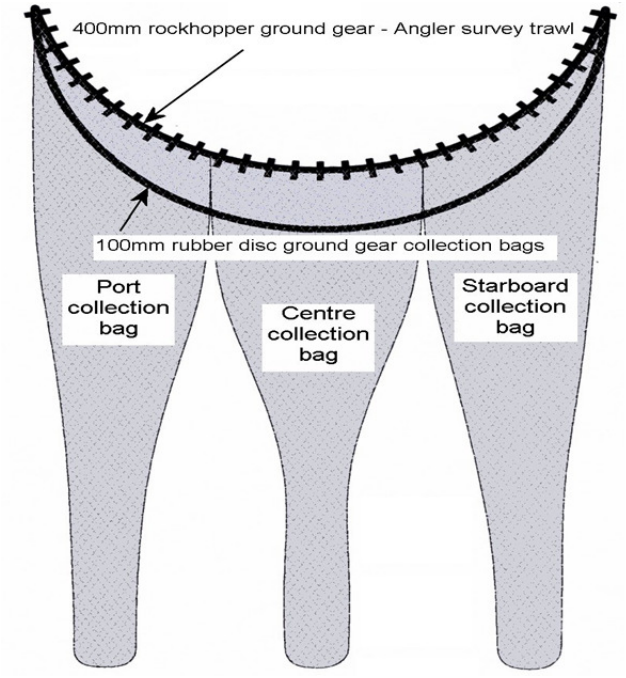


FIGURE 2.1. The collecting bags attached under the main net used in the experimental survey, to collect those fish escaping beneath the footrope of the main net (see Reid *et al.*, 2007b, for more details about the design of the collecting bags).

Given the availability of a fish to haul i, its probability of being exposed to the net is $(v_{1i} + hv_{2i})/(v_{1i} + v_{2i})$. This ratio will be one if all the fish from v_{2i} are herded into the door-swept path, however this is usually not the case. The "herding factor" h is difficult to estimate due to the lack of observations on fish before they respond and when they come abeam of the footrope. In an attempt to estimate the herding factor h, Reid *et al.* (2007a) conducted a survey in which cameras were attached to the doors and wings. Some fish were observed responding. However, the sample size of these camera data was very small (54 sightings from 17 hauls) and it was not possible to follow the fish to see if they were between the wings at the time they came abeam of the footrope.

The herding factor, h, was studied using simulations parameterized based on the available information about anglerfish behaviour, in particular their responsive movement (see Allen, 2006, for details). The empirical distribution of h obtained from this simulation study can be approximated by an exponential distribution with mean

0.017. It is noted that the simulation was based on a few very strong assumptions, and thereafter a new probabilistic model with fewer assumptions is presented in Section 2.4. However, due to a lack of data for fitting the probabilistic model at this stage, no results are presented for this new model.

Therefore, the estimation of capture probability considered here for the anglerfish survey is conditional on the simulated values of h obtained from this study on herding factor. In addition to h, the other component of the sources for imperfect capture probability, the net retention probability, is comprehensively studied using statistical models in Chapter 3. This statistical analysis is built on the data obtained from a sister project, which are referred to as the anglerfish experimental survey data. Full details of these experimental survey data are described in Section 2.2.

2.2 Anglerfish experimental survey data

The experimental survey was carried out in October, 2006 and 2007, and the hauls in the survey were conducted with a pair of nets simultaneously: a main net (see Figure 1.2) and an additional ground net with three collection bags (see Figure 2.1). The ground net is in a position to collect the fish escaping under the footrope of the main net (see Reid *et al.* (2007b) for details), while the main net is the same as the nets used in the abundance survey. The experimental haul locations were chosen to provide clear tows with good expected catches of anglerfish over a range of depths.

In the 2006 and 2007 experimental surveys, the ground collection bags were deployed and recovered undamaged on 36 hauls, and a total of 431 individual fish were captured, out of which a total of 314 fish were captured in the main net and the remaining 117 in the bags below the footrope. Between 2 and 47 fish were captured per haul, with a median of 12. For the combined data from the two years, the minimum length of fish being captured is 9.5 cm, the maximum is 71.5 cm, and the median is 39.6 cm. There were 14 hauls in 2006 with 245 fish captured in total with median length 35.5 cm, and 22 hauls in 2007 with 186 fish captured with median length 45 cm. This histograms of the length of the fish are given in Figure 2.2 for fish that were retained in the cod-end and that escaped under the footrope, respectively.

The information on each individual fish captured in the experiment survey includes

- the length of each fish l,
- the haul number i,
- in which year the fish was captured,
- whether the fish was captured during the day or night, and
- whether the fish was retained in the main net, $y = 1$, or escaped under the footrope, $y = 0$.

The above information shows that the experimental survey data have binary response, y, and this motivates the consideration of logistic regression models in Section 3.1. Given that haul is the sampling unit in the data, the anglerfish survey data are all clustered by hauls, and this clustering structure can be thought of as a spatial aggregation and then modelled as a spatial random effects. The clustering structure further motivates the consideration of the mixed-effects logistic regressions models in Section 3.3. The applications of both fixed-effects and random effects models to anglerfish experimental survey data are presented in Section 3.2 and 3.4, respectively.

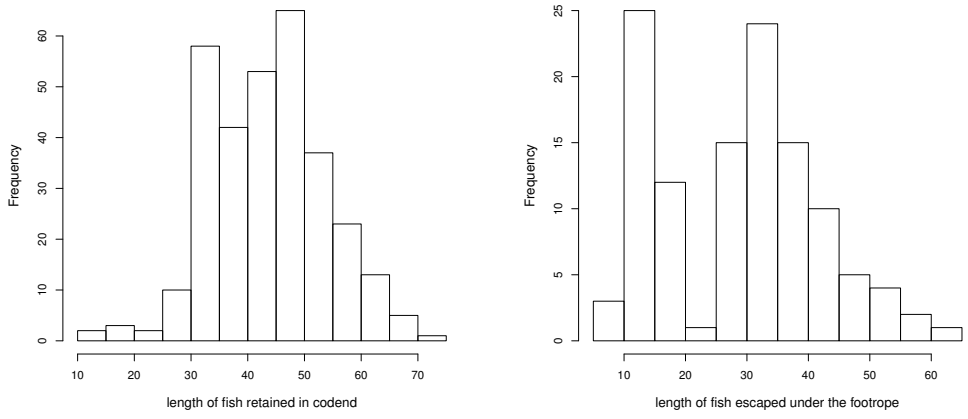


FIGURE 2.2. Histogram of the length of all the anglerfish captured in the 2006-2007 experimental surveys: the left panel gives the histogram of length for the fish that were retained in the cod-end of the trawl net (see Figure 1.2 for a visual description of the net) and the right panel for the fish that escaped under the footrope and then were collected by the auxiliary bags shown in Figure 2.1.

2.3 Literature review on selectivity estimation in fisheries

In the context of trawl surveys, the capture probability of the fishing net is usually referred to as the *gear selectivity* or *catchability*. It has been found in a large number of studies that the gear selectivity is typically a function of fish size, which in most cases is measured by the length of fish (see Munro & Somerton, 2001; Madsen *et al.*, 1999, for example). This section briefly reviews a general statistical methodology for the estimation of size-selection (see Millar & Fryer, 1999, for details).

This section starts with a list of explicit definitions of selection processes, specified with underlying assumptions and limitations. Then a family of logistic regression models is introduced to estimate the gear selectivity as a function of fish length. Note that in this section, only length of fish is considered as a predictor for selectivity. However in practical applications, there might be other predictors depending on the data from the experimental survey.

2.3.1 Definition of selectivity curves

Selectivity curves are used to quantify the probability of a fish being captured given its size. Intuitively, the selectivity curve can be expressed as a function of fish size for the fishing net. The size of a fish is usually measured by its length, because fish length is easy to measure in the field. As a result, the gear selectivity is often modelled as a function of fish length.

The size selectivity can be partitioned into three parts, and each part is defined by a selectivity curve. The three parts of the whole selection process are defined as

- the **population-selection** curve, $s(l)$, is the probability that a fish with length l from the population is captured, which quantifies the differences between the catch and the entire population;

- the **available-selection** curve, $a(l)$, is the probability that a fish with length l is captured given that it was available to the gear, which quantifies the differences between the catch and the fish available to the gear; and

- the **contact-selection** curve, $r(l)$, is the probability that a fish with length l is captured given that it contacted the gear, which quantifies the differences between the catch and the fish coming into contact with the gear.

The above three selection curves differ from each other in terms of the population from which the fish are selected, and the above list of definitions is presented in order so that the population or sub-population which the catch is relative to is decreasing in terms of its range, i.e., the whole population, the sub-population available to the gear and the sub-population contacting the gear. The three selection curves defined above are related by

$$\begin{aligned} a(l) &= r(l) \times \mathbf{P}\{\text{fish of length } l \text{ contacts the gear given} \\ &\qquad\qquad \text{that it is available to the gear}\}, \qquad (2.2) \\ s(l) &= a(l) \times \mathbf{P}\{\text{fish of length } l \text{ is available to the gear}\}. \qquad (2.3) \end{aligned}$$

Note that the probability components $\mathbf{P}\{\}$ on the right-hand sides of (2.2) and (2.3) depend on the fish behaviour, and these probabilities vary for different species (see Millar & Fryer, 1999, p. 92).

The above relations are then illustrated in consideration of the anglerfish survey data. Given the net retention probability defined in Section 2.1, the contact-selection curve for anglerfish survey is the retention probability $r(l)$. Figure 1.2 shows that, in addition to the fish from the haul path between the wings, the fish that contact the gear also include the fish herded by the doors into the path between the wings, i.e., some fish were herded from v_{2i} into v_{1i} in Figure 2.1. Therefore, the available-selection curve $a(l)$ for anglefish survey is $r(l) \times (v_{1i} + hv_{2i})/(v_{1i} + v_{2i})$, where $(v_{1i} + hv_{2i})/(v_{1i} + v_{2i})$ is the probability that a fish of length l contacts the gear given that it is available to the gear. Therefore, the population-selection curve $s(l)$ for anglerfish survey is

$$s(l) = \underbrace{r(l)}_{\text{contact}} \times \underbrace{\frac{v_{1i} + hv_{2i}}{v_{1i} + v_{2i}}}_{\text{available}} \times \underbrace{\frac{v_{1i} + v_{2i}}{A_s}}_{\text{population selectivity}}, \qquad (2.4)$$

where A_s is the surface area of the stratum s.

The population-selection curve for anglerfish survey given in (2.4) is the probability that an anglerfish is captured from the population within stratum s. Note that (2.4) is

also the probability for haul i that a fish located somewhere in stratum s is included in the sample. This probability is the inclusion probability in the Horvitz-Thompson estimator of abundance (Horvitz & Thompson, 1952), which will be used in Part III for the anglerfish abundance estimation. A more general study of the Horvitz-Thompson method for abundance estimation will be presented in Part IV.

2.3.2 Length-based retention curves

As the primary tool used for anglerfish application, a family of logistic regression models have been introduced by Millar & Fryer (1999) for the estimation of the contact-selection curve, $r(l)$. In the context of anglerfish survey data, anglerfish may go beneath the footrope of the towed gear, and the fish finally being captured are those fish that end up with staying in the cod-end of the gear.

In most applications of fisheries research, the contact selectivity is affected by the size and/or the shape of the mesh openings in the cod-end. The larger the fish are, the more easily they are retained in the cod-end. Therefore, it is usually assumed that the contact-selection curve is a monotonically nondecreasing function of fish size, which is usually measured by length of fish. This leads to the usage of logistic curves for a mathematical description of contact-selection curve of a fishing net. The following lists three types of logistic curves for a fish of length l:

- linear logistic is expressed as

$$r(l) = \frac{\exp(\beta_0 + \beta_1 l)}{1 + \exp(\beta_0 + \beta_1 l)}, \tag{2.5}$$

  equivalently

$$\text{logit}(r(l)) = \beta_0 + \beta_1 l, \tag{2.6}$$

  which is symmetric about the median of l with an upper asymptote of unity;

- asymptote-logistic is expressed as

$$r(l) = \gamma \left(\frac{\exp(\beta_0 + \beta_1 l)}{1 + \exp(\beta_0 + \beta_1 l)} \right), \tag{2.7}$$

which can be viewed as an extended form of linear logistic curve given by (2.5), with an extra parameter γ allowing the asymptote to be less than 1; and

- asymmetric logistic is expressed as

$$r(l) \quad = \quad \left(\frac{\exp(\beta_0 + \beta_1 \, l)}{1 + \exp(\beta_0 + \beta_1 \, l)} \right)^{\frac{1}{\kappa}}, \tag{2.8}$$

which is another extended form of the above linear logistic curve with κ modelling the asymmetry of the curve about the median of length l.

The three logistic curves, (2.5), (2.7) and (2.8), are solutions to simpler cases of the Richards curve given by (2.13) in Appendix 2.A. The Richards curve is also known as the generalized logistic curve, and is one of the most flexible functions to model the growth rate for population dynamics. An alternative way to construct flexible logistic regression models is to use a wide parametric class of link functions introduced by Aranda-Ordaz (1981).

When using logistic curves to model contact selectivity as a function of fish size, the situation is usually not as complicated as when modelling the growth rate using (2.13). Therefore, simpler cases of the Richards curve are considered in Appendix 2.A to obtain (2.5), (2.7) and (2.8), from the simplest case to more complicated ones. In this way, we do model assessment step by step to avoid fitting too complex a model to the data we have.

2.4 New probability model of herding factor

This section introduces a probability model of herding factor in the case when a decent sample of movement data is available and these data consist of observations of the number of consecutive moves anglerfish will make when prodded, together with the distance and direction that they move on each occasion. The model is constructed for such data because gathering such data is considered feasible - whereas gathering movement data from direct observations in front of a towed net s not. In the area between the edge of the net and the door, it is assumed that fish are uniformly distributed from the net edge up to distance W, where W is the distance perpendicular to direction of towing from the net edge to the door.

Let d denote the initial distance of a fish. Based on the above uniform assumption, it follows that the pdf of d is

$$g(d) = \frac{1}{W}. \tag{2.9}$$

It is then assumed that a fish moves closer to the net with probability p_{mov} when it encounters the line between the door and net. When it moves closer to the net, it moves a distance x closer, which is assumed to follow a probability distribution with density function $f(x)$. Let $F_k(s_k)$ denote the cumulative distribution function of the sum

$$s_k = \sum_{i=1}^{k} x_i$$

where x_i are i.i.d. random variables from $f(x)$ and k denote how many times that a fish moves. Then the probability that a fish initially located at distance d gets into the net in exactly k moves can be expressed as

$$\begin{aligned} p_k(d) &= \mathbf{P}\{\text{makes } k \text{ moves}\} \\ &\quad \times \mathbf{P}\{\text{total distance moved by } (k-1)\text{th move} < d\} \\ &\quad \times \mathbf{P}\{\text{total distance moved by } k\text{th move} \geq d\} \\ &= p_{mov}^k F_{k-1}(d)[1 - F_k(d)]. \end{aligned} \tag{2.10}$$

Anglerfish have very little stamina and there is a limit K on the number of consecutive moves they will make. The probability that a fish gets into the net if it can move no more than K times is

$$\begin{aligned} P_K(d) &= 1 - \prod_{k=1}^{K} \mathbf{P}\{\text{failed to get into the net in exactly } k \text{ moves}\} \\ &= 1 - \prod_{k=1}^{K} [1 - p_k(d)]. \end{aligned} \tag{2.11}$$

Let $h(W)$ denote the probability that a fish originally located within the distance W from the net edge finally gets into the net. Then $h(W)$ can be evaluated by taking expectation of $P_K(d)$ with respect to d, whose distribution is given by (2.9), i.e.,

$$h(W) = \int_0^W P_K(x) \frac{1}{W} dx. \tag{2.12}$$

Appendices – **Chapter 1**

2.A Derivation of the selectivity curves in Section 2.3.2

The Richards curve, also known as the generalized logistic curve, is a flexible sigmoid function first introduced by Richards (1959) in order to model growth rate. The Richards curve allows variable lower and upper asymptote, as well as asymmetry about the inflexion point of the curve. To start with, let $y(t)$ denote the population size at time t. The most general formulation of $y(t)$ is

$$y(t) = A + \frac{C}{\{1 + T\exp[-B(t-M)]\}^{\frac{1}{T}}}, \tag{2.13}$$

of which the unknowns are

- A, the lower asymptote;
- C, the upper asymptote;
- M, the time at which maximum growth occurs;
- B, the growth rate; and
- T, the time near which asymptote maximum growth occurs.

The rest of this appendix then derives simple cases of (2.13) for the logistic curves given by (2.5), (2.7) and (2.8) introduced in Section 2.3.1.

2.A.1 Logistic curve

The logistic curve models the population growth or decay as a function of time by the following differential equation,

$$\frac{dy(t)}{dt} = Cy(1-y), \tag{2.14}$$

which can be solved step by step as

$$\begin{aligned}
\frac{dy}{dt} &= Cy(1-y) \\
\Leftrightarrow \int \left[\frac{1}{y(1-y)}\right] dy &= \int C dt \\
\Leftrightarrow \int \left[\frac{1}{y} + \frac{1}{1-y}\right] dy &= \int C dt \\
\Leftrightarrow \log|y| - \log|1-y| &= Ct + C_1 \\
\Leftrightarrow \log|\frac{y}{1-y}| &= Ct + C_1 \\
\Leftrightarrow \frac{y}{1-y} &= \exp(Ct + C_1) \\
\Leftrightarrow y &= \frac{\exp(Ct + C_1)}{1 + \exp(Ct + C_1)}
\end{aligned} \tag{2.15}$$

where C_1 is another constant. Therefore, in the context of the application to anglerfish, for a fish with length l, $y(t)$ is analogous to the net retention probability, $r(l)$, as a function of length l. Given the solution in (2.15), it follows that the linear logistic curve is formulated as (2.5) in Section 2.3.2.

2.A.2 Asymptote-logistic curve

The logistic curve allowing a variable asymptote given in (2.7), it is the solution to the differential equation

$$\frac{dy(t)}{dt} = Cy\left(1 - \frac{y}{\gamma}\right) \tag{2.16}$$

for growth rate. Compared with (2.14), it has an extra parameter γ to model the asymmetry.

Let $y^* = y/\gamma$. Then (2.16) can be expressed as

$$\frac{dy^*(t)}{dt} = C\, y^* \left(1 - y^*\right), \tag{2.17}$$

which is the same as (2.14). Based on the solution of (2.14) which has already been given in (2.15), the solution of (2.16) is

$$y^* = \frac{y}{\gamma} = \frac{\exp(Ct + C_1)}{1 + \exp(Ct + C_1)}.$$

Then the solution to (2.16) is

$$y = \gamma \left[\frac{\exp(Ct + C_1)}{1 + \exp(Ct + C_1)} \right]. \tag{2.18}$$

Therefore, for a fish with length l, the contact-selection curve $r(l)$ allowing a variable asymptote is the solution of the differential equation

$$\frac{d\, r(l)}{d\, l} \;=\; C\, r \left(1 - \frac{r}{\gamma} \right), \tag{2.19}$$

and given that (2.18) is the solution to (2.16), the solution to (2.19) is

$$r(l) \;=\; \gamma \left[\frac{\exp(\beta_0 + \beta_1 l)}{1 + \exp(\beta_0 + \beta_1 l)} \right],$$

which gives (2.7) in Section 2.3.2.

2.A.3 Asymmetric logistic curve

The asymmetric logistic curve given in (2.8) is the solution to the differential equation

$$\frac{dy(t)}{dt} \;=\; Cy(1 - y^{\kappa}), \tag{2.20}$$

which is based on (2.14), with an extra parameter κ to model asymmetry.

Equation (2.20) can be solved step by step as,

$$\begin{aligned} \frac{dy}{dt} &= Cy(1 - y^{\kappa}) \\ \Leftrightarrow \frac{dy}{ydt} &= C(1 - y^{\kappa}) \\ \Leftrightarrow \frac{d(\log\, y)}{dt} &= C(1 - y^{\kappa}). \end{aligned} \tag{2.21}$$

Let $y^\kappa = z$, then $\kappa \log y = \log z$. Therefore, (2.21) is equivalent to

$$\begin{aligned}
\frac{d\left(\frac{1}{\kappa}\log z\right)}{dt} &= C(1-z) \\
\Leftrightarrow \frac{1}{\kappa}\frac{1}{z}\frac{dz}{dt} &= C(1-z) \\
\Leftrightarrow \frac{dz}{dt} &= C\kappa z(1-z) \\
\Leftrightarrow \frac{dz}{z(1-z)} &= C\kappa\, dt \\
\Leftrightarrow \int \frac{dz}{z(1-z)} &= \int C\kappa\, dt \\
\Leftrightarrow \log\left|\frac{z}{1-z}\right| &= C\kappa t + C_1 \\
\Leftrightarrow z &= \frac{\exp(C\kappa t + C_1)}{1+\exp(C\kappa t + C_1)} \\
\Leftrightarrow y^\kappa &= \frac{\exp(C\kappa t + C_1)}{1+\exp(C\kappa t + C_1)}, \qquad (2.22)
\end{aligned}$$

where C_1 is just another constant.

Therefore, for a fish with length l, the contact-selection curve $r(l)$ allowing asymmetry is the solution of the differential equation

$$\frac{d\, r(l)}{d\, l} = Cr(1-r^\kappa), \qquad (2.23)$$

and given that (2.22) is the solution to (2.20), the solution of (2.23) is

$$[r(l)]^\kappa = \frac{\exp(\beta_0 + \beta_1 l)}{1+\exp(\beta_0 + \beta_1 l)},$$

which givens the (2.8) in Section 2.3.2.

Chapter 3

Logistic regression models with application to anglerfish

Key Idea: apply logistic regression models to estimate the net retention probability using the 2006-2007 experimental survey data

The probability that animals are detected or captured in a wildlife survey (referred to as *capture probability* here) is a central component in abundance estimation, and sometimes it cannot be estimated from the abundance survey data set itself. In this case, usually a separate survey is designed to collect data for the estimation of capture probability, and this survey is referred to as an experimental survey to be distinguished from abundance surveys, which are designed for abundance estimation.

In the context of the anglerfish application, the estimation of capture probability is focused on the net retention using data from the the anglerfish experimental survey, which has been described in Section 2.2. The net retention probability arises from the fact that anglerfish, especially small ones, can escape beneath the footrope of the net, and the experimental survey records whether or not each individual catch was retained in the main cod-end. These binary response data collected from anglerfish experimental surveys motivate the consideration of logistic regression and its extended forms in this chapter, including both fixed-effects and mixed-effects mod-

els, all of which have been applied to estimate net retention probabilities using the anglerfish experimental survey data.

This chapter starts with the fixed-effects logistic regression model and its extended forms in Section 3.1, and continues with their application to the estimation of net retention probabilities in Section 3.2. Then the mixed-effects logistic regression model and its application to the anglerfish experimental survey data are presented in Sections 3.3 and 3.4, respectively.

3.1 Fixed-effects logistic regression and its extended forms

The data collected from experimental surveys mostly have a binary response denoting whether or not an animal is captured in the experiment. In some cases, the data collected are binomial. However, they can be converted into individual binary data for statistical modelling. The class of models used to fit binary response data is logistic regression models. This section is focused on the fixed-effects logistic regression models whose formulation is given in Section 3.1.1. Then Section 3.1.2 describes parameter estimation for linear logistic regression models to obtain maximum likelihood estimates, continues with a discussion in Section 3.1.3 on the effect of centring regression predictors, and a demonstration in Section 3.1.4 on how to calculate the asymptotic variance-covariance matrix of parameter estimates in the presence of parameter transformation in maximizing the likelihood. Finally, Section 3.1.5 introduces the Hosmer-Lemeshow goodness-of-fit test to assess the logistic regression models in the presence of a continuous predictor. The above consideration of logistic regression models is an essential component in fitting a logistic regression model, and it is of particular interest for the anglerfish application; a more comprehensive study on logistic modelling can be found in Hosmer & Lemeshow (2000).

3.1.1 Model formulation

Let y_i denote the binary response variable for the ith observation in the experimental survey data, and p_i denote its success probability. Therefore,

$$y_i \sim \text{Bernoulli}(p_i)$$

where p_i is the probability that the ith observation is a 'success', i.e. being detected or captured in the survey, and $\text{Var}[y_i] = p_i(1-p_i)$. Furthermore, p_i depends on explanatory variables $\boldsymbol{x}_i$, i.e., $p_i = g(\boldsymbol{x}_i; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter vector. The formulations of linear logistic regression and its extended forms are are listed below:

(a) The linear logistic regression model is a generalized linear model with logit link function, i.e.,

$$\text{logit}(p_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}, \tag{3.1}$$

equivalently,

$$p_i = \frac{1}{1+\exp(-\boldsymbol{x}_i^T \boldsymbol{\beta})}, \tag{3.2}$$

and the parameter vector $\boldsymbol{\theta}$ for a linear logistic regression model is just the regression coefficients $\boldsymbol{\beta}$.

(b) The asymptote-logistic model is an extended form of (3.1) with the additional parameter $\gamma \in (0, 1]$ to allow a variable upper-asymptote, which is formulated as:

$$\text{logit}\left(\frac{p_i}{\gamma}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta}, \tag{3.3}$$

equivalently,

$$p_i = \frac{\gamma}{1+\exp(-\boldsymbol{x}_i^T \boldsymbol{\beta})}; \tag{3.4}$$

and the parameter vector $\boldsymbol{\theta}$ for an asymptote-logistic regression is the regression coefficients $\boldsymbol{\beta}$ together with the asymptote parameter γ, i.e., $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \gamma)^T$.

(c) The asymmetric logistic model is an extended form of (3.1) with the additional parameter $\kappa \in (0, +\infty)$ to allow the asymmetry of the logistic curve, which is formulated as:

$$\text{logit}\left(p_i^{\kappa}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta}, \tag{3.5}$$

equivalently,

$$p_i = \left(\frac{1}{1+\exp(-\boldsymbol{x}_i^T\boldsymbol{\beta})}\right)^{\frac{1}{\kappa}}; \tag{3.6}$$

and the parameter vector $\boldsymbol{\theta}$ for an asymmetric logistic regression is the regression coefficients $\boldsymbol{\beta}$ together with the parameter for asymmetry, i.e., $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \kappa)^T$.

Note that the models given by (3.4) and (3.6) cannot be generalized to linear models, and so the algorithm presented in the forthcoming section for linear logistic regression cannot be applied to fit the extended logistic regressions of the forms (3.4) and (3.6).

3.1.2 Parameter estimation

Given the parameter vector $\boldsymbol{\theta}$, $p_i = g(\boldsymbol{x}_i;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ depends on the choice of models formulated in Section 3.1.1. In more detail, for linear logistic regression, $\boldsymbol{\theta} = \boldsymbol{\beta}$; for asymptote-logistic regression, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \gamma)^T$; and for asymmetric logistic regression, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \kappa)^T$.

Given the choice of model for p_i and assuming independent trials, the likelihood of $\boldsymbol{\theta}$ can be expressed as

$$\begin{aligned} L(\boldsymbol{\theta};\, \boldsymbol{y}, \boldsymbol{x}) &= \prod_{i=1}^{n} p_i^{y_i}[1-p_i]^{1-y_i} \\ &= \prod_{i=1}^{n} g(\boldsymbol{x}_i;\boldsymbol{\theta})^{y_i}\,[1-g(\boldsymbol{x}_i;\boldsymbol{\theta})]^{1-y_i}\,. \end{aligned} \tag{3.7}$$

The log-likelihood is

$$l(\boldsymbol{\theta};\, \boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{n}[\, y_i \,\mathrm{logit}(p_i) + \log(1-p_i)\,], \tag{3.8}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ is the observation vector. Among these three models, the linear logistic regression (3.2) is one of the most widely used statistical models, and it can be easily fitted by functions provided by statistical software tools, such as

`glm()` in `R` which is a generic function to fit generalized linear models. The function `glm()` implements the iteratively re-weighted least squares algorithm to obtain the maximum likelihood estimates, which is referred to as the IRLS algorithm. For the linear regression model, it can be easily shown that the least squares estimates are also the maximum likelihood estimates (see Appendix 3.B.1 for details). However, it is not apparent why the IRLS algorithm produces maximum likelihood estimates, and the following gives an explanation for this question in the setting of a linear logistic regression.

To start with, re-write (3.2) in more detail as

$$\eta_i \;=\; \text{logit}(p_i) = \beta_0 + x_{i1}\beta_1 + \ldots + x_{i\,q-1}\beta_{q-1}, \tag{3.9}$$

and equivalently,

$$p_i \;=\; \text{logit}^{-1}(\eta_i) = \frac{1}{1+\exp(-\eta_i)}, \tag{3.10}$$

where $\boldsymbol{x}_i$ is the q-dimensional vector of predictors for the ith observation, i.e. $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{i\,q-1})^T$ with corresponding parameter $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{q-1})^T$. Then η_i can be expressed as $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$.

In order to derive the first-order partial derivative of the log-likelihood, the following derivatives of p_i based on (3.9) and (3.10) are obtained

$$\frac{dp_i}{d\eta_i} \;=\; \frac{\exp(-\eta_i)}{(1+\exp(-\eta_i))^2} = (1-p_i)\,p_i, \tag{3.11}$$

$$\frac{\partial p_i}{\partial \beta_r} \;=\; \frac{dp_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_r} = \frac{dp_i}{d\eta_i}\,x_{ir}, \quad r \in \{0, 1, \ldots, q-1\}. \tag{3.12}$$

Then the first-order partial derivative of the log-likelihood (3.8) with respect to p_i is:

$$\frac{\partial l}{\partial p_i} \;=\; \frac{y_i - p_i}{p_i(1-p_i)},$$

which is used in the chain rule together with (3.11) and (3.12) for the first-order partial derivative of the log-likelihood (3.8):

$$\begin{aligned}\frac{\partial l}{\partial \beta_r} &= \sum_{i=1}^{n} \frac{\partial l}{\partial p_i}\frac{\partial p_i}{\partial \beta_r} = \sum_{i=1}^{n} \frac{(y_i - p_i)}{p_i(1-p_i)}\frac{\partial p_i}{\partial \beta_r} \\ &= \sum_{i=1}^{n} \frac{(y_i - p_i)}{p_i(1-p_i)}\frac{dp_i}{d\eta_i}\, x_{ir}. \end{aligned} \tag{3.13}$$

Let $\boldsymbol{W}$ denote the diagonal matrix with the ith diagonal element being

$$\boldsymbol{W}_{ii} = p_i(1-p_i) = \text{Var}[y_i]. \tag{3.14}$$

Based on the first-order partial derivative given in (3.13), the second-order partial derivative of log-likelihood is obtained as

$$\begin{aligned}\frac{\partial^2 l}{\partial \beta_r\, \partial \beta_s} &= \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial p_i}\left[\frac{y_i - p_i}{p_i(1-p_i)}\right]\frac{\partial p_i}{\partial \beta_s}\frac{dp_i}{d\eta_i}x_{ir} \right. \\ &\quad \left. + \left[\frac{y_i - p_i}{p_i(1-p_i)}\right]\frac{d}{d\eta_i}\left(\frac{dp_i}{d\eta_i}\right)\frac{\partial \eta_i}{\partial \beta_s}x_{ir} \right\} \\ &= \sum_{i=1}^{n} \left\{ \left(\frac{-1}{p_i(1-p_i)} + \frac{(y_i - p_i)(2p_i - 1)}{p_i^2(1-p_i)^2}\right)\left(\frac{dp_i}{d\eta_i}\right)^2 x_{is}x_{ir} \right. \\ &\quad \left. + \frac{(y_i - p_i)}{p_i(1-p_i)}\frac{d^2 p_i}{d\eta_i^2}\frac{\partial \eta_i}{\partial \beta_s}x_{ir} \right\}, \end{aligned} \tag{3.15}$$

where $r,\ s \in \{0, 1, \ldots, q-1\}$.

Note that $\mathbf{E}[y_i] = p_i$. Then the components including $(y_i - p_i)$ in (3.15) disappear after taking expectation. The Fisher information matrix is the expectation of (3.15) and its $(r,\ s)$th (rth row, sth column) element is defined as

$$\begin{aligned}[\boldsymbol{I}(\boldsymbol{\beta})]_{rs} = -\mathbf{E}\left[\frac{\partial^2 l}{\partial \beta_r\, \partial \beta_s}\right] &= \sum_{i=1}^{n} \frac{1}{p_i(1-p_i)}\left(\frac{d\, p_i}{d\, \eta_i}\right)^2 x_{is}x_{ir} \\ &= \sum_{i=1}^{n} x_{ir}\frac{(d\, p_i/d\, \eta_i)^2}{p_i(1-p_i)}\, x_{is} \\ &= \sum_{i=1}^{n} x_{ir}p_i(1-p_i)\, x_{is}, \end{aligned} \tag{3.16}$$

which can be written in matrices as $\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} \in \mathbb{R}^{q\times q}$.

Then the maximum likelihood estimates, $\hat{\boldsymbol{\beta}}$, can be obtained by solving the first-order partial derivatives of the log-likelihood with respect to $\boldsymbol{\beta}$, which is called 'the score' and denoted as $\boldsymbol{S}(\boldsymbol{\beta})$, i.e.,

$$\boldsymbol{S}(\boldsymbol{\beta}) = (S_0(\boldsymbol{\beta}), S_1(\boldsymbol{\beta}), \ldots, S_{q-1}(\boldsymbol{\beta}))^T,$$

with the jth element being

$$S_j(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{y}, \boldsymbol{\beta})}{\partial \beta_j}, \quad j = 0, 1, \ldots, q-1.$$

The Hessian matrix is defined as

$$\boldsymbol{H}(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial S_0}{\partial \beta_0} & \cdots & \frac{\partial S_0}{\partial \beta_{q-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial S_{q-1}}{\partial \beta_0} & \cdots & \frac{\partial S_{q-1}}{\partial \beta_{q-1}} \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 l}{\partial \beta_0 \partial \beta_0} & \cdots & \frac{\partial^2 l}{\partial \beta_0 \partial \beta_{q-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial \beta_{q-1} \partial \beta_0} & \cdots & \frac{\partial^2 l}{\partial \beta_{q-1} \partial \beta_{q-1}} \end{pmatrix}. \tag{3.17}$$

The $(r,\ s)$th element of the Hessian matrix is already given in (3.15), together with (3.16). It then follows that the Fisher information matrix can also be expressed as $\boldsymbol{I}[\boldsymbol{\beta}] = -\mathbf{E}[\boldsymbol{H}(\boldsymbol{\beta})]$.

Finally, given the Fisher information matrix (3.16), the Fisher scoring method is applied to obtain $\hat{\boldsymbol{\beta}}$ by iteratively solving the score equation $S_j(\boldsymbol{\beta}) = 0$, where $j = 0, 1, \ldots, q-1$. The following illustrates the connection between Fisher scoring method and the IRLS algorithm.

To start with, the Newton-Raphson algorithm, expressed as (3.73) in Appendix 3.A, is applied to solve the score equation. Given the estimates at iteration t, $\hat{\boldsymbol{\beta}}^{(t)}$, the updated solution at iteration $t+1$ is

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - \left[\boldsymbol{H}\left(\hat{\boldsymbol{\beta}}^{(t)}\right)\right]^{-1} \boldsymbol{S}\left(\hat{\boldsymbol{\beta}}^{(t)}\right). \tag{3.18}$$

To stabilise the algorithm, the Fisher scoring method replaces the Hessian matrix evaluated at $\hat{\boldsymbol{\beta}}^{(t)}$ in (3.18) by its expectation, i.e. replacing $\boldsymbol{H}(\boldsymbol{\beta}^{(t)})$ by $-\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}^{(t)}\right)$.

The algorithm becomes

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(t+1)} &= \hat{\boldsymbol{\beta}}^{(t)} - \left[-\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}^{(t)}\right)\right]^{-1} \boldsymbol{S}\left(\hat{\boldsymbol{\beta}}^{(t)}\right) \\ &= \hat{\boldsymbol{\beta}}^{(t)} + \left[\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}^{(t)}\right)\right]^{-1} \boldsymbol{S}\left(\hat{\boldsymbol{\beta}}^{(t)}\right), \end{aligned} \tag{3.19}$$

equivalently,

$$\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}^{(t)}\right) \hat{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{I}\left(\hat{\boldsymbol{\beta}}^{(t)}\right) \hat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{S}\left(\hat{\boldsymbol{\beta}}^{(t)}\right) \quad \in \mathbb{R}^q. \tag{3.20}$$

Then it can be shown that the estimates obtained by Fisher scoring method given in (3.19) can be thought of as iteratively re-weighted least squares (IRLS) estimates. To start with, working on the rth element of the q-dimensional vector on the left-hand side of (3.20), we have

$$\begin{aligned}\left[\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}^{(t)}\right) \hat{\boldsymbol{\beta}}^{(t+1)}\right]_r &= \sum_{j=0}^{q-1} \left[\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}^{(t)}\right)\right]_{rj} \hat{\beta}_j^{(t+1)} \\ &= \sum_{j=0}^{q-1} \left[\sum_{i=1}^{n} \boldsymbol{W}_{ii}^{(t)} x_{ir} x_{ij}\right] \hat{\beta}_j^{(t+1)} \\ &= \sum_{i=1}^{n} \left(\boldsymbol{W}_{ii}^{(t)} x_{ir} \left[\sum_{j=0}^{q-1} x_{ij} \hat{\beta}_j^{(t+1)}\right]\right) \\ &= \sum_{i=1}^{n} \boldsymbol{W}_{ii}^{(t)} x_{ir} \eta_i^{(t+1)}, \end{aligned} \tag{3.21}$$

while on the other side of (3.20)

$$\begin{aligned}&\left[\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}^{(t)}\right) \hat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{S}\left(\hat{\boldsymbol{\beta}}^{(t)}\right)\right]_r \\ =\ & \sum_{j=0}^{q-1} \left[\boldsymbol{I}\left(\hat{\boldsymbol{\beta}}^{(t)}\right)\right]_{rj} \hat{\beta}_j^{(t)} + S_r\left(\hat{\boldsymbol{\beta}}^{(t)}\right) \\ =\ & \sum_{j=1}^{q-1} \left[\sum_{i=1}^{n} \boldsymbol{W}_{ii}^{(t)} x_{ir} x_{ij}\right] \hat{\beta}_j^{(t)} + \sum_{i=1}^{n} \boldsymbol{W}_{ii}^{(t)} \left(y_i - p_i^{(t)}\right) \left.\frac{\partial \eta_i}{\partial p_i}\right|_{\hat{\boldsymbol{\beta}}^{(t)}} x_{ir} \\ =\ & \sum_{i=1}^{q-1} \boldsymbol{W}_{ii}^{(t)} x_{ir} \sum_{j=1}^{q-1} \left(x_{ij} \hat{\beta}_j^{(t)}\right) + \sum_{i=1}^{n} \boldsymbol{W}_{ii}^{(t)} \left(y_i - p_i^{(t)}\right) \left.\frac{\partial \eta_i}{\partial p_i}\right|_{\hat{\boldsymbol{\beta}}^{(t)}} x_{ir} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{q-1} \boldsymbol{W}_{ii}^{(t)} x_{ir} \eta_i^{(t)} + \sum_{i=1}^{n} \boldsymbol{W}_{ii}^{(t)} x_{ir} \left(y_i - p_i^{(t)} \right) \left. \frac{\partial \eta_i}{\partial p_i} \right|_{\hat{\boldsymbol{\beta}}^{(t)}} \\
&= \sum_{i=1}^{n} \boldsymbol{W}_{ii}^{(t)} x_{ir} \underbrace{\left[\eta_i^{(t)} + \left(y_i - p_i^{(t)} \right) \left. \frac{\partial \eta_i}{\partial p_i} \right|_{\hat{\boldsymbol{\beta}}^{(t)}} \right]}_{z_i^{(t)}},
\end{aligned} \tag{3.22}
$$

where $z_i^{(t)}$ is considered as a adjusted response variable and is the working response in the estimation process.

Equating (3.21) and (3.22), it follows that

$$
\sum_{i=1}^{n} \boldsymbol{W}_{ii}^{(t)} x_{ir} \eta_i^{(t+1)} = \sum_{i=1}^{n} \boldsymbol{W}_{ii}^{(t)} x_{ir} z_i^{(t)},
$$

which can be written in matrix form as

$$
\boldsymbol{X}^T \boldsymbol{W}^{(t)} \boldsymbol{X} \hat{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{X}^T \boldsymbol{W}^{(t)} \boldsymbol{Z}^{(t)}, \tag{3.23}
$$

since $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$.

Therefore,

$$
\hat{\boldsymbol{\beta}}^{(t+1)} = \left(\boldsymbol{X}^T \boldsymbol{W}^{(t)} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{W}^{(t)} \boldsymbol{Z}^{(t)}, \tag{3.24}
$$

and

$$
\boldsymbol{Z}^{(t)} = \boldsymbol{X}^T \hat{\boldsymbol{\beta}}^{(t)} + \left(\boldsymbol{y} - \boldsymbol{p}^{(t)} \right) \left. \frac{d\boldsymbol{\eta}}{d\boldsymbol{p}} \right|_{\hat{\boldsymbol{\beta}}^{(t)}}, \tag{3.25}
$$

which is a linear approximation to $\boldsymbol{y}$.

Finally, given the estimator (3.24) by using the Fisher scoring method, the value to which $\hat{\boldsymbol{\beta}}^{(t)}$ converges is the maximum likelihood estimate of $\boldsymbol{\beta}$, from which it can be seen that (3.24) is analogous to the weighted least squares estimator given by (3.80) in Appendix 3.B, but in an iterative way with working response $\boldsymbol{Z}^{(t)}$ and weights $\boldsymbol{W}^{(t)}$, both of which depend on the estimates $\hat{\boldsymbol{\beta}}^{(t)}$. This dependence leads to a further complication in the algorithm in that $\boldsymbol{W}^{(t)}$ and $\boldsymbol{Z}^{(t)}$ are updated in each iteration to

give the current estimate of $\boldsymbol{\beta}$. This process is then called the iteratively re-weighted least squares algorithm (IRLS) with the following steps for iteration $t+1$,

1. start with the estimates $\hat{\boldsymbol{\beta}}^{(t)}$;
2. update the working responses $\boldsymbol{Z}^{(t)}$ according to (3.25) with $\boldsymbol{\eta}^{(t)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(t)}$ and then $\hat{\boldsymbol{p}}^{(t)} = g^{-1}(\boldsymbol{\eta}^{(t)})$;
3. update the weight matrix $\boldsymbol{W}^{(t)}$;
4. calculate $\hat{\boldsymbol{\beta}}^{(t+1)}$ by the weighted least squares estimator (3.24); and
5. repeat the above steps with $\hat{\boldsymbol{\beta}}^{(t+1)}$ in step 1 until the series of $\hat{\boldsymbol{\beta}}$ converges.

The $\hat{\boldsymbol{\beta}}$, the value to which $\hat{\boldsymbol{\beta}}^{(t)}$ obtained by this iterative process converges, is the MLE of a linear logistic regression model, and this is how `glm` fits generalized linear models. The function `glm` is used to fit the linear logistic regression model (3.1) for the anglerfish experimental survey data, and the estimation results are presented in Section 3.2. However, as the extended logistic regression models (3.3) and (3.5) cannot be transformed into generalized linear models, `optim()` is used to maximize their log-likelihood functions defined in (3.8) for parameter estimation.

3.1.3 Centring the predictor

In practice, the predictor variables are centred by their sample means and this is referred to as the mean centring approach. It can be shown that the mean centring approach does not change the overall fit of the regression lines, but only the intercept estimate and its interpretation (see Figure 3.11 in Appendix 3.C for an illustration). In a linear logistic regression model, without centring the predictor, the meaning of the intercept estimate is the logarithm of the predicted odds ratio at the zero point of the predictor. However, in many applications, such as the anglerfish survey data, the predictor has no meaningful zero point, i.e., there is no fish of length zero. After centring the data, the intercept estimate is the logarithm of the predicted odds ratio at the mean of predictors, and this makes the intercept easier to interpret: the intercept estimates the logarithm of the expected odds ratio of the studied event when the predictor variables are all equal to their means.

In addition, centring eliminates the covariance between the estimates of intercept and slopes in regression models. Let $\text{Var}\,[\hat{\boldsymbol{\beta}}]$ denote the variance-covariance matrix of regression coefficient estimates. For a linear logistic regression model given as (3.1) with a large sample size n, based on (3.24), analogous to (3.81) given in Appendix 3.B.2, $\text{Var}\,[\boldsymbol{\beta}]$ can be estimated by

$$\widehat{\text{Var}}\,[\hat{\boldsymbol{\beta}}] \;=\; (\boldsymbol{X}^T\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}, \tag{3.26}$$

where $\widehat{\boldsymbol{W}}$ is the weight matrix evaluated at $\hat{\boldsymbol{\beta}}$ and it was given in (3.14) (see McCullagh & Nelder, 1989, p. 119 for details).

If correlation between elements of $\hat{\boldsymbol{\beta}}$ is high, then different sets of parameter estimates might result in a very similar fit and likelihood. This makes it harder to interpret the parameter estimates. Note that $\widehat{\boldsymbol{W}}$ given in (3.14) is diagonal for a linear logistic regression model. By centring $\boldsymbol{x}$ by its mean $\bar{\boldsymbol{x}}$, the correlation between estimates of intercept and slopes can be effectively eliminated. For a linear logistic regression model, it is not feasible to give an analytical explanation of why centring predictors can eliminate the correlation between estimates of intercept and slopes. However, the benefits of centring predictors are analytically illustrated in the setting of a simple linear regression model in Appendix 3.C.

3.1.4 Asymptotic variance-covariance matrix

Given the parameter estimates, it is of great importance to obtain a knowledge of their uncertainty. For the maximum likelihood method, the inverse of the observed Fisher information matrix is used as an estimate of the asymptotic variance-covariance matrix of maximum likelihood estimates (see Appendix 3.D for details). Let $\boldsymbol{\theta}$ denote the parameter vector of interest and $\boldsymbol{i}^{-1}(\hat{\boldsymbol{\theta}})$ denote the inverse of observed Fisher information matrix $-\boldsymbol{H}(\hat{\boldsymbol{\theta}})$, i.e., the negative Hessian matrix evaluated at the maximum likelihood estimate of $\boldsymbol{\theta}$.

When fitting the extended logistic regression models (3.3) and (3.5), there are constraints on the parameters other than the regression coefficients $\boldsymbol{\beta}$, namely $\gamma \in (0,\, 1]$ for (3.3) and $\kappa \in (0,\, \infty)$ for (3.5). In order to avoid putting constraints on optimization for maximum likelihood estimates, the additional parameters γ and κ, are transformed to parameter values that take from $-\infty$ to $+\infty$. Let $\boldsymbol{\theta}^*$ denote the transformed parameter vector and $\boldsymbol{\theta} = \boldsymbol{t}(\boldsymbol{\theta}^*)$ for a suitable function $\boldsymbol{t}$.

Given the transformation $\boldsymbol{\theta} = \boldsymbol{t}(\boldsymbol{\theta}^*)$, the observed $\boldsymbol{i}^{-1}(\hat{\boldsymbol{\theta}}^*)$ needs to be adjusted for the estimation of the asymptotic variance-covariance matrix of $\boldsymbol{\theta}$. The general rule to obtain an estimate of the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ given $\boldsymbol{i}(\hat{\boldsymbol{\theta}}^*)$ is

$$\boldsymbol{i}^{-1}(\hat{\boldsymbol{\theta}}) \;=\; \left(\frac{\partial \boldsymbol{t}}{\partial \boldsymbol{\theta}^*}\right)^T\Bigg|_{\hat{\boldsymbol{\theta}}^*} \boldsymbol{i}^{-1}(\hat{\boldsymbol{\theta}}^*) \left(\frac{\partial \boldsymbol{t}}{\partial \boldsymbol{\theta}^*}\right)\Bigg|_{\hat{\boldsymbol{\theta}}^*}, \tag{3.27}$$

and in other words,

$$\widehat{\text{Var}}\,[\hat{\boldsymbol{\theta}}] = \left(\frac{\partial \boldsymbol{t}}{\partial \boldsymbol{\theta}^*}\right)^T\Bigg|_{\hat{\boldsymbol{\theta}}^*} \widehat{\text{Var}}\,[\hat{\boldsymbol{\theta}}^*] \left(\frac{\partial \boldsymbol{t}}{\partial \boldsymbol{\theta}^*}\right)\Bigg|_{\hat{\boldsymbol{\theta}}^*}.$$

Then based on (3.27), to obtain $\boldsymbol{i}(\hat{\boldsymbol{\theta}})$ for the $\hat{\boldsymbol{\theta}}$ in the extended logistic regression models, $\partial \boldsymbol{t}/\partial \boldsymbol{\theta}^*$ for (3.3) and (3.5) can be obtained as follows:

- For the asymptote-logistic model given by (3.3), the original parameter vector and its transformation for optimization are:

$$\begin{aligned} \boldsymbol{\theta} &= (\boldsymbol{\beta}^T, \gamma)^T, \\ \boldsymbol{\theta}^* &= (\boldsymbol{\beta}^T, \gamma^*)^T, \end{aligned}$$

  and,

$$\boldsymbol{\theta} = \boldsymbol{t}(\boldsymbol{\theta}^*) = (I(\beta_0), \ldots, I(\beta_k), \text{logit}^{-1}(\gamma^*))^T,$$

  where I denotes the identity function. Then

$$\text{logit}^{-1}(\gamma^*) = (1 + \exp(-\gamma^*))^{-1} = \gamma.$$

  Therefore, the first-order derivative of $\boldsymbol{\theta}$ with respect to $\boldsymbol{\theta}^*$ for the asymptote-logistic regression model is

$$\frac{\partial \boldsymbol{t}}{\partial \boldsymbol{\theta}^*} = \begin{pmatrix} \boldsymbol{I}_{q\times q} & 0 \\ 0 & \dfrac{\exp(\gamma^*)}{[1+\exp(\gamma^*)]^2} \end{pmatrix}, \tag{3.28}$$

  where $\boldsymbol{I}_{q\times q}$ denotes the $q \times q$ identity matrix.

- For the asymmetric-logistic regression model, the parameter vector and its transformed parameter vector are:

$$\begin{aligned} \boldsymbol{\theta} &= (\boldsymbol{\beta}^T, \kappa)^T, \\ \boldsymbol{\theta}^* &= (\boldsymbol{\beta}^T, \kappa^*)^T, \end{aligned}$$

and

$$\boldsymbol{t} = (I(\beta_0), \ldots, I(\beta_k), \exp(\kappa^*))^T,$$

where $\exp(\kappa^*) = \kappa$.

Therefore, the first-order derivative of $\boldsymbol{t}$ with respect to $\boldsymbol{\theta}^*$ is

$$\frac{\partial \boldsymbol{t}}{\partial \boldsymbol{\theta^*}} = \begin{pmatrix} \boldsymbol{I}_{q\times q} & 0 \\ 0 & \exp(\kappa^*) \end{pmatrix}. \tag{3.29}$$

3.1.5 Hosmer-Lemeshow goodness-of-fit test

Assessment of model fit is an important step of a modelling procedure, and there are many statistical tools for it, including both graphical and numerical ones. However, graphical analysis is usually difficult for logistic regression with binary response data, such as the graphical residual analysis given in Figure 3.1. In this case, numerical methods, such as the Pearson chi-squared test and the deviance test, provide a different type of tools to assess the fitted model. Tests of this type are usually carried out by measuring the discrepancy between observed data and predicted or expected outcomes based on the fitted model. For logistic regression with binary response data, Hosmer *et al.* (1997) compared different goodness-fit-tests and deduced the superiority of performance of the Pearson chi-squared test and deviance test in the case when only categorical predictor variables are involved in the logistic regression.

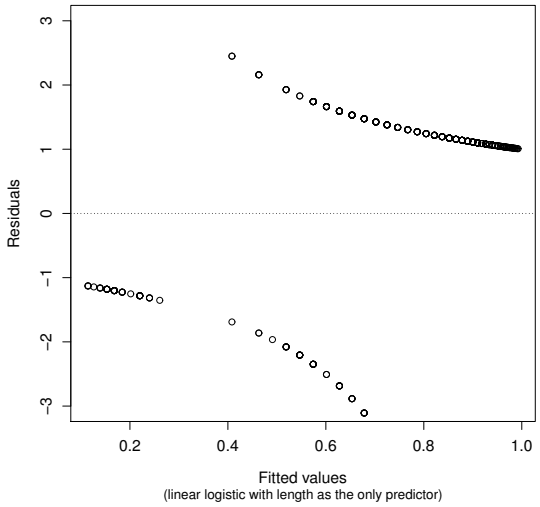


FIGURE 3.1. Plot of the residuals versus fitted values for the fitted linear logistic model with length as the only predictor, i.e. model described by (3.32) in Section 3.2.1.

However, in the case of logistic regression with continuous predictors and binary response, the test statistics of the commonly used tests, Pearson chi-squared or deviance test, do not have approximate chi-squared distributions under the null hypothesis that the fitted model is the correct model. This is due to the very small expected cell sizes resulting from a contingency table with the number of rows equal to the total number of individual subjects in the observed data. Taking the Pearson chi-squared test for binary response data as an example, the test statistic is

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)},$$

and the square root of the contribution from the ith observation (i.e., the Pearson residual) is

$$\frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}.$$

The distribution of this cannot be approximated by a standard normal distribution, as the normal approximation for a binomial distribution works only when the number of trials for the ith observation (n_i) is large (the rule of thumb is $\min\{n_i p_i,\ n_i(1-p_i)\} > 5$). However, this is not the case for binary response data as $n_i = 1$.

To compensate for this, Hosmer & Lemeshow (1980) introduced a goodness-of-fit test (HL-GOF test) which groups the data with respect to predicted success probabilities based on the fitted model, and then compares the observed to the expected counts for both successes and failures of Bernoulli response. The cut-points of predicted probability for each cell are chosen in a way that the total number of observations in each cell is about the same. This grouping strategy allows sufficient cell size to perform a chi-squared goodness-of-fit test, which is reviewed in Appendix 3.E.

The total number of cells, k, lies between 6 and 10 in most cases. Table 3.1 is a contingency table for performing a HL-GOF test, with the ith row consisting of the cut-points of the cell $(\hat{p}_{i-1},\, \hat{p}_i]$, its total number of observations (N_i), the observed counts of failure and of success (denoted as O_{i0} and O_{i1} respectively), and the predicted counts of failure and success (denoted as E_{i0} and E_{i1} respectively).

TABLE 3.1. Partition for the Hosmer-Lemeshow GOF test.

			$\delta = 0$		$\delta = 1$	
Cell	$\hat{p}$	Total	Observed	Expected	Observed	Expected
1	$(0, \hat{p}_1]$	N_1	O_{10}	E_{10}	O_{11}	E_{11}
2	$(\hat{p}_1, \hat{p}_2]$	N_2	O_{20}	E_{20}	O_{21}	E_{21}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	$(\hat{p}_{i-1}, \hat{p}_i]$	N_i	O_{i0}	E_{i0}	O_{i1}	E_{i1}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	$(\hat{p}_{k-1}, 1]$	N_k	O_{k0}	E_{k0}	O_{k1}	E_{k1}

In summary, the HL-GOF test is conducted with the following steps

1. order the fitted values $\hat{\boldsymbol{p}}$ for all individual subjects in the data;

2. group the fitted values into k cells (mostly 10, but usually between 6 and 10) so that the size of each cell is roughly the same;

3. calculate the observed and expected number for each cell in the cases of both success and failure of the binary response; and

4. perform a chi-squared GOF test whose detail is given below.

The test statistic of the chi-squared test in the above step 4 is calculated as

$$\chi^2 = \sum_{i=1}^{k} \sum_{\delta=0}^{1} \frac{(O_{i\,\delta} - E_{i\,\delta})^2}{E_{i\,\delta}}, \tag{3.30}$$

where δ stands for the binary response with $\delta = 0$ for the ith observation being a failure and $\delta = 1$ otherwise. For the anglerfish experimental survey data, $\delta = 1$ means that the fish was retained in the main cod-end and $\delta = 0$ that the fish escaped beneath the footrope.

Unlike a chi-squared GOF test with k cells for which the degrees of freedom equal to $k-1$ (see Appendix 3.E for details), for a HL-GOF test, the degrees of freedom of the chi-squared distribution under the null hypothesis is $k-2$. The intuitive explanation for this decrease in the degrees of freedom is the constraint of a fixed total number of observations in each cell.

Finally, for a given significance level α, the test statistic χ^2 obtained by (3.30) for a HL-GOF test is compared to a critical value, $\chi_{k-2,\alpha}$, which is the $(1-\alpha) \times 100\%$ percentile of a chi-squared distribution with $k-2$ degrees of freedom. If χ^2 is larger than the upper critical value $\chi_{k-2,\alpha}$, then the null hypothesis of no lack of fit is rejected at significance level α.

3.2 Application of fixed-effects logistic regression to anglerfish

Returning to the anglerfish experimental survey data described in Section 2.2, all the models formulated in Section 3.1.1 are applied in this section in order to provide an estimate of the net retention probability, which is essential for the anglerfish abundance estimation in Chapter 4. The linear logistic regression (3.1) is fitted by `glm()` and its two extended forms, (3.3) and (3.5), are fitted by maximum likelihood estimation using `optim()`. All the estimation results, model comparison and diagnostics are presented in Section 3.2.1. Section 3.2.2 gives a discussion on the choice of fixed-effects logistic regression model for abundance estimation, and some potential issues which lead to the considerations in Section 3.3 of mixed-effects logistic regression models.

3.2.1 Estimation results

Given the anglerfish experimental survey data described in Section 2.2, this section starts with fitting linear logistic regression models given as (3.1) and selecting important covariates. Then based on the final selected linear logistic regression model, its extended forms are fitted and their parameter estimates together with their estimated asymptotic variance estimates are presented.

The information included in the 2006-2007 anglerfish experimental survey data shows that there are four potential predictors: haul, day, year and fish length (see Section 2.2 for more detail). Note that, strictly speaking, the predictor fish length is the length class instead of the exact length of each captured fish. However, given that the length varies from 9.5 cm up to 71.5 cm with the unit size being 1 cm, the length is still considered as a continuous predictor and as discussed in Section 3.1.3, the predictor length is centred by the sample mean 39.6 cm in fitting logistic regression models (3.1), and its extended forms (3.3) and (3.5).

Starting with the linear logistic regression model, all the available predictors are included and the fitted model is referred to as 'linear logistic 4' in Table 3.2, with the number '4' referring to the number of predictors. Then the models with one predictor dropping out at each step are fitted and these models are referred to as 'linear logistic 3' and 'linear logistic 2' sequentially. In this stepwise process, the predictor to be omitted is chosen if there is little evidence for its significance using a likelihood ratio test. The simplest model with only length of fish as predictor is referred to as 'linear logistic' in Table 3.2. The first four models in Table 3.2 show the stepwise process for selecting the predictors for linear regression models. The table also gives the number of parameters ($\dim(\boldsymbol{\theta})$) for each model, and Akaike's information criterion (AIC Akaike, 1973). Note that the 'linear logistic 4' has $\dim(\boldsymbol{\theta}) = 37$, this is because the predictor 'haul' is a factor and there are 36 hauls in total in the data.

Based on p-values of the HL-GOF test in Table 3.2, it can be seen that there are no p-values less than 5%, which is considered as statistically insignificant evidence for lack-of-fit of any model fitted using anglerfish experimental survey data. Then the question in model selection becomes a problem of choosing a valid parsimonious regression model. The AIC criterion and the likelihood ratio test for each predictor are applied for this purpose. The linear logistic regression with length as the only predictor is finally chosen among all fitted fixed-effects logistic regression models.

It is noted that the asymmetric logistic regression model has been eliminated from the selection process due to the failure to fit a unique model, i.e., the likelihood surface for the asymmetric logistic regression model possess a flat ridge.

The predictor 'Day' is dropped out in the stepwise model selection for practical reasons, even though the likelihood ratio test shows some evidence for its significance at 5%, but not at 1%. The practical reason is that there is no available information for Day or Night in the abundance survey, therefore the inclusion of Day as a predictor will complicate the prediction of retention probability in abundance survey, which is not worthwhile in the light of its marginal statistical significance and the contribution in improving the fitting of the model by including it.

Among all the linear logistic models in Table 3.2, the 'linear logistic 2' has the lowest AIC, with 'linear logistic 3' and 'linear logistic' having their AICs within 2 of the minimum. Therefore, three models are considered further here. Likelihood ratio tests have been done for the significance of predictors Day, Year and Length, and the effect Year and Day are not significant at the 5% level, but there is extremely strong evidence for Length. This means that the predictor length of fish explains substantial structural variation in the data.

Therefore, 'linear logistic (Length)' is selected among all linear logistic regression models and its extended forms, asymptote-logistic and asymmetric logistic, are further fitted to check the asymptote and asymmetry of the logistic curve as a function of length. AICs are also given in Table 3.2. In addition, the goodness of fit of all the models is assessed by the HL-GOF test described in Section 3.1.5, and no significant lack of fit is found for any of the fitted models presented in Table 3.2.

The contingency table for performing the HL-GOF test for 'linear logistic (Length)' is given in Table 3.3. It presents the ordered and grouped fitted values for each cell based on the fitted linear regression model given by (3.36). Based on the table, the test statistic is then calculated by (3.30), $\chi^2 = 11.036 < 15.507 = \chi^2_{8,0.05}$. Therefore there is no evidence against the null hypothesis that the model is a good fit to the data at 5% significant level.

TABLE 3.2. Anglerfish application: linear logistic and its extended forms. ΔAIC is the difference between the AIC of the model in question and the model with the lowest AIC, i.e., the asymptote-logistic given by (3.37). The number after 'Linear logistic' in the table are the number of the explanatory variables, which are given in brackets for each linear logistic regression model.

Model	Dim($\boldsymbol{\beta}$)	AIC	ΔAIC	HL-GOF p-value
Linear logistic 4 (Haul, Day, Year, Length)	37	407.25	21.84	0.811
Linear logistic 3 (Day, Year, Length)	4	387.52	2.11	0.445
Linear logistic 2 (Day, Length)	3	386.48	1.07	0.279
Linear logistic (Length)	2	388.49	3.08	0.200
Asymptote-logistic (Length)	3	385.41	0	0.744
Asymmetric logistic (Length)	3	385.13 [a]	NA [a]	0.513 [a]

[a] the fitted asymmetric logistic model, given by (3.44), is eliminated from model selection due to the failure of convergence when maximizing its likelihood function.

TABLE 3.3. The partition of the Hosmer-Lemeshow GOF test for the fitted linear logistic model with length of fish as the only predictor, given in (3.36). 'Cod-end=0' means that fish was retained in cod-end and 'Cod-end=1'means that fish escaped beneath the footrope. $\hat{p}$ gives the cut-points of each cell and 'Total' gives the total number of observations for each cell. The last column 'χ^2' gives the contribution for χ^2 statistic for each cell.

			Cod-end=0		Cod-end=1		
Cell	$\hat{p}$	Total	Observed	Expected	Observed	Expected	χ^2
1	$(0, 0.24]$	44	39	36.17	5	7.83	1.24
2	$(0.24, 0.574]$	47	24	22.32	23	24.68	0.24
3	$(0.574, 0.653]$	48	12	17.67	36	30.33	2.88
4	$(0.653, 0.725]$	44	12	13.12	32	30.88	0.14
5	$(0.725, 0.805]$	39	10	8.76	29	30.25	0.23
6	$(0.805, 0.852]$	37	3	5.86	34	31.14	1.66
7	$(0.852, 0.89]$	45	7	5.51	38	39.49	0.46
8	$(0.89, 0.918]$	41	3	3.77	38	37.23	0.17
9	$(0.918, 0.957]$	45	4	2.71	41	42.29	0.65
10	$(0.957, 1]$	41	3	1.10	38	39.90	3.36

The following gives a more detailed description of the final selected linear logistic regression model 'linear logistic (Length)' and its extended forms. These details will be used later in Part III for the anglerfish abundance estimation. Let y_i denote the response variable of whether or not the ith sampled fish was retained in the main cod-end, where $y_i = 1$ denotes the ith sample fish being retained and $y_i = 0$ otherwise, where $i = 1, 2, \ldots, 341$, as there are 341 individual fish captured in the 2006-2007 anglerfish experimental survey data. Then a Bernoulli distribution with success probability equal the net retention probability r is assumed for y_i, where r is estimated as a function of length for the final selected linear regression model ('linear logistic (length)') and its extended forms (asymptote-logistic and asymmetric logistic). Therefore, for the ith sample fish, $y_i \sim$ Bernoulli$(r(l_i; \boldsymbol{\theta}))$, and the likelihood is

$$L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{l}) = \prod_{i=1}^{341} \left\{ r(l_i; \boldsymbol{\theta})^{y_i} \left[1 - r(l_i; \boldsymbol{\theta})\right]^{1-y_i} \right\}, \tag{3.31}$$

where the parameter vector $\boldsymbol{\theta}$ depends on the type of logistic regression models and so on the retention probability function $r(l_i)$. The detailed formulae for $r(l,\, \boldsymbol{\theta})$ in (3.31) for each of the final three models in Table 3.2 are listed below:

1. For the 'linear logistic (Length)' in Table 3.2, $\boldsymbol{\theta}$ in (3.31) is just the regression coefficients $\boldsymbol{\beta}$, and thereafter

$$r(l_i;\, \boldsymbol{\beta}) \;=\; \frac{1}{1+\exp[-\beta_0 \,-\, \beta_1\,(l_i - \bar{l})]}, \tag{3.32}$$

where $\bar{l} = \sum_{i=1}^{341} l_i/341 = 39.6\,\text{cm}$, is the mean of length of all fish captured in the experimental survey data. The following lists the MLE of $\boldsymbol{\beta}$ estimated by the IRLS algorithm described in Section 3.1.2,

$$\hat{\beta}_0 \;=\; 1.315, \tag{3.33}$$

$$\hat{\beta}_1 \;=\; 0.112, \tag{3.34}$$

with the estimated asymptotic variance-covariance matrix

$$\begin{array}{ccc} & \hat{\beta}_0 & \hat{\beta}_1 \\ \hat{\beta}_0 & 0.0194 & 0.0007 \\ \hat{\beta}_1 & 0.0007 & 0.0002 \end{array} \tag{3.35}$$

Therefore, for a fish of length l, its net retention probability predicted using the selected linear logistic regression model is,

$$\hat{r}(l) \;=\; [1+\exp(3.1202 \,-\, 0.112\times l)]^{-1}. \tag{3.36}$$

2. For the 'asymptote-logistic' model in Table 3.2, $\boldsymbol{\theta}$ is the regression coefficient $\boldsymbol{\beta}$ and the asymptote parameter $\gamma \in [0,\, 1]$, so that the retention probability is allowed to increases monotonically to an non-unity asymptote,

$$r(l_i;\boldsymbol{\theta}) \;=\; \gamma\left\{\frac{1}{1+\exp[-\beta_0 - \beta_1\,(l_i - \bar{l})]}\right\}. \tag{3.37}$$

The MLE of each element in $\boldsymbol{\theta}$ is

$$\hat{\gamma} = 0.933, \tag{3.38}$$
$$\hat{\beta}_0 = 1.986, \tag{3.39}$$
$$\hat{\beta}_1 = 0.156, \tag{3.40}$$

and the estimated asymptotic variance-covariance matrix for $(\hat{\gamma}, \hat{\beta}_0, \hat{\beta}_1)^T$ is calculated based on (3.27) given in Section 3.1.4,

$$\begin{array}{ccccc} & \hat{\gamma} & & \hat{\beta}_0 & \hat{\beta}_1 \\ \hat{\gamma} & 0.00088 & & -0.0088 & -0.00049 \\ \hat{\beta}_0 & -0.0088 & & 0.1504 & 0.0087 \\ \hat{\beta}_1 & -0.00049 & & 0.0088 & 0.00072 \end{array} \tag{3.41}$$

The standard errors of each parameter estimates are se$(\hat{\gamma}) = 0.030$, se$(\hat{\beta}_0) = 0.338$ and se$(\hat{\beta}_1) = 0.027$. The estimated correlation matrix is

$$\begin{array}{ccccc} & \hat{\gamma} & & \hat{\beta}_0 & \hat{\beta}_1 \\ \hat{\gamma} & 1 & & -0.766 & -0.613 \\ \hat{\beta}_0 & -0.766 & & 1 & 0.833 \\ \hat{\beta}_1 & -0.613 & & 0.833 & 1 \end{array} \tag{3.42}$$

Therefore, for a fish of length l, the net retention probability predicted by the extended linear logistic regression model with non-unity asymptote is,

$$\hat{r}(l) = 0.933[1 + \exp(-4.1916 - 0.156 \times l)]^{-1}. \tag{3.43}$$

3. For the 'asymmetric logistic' model in Table 3.2, $\boldsymbol{\theta}$ is the regression coefficient $\boldsymbol{\beta}$ and the parameter for asymmetry $\kappa \in [1, +\infty]$, so that the retention probability is allowed to be non-symmetric about the median of length,

$$r(l_i; \boldsymbol{\theta}) = \left\{ \frac{1}{1 + \exp[-\beta_0 - \beta_1 (l_i - \bar{l})]} \right\}^{1/\kappa}. \tag{3.44}$$

The MLE of each element in $\boldsymbol{\theta}$ is

$$\begin{aligned} \hat{\kappa} &= 0.065, \\ \hat{\beta}_0 &= 4.207, \\ \hat{\beta}_1 &= 0.087. \end{aligned}$$

The standard errors of the parameter estimates are $\text{se}(\hat{\kappa}) = 0.361$, $\text{se}(\hat{\beta}_0) = 5.574$ and $\text{se}(\hat{\beta}_1) = 0.00876$, and the correlation matrix of the parameter estimates is (rounded to three decimal places)

$$\begin{array}{c|ccc} & \hat{\kappa} & \hat{\beta}_0 & \hat{\beta}_1 \\ \hline \hat{\kappa} & 1 & -1 & 0.249 \\ \hat{\beta}_0 & -1 & 1 & -0.237 \\ \hat{\beta}_1 & 0.249 & -0.237 & 1 \end{array} \tag{3.45}$$

Note that the estimated correlation between $\hat{\kappa}$ and $\hat{\beta}_0$ is -1, the interpretation of $\text{se}(\hat{\kappa})$ and $\text{se}(\hat{\beta}_0)$ then becomes problematic.

3.2.2 Discussion

Figure 3.2 shows the fitted logistic curves for the last three models given in Table 3.2. It can be seen that the fitted linear logistic (with length of fish as the only predictor) is very similar to the asymmetric logistic curve. However, it is important to note from (3.45) that the estimated correlation between $\hat{\kappa}$ and $\hat{\beta}_0$ is -1, which means a perfect negative linear relationship. Hence, it is not possible to estimate κ and β_0 uniquely from the anglerfish experimental survey data. This has been confirmed by setting out different initial values in `optim()` when maximizing the likelihood function for the MLE of $(\kappa, \beta_0, \beta_1)^T$, and the failure of convergence means that there is not enough information in the data for fitting the asymmetric logistic regression model. Therefore, the asymmetric logistic regression model is excluded in further considerations for the anglerfish abundance estimation in Part 4.

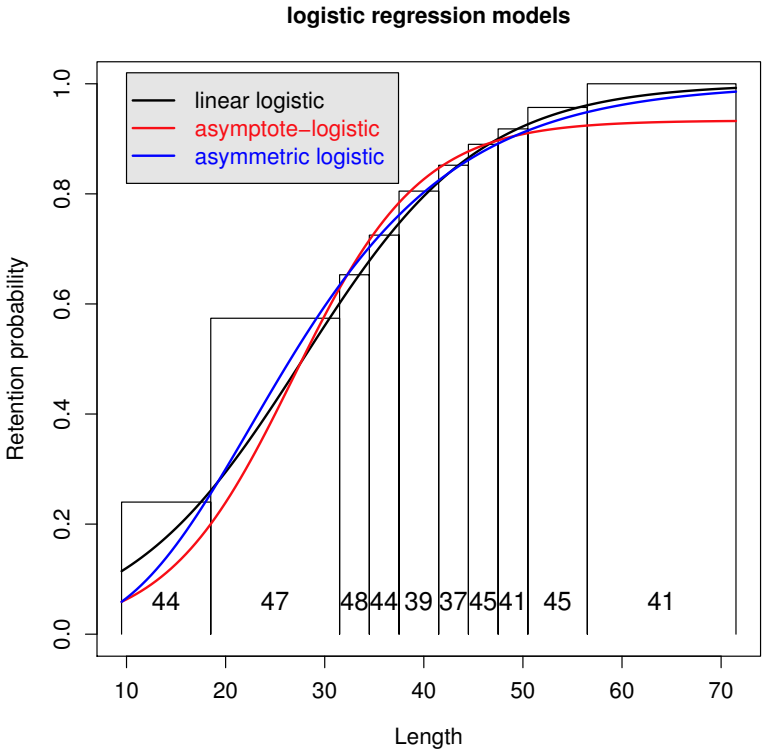


FIGURE 3.2. Plot of the estimated net retention probability $\hat{r}(l)$ in the form of (3.32), (3.37) and (3.44). The break-points of each bin are the grouping cut-points chosen for the HL–GOF test of the linear logistic model with only length as predictor. The height of each bin stands for the predicted retention probability and the number at the bottom of each bin is the corresponding number of observed individual subjects for each cell. Note that the grouping presented in the figure is chosen in testing fitted the linear logistic model (3.32), and the cut–points are slightly different for the other two models. Note that the bins in the figure show the cell size of the HL-GOF test for the linear logistic regression model with length as the only predictor.

For the asymptote-logistic curve, Figure 3.2 also shows its similarity to the linear logistic curve with length as the only predictor, but with an asymptote less than 1 which is consistent with the expression given by (3.43). If a likelihood ratio test were conducted to check the significance of γ, then the null hypothesis would be $\gamma = 1$ against the alternative $\gamma < 1$. This means that the null hypothesis is on the boundary of the parameter space and hence the test statistic does not have a chi-squared distribution. Instead, a 95% confidence interval is calculated to check the significance of the asymptote parameter γ. Given that $\hat{\gamma} = 0.933$ and $\text{se}(\hat{\gamma}) = 0.027$ in the previous section and based on normality assumption, the upper bound of the 95% confidence interval is 0.986. The confidence interval with an upper bound being

less than 1 gives some evidence for the significance of the asymptote γ, which is also suggested by comparing the AICs of the models 'Linear logistic' and 'Asymptotic-logistic' in Table 3.2. It is then concluded that there is some evidence of a non-unity asymptote and the fitted asymptote-logistic regression model (3.43) is further applied in Chapter 4 for the anglerfish abundance estimation.

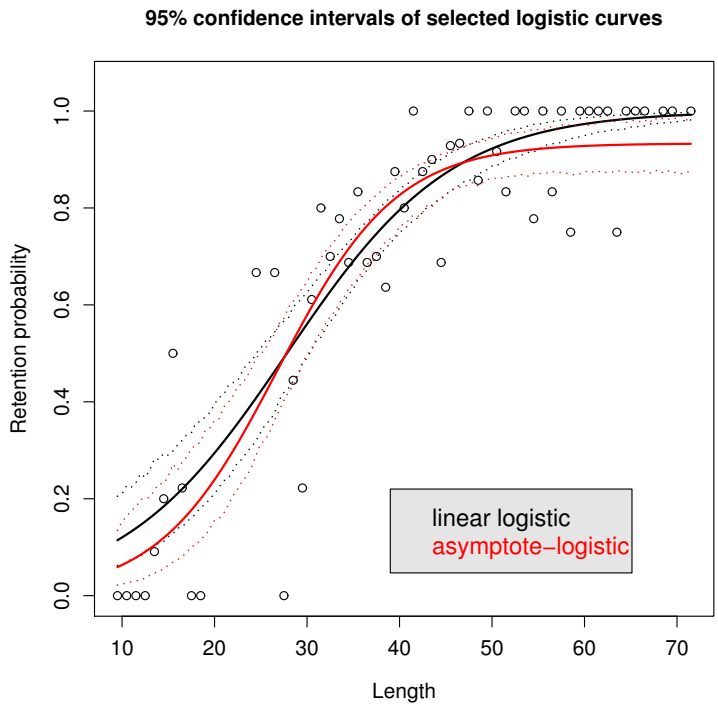


FIGURE 3.3. Plot of estimated logistic regression models with 95% confidence interval: the black curve is for $\hat{r}(l)$ estimated by a linear logistic regression model given by (3.36) and the red curve is for the $\hat{r}(l)$ estimated by an asymptote-logistic regression given by (3.43), together with the confidence interval for each curve given by the dotted line. The circles representing the observed proportions of the fish retained in the main cod-end from the 2006-2007 experimental survey data.

In summary, for abundance estimation, the net retention probabilities are estimated using two fitted models among fixed-effects logistic regression models: the linear logistic with length as the only predictor and its extended form allowing the asymptote to be less than unity. The confidence intervals for the $\hat{r}(l)$ of these two selected models are presented in Figure 3.3: the black colour for the linear logistic with length as predictor ('linear logistic') and the red color for its extended form ('asymptotic-logistic'). The 95% confidence intervals in dotted lines are obtained by parametric bootstrapping based on their estimated asymptotic distribution with mean

and variance-covariance given in Section 3.2.1: (3.33)–(3.35) for the selected linear logistic model with length as the only predictor; (3.38)–(3.41) for the asymptote-logistic regression model.

In bootstrapping, samples are drawn 999 times and 2.5% and 97.5% quantiles are used for the interval bounds (dotted lines), and the median (50%) is plotted as the solid lines. From the plot it can be seen that the width of the confidence intervals for $\hat{r}(l)$ at length greater than 60 cm becomes very small for the estimated linear logistic model (with length as the only predictor). However, this is not the case for asymptote-logistic in which the allowance of a variable asymptote makes the capture less certain for larger fishes. This may cause a problem in the presence of extrapolation in predicting net retention probabilities for larger fish, as there are more larger fish in the abundance survey data than in the experimental survey data. There are 10% fish larger than 70 cm in the 2007 abundance survey, and using the asymptote-logistic regression model to make predictions for these fish may cause extrapolation. The effect of the extrapolation for larger fish will be discussed in Section 4.6.

In the 2006-2007 experimental survey data, a total of 431 individual fish were captured with minimum length of fish 9.5 cm, and the maximum 71.5 cm, while the length range of fish captured in the 2007 anglerfish abundance survey is from 12cm to 126 cm. Comparing length ranges of the above two data sets, it is noted that the prediction of net retention probabilities for abundance estimation involves extrapolation beyond the length range of the experimental data. For the linear logistic regression model with length as predictor, prediction of the retention probability for fish larger than 70 cm is unlikely to be problematic, as the estimated retention probability is very close to 1 with little uncertainty (as shown in Figure 3.3). However, for the asymptote-logistic, it might be problematic to extrapolate given the uncertainty of prediction for fish beyond 70 cm. The extrapolation of the estimated asymptote-logistic regression model can be addressed only if there are more data about larger fish obtained from experimental surveys, and the new data must have larger fish with length varying between 71.5 cm and 126 cm. Given that there is no belief that larger fish must be retained in the cod-end from fisheries expert opinions, the extended logistic regression with varying asymptote is considered further for abundance estimation, and it turns out that the extrapolation at larger length classes has little effect on the abundance estimator used in Chapter 4; Section 4.6 will discuss this in full detail.

Checking the assumptions made in the fixed-effects logistic regression models, there is no evidence found for over-dispersion in the data and the only issue is the independence assumption of retention probability for the fish captured by the same haul. The anglerfish experimental survey data are grouped by hauls and each haul is treated as one cluster. The anglerfish caught in the same haul share the common living environment or biologically speaking, they come from the same school of fish. Therefore, the analysis in Section 3.2.1 that assumes they are independent from each other within the same haul might not be appropriate. To incorporate the haul effect (avoiding the independence assumption), fixed-effects logistic regression models can include haul as a factor, and then a fixed unknown constant for each haul is estimated. However, treating haul as a fixed effect will make it problematic in prediction for the abundance estimation, because the sample of hauls in the anglerfish abundance survey is different from that in the experimental surveys, though it is assumed that they both come from the same population of hauls.

In this case, the use of random effect is common and convenient for modelling such clustering structure of the data. Instead of an unknown constant for each haul in fixed-effects logistic regression, the haul effect is assumed to be a realized value of a random variable, and as a result, the haul effect becomes a random effect and the parameters of its distribution are then estimated. The incorporation of random effects provides a more flexible and parsimonious framework for analysing clustered data. More importantly, the incorporation of random effects allows for dependence of observations within the same cluster, i.e., the capture of fish within the same haul, and the estimated distribution of random effects provides the basis of prediction for a different sample of random effects.

The above considerations motivate including haul as a random effect in estimating net retention probability for anglerfish surveys. Therefore, mixed-effects models are pursued further. These models are also referred to as multilevel models, hierarchical models, or random-effects models in the literature. Given the binary response data from anglerfish experimental surveys, the next section concentrates on the generalized linear mixed-effects model with the Bernoulli distribution assumed for the response variable and the canonical link function being logit. This particular type of model is referred to as a mixed-effects logistic regression model in the following sections.

3.3 Mixed-effects logistic regression model

As discussed in Section 3.2.2, the linear logistic regression in (3.1) is extended by incorporating random effects to allow dependence within clusters, which could be repeated measures over time, space or experimental subject. This type of clustered data can be analysed by mixed-effects models, in which the data hierarchy consists of lower-level observations clustered within a higher level, and an error term can then be specified for each level. This explains why the mixed-effects models are also referred to as hierarchical models or multilevel models in the literature. The lowest level is referred to as level-one, while the higher levels are referred to as level-two, level-three, and so on in accordance with their hierarchical ordering. Figure 3.4 plots a two-level mixed-effects model using a tree diagram, which explicitly shows that the lowest level observations are defined as being the level-one observations.

In the context of the anglerfish application, the survey data are clustered by hauls as shown in Figure 3.4. Such a clustering structure can be considered as a spatial aggregation in the data due to the sampling process in the trawl surveys. With haul being included as a random effect in estimating the footrope net retention probability, the captures of fish within the same haul are no longer independent and the resulting model may capture the potential spatial heterogeneity in capture probabilities that is not explained by fixed-effects predictors.

As the mixed-effects models allow for correlation among repeated measurements made on individual clusters by incorporating appropriate random effects, the questions then arise of how to incorporate the random effects, how to estimate the random-effects parameters, and whether or not random effects are incorporated appropriately. To answer these questions, this section starts with the formulation of mixed-effects logistic regression in Section 3.3.1, which leads to Sections 3.3.2 and 3.3.3 describing two approximation methods used in parameter estimation. Then Section 3.3.4 moves on to answer the questions about whether or not the random effects are incorporated properly in the model and how to do the model selection. In addition, Section 3.3.5 addresses some issues about centring a predictor in multilevel mixed-effects models, and illustrates the idea by a simple random-intercept-only two-level mixed-effects model.

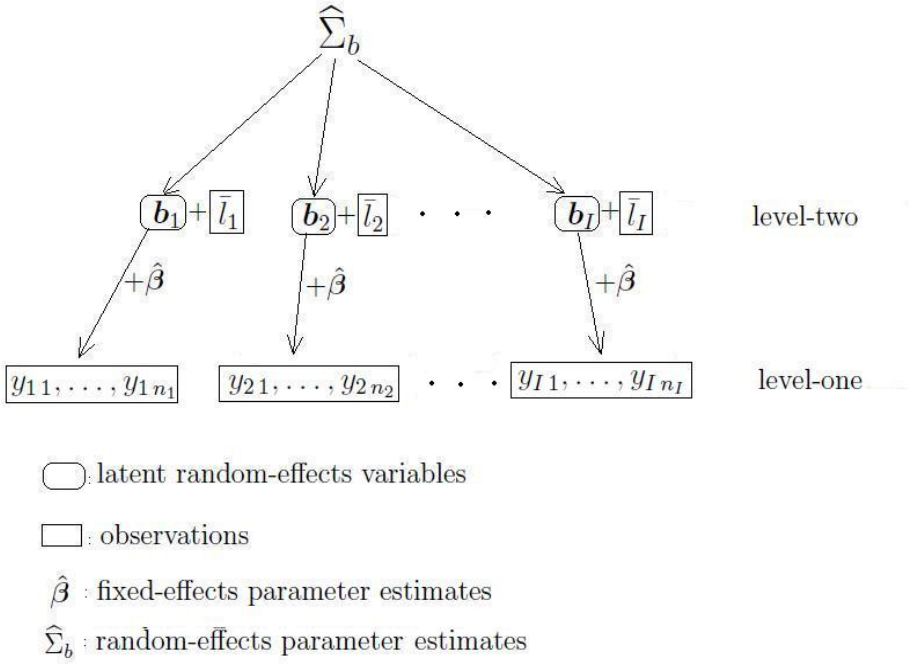


FIGURE 3.4. The tree diagram of the mixed-effects logistic regression model used in estimation of the net retention probability. The level-two predictor is the group-mean length for each haul and the level-one predictor is the group-mean-centred length of each individual fish.

3.3.1 Model formulation

Here is a list of extended notation for two-level mixed-effects logistic regression models (consistent with the notation used in Section 3.1.1 for fixed-effects logistic regression models):

- i: the ith group out of a total of I groups, $i = 1, 2, \ldots, I$;
- j: the jth observation within a given group, also referred to as a level-one unit nested within each group (i.e., level-two); for group i, $j = 1, 2, \ldots, n_i$;
- $\boldsymbol{\beta}$: the q-dimensional fixed-effects parameter vector;
- $\boldsymbol{x}_{ij}$: the $q \times 1$ fixed-effects model vector of the jth observation within the ith group, with the first element being 1 if the model has an intercept, and the number of fixed-effects explanatory variables is $q - 1$;
- $\boldsymbol{b}_i$: the m-dimensional random-effects coefficient vector of the ith group; and

- $\boldsymbol{z}_{ij}$: the corresponding $m \times 1$ random-effects model vector where m is the number of random-effects coefficients, and the first element of $\boldsymbol{z}_{ij}$ is 1 if the model has a random intercept.

Note that $\boldsymbol{b}_i$ is a 'coefficient' vector, instead of being a 'parameter' vector like $\boldsymbol{\beta}$, because $\boldsymbol{b}_i$ is a vector of unobserved random values from the underlying distribution of random effects, which is also referred to as latent observations in mixed-effects model terminology. The random effects are governed by a probability distribution, $f_b(\boldsymbol{b})$, which is usually assumed to be a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix Σ_b. The parameters for random effects that are of interest in estimation are the variance-covariance matrix Σ_b.

Let y_{ij} denote the binary response variable, and in the case of anglerfish application, it indicates whether or not the jth fish captured by haul i was retained in the main cod-end, and with probability p_{ij}, $y_{ij} = 1$ meaning that the fish was retained in the cod-end. With probability $1 - p_{ij}$, $y_{ij} = 0$ meaning that the fish escaped beneath the footrope. In more detail, given the random effects $\boldsymbol{b}_i$ of group i, it is assumed that $y_{ij}|\boldsymbol{b}_i \sim \text{Bernoulli}(p_{ij|\boldsymbol{b}_i})$, and

$$\text{logit}(p_{ij|\boldsymbol{b}_i}) \quad = \quad \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i. \tag{3.46}$$

Equivalently, the conditional mean $\mathbf{E}[y_{ij}|\boldsymbol{b}_i]$, denoted by $p_{ij|\boldsymbol{b}_i}$, is

$$p_{ij|\boldsymbol{b}_i} \quad = \quad \frac{1}{1 + \exp(-\boldsymbol{x}_{ij}^T\boldsymbol{\beta} - \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)}, \tag{3.47}$$

where the random-effects coefficient vector, $\boldsymbol{b}_i$, is assumed to follow a m-variate normal distribution $N(\mathbf{0},\, \Sigma_b)$. Its probability density function is

$$f_b(\boldsymbol{b}_i) = \frac{1}{(2\pi)^{m/2}|\Sigma_b|^{1/2}} \exp\left(-\frac{\boldsymbol{b}_i^T\Sigma_b\boldsymbol{b}_i}{2}\right). \tag{3.48}$$

From the above it can be seen that the mixed-effects model of y_{ij} is defined hierarchically: the logistic regression of p_{ij} is defined conditionally on the values of random effects for group i, $\boldsymbol{b}_i$, which themselves are assumed to have a probabilistic distribution with density $f_b(\boldsymbol{b})$ in terms of further parameters Σ_b, which are also referred to as hyper-parameters. Given the random effects $\boldsymbol{b}_i$, y_{ij} are assumed to be independent for different groups. In other words, y_{ij} are conditionally independent

across different groups given the random effects. This hierarchy model structure is illustrated by Figure 3.4 with I denoting the total number of hauls towed in the experimental survey, and $\boldsymbol{b}_i$ denoting the random-effects coefficients for haul i. The model shown in Figure 3.4 will be discussed in detail in Section 3.4. Note that Figure 3.4 is for a mixed-effects model with fish length l as the only predictor. The formulation presented here is considered for the more general case of mixed-effects logistic regression models, with explanatory variables denoted by $\boldsymbol{x}$ and $\boldsymbol{z}$.

Therefore, based on the conditional mean given in (3.47), the conditional likelihood for the jth observation within the ith cluster can be expressed as

$$\begin{aligned} & L_{ij}(\boldsymbol{\beta}\,|\boldsymbol{b}_i;\; y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) \\ = & \left\{ \frac{\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)}{\left[1+\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)\right]} \right\}^{y_{ij}} \left\{ 1 - \frac{\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)}{\left[1+\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)\right]} \right\}^{1-y_{ij}} \\ = & \frac{\exp(y_{ij}\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + y_{ij}\boldsymbol{z}_{ij}^T\boldsymbol{b}_i)}{\left[1+\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)\right]}. \end{aligned} \tag{3.49}$$

The conditional likelihood given above can be thought of as the conditional probability mass function of y_{ij} given $\boldsymbol{b}_i$. Together with the (unconditional) density of the random effects defined in (3.48), the joint distribution of y_{ij} and $\boldsymbol{b}_i$ can be further specified as the production of the L_{ij} in (3.49) and $f_b(\boldsymbol{b}_i)$ in (3.48). Therefore, the marginal likelihood (also the marginal probability mass function of y_{ij}) can be derived by integrating out $\boldsymbol{b}_i$ from the joint distribution. The marginal likelihood for the ith cluster is

$$\begin{aligned} & L_i(\boldsymbol{\beta}, \Sigma_b; \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) \\ = & \int \cdots \int_{\mathbb{R}^m} \prod_{j=1}^{n_i} L_{ij}(\boldsymbol{\beta}\,|\boldsymbol{b}_i;\; y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})\; f_b(\boldsymbol{b}_i) d\,\boldsymbol{b}_i \\ = & \int \cdots \int_{\mathbb{R}^m} \prod_{j=1}^{n_i} \left\{ \frac{\exp(y_{ij}\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + y_{ij}\boldsymbol{z}_{ij}^T\boldsymbol{b}_i)}{\left[1+\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)\right]} \right\} \frac{\exp\left(-\frac{1}{2}\,\boldsymbol{b}_i^T\Sigma_b\boldsymbol{b}_i\right)}{(2\pi)^{m/2}|\Sigma_b|^{1/2}} d\,\boldsymbol{b}_i\,, \end{aligned} \tag{3.50}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$, $\boldsymbol{X}_i$ is an $n_i \times q$ matrix with rows $\boldsymbol{x}_{ij}$ and $\boldsymbol{Z}_i$ is an $n_i \times m$ matrix with rows $\boldsymbol{z}_{ij}$.

Therefore, the marginal likelihood of the response given the data of all groups is

$$\begin{aligned} & L(\boldsymbol{\beta}, \Sigma_b; \text{data}) \\ = & \prod_{i=1}^{I} L_i(\boldsymbol{\beta}, \Sigma_b; \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) \\ = & \prod_{i=1}^{I} \int \cdots \int_{\mathbb{R}^m} \prod_{j=1}^{n_i} \frac{\exp\left[y_{ij}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i) - \boldsymbol{b}_i^T\Sigma_b\boldsymbol{b}_i/2\right]}{(2\pi)^{m/2}|\Sigma_b|^{1/2}[1+\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)]}\, d\,\boldsymbol{b}_i. \quad (3.51) \end{aligned}$$

The integral in (3.51) has no closed form, and this means that the exact likelihood of mixed-effects logistic regression model is not available and therefore, in order to obtain maximum likelihood estimates of parameters, numerical integration is required to approximate the marginal likelihood, so that the numerical optimization can be carried out on the approximate marginal likelihood. The next section describes numerical integration methods that have been implemented in the `R` package `lme4` (Bates, D. and Maechler. M and Bolker, B., 2011) to approximate the marginal likelihood in mixed-effects models.

3.3.2 Numerical integration of marginal likelihood

There are two numerical integration methods implemented in `lme4`: the Laplace method and adaptive Gauss-Hermite quadrature methods; the latter is referred to as the AGQ method. Laplace approximation is one of the most commonly used numerical approximation methods, which has been widely applied in Bayesian inference for estimating posterior moments and marginal densities (see Tierney & Kadane, 1986, for details). The approximation for posterior moments or marginal densities is analogous to obtaining the marginal likelihood by integrating out the random effects. Note that as an alternative to the maximum likelihood approach using numerical integration methods, the mixed-effects logistic regression models can also be fitted under a Bayesian framework using a *data augmentation* algorithm (Tanner & Wong, 1987) within a *Markov chain Monte Carlo* method (Brooks, 1998), or the *integrated nested Laplace approximation* (Rue *et al.*, 2009).

Most applications use the Laplace approximation, using a second-order Taylor expansion with higher-order terms in Taylor expansion being neglected. Such neglect has an effect on the accuracy of approximation, and the extent of this effect depends on the sample sizes and the complication of the random-effects structure in the data.

For complicated random-effects structure, Raudenbush *et al.* (2000) developed a higher-order Laplace approximation method for increased accuracy of approximation, and illustrated this method using a mixed-effects logistic regression with nested random effects. Their research shows that the higher-order Laplace approximation evaluates the marginal likelihood more accurately when there are nested random effects in the data, but at the expense of more analytical and computational complications. For the simple random-effects structure of the anglerfish experimental survey data shown in Figure 3.4, the higher-order Laplace approximation method is not considered here and the method of Laplace approximation with second-order Taylor expansion is then illustrated below for approximation of the integral given by (3.50).

Given the values of $\boldsymbol{\beta}$ and Σ_b in (3.50), the problem of evaluating the integral can be simplified to a problem of approximating an integral of the form

$$\int \cdots \int_{\mathbb{R}^m} \exp[l(\boldsymbol{b})] d\,\boldsymbol{b}, \tag{3.52}$$

where $l(\boldsymbol{b})$ is bounded and unimodal function of a m-dimensional vector $\boldsymbol{b}$. To illustrate the idea, let $\hat{\boldsymbol{b}}$ be the value of $\boldsymbol{b}$ at which $l(\boldsymbol{b})$ reaches its maximum. Then the second-order Taylor expansion of $l(\boldsymbol{b})$ is

$$l(\boldsymbol{b}) \approx l(\hat{\boldsymbol{b}}) + (\boldsymbol{b}-\hat{\boldsymbol{b}})^T \frac{\partial l(\boldsymbol{b})}{\partial \boldsymbol{b}}\bigg|_{\hat{\boldsymbol{b}}} + \frac{1}{2}(\boldsymbol{b}-\hat{\boldsymbol{b}})^T \frac{\partial^2 l(\boldsymbol{b})}{\partial \boldsymbol{b}^2}\bigg|_{\hat{\boldsymbol{b}}} (\boldsymbol{b}-\hat{\boldsymbol{b}}), \tag{3.53}$$

of which the second term is zero, therefore,

$$l(\boldsymbol{b}) \approx l(\hat{\boldsymbol{b}}) + \frac{1}{2}(\boldsymbol{b}-\hat{\boldsymbol{b}})^T\, l''(\hat{\boldsymbol{b}})(\boldsymbol{b}-\hat{\boldsymbol{b}}), \tag{3.54}$$

where $l''(\hat{\boldsymbol{b}})$ is the Hessian matrix evaluated at $\hat{\boldsymbol{b}}$; see (3.17) for the definition of the Hessian matrix. Therefore, based on the approximation of $l(\boldsymbol{b})$ given in (3.54) and thinking of the second term in (3.54) as an exponent of a m-variate normal density function like (3.48), but with variance-covariance matrix $|-l''(\hat{\boldsymbol{b}})|^{-1}$, the integral can be approximated by

$$\int \cdots \int_{\mathbb{R}^m} \exp[l(\boldsymbol{b})] d\,\boldsymbol{b} \approx \exp[l(\hat{\boldsymbol{b}})]\,(2\pi)^{m/2} |-l''(\hat{\boldsymbol{b}})|^{-1/2}, \tag{3.55}$$

where $-l''(\hat{\boldsymbol{b}})$ is the observed information matrix evaluated at $\hat{\boldsymbol{b}}$. Then taking the logarithm on both sides of (3.55), equivalently, the Laplace approximation can be

expressed as

$$\int \cdots \int_{\mathbb{R}^m} l(\boldsymbol{b})\, d\,\boldsymbol{b} \approx l(\hat{\boldsymbol{b}}) - \frac{1}{2}\log|-l''(\hat{\boldsymbol{b}})| + \frac{m}{2}\log(2\pi). \tag{3.56}$$

The above illustration for the Laplace approximation is set up in an ideal case when the log-likelihood is a function of random effects $\boldsymbol{b}$ only. However, this is not the case for mixed-effects models given the unknown parameters $\boldsymbol{\beta}$ and Σ_b.

This expression shows that $\hat{\boldsymbol{b}}$, around which the Laplace approximation is expanded, also depends on the unknown parameters $\boldsymbol{\beta}$ and Σ_b. This leads to an iterative process for obtaining the conditional mode of $\boldsymbol{b}$, which is the value $\tilde{\boldsymbol{b}}(\boldsymbol{\beta}, \Sigma_b)$ of $\boldsymbol{b}$ that maximizes the integrand for the marginal likelihood (3.51) given the values of $\boldsymbol{\beta}$ and Σ_b.

The algorithm used to obtain $\tilde{\boldsymbol{b}}(\boldsymbol{\beta}, \Sigma_b)$ is called the penalized iteratively re-weighted least squares algorithm (P-IRLS), which combines the IRLS algorithm for fitting generalized linear models (see Section 3.1.2 for details) and the penalized least squares algorithm for fitting linear mixed-effects models with a penalty term of random-effects (see Bates & DebRoy, 2004, for details).

Given the random effects $\boldsymbol{b}_i$ and the observed data, followed by the likelihood given in (3.50), the conditional log-likelihood for group i is

$$l_i(\boldsymbol{\beta}|\boldsymbol{b}_i) \;=\; \sum_{j=1}^{n_i} \left\{ y_{ij}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i) - \log[1 + \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)] \right\}$$

based on which the marginal likelihood is further approximated using the Laplace method given in (3.56) as

$$\begin{aligned}
& l_i(\boldsymbol{\beta}, \Sigma_b; \text{data}_i) = \int \cdots \int_{\mathbb{R}^m} l_i(\boldsymbol{\beta}|\boldsymbol{b}_i) f_b(\boldsymbol{b}_i) d\boldsymbol{b}_i \\
= \;& \int \cdots \int_{\mathbb{R}^m} \underbrace{\left[l_i(\boldsymbol{\beta}|\boldsymbol{b}_i) - \frac{1}{2}\boldsymbol{b}_i^T\Sigma_b\boldsymbol{b}_i - \frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_b| \right]}_{\text{thought of as } l(\boldsymbol{b}) \text{ in (3.56)}} d\boldsymbol{b}_i \\
\approx \;& l_i(\boldsymbol{\beta}|\tilde{\boldsymbol{b}}_i) - \frac{1}{2}\tilde{\boldsymbol{b}}_i^T\Sigma_b\tilde{\boldsymbol{b}}_i - \frac{1}{2}\log|\Sigma_b| - \frac{1}{2}\log|-l''(\tilde{\boldsymbol{b}}_i)|
\end{aligned}$$

of which the last term $-l''(\tilde{\boldsymbol{b}}_i)$ can be viewed as an estimate of the conditional variance of $\boldsymbol{b}$ given the parameter values and the data, and $\tilde{\boldsymbol{b}}_i$ is the conditional mode of $\boldsymbol{b}_i$ given values of $\boldsymbol{\beta}$ and Σ_b and the data observed for group i.

On the other hand, given that the distribution of the random effects is usually assumed to be normal, the marginal likelihood (3.51) has a Gaussian density factor in the integrand. This leads to the consideration of the Gauss-Hermite quadrature, which is a numerical integration method for an integrand with a Gaussian density factor. Illustrating the method in the univariate case, this quadrature method evaluates an integral of the form

$$\int_{-\infty}^{+\infty} l(x)\phi(x)d\,x,$$

by the following summation

$$\int_{-\infty}^{+\infty} l(x)\phi(x)d\,x, \approx \sum_{i=1}^{n} w_i l(x_i),$$

where $\phi(x)$ is the standard normal density function, x_1, $x_2, \ldots,$ x_n are the quadrature points with corresponding weights w_i, and n is the total number of quadrature points. The weights w_i only depend on the number of quadrature points n and the normal density ϕ, which is obtained by an equivalent of importance sampling in the context of quadrature methods (see Liu & Pierce, 1994; Pinheiro & Bates, 1995, for details on the calculation of w_i).

However, the performance of this quadrature method is highly related to the number of quadrature points in most cases, and for complicated integrands, the approximation is not accurate unless a high number of quadrature points is used. In order to improve the approximation accuracy, Liu & Pierce (1994) introduced the adaptive Gauss-Hermite quadrature (AGQ) method. In the context of nonlinear mixed-effects models, Pinheiro & Bates (1995) introduced the AGQ method as an importance sampling version of the ordinary Gauss-Hermite quadrature method, meaning that unlike the ordinary quadrature methods, the quadrature points in the AGQ approximation are chosen in the light of the behaviour of integrands. This difference is shown explicitly in Figure 3.5. It is also important to note that the Laplace approximation method with second-order Taylor expansion is equivalent to the one-point AGQ method (see

Pinheiro & Chao, 2006, for a full discussion). On the other hand, the AGQ method can be viewed as a higher-order Laplace approximation.

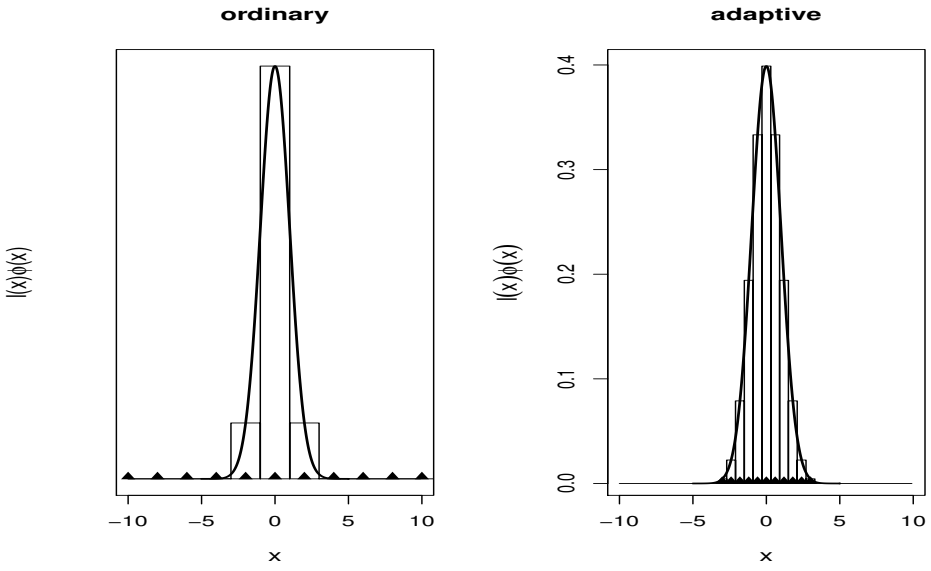


FIGURE 3.5. A plot to show the difference between an ordinary Gauss-Hermite quadrature and an adaptive Gauss-Hermite with 10 quadrature points to estimate the integral of the standard normal density function. The left panel illustrates the ordinary quadrature method with fixed points over-dispersed with regard to the density. The right panel shows the more accurate approximation obtained by the adaptive quadrature points conditional on the shape of the density (i.e., covering the interval of interest). By comparing these two panels it can be seen that the adaptive quadrature method produces more accurate and efficient approximation to the integral than the ordinary method.

The Laplace and AGQ approximation methods have been implemented in the R-package `lme4` (Bates, D. and Maechler. M and Bolker, B., 2011) to fit generalized linear mixed-effects models; see Pinheiro & Chao (2006) for a full description of the algorithm. The corresponding function used for mixed-effects logistic regression is called `glmer()`, and the number of AGQ points is set up by giving a positive integer to the argument 'nAGQ'. The default value of nAGQ is 1, meaning that the default approximation method of `glmer()` is the Laplace approximation. Pinheiro & Chao (2006) suggested that in fitting generalized linear mixed-effects models, nAGQ ≤ 7 is usually sufficient for estimation of the marginal likelihood, and the one-point AGQ method, which is equivalent to the Laplace approximation, often gives a reasonable approximation to the integral. However, even though the adaptive quadrature method usually produces accurate approximation, Lesaffre & Spiessens (2001) sug-

gested checking the convergence carefully for the AGQ method. Therefore, in the later application to anglerfish data, different numbers of quadrature points are tried in order to be sure of the convergence in approximating the marginal likelihood.

3.3.3 Quasi-likelihood methods

Note that in the case of linear mixed effects models, the likelihood given in (3.50) is the integral of an exponential function of a quadratic in $\boldsymbol{b}_i$. As a result, the marginal likelihood of a linear mixed-effects model has a closed form. This leads to another way of fitting generalized linear models. Instead of approximating the marginal likelihood based on the original data using the methods decribed in Section 3.3.2, this class of methods starts with transforming the original data by a linear expansion, so that a linear mixed-effects model is then fitted based on the resulting pseudo-data from the first step. This class of methods is referred to as *quasi-likelihood* methods, since the function being optimized in parameter estimation is not a true likelihood. This section starts by introducing the penalized quasi-likelihood (PQL) methods in the setting of a mixed-effects logistic regression model, then moves on to an alternative – the marginal quasi-likelihood (MQL) method, and finally gives a discussion of these two methods in the light of the anglerfish application for estimating net retention probabilities.

In order to illustrate the linear transformation, a mixed-effects logistic regression with binary response data is formulated as

$$y_{ij} = \underbrace{\text{logit}^{-1}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)}_{\hat{p}_{ij|\boldsymbol{b}_i}} + \epsilon_{ij}, \tag{3.57}$$

where ϵ_{ij} is an error term following a shifted Bernoulli distribution (i.e., a Bernoulli distribution that is shifted to have mean zero for the error term), and $\hat{p}_{ij|\boldsymbol{b}_i}$ is the conditional mean of y_{ij} given $\boldsymbol{b}_i$, which is also the success probability of the Bernoulli distribution conditional on $\boldsymbol{b}_i$. Then it can easily be shown that

$$\text{Var}\left[y_{ij}|\boldsymbol{b}_i\right] = \text{Var}\left[\epsilon_{ij}\right] = \frac{\exp(-\boldsymbol{x}_{ij}^T\boldsymbol{\beta} - \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)}{[1+\exp(-\boldsymbol{x}_{ij}^T\boldsymbol{\beta} - \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)]^2}. \tag{3.58}$$

Let $\widehat{V}_{ij}$ denote Var $[y_{ij}|\boldsymbol{b}_i]$ evaluated at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}_i$, i.e.,

$$\widehat{V}_{ij} = \frac{\exp(-\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}} - \boldsymbol{z}_{ij}^T\hat{\boldsymbol{b}}_i)}{[1+\exp(-\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}} - \boldsymbol{z}_{ij}^T\hat{\boldsymbol{b}}_i)]^2}.$$

Applying linear Taylor expansion to (3.57) around the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}_i$ gives

$$\begin{aligned} y_{ij} &\approx \text{logit}^{-1}(\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^T\hat{\boldsymbol{b}}_i) + \left.\frac{\partial y_{ij}}{\partial \boldsymbol{\beta}}\right|_{(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{b}}_i)}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \\ &\qquad + \left.\frac{\partial y_{ij}}{\partial \boldsymbol{b}_i}\right|_{(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{b}}_i)}(\boldsymbol{b}_i-\hat{\boldsymbol{b}}_i) + \epsilon_{ij} \\ &= \hat{p}_{ij|\boldsymbol{b}_i} + \epsilon_{ij} + \widehat{V}_{ij}\boldsymbol{x}_{ij}^T(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + \widehat{V}_{ij}\boldsymbol{z}_{ij}^T(\boldsymbol{b}_i-\hat{\boldsymbol{b}}_i) \\ &= \hat{p}_{ij|\boldsymbol{b}_i} + \epsilon_{ij} + \widehat{V}_{ij}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i) - \widehat{V}_{ij}(\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^T\hat{\boldsymbol{b}}_i). \end{aligned} \tag{3.59}$$

Multiplying (3.59) by $\widehat{V}_{ij}^{-1}$ on both sides and re-arranging the resulting equation gives

$$\underbrace{\widehat{V}_{ij}^{-1}(y_{ij}-\hat{p}_{ij|\boldsymbol{b}_i}) + (\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^T\hat{\boldsymbol{b}}_i)}_{y_{ij}^*} \approx \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i + \underbrace{\widehat{V}_{ij}^{-1}\epsilon_{ij}}_{\epsilon_{ij}^*}, \tag{3.60}$$

with y_{ij}^* being the pseudo response data and ϵ_{ij}^* being the transformed error term for PQL methods. Therefore, based on the pseudo-data obtained by (3.60), the original model (3.57) is approximately transformed to

$$y_{ij}^* \approx \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i + \epsilon_{ij}^*, \tag{3.61}$$

which is a linear mixed effects model with fixed-effects parameter $\boldsymbol{\beta}$, mixed-effects coefficients $\boldsymbol{b}_i$, and the pseudo-response data y_{ij}^* and the transformed error ϵ_{ij}^*. The algorithm used to fit linear mixed-effects models can then be applied to estimate the parameters in the original mixed-effects logistic regression model. The resulting parameter estimates are referred to as the penalized quasi-likelihood estimates because these parameter estimates can be obtained by optimizing a quasi-likelihood function including a penalty term of random effects (see Breslow & Clayton, 1993, for details).

As an alternative to the PQL method described above, the marginal quasi-likelihood method (MQL) differs from PQL in that the linear Taylor expansion of y_{ij} is around

$\hat{\boldsymbol{\beta}}$ and $\boldsymbol{b}_0$, instead of $\hat{\boldsymbol{b}}_i$ in (3.59), i.e.,

$$\begin{aligned} y_{ij} &\approx \text{logit}^{-1}(\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}}) + \left.\frac{\partial y_{ij}}{\partial \boldsymbol{\beta}}\right|_{(\hat{\boldsymbol{\beta}},\mathbf{0})}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + \left.\frac{\partial y_{ij}}{\partial \boldsymbol{b}_i}\right|_{(\hat{\boldsymbol{\beta}},\mathbf{0})}(\boldsymbol{b}_i-\mathbf{0}) + \epsilon_{ij} \\ &= \hat{p}_{ij|\boldsymbol{b}_i=\mathbf{0}} + \epsilon_{ij} + \widehat{V}_{ij}^{(m)}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i) - \widehat{V}_{ij}^{(m)}(\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}}), \end{aligned} \tag{3.62}$$

where $\hat{p}_{ij|\boldsymbol{b}_i=\mathbf{0}} = \text{logit}^{-1}(\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}})$ and $\widehat{V}_{ij}^{(m)}$ is the conditional variance of y_{ij}, the expression given in (3.58), evaluated at $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{b}_i = 0$,

$$\widehat{V}_{ij}^{(m)} = \frac{\exp(-\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}})}{[1+\exp(-\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}})]^2}.$$

Similarly, multiplying (3.62) by $\widehat{V}_{ij}^{(m)\,-1}$ on both sides, and then re-arranging the resulting equation, the transformed response data and error terms are given in the following expression,

$$\underbrace{\widehat{V}_{ij}^{(m)\,-1}\left(y_{ij} - \hat{p}_{ij|\boldsymbol{b}_i=\mathbf{0}}\right) + \boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}}}_{y_{ij}^{*m}} \approx \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i + \underbrace{\widehat{V}_{ij}^{(m)\,-1}\epsilon_{ij}}_{\epsilon_{ij}^{*m}}. \tag{3.63}$$

Therefore, the approximate linearization for MQL is

$$y_{ij}^{*m} \approx \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i + \epsilon_{ij}^{*m}. \tag{3.64}$$

Comparing (3.60) with (3.63), the difference between PQL and MQL is that MQL completely ignores the random effects in both the conditional mean and conditional variance of y_{ij}. As a result, the approximate linearization of the binary response data is reasonable only if the random effects are small. Therefore, the performance of MQL method depends on the size of the random effects (see Breslow & Clayton, 1993, for more details).

In summary, both PQL and MQL methods start with approximately linearizing the data, and then fit linear mixed-effects models based on the resulting pseudo-data. The advantage of these two methods is that they are easy to implement given the available software tools for fitting linear mixed effects models. For example, the R-function `glmmPQL()` fits a generalized linear mixed-effects model using the PQL method by repeated calls to another function `lme()` in R, which is a generic function fitting linear mixed-effects models (Venables & Ripley, 2002). We started with

using `glmmPQL()` to fit logistic mixed-effects regression models to the 2006-2007 anglerfish experimental survey data as the maximum likelihood approach for fitting non-linear mixed-effects models was not available. The results obtained by using `glmmPQL()` (i.e., the penalized likelihood approach) were very similar to those obtained by using `glmer()` (i.e., the maximum likelihood approach) provided by the package `lme4` (Bates, D. and Maechler. M and Bolker, B., 2011), with negligible difference in the random-effects parameter estimates.

However, the estimates produced by both PQL and MQL methods are potentially biased in practical applications, of which the case giving most concern is that of binary response data with few observations in each cluster (see Rodriguez & Goldman, 1995, for a full investigation). Furthermore, all likelihood-based inferences are inappropriate for these two methods, because the functions being optimized are based on the pseudo-data and therefore do not reflect the true marginal likelihood. For example, likelihood-based model comparison tools, such as AIC and likelihood ratio test, are not applicable for PQL and MQL methods.

Finally, in the anglerfish application, the experimental survey data have binary response and there are 18 hauls out of the total 34 that captured fewer than 10 fish in total. As discussed above, in the case of binary response data with small sample sizes, PQL and MQL methods are not reliable. In this case, the numerical integration methods described in Section 3.3.2 produce a more accurate approximation of the marginal likelihood for more reliable parameter estimation, and additionally make the likelihood-based inferences feasible. Therefore, the approximate maximum likelihood method described in Section 3.3.2 is preferred for estimating anglerfish net retention probabilities.

3.3.4 Model selection and assessment

As stated in Section 3.3.3, quasi-likelihood methods are potentially biased, and therefore this section considers model selection and assessment only for the maximum likelihood methods. Zuur *et al.* (2009) suggested a top-down strategy to fit mixed-effects models, and their general guidance for this strategy is summarized by the following steps:

- step 1: start with a model containing as many fixed-effects explanatory variables as indicated by the data and potential interactions, which is referred to as a *'beyond-optimal'* model (compare to the final selected *'optimal'* model);

- step 2: decide the structure of the random effects based on the beyond-optimal model from step 1;

- step 3: given the random-effects structure decided in step 2, select the fixed-effects predictors; and

- step 4: assess the selected model with the random-effects variance structure from step 2 and fixed-effects predictor from step 3 in terms of its goodness-of-fit, normality assumption of random effects, and other related aspects.

It is important to note that the above steps are considered only as a general protocol for fitting mixed-effects models, and that these steps usually need to be adjusted according to the observed data together with the prediction inference of the statistical analysis, if that is the practical objective of the statistical analysis. The following lists a further discussion of the statistical tools or issues within each step:

- In step 1: it is important to note that it should be feasible to fit the beyond-optimal model by the chosen statistical software. In the case of large data set with lots of explanatory variables and potential interactions, plotting the data usually gives a good idea of how to make a selection of predictors that contributes to explaining the variation of the responses.

- In step 2: given the beyond-optimal model fitted in step 1, it is expected that all the fixed-effects predictors make their contributions to explaining variation of responses such that the random-effects coefficients do not contain any information from fixed effects predictors. Then the problem becomes that of testing whether or not a random-effects variance component is significant. Likelihood ratio tests can be applied here, but they are not appropriate for testing the standard deviations of random coefficients. To explain this, let σ_0^2 denote the variance of a random intercept. Then the null hypothesis is $\mathrm{H}_0 : \sigma_0^2 = 0$ for the simpler model against the alternative hypothesis $\mathrm{H}_1 : \sigma_0^2 > 0$ for the more complicated model. In this case, the null hypothesis of the likelihood ratio test on σ_0^2 is on the boundary of the feasible parameter space of the random-effects

variance, so the test is referred to as a constrained likelihood ratio test. For constrained likelihood ratio tests, the distribution of the test statistic under the null hypothesis is no longer a chi-squared distribution; instead, it is a mixture of chi-squared distributions whose form depends on the specific case (see Self & Liang, 1987; Molenberghs & Verbeke, 2007, for more discussion). Pinheiro & Bates (2000) pointed out that likelihood ratio tests on random-effects variance are conservative and in the case of a single random-effects standard deviation, the p-value of the likelihood ratio test is approximately twice as large as it should be. Unlike the constrained case for testing random-effects variances, the p-values are correct for testing random-effects correlations using the likelihood ratio test.

The estimates of random-effects variance may be zero, even in the case when the true variance is not zero. In the case where the variance estimates are very small or the absolute values of correlations are very close to one, this indicates that the assumed random-effects variance structure cannot be identified or over-fitted given the observed data set. This could be a problem caused by lack of information contained in the data for fitting the intended model with the complex random-effects variance structure.

- In step 3: once the random-effects structure is decided in Step 2, the next step is to select fixed-effects predictors. Wald tests are usually used for this purpose in fixed-effects models, however, the p-value for the tests with $\mathrm{H}_0 : \beta_s = 0$ versus $\mathrm{H}_1 : \beta_s \neq 0$ is not as straightforward as the one in fixed-effects models. First, the test statistic does not have a t-distribution under the null hypothesis, because the independence of observations within each group is no longer assumed. Second, the degrees of freedom of the denominator for penalizing certainty are unknown for mixed-effects models, as the degrees of freedom for a random-effects parameter can be counted as 1, or some value between 1 and the total number of level-two units (see Hodges & Sargent, 2001, for more discussion).

  Therefore, for testing the significance of a particular explanatory variable, the likelihood ratio test is then suggested for comparing two nested models with the same random-effects variance structure: one with the predictor of interest and the other one without. It is important to note that the corresponding p-value is only a guide to the significance of the particular predictor. If the research question is about the significance of this predictor, then it is strongly

suggested that Monte Carlo based methods or the parametric bootstrap be used for drawing a conclusion about the effect of this predictor.

- Using AIC in step 2 and 3:

  AIC is widely used for model selection in terms of the relative goodness of fit of a model. It is defined as the maximized log-likelihood penalized by the degrees of freedom of the fitted model. Such a definition raises two issues of using AIC for mixed-effects models.

  – First, the likelihood for random-effects logistic regression models is approximated by numerical integration methods, such as Laplace approximation or adaptive Gaussian quadrature. The latter can be thought of as a higher-order Laplace approximation in the case of multiple quadrature points. Therefore, if the fitted models being compared are not approximated to the same order in the numerical integration, then we cannot be sure whether the difference in their likelihoods is caused by the different model structure or the different accuracy level in the numerical integration. It is important to note that the AIC of a fixed-effects model is not commensurate with the AIC of a corresponding random-effects model with the same fixed-effects component. Taking a random-intercept mixed-effects model for example, the AIC of the random intercept model should not be compared with a fixed-effects model (without the random intercept) when deciding the significance of the random intercept. Similarly, it is not suggested to compare AICs of mixed-effects models when the models use the adaptive Gaussian quadrature method with different quadrature points.

  – Second, counting the degrees of freedom for random-effects variance parameters is another issue when using AIC for mixed-effects models. There have been some adjusted forms of AIC for mixed-effects models, such as marginal AIC and conditional AIC, but these developments are considered only for linear mixed-effects models. Florin & Blanchard (2005) proposed a conditional AIC to compare linear mixed-effects models with different random-effects structures, which can be viewed as a finite-size correction for AIC. Greven & Kneib (2010) did a simulation study for both marginal AIC and conditional AIC for linear mixed effects models.

More importantly, model selection for mixed-effects models should be considered in the light of the final inferences or prediction. If further inferences, such as prediction, are only of interest at the population-level, then marginal AIC is suggested; if at a particular group or cluster, then the conditional AIC is recommended (see Greven & Kneib (2010) for more discussion). More specifically, if the statistical inferences are based on the mode of the random-effects coefficient $\boldsymbol{b}_i$ for the ith group, then the degrees of freedom of random effects should be counted as the total number of groups, and model selection should be based on the conditional AIC. On the other hand, if the statistical inferences are based on the estimated distribution of $\boldsymbol{b}_i$, equivalently $\widehat{\Sigma}_b$, then the degrees of freedom of the random-effects parameters should be counted as the number of parameters in $\widehat{\Sigma}_b$, and the model selection should be based on the marginal AIC.

Consider the anglerfish application, incorporating haul as a random effect with 36 levels (i.e., 36 individual hauls in the data), and fitting a random-intercept logistic regression model. Then the haul effect is taken into account at a cost of one parameter (variance of the haul-specific random intercept). However, for fixed-effects models, the cost of incorporating the haul effect is 35 parameters. This shows that if we compare the random-effects model with fixed-effects models, it is not appropriate to count the degrees of freedom for the random-effects variance parameter as 1, as it would probably give too small a degree of freedom.

- In step 4: for the final selected model from step 3, it is sensible to check this model in an absolute sense, e.g. using the goodness-of-fit test, checking the normality assumption of random effects, and looking for over-dispersion. Graphical tools can be very useful for understanding the fitted model. The normality assumption of the random-effects distribution can be checked by plotting the conditional modes of the random effects, which can be thought of as the MLE of $\boldsymbol{b}$ obtained in the iterative process of the P-IRLS algorithm (i.e., the $\tilde{\boldsymbol{b}}(\boldsymbol{\beta}, \Sigma_b)$ described in Section 3.3.2). In most cases, when the number of groups or clusters is large, the normally assumption of the random effects is usually reasonable. However, when the number of groups is small, the normality assumption could be problematic. This assumption can be loosened by

more complicated models, e.g. a mixture of normal distributions for the random effects, see Komárek & Lesaffre (2008) for example.

The Hosmer-Lemeshow goodness-of-fit test described in Section 3.1.5 can be extended to the mixed-effects logistic regression models, with $\hat{\boldsymbol{p}}$ in step 1 calculated as

$$\hat{\boldsymbol{p}}_{ij} \;=\; \int \cdots \int_{\mathbb{R}^m} p_{ij|\boldsymbol{b}_i} \hat{f}_b(\boldsymbol{b}_i)\, d\, \boldsymbol{b}_i, \tag{3.65}$$

where $p_{ij|\boldsymbol{b}_i} \hat{f}_b(\boldsymbol{b}_i)$ is the conditional success probability which has been given in (3.47). Based on these $\hat{\boldsymbol{p}}_{ij}$, the expected counts $E_{i\,\delta}$ in (3.30) are then calculated for mixed-effects logistic regression.

3.3.5 Centring the predictor in mixed-effects models

There are many different ways of scaling the variables in fitting a model, such as using the mean or median and possibly the standard deviation. Each of these methods leads to a different interpretation of the parameter estimates. Centring by the mean of the observations of a given predictor is a quite common approach used in practice, as the mean is a very important and useful statistical summary of data and is more stable than the median. As described in Section 3.1.3, centring by the mean in fixed-effects regression models does not change the regression coefficient estimates except the intercept. In practice, centring by the mean of a predictor is often used for a more interpretable intercept estimate. In the context of the anglerfish application, the intercept becomes the logarithm of the estimated odds ratio of a fish was successfully retained in the cod-end when its length is equal to the mean length of the observed sample in the survey.

In a multilevel random-effects modelling analysis, this situation is more complicated because of the hierarchical structure of the data. Taking a two-level data analysis for example, the hierarchical structure of the data leads to several means: the grand-mean of all the level-one observations and the group-mean for each unit at level-two (see Figure 3.4 for an illustration). A random-effects model with predictors centred around their grand means is referred to as a *grand-mean model*, while a model with predictors centred around their group means is referred to as a *group-mean model*. For comparison, a model fitted without centring predictors is referred to as a *raw-data model*. In the light of the simple random effect in the anglerfish application,

this section discusses the centring methods in the context of mixed-effects models for a data set with a two-level hierarchical structure. Without loss of generality, only one explanatory variable is considered. We use the upper case to denote the random variables and lower case to denote their observations.

From the discussion of centring in a fixed-effects ordinary regression model, the mean centring approach in an ordinary regression model does not change the estimates of all the coefficients but only that of the intercept. Figure 3.11 in Appendix 3.C showed that centering the predictor by its sample mean does not change the fitted model in terms of the fitted regression line and its confidence intervals. In the context of random-effects models, Kreft *et al.* (1995) considered two models with different centering methods as equivalent to each other if they generate the same fit, and the same prediction together with its uncertainty. However, the parameter estimates of two equivalent models need not necessarily be the same, and even the numbers of parameters can be different. In terms of the expected value and covariance matrix of the dependent variable (i.e., $\boldsymbol{y}$), two models are defined as equivalent if they produce the same $\mathbf{E}[\boldsymbol{y}]$ and $\text{Var}[\boldsymbol{y}]$. It can be shown that the raw-data model is equivalent to the grand-mean model, but that these two models are not equivalent to the group-mean model.

Without loss of generality, the centring issue is discussed in this section in terms of a full (i.e., random intercept and slope) mixed-effects logistic regression model. Given the observed explanatory variable, x_{ij}, for the jth observation in group i, the full logistic regression model is formulated as

$$y_{ij} = \text{logit}^{-1}[\beta_0 + b_{0i} + (\beta_1 + b_{1i})\, x_{ij}] + \epsilon_{ij}, \tag{3.66}$$

where the random intercept b_{0i} and random slope b_{1i} for haul i are assumed to follow a bivariate normal distribution with mean zero and variance-covariance matrix Σ_b. Then the corresponding grand-mean centred model is obtained by replacing x_{ij} in (3.66) by $x_{ij} - \bar{x}$, where $\bar{x} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} x_{ij}/n$ is the grand-mean (n is the total number of observations at level-one). Similarly the group-mean centred model is obtained by replacing x_{ij} by $x_{ij} - \bar{x}_i$, where $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$ is the group mean for the ith unit at level-two. Using the superscript $*$ for the grand-mean model and $**$ for the group-mean model, the grand-mean model can be expressed as

$$y_{ij} = \text{logit}^{-1}[\beta_0^* + b_{0i}^* + (\beta_1^* + b_{1i}^*)\, (x_{ij} - \bar{x})] + \epsilon_{ij}, \tag{3.67}$$

and the group-mean model as

$$y_{ij} = \text{logit}^{-1}[\beta_0^{**} + b_{0i}^{**} + (\beta_1^{**} + b_{1i}^{**})\,(x_{ij} - \bar{x}_i)] + \epsilon_{ij}. \tag{3.68}$$

For simplicity, the model fitted to the raw data given by (3.66) is referred to as RAW_1, the grand-mean centred model given by (3.67) is referred to as GDM_1, and the group-mean model given by (3.68) is referred to as GPM_1, with the subscript 1 meaning that the model has only level-one explanatory variables. It can be seen from the GPM_1 model that after centring by the group mean $\bar{X}_i$, the between-group variation contained in the explanatory variable X_{ij} is not included in (3.68) any more. This loss of information resulting from group-mean centring may inflate the variance estimates of the random coefficients, i.e., Σ_b. In order to re-introduce the source of between-group variation in X_{ij} in the GPM_1 given by (3.68), Burstein (1980) suggested including $\bar{X}_i$ as a level-two predictor. A model with level-two predictor is referred to as a *two-level mixed-effects model*,

$$y_{ij} = \text{logit}^{-1}[\beta_0 + b_{0i} + \beta_{10}\bar{x}_i + (\beta_{11} + b_{1i})\,x_{ij}] + \epsilon_{ij}. \tag{3.69}$$

which is referred to as RAW_2 with subscript 2 denoting a two-level mixed-effects model. For more complicated data sets with more than two levels, the multilevel model refers to mixed-effects models with level-specific predictors (higher than level one). Then as with the mixed-effects models with only level-one predictor, i.e., RAW_1, GDM_1, and GPM_1, the GDM_2 is obtained by replacing x_{ij} by $x_{ij} - \bar{x}$, and the GPM_2 is obtained by replacing x_{ij} by $x_{ij} - \bar{x}_i$.

Given the hierarchical structure of the data, it is very likely that the grand mean is not within the observations for a level-two unit. Taking the anlgerfish experimental survey data for example, Figure 3.6 gives the boxplots for all the hauls (the level-two units) and the grand mean 39.6 cm by a dotted line. As shown in this figure, there is one haul (haul ID 419 in Figure 3.6) with fish all larger than the grand-mean. In this case, the estimate of β_0 is unreliable because of extrapolating $x_{ij} - \bar{x}$ in (3.67) outside the observed range for haul i (see Enders & Tofighi, 2007, p. 126). Therefore, centring by the group mean within each level-two unit seems to be a more natural approach for a multilevel dataset.

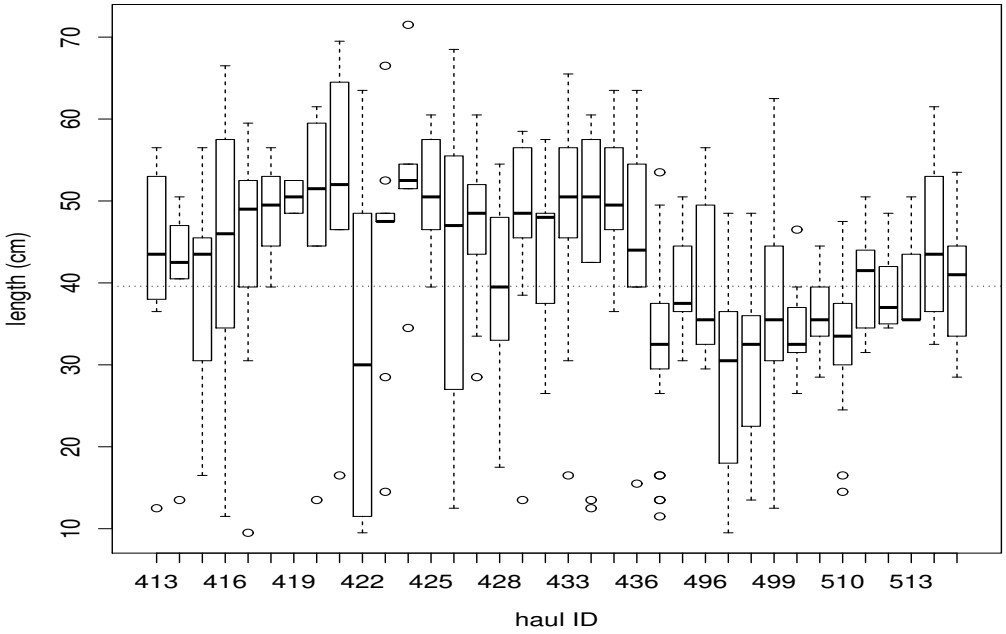


FIGURE 3.6. Boxplot of the observed length of all captured fish with respect to each haul with the dotted line for the grand mean length of all fish in 2006-2007 experimental survey data.

In addition, the explanatory variables $\bar{X}_i$ and X_{ij} are usually correlated with each other in the RAW_2 model given by (3.69), and this results in collinearity problem in the RAW_2 model. This collinearity problem usually causes bias in variance-covariance parameter estimates and standard errors of the fixed-effects regression coefficient estimates (Aitkin & Longford, 1986; Bonate, 1999). The group-mean centring approach can solve this problem, since $X_{ij} - \bar{X}_i$ and $\bar{X}_i$ are orthogonal. Centring X by its group mean in (3.69) results the following GPM_2 model,

$$y_{ij} = \text{logit}^{-1}[\beta_0 + b_{0i} + \beta_1^B \bar{x}_i + (\beta_1^W + b_{1i})\,(x_{ij} - \bar{x}_i)] + \epsilon_{ij}, \tag{3.70}$$

where the superscript B in β_1^B denotes the between-group effect and W in β_1^W denotes the within-group effect.

By group-mean centring, the GPM_2 model divides the effect of X_{ij} into two components: the between-group part meaured by $\bar{x}_i$ and the within-group part by $(x_{ij} - \bar{x}_i)$. Such a separation of the effect of X is suggested by Kreft & de Leeuw (1998) to distinguish the level-two effect from level-one characteristics, which means to dis-

tinguish the haul effect from the length of each individual fish in the context of anglerfish survey data.

The group-mean centred two-level mixed-effects model, i.e., GPM_2 given by (3.70) is a natural choice suggested by Burstein (1980) to model the between- and within-group effect in sociological and social-psychological researches, in which the group or clustering has an effect on individual behaviour, and the individual effect on the response variable should be considered in the context of the group to which that individual observation belongs. In addition, Kreft *et al.* (1995) suggested the group-mean centring model for the attenuation of the correlation between between b_{0i} and b_{1i}. The presence of this correlation may lead to unstable prediction, which will be illustrated by the woodrats survey data studied in Chapter 7. In addition, Raudenbush & Bryk (2002) concluded that the group-mean centring approach usually produces the most accurate estimate of the variance of random slope.

As discussed by Longford (1989); Plewis (1989); Hofmann & Gavin (1998); Kreft *et al.* (1995); Paccagnella (2006), centring explanatory variables in a multilevel mixed-effects data analysis is not as simple an issue as it may appear at first sight and the decision of centring explanatory variables should be made with caution. The group-mean model is usually different from the raw-data model or the grand-mean centred model. This difference is in terms of prediction and its uncertainty estimation. Kreft *et al.* (1995) have shown that GDM_1 and RAW_1 are equivalent, but they are not equivalent to GPM_1; GPM_2 is not equivalent to GDM_2 and RAW_2, while GDM_2 and RAW_2 are equivalent. The question in multilevel data analysis is whether or not to centre the predictor and which centring approach to use. The answer to these two questions should be considered in the light of the specific data set, its ecological and biological meaning and the purpose of the analysis. Last but not the least, after centring the length by group means within each haul, the interpretation of the within-haul slope, $\hat{\beta}_1^W$, is the expected change in logarithm of the odds of being retained when the fish size is 1 cm larger compared with the average size of fish captured within each haul. This interpretation is more meaningful than the slope of length in the raw-data and grand-mean models.

For the anglerfish survey data, one important assumption is that the success of retention is not considered purely as an individual effect. Instead, the length of each individual fish is compounded with the haul effect given the sampling process. The length of an individual fish is not just an individual characteristic, and it is also re-

lated to the haul effect, such as the sea environment of the swept area and the weather condition of when the haul was towed. It seems to us that a group-mean centring approach including the level-two predictor is more appropriate to distinguish the haul effect from the individual effect, which will give us a more precise estimate of the effect of length on success of retention. Therefore, the GPM_2 model is applied to the experimental survey data in Section 3.4 to estimate the net retention probability.

In Part IV, the group-mean centring approach will be applied again in estimating capture probability for one of the applications studied in Chapter 7. However, the group-mean centring approach is used again but for a different reason from the anglerfish. Unlike the anglerfish application, the choice of centring in Section 7.1 is made to stabilize the prediction in abundance estimation by eliminating the correlation between random intercept and random slope. This issue will be fully discussed in Section 7.1.1.

3.4 Application of mixed-effects logistic regression to anglerfish

This section applies the mixed-effects logistic regression model to the 2006-2007 anglerfish experimental survey data to incorporate the haul effect in estimating net retention probabilities. The experimental survey data were described in Section 2.2. The model considered for this purpose is a two-level mixed-effects logistic regression model, with haul being a level-two unit and each individual fish being a level-one unit (see Figure 3.4 for the hierarchical structure of the anglerfish survey data).

In the anglerfish experimental data used here, there are 431 anglerfish captured in total by 36 hauls, with between 2 and 47 fish per haul. The grand-mean length is 39.6 cm. The sample standard deviation of the individual fish length is 12.7 cm, the haul-specific sample has sample standard deviation that varies from 2.8 to 21.4. The standard deviation of the haul-specific mean lengths over 36 hauls is 3.4. This shows that the fish captured by different hauls vary in sizes. If haul is treated as the random effect, then the variation in the net retention probability due to length is partitioned in two parts by applying the GDM_2 given by (3.70). The between- and within-haul parts are denoted by superscript B and W, respectively. The between-haul part represents the haul effect, while the within-haul part represents the length effect of each fish in relation to its peers in the same haul. We incorporated the

between-haul effect by including the average length of all fish captured by each haul, referred to as group-mean length, and the within-haul effect via each fish's length centred on its group-mean length.

These two parts of variation in the detection probability can be modeled by a two-level mixed-effects logistic model (Kreft *et al.*, 1995): the group-mean length is a predictor at the haul-level (level-two) and the group-mean-centred length for each individual is an individual-level (level-one) predictor. A fish's success of being retained in the cod-end is determined by it being a big or small fish compared with its peers captured within the same haul, and the average size of fish captured within the same haul. Note that the total number of fish captured within each haul is also a potential level-two predictor, which can be considered as a measure of the size of each haul.

3.4.1 Estimation results

Following the top-down strategy described in Section 3.3.4, we fit a two-level mixed-effects model of the form (3.70) with all the potential explanatory variables to the 2006-2007 experimental survey data. These explanatory variables include

- a level-one predictor: the group-mean-centred length; and
- level-two predictors: the group-mean length, the total number of fish within each haul, and day or night when each haul was towed.

We use the `glmer()` function provided by the package `lme4` (Bates, D. and Maechler. M and Bolker, B., 2011) to fit mixed-effects logistic regression models. A model is fitted with all the potential explanatory variables listed above, and we refer this model as the full model. Then models with a subset of the above listed explanatory variables are fitted, and we refer to these models as nested models compared with the full model. All the models are fitted with the same random-effects structure, and then likelihood ratio tests are used to test the significance of potential predictors. The tests showed significance at 5% for the group-mean length and group-mean centred length only. Therefore, the selected random-effects model is of the following form

$$\hat{r}(l, \bar{l}_i \,|\, b_{0i}, b_{1i}) \;=\; \text{logit}^{-1}[(\beta_0 + b_{0i}) + \beta_1^B \, \bar{l}_i + (\beta_1^W + b_{1i}) \, (l - \bar{l}_i)] \quad (3.71)$$

where $\bar{l}_i$ is the sample mean length for haul i, β_0, β_1^B and β_1^W are the fixed-effects parameters, and b_{0i} and b_{1i} are the random-effects coefficients that follow a bivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix Σ_b.

The estimated variance of the random slope in (3.71) is essentially zero when assuming no correlation between random slope and intercept, i.e., a diagonal Σ_b. Note that the random slope can be thought of as an interaction effect between the length and the haul effect. Another way to model this interaction is to include an interaction term in fixed-effects regression analysis. If the random slope b_{1i} is thought of as a measure of interaction effect between haul and size of fish, then instead of being a fixed unknown number as in fixed-effects models, this quality b_{1i} is a random number from a normal distribution with mean zero and unknown variance in mixed-effects models. As described in Section 3.3.4 that a standard likelihood ratio test should not be used to test a random-effects variance parameter, we test the significance of the random slope by fitting a fixed-effects logistic regression with an interaction term between length and haul, and the fitted model shows no evidence for this interaction term at 5% significant level.

Therefore, a two-level mixed-effects model with only random intercept (b_{0i}) is the chosen model from the experimental survey data. Given the random intercept for haul i (b_{0i}), and the average length of all the fish captured by haul i ($\bar{l}_i$), the net retention probability of a fish with length l and captured by haul i is estimated by

$$\hat{r}(l, \bar{l}_i \,|\, b_{0i}) \;=\; \text{logit}^{-1}[(-3.606 + b_{0i}) + 0.125\,\bar{l}_i + 0.110\,(l - \bar{l}_i)] \quad (3.72)$$

where the estimated standard deviation of the random intercept b_{0i} is 0.298. As suggested in Section 3.3.2, adaptive Gauss-Hermite quadrature is used for more accurate approximation of the marginal likelihood, and a variety of quadrature points were tried. After trying 5, 10, 15, 20 and 25 quadrature points, estimates were found to converge at 10 quadrature points. Therefore estimates are based on the approximated likelihood from 10-points adaptive Gauss-Hermite quadrature.

The absolute goodness-of-fit of the model is measured by the HL-GOF test with the expected number of observations for each cell in Table 3.1 obtained from the expected net retention probability given in (3.65). Figure 3.7 gives the cut-points of the 10 cells for the HL-GOF test performed on the fitted model given by (3.72). The

test statistic of the HL-GOF test is 8.08 and the corresponding p-value is 0.426, which is statistically insignificant evidence for lack-of-fit of the fitted mixed-effects model given in (3.72).

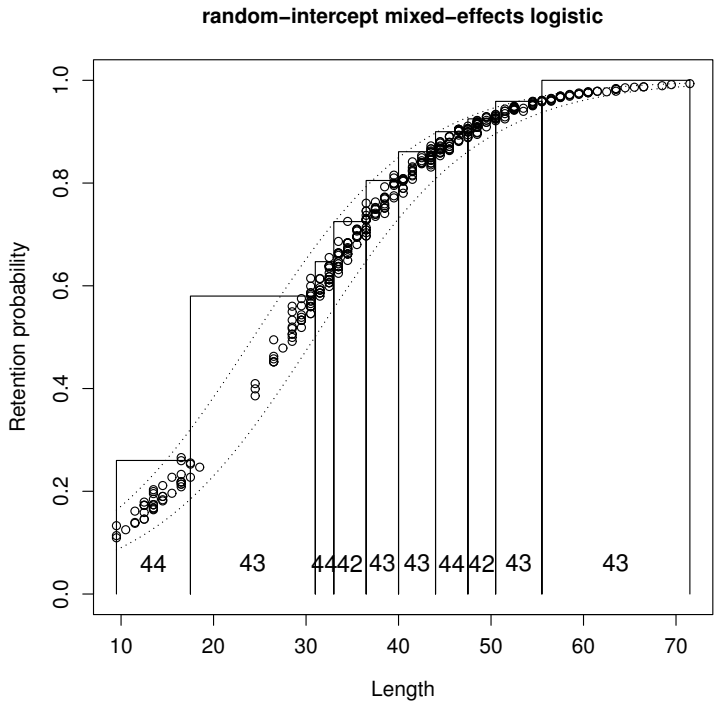


FIGURE 3.7. Plot of the fitted random-intercept mixed-effects logistic model with the circles representing the estimated net retention probabilities given the conditional modes of random intercepts, b_{0i}. The dotted line represents the minimum and maximum of the estimated retention probabilities over all hauls in the experimental survey data. The HL-GOF test described by (3.65) is also illustrated in the plot – the number at the bottom of each bin represents the size of each cell for the HL-GOF test.

Figure 3.7 also gives the estimated proportion of fish being retained in the cod-end over all hauls in the 2006-2007 experimental survey data (the circles in the figure), together with the boundaries of fitted logistic curves over all hauls (the dotted lines). The fitted logistic curves for all 36 hauls are given in Figure 3.8, where each curve is conditioned on the conditional modes of b_{0i} given the observed data. The uncertainty of the conditional modes of b_{0i} is given in Figure 3.9. These confidence intervals are based on the estimated marginal distribution of the random intercept, which is a univariate normal distribution with mean being the estimated conditional mode of b_{0i}, and its estimated standard deviation $\hat{\sigma}_0$. Figure 3.9 gives the 95% CI for each haul in the experimental survey data.

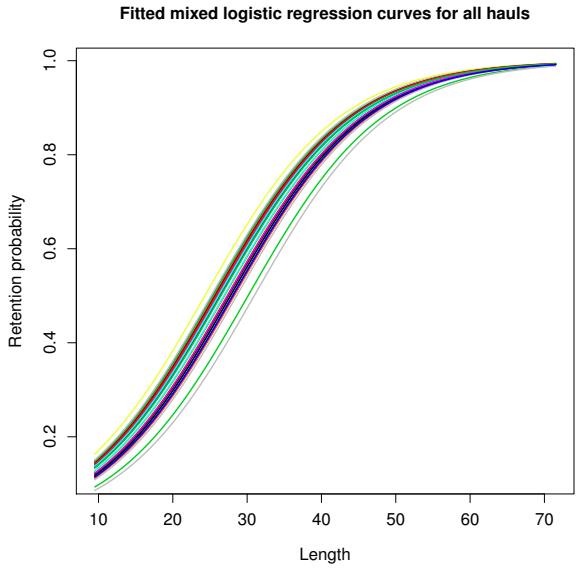


FIGURE 3.8. Plot of the fitted mixed-effects logistic regression curves for all the hauls in 2006-2007 experimental survey (each colour represents one haul), using the conditional modes of the random effects (i.e., the random intercept b_{0i}).

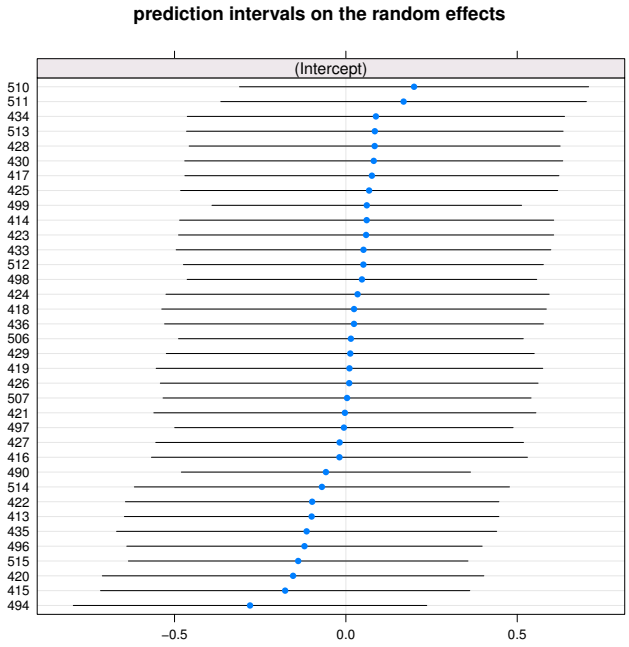


FIGURE 3.9. Plots of the conditional modes of the random intercept $\hat{b}_0$ from the fitted random-intercept multilevel mixed-effects logistic regression model given by (3.72). The uncertainty of the conditional distribution of the random effects b_0 is indicated by a line that extends $+/-1.96$ conditional standard errors in each direction from the conditional mode (shown as a blue dot).

As the sampling distribution of random effects variance estimates is usually highly skewed with unknown distribution, the standard error of $\hat{\sigma}_0$ might be a poor characterization of the uncertainty of $\hat{\sigma}_0$. Therefore, the package `lme4` (Bates, D. and Maechler. M and Bolker, B., 2011) does not present standard errors of random-effects variance estimates in its output list. When estimating the uncertainty of predicted binary response variable, there are studies conditional on the estimated variance-covariance matrix of random effects, such as Skrondal (2009). In this case, the uncertainty of the random-effects parameter estimates is not included in the uncertainty estimation of predictions. Therefore, in order to incorporate the uncertainty of the random-effects parameter estimates in the anglerfish abundance estimation described in Part III, the 2006-2007 experimental survey data are re-sampled with haul being the sampling unit, and a model of the form (3.71) is then fitted to the re-sampled experimental survey data.

3.4.2 Discussion

Given the two-level mixed-effects logistic regression model (3.72) in the previous section, in this section, we check the normality assumption of the random effects and discuss the potential sample size issues for a reliable random-effects variance estimation.

Deviations from the normality assumption of random effects might have more impact on inferences on generalized mixed-effects models than on linear mixed-effects models (Agresti *et al.*, 2004). Litière *et al.* (2008) studied this impact in particular for a mixed-effects logistic regression model, and they found that mis-specification of the random-effects distribution might lead to inconsistent maximum likelihood estimators of both random-effects and fixed-effects parameters. However, this might not be the case when random-effect variance is small, which was suggested by simulation studies. For the anglerfish experimental survey data, we check the normality assumption of the two-level mixed-effects logistic regression model by graphical analysis. Figure 3.10 gives the density plot of the conditional modes of b_{0i} together with the corresponding qqplot, both of which show that the normality assumption of the random intercept is not unreasonable. In addition, given that (3.72) has random intercept only and its estimated standard deviation, $\hat{\sigma}_0 = 0.298$, is not high, it seems to us that the parameter estimation is not likely to be affected by the small deviation from normality shown in the plots.

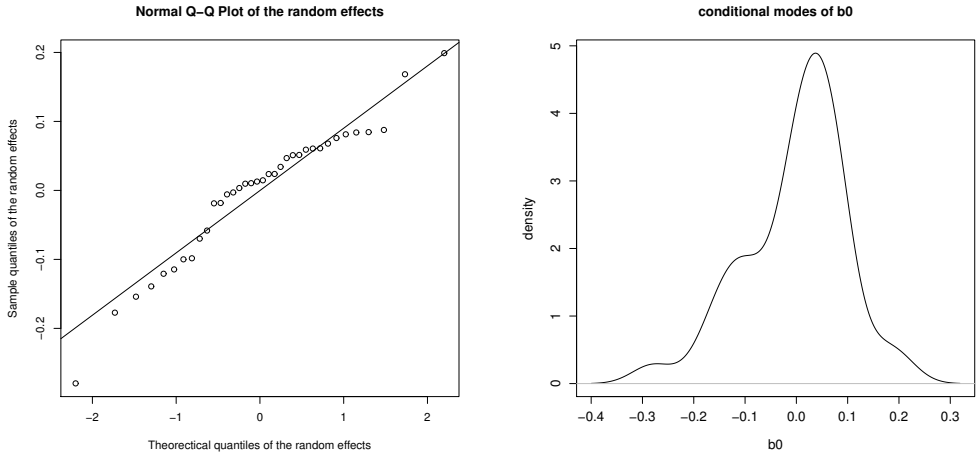


FIGURE 3.10. Plots to check the normality assumption of random effects: the left panel gives the quantile-quantile distribution plot (qq-plot) of the conditional modes for the random effects and the right panel plots the kernel density estimate of the conditional modes of b_0 for the estimated two-level mixed-effects random-intercept logistic regression model for retention probability estimation.

As described in Section 3.3.4, closed forms of the marginal likelihood for mixed-effects logistic regression models are not available. Unlike the power analysis for mixed-effects linear models (Snijders & Bosker, 1993, 1999; Maas & Hox, 2005), simulation studies are usually conducted for the effect of sample size (at both level two and level one) for mixed-effects logistic regression (see Moineddin *et al.*, 2007, for example). Their studies show that usually a larger sample size is required for mixed-effects logistic regression models than that for mixed-effect linear regression models, and the estimation of fixed-effects parameters are less sensitive to sample size than that of the random-effects variance parameters. In addition, the sample size (both level-one and level-two) should be adjusted when the probability of the binary response being 1 is low and one rule of thumb is that the expected number of successes (i.e., the binary response being one) should be larger than one.

In terms of the random-effects variance estimation, groups or clusters with few observations contribute little to estimation of random-effects variance parameters. The simulation of Moineddin *et al.* (2007) shows that for a level-one sample size of at least 30, the random-effects variance parameters are still consistently underestimated regardless of the level-two sample size. As all the hauls in the 2006-2007 experimental survey data captured less than 30 fish, and 50% of them captured less than 10 fish, it becomes apparent to us that the standard error of haul effect (i.e., $\hat{\sigma}_0$ and $\hat{\sigma}_1$) might

be under-estimated due to the small sample size within each haul. Therefore, for more reliable estimation of the haul effect, we suggest that the hauls in the experimental survey should be towed for longer and in the area where the fish density is expected to be high.

3.5 Fixed-effects vs mixed-effects models

As described in Section 3.3.4, AIC should not be used to make a choice between a mixed-effects model and the corresponding fixed-effects model. If the random effect is a part of the experimental design or sampling process, as in the anglerfish experimental survey, the observations are no longer independent within the same level-two unit, i.e., a group or cluster. In addition, if the purpose of statistical analysis is to make predictions for a different sample of the random effect, then the fixed-effects model cannot be used for this purpose because it cannot be used to draw inferences at levels of factors other than those used to fit the model.

The modelling process depends on the amount of information contained in the data, the nature of the survey designed to collect the data and last but not least, the essential biological or ecological questions of the study. In other words, we are not only looking for a model that can adequately explain the underlying process that generates the observed data, but also a model that can be used to serve our research objectives. McCullagh & Nelder (1989), p. 8, suggested a few principles to guide the modellers and the first one is that "all models are wrong; some, though, are more useful than others and we should seek those", which is the title of a section of the paper written by Box (1979). Mixed-effects models with haul as a random effect are more useful than fixed-effects models for the estimation of anglerfish abundance from the survey data described here.

Appendices – **Chapter 3**

3.A Newton-Raphson algorithm involved in Fisher scoring method

To start with, here is a reminder of the first-order Taylor expansion, which gives an approximation of $f(x)$ by

$$\begin{aligned} f(x) &\approx f(x_0) + (x - x_0) \left.\frac{df(x)}{dx}\right|_{x=x_0} \\ &= f(x_0) + f'(x_0)(x - x_0). \end{aligned}$$

Here the aim is to solve $f(x) = 0$. Say x^* is the solution, it then follows

$$\begin{aligned} 0 &= f(x^*) \approx f(x_0) + \left.\frac{df(x)}{dx}\right|_{x=x_0} (x^* - x_0) \\ 0 &= f(x_0) + (x^* - x_0) f'(x_0). \end{aligned}$$

Therefore,

$$x^* = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

The Newton-Raphson method is an iterative algorithm, with x_0 as a starting value and for the tth iteration,

$$x_{t+1} = x_t - f(x_t)/f'(x_{t-1}). \tag{3.73}$$

If the series does converge, then this algorithm is repeated until the series converges and the final value is taken as the solution of $f(x) = 0$.

3.B Review of least squares estimation

In order to illustrate the idea of the iteratively re-weighted least squares estimator (3.24) in Section 3.1.2, this section starts with the derivation of ordinary least squares estimator (OLS) of an ordinary linear model, based on which a more efficient estima-

tor in the case of constant error variance, the weighted least squares estimator (WLS), is introduced. Then for a generalized linear model, the IRLS estimator can be considered as an iterative WLS approximation of the model with a linearized response variable, which is referred to as the adjusted response variable, or working response variable.

3.B.1 Ordinary least squares estimation

For an ordinary linear model given as

$$\underset{n\times 1}{\boldsymbol{Y}} \;=\; \underset{n\times q}{\boldsymbol{X}}\;\underset{q\times 1}{\boldsymbol{\beta}} + \;\underset{n\times 1}{\boldsymbol{\epsilon}}\,, \tag{3.74}$$

the residual is $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{\epsilon} \sim$ normal $(\mathbf{0},\ \boldsymbol{I}_n\sigma^2)$. The least squares approach is to minimize the sum of squared residuals to obtain a line that is a best 'fit' for the data. Written in terms of matrices, the sum of squared residuals is $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$, and $\boldsymbol{\epsilon}^T\,\boldsymbol{\epsilon} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. Then applying the least squares criterion, in order to find out the value of $\boldsymbol{\beta}$ that minimizes $\boldsymbol{\epsilon}^T\,\boldsymbol{\epsilon}$, the first-order partial derivative of $\boldsymbol{\epsilon}^T\,\boldsymbol{\epsilon}$ with respect to $\boldsymbol{\beta}$ is derived as follows

$$\frac{\partial \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}}{\partial \boldsymbol{\beta}} = 2\boldsymbol{\epsilon}^T\frac{\partial \boldsymbol{\epsilon}}{\partial \boldsymbol{\beta}} = 2(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(-\boldsymbol{X}),$$

of which the right side is set to zero, i.e, $2(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(-\boldsymbol{X}) = \mathbf{0}$. It follows that the OLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{ols} \;=\; (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \quad \text{if } (\boldsymbol{X}^T\boldsymbol{X})^{-1} \text{ exists.} \tag{3.75}$$

Then it can be easily shown that

$$\text{Var}\,[\hat{\boldsymbol{\beta}}_{ols}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{Var}\,[\boldsymbol{Y}]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \tag{3.76}$$

To check that $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$ reaches a minimum at $\hat{\boldsymbol{\beta}}_{ols}$ we check that the second-order derivative at $\hat{\boldsymbol{\beta}}_{ols}$,

$$\left.\frac{\partial^2(\boldsymbol{\epsilon}^T\boldsymbol{\epsilon})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right|_{\hat{\boldsymbol{\beta}}_{ols}} \;=\; 2\boldsymbol{X}^T\boldsymbol{X} \quad \text{is positive-definite for } \forall\ \boldsymbol{X}.$$

Therefore, (3.75) is the least squares estimator of an ordinary linear model. Furthermore, for an ordinary linear model, it can be shown that the OLS estimator, $\hat{\boldsymbol{\beta}}_{ols}$, is also the maximum likelihood estimator (MLE). For a linear model, $\boldsymbol{Y} \sim$ Normal$(\boldsymbol{X}\boldsymbol{\beta},\, \sigma^2\boldsymbol{I})$, and the likelihood function is

$$L(\boldsymbol{\beta}; \boldsymbol{Y}) \;\; = \;\; (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\right].$$

Then the the log-likelihood function as

$$l(\boldsymbol{Y}; \boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}, \tag{3.77}$$

where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ and n is the number of observation.

Therefore, the MLE can be obtained by maximizing (3.77) as a function of $\boldsymbol{\beta}$ for given $\boldsymbol{X}$ and $\boldsymbol{Y}$. Take the first-order partial derivative of (3.77) with respect to $\boldsymbol{\beta}$ to obtain

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{2\sigma^2}\frac{\partial(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{2\sigma^2}\frac{\partial \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}}{\partial \boldsymbol{\beta}},$$

from which it follows that MLE of $\boldsymbol{\beta}$ also minimizes the sum of squared residuals. Therefore $\hat{\boldsymbol{\beta}}_{ols}$ is also the MLE of $\boldsymbol{\beta}$.

3.B.2 Weighted least squares estimation

If an observation y_i is more important than the others in terms of more accurate information contained in y_i, then associating an appropriate weight w_i to the sum of squared residuals of y_i leads to a more efficient estimator of $\boldsymbol{\beta}$. The least squares criterion is applied to minimize the following weighted sum of the squared residuals,

$$\sum_{i=1}^{n} w_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 \tag{3.78}$$

where the weight is assumed known and $w_i = 1/\text{Var}\,[y_i]$, i.e., the more information contained in the ith observation, the more weight is assigned to it. Let $\boldsymbol{W}$ denote a diagonal matrix with w_i as the ith element on the diagonal. It follows that $\boldsymbol{W}^T = \boldsymbol{W}$ and $\boldsymbol{W} \;=\; \boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2}$. Then the weighted sum of the squared residuals can be

written as

$$\begin{aligned} & \sum_{i=1}^{n} w_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \\ = & \sum_{i=1}^{n} (w_i^{1/2} y_i - w_i^{1/2} \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \\ = & \left(\underbrace{\boldsymbol{W}^{1/2}\boldsymbol{Y}}_{\boldsymbol{Y}_{\boldsymbol{w}}} - \underbrace{\boldsymbol{W}^{1/2}\boldsymbol{X}}_{\boldsymbol{X}_{\boldsymbol{w}}} \boldsymbol{\beta} \right)^T \left(\underbrace{\boldsymbol{W}^{1/2}\boldsymbol{Y}}_{\boldsymbol{Y}_{\boldsymbol{w}}} - \underbrace{\boldsymbol{W}^{1/2}\boldsymbol{X}}_{\boldsymbol{X}_{\boldsymbol{w}}} \boldsymbol{\beta} \right). \end{aligned}$$

Let $\boldsymbol{Y}_{\boldsymbol{w}} = \boldsymbol{W}^{1/2}\boldsymbol{Y}$ and $\boldsymbol{X}_{\boldsymbol{w}} = \boldsymbol{W}^{1/2}\boldsymbol{X}$ as shown above. Then the weighted sum of the squared residuals is

$$(\boldsymbol{Y}_{\boldsymbol{w}} - \boldsymbol{X}_{\boldsymbol{w}}\boldsymbol{\beta})^T (\boldsymbol{Y}_{\boldsymbol{w}} - \boldsymbol{X}_{\boldsymbol{w}}\boldsymbol{\beta}). \tag{3.79}$$

Applying the OLS estimator (3.75) to the above weighted sum treating $\boldsymbol{Y}_{\boldsymbol{w}}$ and $\boldsymbol{X}_{\boldsymbol{w}}$ as the corresponding variables, it follows that the weighted least squares estimator of $\boldsymbol{\beta}$ is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{wls} &= \left(\boldsymbol{X}_{\boldsymbol{w}}{}^T \boldsymbol{X}_{\boldsymbol{w}}\right)^{-1} \boldsymbol{X}_{\boldsymbol{w}}{}^T \boldsymbol{Y}_{\boldsymbol{w}} \\ &= \left(\boldsymbol{X}^T \boldsymbol{W}^{1/2} \boldsymbol{W}^{1/2} \boldsymbol{X}_{\boldsymbol{w}}\right)^{-1} \boldsymbol{W}^{1/2} \boldsymbol{X}^T \boldsymbol{W}^{1/2} \boldsymbol{Y} \\ &= \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Y}, \end{aligned} \tag{3.80}$$

from which it can be easily shown that

$$\begin{aligned} \text{Var}\,[\hat{\boldsymbol{\beta}}_{wls}] &= \left\{ \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \right\} \text{Var}\,[\boldsymbol{Y}] \left\{ \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \right\}^T \\ &= \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \cancel{\boldsymbol{W}\boldsymbol{W}^{-1}} \boldsymbol{W} \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \cancel{\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1}} \\ &= \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1}. \end{aligned} \tag{3.81}$$

Let $\boldsymbol{V}$ denote a diagonal matrix with the ith element of the diagonal $v_i = \text{Var}\,[y_i]$. Then $\boldsymbol{V}^{-1} = \boldsymbol{W}$ and $\hat{\boldsymbol{\beta}}_{wls}$ can also be written as

$$\hat{\boldsymbol{\beta}}_{wls} = \left(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{Y}$$

and

$$\text{Var}\,[\hat{\boldsymbol{\beta}}_{wls}] = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}.$$

Compared with $\text{Var}\,[\hat{\boldsymbol{\beta}}_{ols}]$ given in (3.76), for known $\boldsymbol{V}$, $\hat{\boldsymbol{\beta}}_{wls}$ is more efficient than $\hat{\boldsymbol{\beta}}_{ols}$ in most cases.

Consider a generalized linear model in the form of $\boldsymbol{Y} = g^{-1}(\boldsymbol{\eta}) + \boldsymbol{\epsilon}$, where g^{-1} is the inverse of the link function such that $\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}$. Then the weight matrix $\boldsymbol{W}$ is defined as

$$\boldsymbol{W} = \boldsymbol{V}^{-1}\left(\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}}\right)^2,$$

where $\boldsymbol{V}$ is a diagonal matrix with the ith element on the diagonal $v_i = \text{Var}\,[y_i]$. The complication for generalized linear models lies in that $\boldsymbol{W}$ depends on $\boldsymbol{\beta}$, and so an iterative process is required to obtain an estimate of $\boldsymbol{\beta}$, and at each iteration, the weight matrix is calculated using the estimates of $\boldsymbol{\beta}$ from the previous iteration. The estimator obtained by this process is called the *iteratively re-weighted least squares* estimator, and the algorithm is known as the IRLS algorithm. The derivation showing that the IRLS algorithm produces the maximum likelihood estimator for a generalized linear model is given in Section 3.1.2. Note that although the proof is for a linear logistic regression model, for generalized linear models with other link functions, the IRLS algorithm also produces maximum likelihood estimates.

3.C Centring predictors in simple linear regression

This section is focused on some analytical discussion of centring a predictor variable by its mean (the mean-centring approach) in a simple linear model, together with a list of practical reasons for the mean-centring approach. Linear models (LMs) are for continuous normal data with constance variance: $Y_k \sim N(\mu_k, \sigma^2)$ and

$$\mu_k = \sum_{j=1}^{q} x_{kj}\,\beta_j = \boldsymbol{x}_k^T\boldsymbol{\beta},$$

where $\boldsymbol{x}_k = (1, x_{k1}, \ldots, x_{k\,q-1})^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{q-1})^T$. It is assumed that Y_k are independent from each other, $k = 1, 2, \ldots, n$. Let the superscript * denote the centred LM. This appendix shows why centring the predictor reduces the correlation

between the intercept estimate ($\hat{\beta}_0$) and slope estimates ($\hat{\beta}_j$, $j = 1, 2, \ldots, q-1$) to zero.

The non-centred LM and centred LM are give as follows:

$$y_k = \boldsymbol{x}_k^T \boldsymbol{\beta} + \epsilon_k, \tag{3.82}$$

$$y_k = \boldsymbol{x}_k^{*T} \boldsymbol{\beta}^* + \epsilon_k, \tag{3.83}$$

for $k = 1, 2, \ldots, n$, and

$$\boldsymbol{x}_k^{*T} = (1, x_{k1} - \bar{x}_1, \ldots, x_{k\,q-1} - \bar{x}_{q-1}),$$

where $\bar{x}_p = \frac{1}{n} \sum_{k=1}^{n} x_{k\,p}$, $p = 1, 2, \ldots, q-1$.

We use the following notation for the model matrix,

$$\boldsymbol{x} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1\,q-1} \\ 1 & x_{21} & \ldots & x_{2\,q-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{n\,q-1} \end{pmatrix},$$

and

$$\boldsymbol{x}^* = \begin{pmatrix} 1 & x_{11} - \bar{x}_1 & \ldots & x_{1\,q-1} - \bar{x}_{q-1} \\ 1 & x_{21} - \bar{x}_1 & \ldots & x_{2\,q-1} - \bar{x}_{q-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} - \bar{x}_1 & \ldots & x_{n\,q-1} - \bar{x}_{q-1} \end{pmatrix}.$$

For the non-centred model, $\text{Var}\,[\hat{\boldsymbol{\beta}}] = \sigma^2 \, (\boldsymbol{x}^T \boldsymbol{x})^{-1}$. However, σ^2 is generally unknown and an estimate of it, s^2, is used. Therefore, $\text{Var}\,[\hat{\boldsymbol{\beta}}]$ is estimated by

$$\widehat{\text{Var}}\,[\hat{\boldsymbol{\beta}}] = s^2 \, (\boldsymbol{x}^T \boldsymbol{x})^{-1} \tag{3.84}$$

where s^2 is an unbiased estimate of σ^2 and

$$s^2 = \frac{(\boldsymbol{y} - \boldsymbol{x}\hat{\boldsymbol{\beta}})^T (\boldsymbol{y} - \boldsymbol{x}\hat{\boldsymbol{\beta}})}{n - q}$$

where $q = \dim(\boldsymbol{\beta})$ and s^2 is a scalar, and given (3.84), the variance-covariance structure of $\hat{\boldsymbol{\beta}}$ is determined by $\boldsymbol{x}^T\boldsymbol{x}$. Similarly, for the centred LM, the variance-covariance structure of $\hat{\boldsymbol{\beta}}$ is determined by $\boldsymbol{x}^{*T}\boldsymbol{x}^*$.

Given the above expressions for the non-centred LM, the variance-covariance structure of the parameter estimates $\hat{\boldsymbol{\beta}}^*$ for the centred LM is

$$
\begin{aligned}
&\boldsymbol{x}^{*T}\boldsymbol{x}^* \\
&= \begin{pmatrix}
n & \sum_{k=1}^n (x_{k1} - \bar{x}_1) & \cdots & \sum_{k=1}^n (x_{k\,q-1} - \bar{x}_{q-1}) \\
\sum_{k=1}^n (x_{k1} - \bar{x}_1) & \sum_{k=1}^n (x_{k1} - \bar{x}_1)^2 & \cdots & \sum_{k=1}^n (x_{k1} - \bar{x}_1)(x_{k\,q-1} - \bar{x}_{q-1}) \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{k=1}^n (x_{k\,q-1} - \bar{x}_{q-1}) & \sum_{k=1}^n (x_{k1} - \bar{x}_1)(x_{k\,q-1} - \bar{x}_{q-1}) & \cdots & \sum_{k=1}^n (x_{k\,q-1} - \bar{x}_{q-1})^2
\end{pmatrix} \\
&= \begin{pmatrix}
n & 0 & \cdots & 0 \\
0 & \sum_{k=1}^n (x_{k1} - \bar{x}_1)^2 & \cdots & \sum_{k=1}^n (x_{k1} - \bar{x}_1)(x_{k\,q-1} - \bar{x}_{q-1}) \\
\vdots & \vdots & \ddots & \vdots \\
0 & \sum_{k=1}^n (x_{k1} - \bar{x}_1)(x_{k\,q-1} - \bar{x}_{q-1}) & \cdots & \sum_{k=1}^n (x_{k\,q-1} - \bar{x}_{q-1})^2
\end{pmatrix}
\end{aligned}
$$

which shows the block diagonal structure of $\boldsymbol{x}^{*T}\boldsymbol{x}^*$. In the case of only one explanatory variable, then $\boldsymbol{x}^{*T}\boldsymbol{x}^*$ becomes diagonal. Therefore, the covariance between $\hat{\beta}_0^*$ and $\hat{\beta}_i^*$, $i = 1, 2, \ldots, q-1$ is zero.

Further notes for generalized linear models: given sample size n, the variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ is (see McCullagh & Nelder, 1989, p.119)

$$\text{Var}\,[\hat{\boldsymbol{\beta}}] \;=\; (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\{1 + O(n^{-1})\}, \tag{3.85}$$

where $\boldsymbol{W}$ is a diagonal matrix of weights, and for a generalized linear model with binary response data, $\boldsymbol{W}$ was given in (3.14). Because $\boldsymbol{W}$ is a diagonal matrix, the covariance structure (i.e., the off-diagonal elements of the variance matrix) of $\hat{\boldsymbol{\beta}}$ is also determined by the product $\boldsymbol{X}^T\boldsymbol{X}$. It follows that, as for the linear models, if the predictors X_i are centred by their mean $\bar{X}_i$, $1 = 1, 2, \ldots, q-1$, then the correlations between the estimates of intercept and slopes become zero.

The above gives some analytical considerations of the effect of centring. It shows that the mean-centring approach does not change the overall fit of the model for a given sample (identical lines for the fitted centred and non-centred modes in Figure 3.11), neither does it change the meaning or magnitude of the slope (only changes the intercept estimates). It removes the correlation between the estimates of intercept

and slopes. In addition, there are practical reasons for the mean-centring approach, such as a more meaningful interpretation of the intercept estimates in practice, and usually faster convergence when using IRLS algorithm.

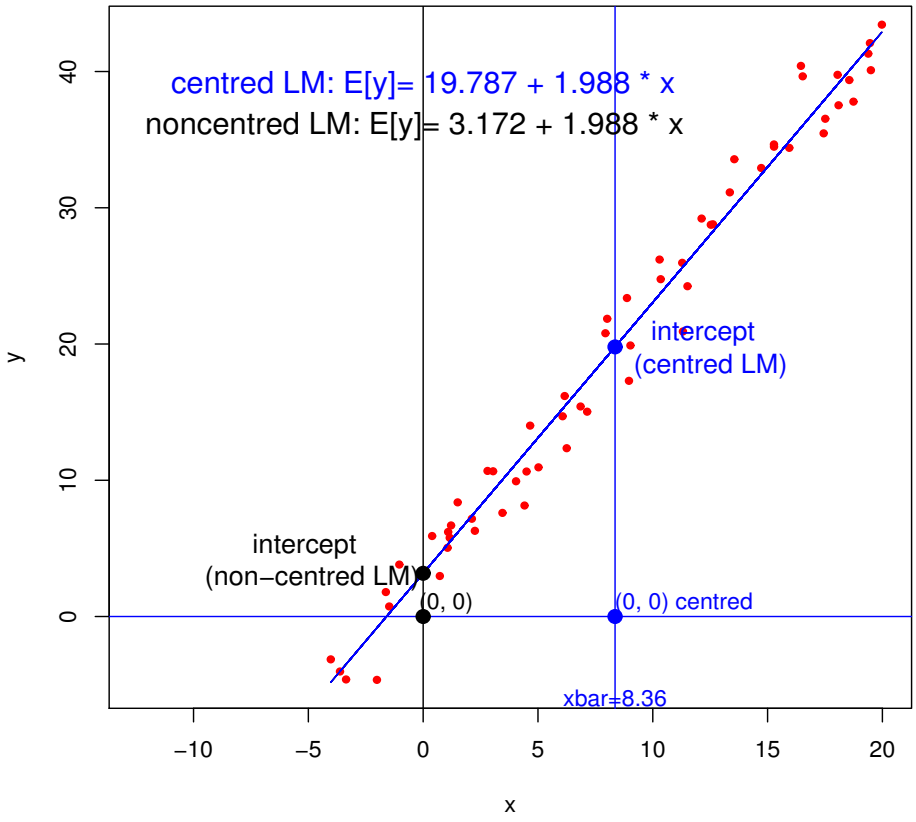


FIGURE 3.11. Plot of fitted LMs without centring (black line) and with centering (blue line) on the simulated data (red dots). Note that the blue line is on top of the black line and so only a blue line can be seen in the plot. After centering, the x-axis stays the same, but the y-axis moves rightward by $\bar{x}$. It follows that the fitted centred-LM always goes through the value of y predicted at $\bar{x}$, which can be thought of as $\bar{y}$, an estimate of population mean, μ_y. Therefore, given another sample of x, the estimated slope will pivot around about $\bar{y}$ (see Figure 3.12 for an illustration). This will lead to the estimated covariance between $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ being very close to zero (8.15×10^{-17}), which means the theoretical value is zero. However, before centring the estimated correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$ is -0.76.

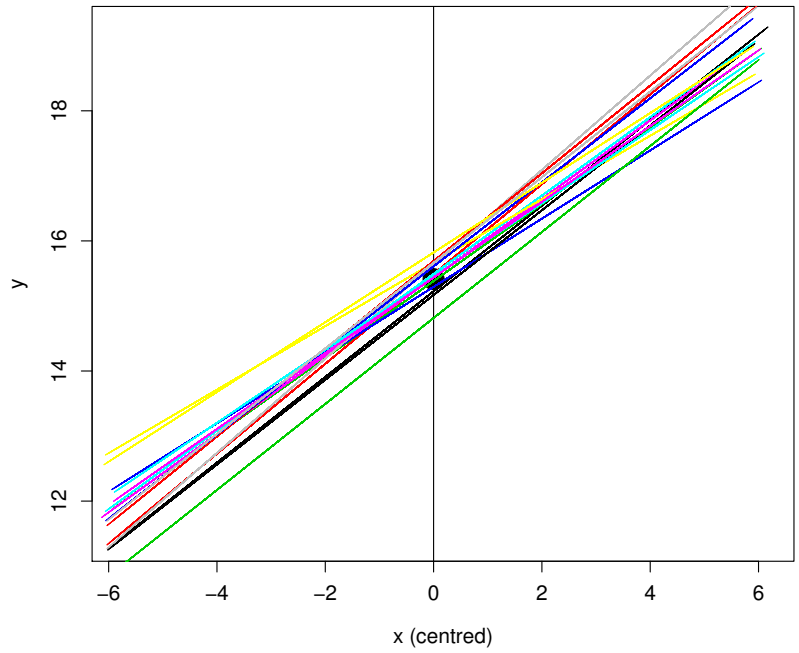


FIGURE 3.12. A simulation study to show the estimated linear regression curves pivot about the population mean of the response variable, which is represented by the solid black dot in the plot.

3.D Fisher information matrix

Let $\boldsymbol{\theta}$ denote the parameter vector in the log-likelihood function denoted by l. For simplicity, the derivation is considered for the case with only one data point, i.e., y is a scalar in the likelihood $l(\boldsymbol{\theta}; y)$. For simplicity, the random variable Y is a continuous random variable varying from $-\infty$ to $+\infty$.with probability density function $f(y; \boldsymbol{\theta})$ for a single observation y.

However, for the case with n-dimensional observation vector, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ where n is the sample size, the results shown in this section are also valid given that the likelihood L of $\boldsymbol{\theta}$ based on $\boldsymbol{y}$ is the product of the likelihoods for each observation y_i, i.e.,

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} L(\boldsymbol{\theta}; y_i).$$

Based on this all the derivation shown below can easily be extended for the case with $\boldsymbol{y}$. To start with, some regularity conditions of the probability density function $f(y;\boldsymbol{\theta})$ are assumed, and these conditions hold for most distributions. The score vector is a random vector defined as

$$\boldsymbol{S}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}}.$$

This section starts with the derivation of some important properties of the score vector, which are later used in the proof of Cramér-Rao inequality. The Cramér-Rao lower bound is used as a variance estimate of the maximum likelihood estimator.

3.D.1 The score vector and its properties

Note that for a single observation y, the density function $f(y;\boldsymbol{\theta})$ and likelihood function $L(\boldsymbol{\theta};y)$ are interchangeable, therefore, for continuous y defined on $-\infty$ to ∞,

$$\int_{-\infty}^{+\infty} L(\boldsymbol{\theta};y)d\,y = 1. \tag{3.86}$$

Then take the derivative of the above equation with respect to $\boldsymbol{\theta}$,

$$\int_{-\infty}^{+\infty} \frac{\partial L(\boldsymbol{\theta};y)}{\partial \boldsymbol{\theta}} d\,y = \mathbf{0}. \tag{3.87}$$

Multiply and divide (3.87) by $f(y;\boldsymbol{\theta})$,

$$\int_{-\infty}^{+\infty} \frac{1}{f(y;\boldsymbol{\theta})} \frac{\partial L(\boldsymbol{\theta};y)}{\partial \boldsymbol{\theta}} f(y;\boldsymbol{\theta})d\,y = \mathbf{0},$$

where $f(y;\boldsymbol{\theta}) = L(\boldsymbol{\theta};y)$, and it follows that

$$\begin{aligned} \mathbf{0} &= \int_{-\infty}^{+\infty} \frac{\partial \log\left[L(\boldsymbol{\theta};y)\right]}{\partial \boldsymbol{\theta}} f(y;\boldsymbol{\theta})d\,y \\ &= \mathbf{E}\left[\frac{\partial l(\boldsymbol{\theta};y)}{\partial \boldsymbol{\theta}}\right]. \end{aligned} \tag{3.88}$$

Therefore, the mean of the score vector is zero,

$$\mathbf{E}[\boldsymbol{S}(\boldsymbol{\theta})] = \mathbf{0}. \tag{3.89}$$

Finally, taking the partial derivative of (3.88) with respect to $\boldsymbol{\theta}$,

$$\begin{aligned}
\mathbf{0} &= \int_{-\infty}^{+\infty} \left\{ f(y;\boldsymbol{\theta}) \left[\frac{\partial l(\boldsymbol{\theta};y)}{\partial \boldsymbol{\theta}}\right] \left[\frac{\partial l(\boldsymbol{\theta};y)}{\partial \boldsymbol{\theta}^T}\right] + f(y;\boldsymbol{\theta}) \frac{\partial^2 l(\boldsymbol{\theta};y)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \right\} d\,y \\
&= \int_{-\infty}^{+\infty} \boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})^T f(y;\boldsymbol{\theta}) d\,y + \int_{-\infty}^{+\infty} \frac{\partial^2 l(\boldsymbol{\theta};y)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} f(y;\boldsymbol{\theta}) d\,y \\
&= \mathbf{E}\left[\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})^T\right] + \mathbf{E}\left[\frac{\partial^2 l(\boldsymbol{\theta};y)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right]. \qquad (3.90)
\end{aligned}$$

The Fisher information $\boldsymbol{I}(\boldsymbol{\theta})$ is defined as

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\mathbf{E}\left[\frac{\partial^2 l(\boldsymbol{\theta};y)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right]. \qquad (3.91)$$

Then (3.90) can be written as

$$\mathbf{E}\left[\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})^T\right] = \boldsymbol{I}(\boldsymbol{\theta}), \qquad (3.92)$$

and so, using (3.89), it follows that

$$\text{Var}\left[\boldsymbol{S}(\boldsymbol{\theta})\right] = \mathbf{E}\left[(\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})^T\right] = \boldsymbol{I}(\boldsymbol{\theta}) \qquad (3.93)$$

If we define the Hessian matrix as the matrix of second-order derivatives of the log-likelihood with respect to $\boldsymbol{\theta}$,

$$\boldsymbol{H}(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta};y)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \qquad (3.94)$$

then the Fisher information matrix is

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\mathbf{E}[\boldsymbol{H}(\boldsymbol{\theta})],$$

which is also known as the expected Fisher information matrix. For a sample of n observations, i.e., $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, the Fisher information matrix is

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = -\mathbf{E}\left[\frac{\partial^2 l(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right]. \qquad (3.95)$$

Note that it is not always possible to calculate $\boldsymbol{I}_n(\boldsymbol{\theta})$ due to the difficulty in obtaining the expectation in (3.95). However, the *observed Fisher information* matrix is usually easier to calculate. Let $\boldsymbol{i}_n(\boldsymbol{\theta})$ denote the observed Fisher information matrix for a

sample of size n. $\boldsymbol{i}_n(\boldsymbol{\theta})$ is defined as

$$\boldsymbol{i}_n(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}.$$

However, the parameter $\boldsymbol{\theta}$ is unknown, and then the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$, an estimate of $\boldsymbol{\theta}$, is

$$\boldsymbol{i}_n(\hat{\boldsymbol{\theta}}) = -\left.\frac{\partial^2 l(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\hat{\boldsymbol{\theta}}}. \tag{3.96}$$

The next section explains the reason for using the inverse of (3.96) as an estimate of the variance-covariance matrix for a maximum likelihood estimator, which has been applied in Section 3.1.4 in the extended fixed-effects logistic regression models.

3.D.2 Cramér-Rao inequality and Fisher information

Assume that $\hat{\boldsymbol{\theta}}(y)$ is an unbiased estimator of $\boldsymbol{\theta}$, i.e.,

$$\mathbf{E}[\hat{\boldsymbol{\theta}}(y)] = \boldsymbol{\theta}, \quad \text{i.e.,} \int \hat{\boldsymbol{\theta}}(y) f(y; \boldsymbol{\theta})\, d\, y = \boldsymbol{\theta}. \tag{3.97}$$

Note that $\hat{\boldsymbol{\theta}}(y)$ is a function of data which is independent of $\boldsymbol{\theta}$. Then taking the derivative of (3.97) on both sides with respect to $\boldsymbol{\theta}$ gives

$$\int \hat{\boldsymbol{\theta}}(y) \frac{\partial f(y; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\, d\, y = \mathbf{1}.$$

Multiplying and dividing the above equation by $f(y; \boldsymbol{\theta})$ gives

$$\begin{aligned} \mathbf{1} &= \int \hat{\boldsymbol{\theta}}(y) \frac{1}{f(y; \boldsymbol{\theta})} \frac{\partial f(y; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y; \boldsymbol{\theta})\, d\, y \\ &= \int \hat{\boldsymbol{\theta}}(y) \frac{\partial \log f(y; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y; \boldsymbol{\theta})\, d\, y \\ &= \mathbf{E}\left[\hat{\boldsymbol{\theta}}(y) \boldsymbol{S}(\boldsymbol{\theta})\right]. \end{aligned} \tag{3.98}$$

Applying the property (3.89) of the score vector, the covariance between the unbiased estimator and the score vector can be derived as

$$\begin{aligned} \text{cov}\left[\hat{\boldsymbol{\theta}}(y),\, \boldsymbol{S}(\boldsymbol{\theta})\right] &= \mathbf{E}[\hat{\boldsymbol{\theta}}(y)\, \boldsymbol{S}(\boldsymbol{\theta})] - \mathbf{E}[\hat{\boldsymbol{\theta}}(y)]\mathbf{E}[\boldsymbol{S}(\boldsymbol{\theta})] \\ &= \mathbf{E}[\hat{\boldsymbol{\theta}}(y)\, \boldsymbol{S}(\boldsymbol{\theta})] = \mathbf{1} \end{aligned} \tag{3.99}$$

Applying the Cauchy-Schwarz inequality

$$[\text{cov}(X,\, Y)]^2 \leq \text{Var}\,[X]\text{Var}\,[Y]$$

to cov $\left[\hat{\boldsymbol{\theta}}(y),\, \boldsymbol{S}(\boldsymbol{\theta})\right]$, together with (3.99) and (3.93) gives

$$\mathbf{1} = \text{cov}\left[\hat{\boldsymbol{\theta}}(y),\, \boldsymbol{S}(\boldsymbol{\theta})\right] \leq \text{Var}\,[\hat{\boldsymbol{\theta}}(y)]\ \text{Var}\,[\boldsymbol{S}(\boldsymbol{\theta})]. \tag{3.100}$$

Therefore,

$$\text{Var}\,[\hat{\boldsymbol{\theta}}(y)] \geq \boldsymbol{I}(\boldsymbol{\theta})^{-1},$$

which is known as the Cramér-Rao inequality and the inverse Fisher information matrix is the Cramér-Rao lower bound. In practice, for a sample of size n, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, where y_i are i.i.d. distributed, the Cramér-Rao inequality is

$$\text{Var}\,[\hat{\boldsymbol{\theta}}(\boldsymbol{y})] \geq \boldsymbol{I}(\boldsymbol{\theta})^{-1}/n \tag{3.101}$$

Further, an unbiased estimator is called an efficient estimator if this variance reaches the Cramér-Rao lower bound. Note that a maximum likelihood estimator is not necessarily efficient. However, one important property of the maximum likelihood estimator is that it is asymptotically efficient. In other words, a maximum likelihood estimator reaches the Cramér-Rao lower bound asymptotically. Given a large sample size n, the Cramér-Rao lower bound (i.e., the inverse of the Fisher information matrix $\boldsymbol{I}_n(\boldsymbol{\theta})^{-1}$) is used as an estimate of parameter variance. In practice, the calculation of the Fisher information involves the unknown parameter values, which are not obtainable in most cases. However, as n approaches infinity, the limit of $\boldsymbol{i}_n(\hat{\boldsymbol{\theta}})/n$ is $\boldsymbol{I}(\boldsymbol{\theta})$, where $\boldsymbol{i}_n(\hat{\boldsymbol{\theta}})$ is given by (3.96). Therefore, the observed information matrix is usually used as an approximation to the Fisher information matrix, and the inverse of the observed Fisher information matrix is then used as an estimate of variance-covariance matrix for a maximum likelihood estimator.

3.E Chi-squared goodness-of-fit test

The null hypothesis of the chi-squared goodness-of-fit (GOF) test is that the observed frequencies of the events for sampled data are consistent with a specified distribution, and the alternative hypothesis is they are not. The test proceeds by 1) specifying k mutually exclusive classes of the random variable, 2) finding out the observed frequencies (O_i) of the events falling in the class i, and 3) calculating the expected frequencies (E_i) when the null hypothesis is true. The test statistic is calculated as

$$\chi^2 \;=\; \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

Let p_i denote the probability that an event falls into class i and $\boldsymbol{X} = (O_1, \ldots, O_k)^T$, then we have

$$\boldsymbol{X} \sim \text{Multinomial}(N,\, \boldsymbol{p}),$$

where $\boldsymbol{p} = (p_1, \ldots, p_k)^T$ and $N = \sum_{i=1}^{k} O_i$. The χ^2 test statistic can be written as

$$\chi^2 \;=\; \sum_{i=1}^{k} \frac{(O_i - Np_i)^2}{Np_i}.$$

The test statistic asymptotically (i.e., as N approaches infinity) follows a χ^2 distribution with $k - 1$ degree of freedom if the null hypothesis is true.

Part III

Anglerfish abundance estimation

Chapter 4

Anglerfish abundance estimation

Key Idea: use a HT-like estimator to estimate the anglerfish abundance with the estimated capture probability

Abundance estimation is a key component in stock assessment and the abundance estimates by age class provide information required to understand further the age structure, population growth and other details of the fish stock studied. Data collected from bottom trawl surveys can provide a reliable indication of stock size and distribution (Chen *et al.*, 2004; Rago, 2005; Belcher & Jennings, 2009), and therefore play an important role in monitoring commercial species, and setting total allowable catch (TAC) limits.

In this chapter, abundance estimators for the anglerfish survey data are developed using a Horvitz-Thompson-like estimator, and the 2007 abundance survey data are used to illustrate the abundance estimation method. The estimators use the estimated capture probability in Part II, which includes the herding factor (described in Chapter 2) and the footrope net retention probability (studied in Chapter 3). To start with, Section 4.1 describes the 2007 abundance survey data used throughout this chapter, and Section 4.2 introduces the Horvitz-Thompson-like estimator. Section 4.3 gives the anglerfish abundance estimators with different forms of the net retention probability. Section 4.4 deals with the data with missing ages. Finally, the estimation of

uncertainty of the resulting abundance estimators is addressed in Section 4.5 and all the results of abundance estimation for the 2007 abundance survey data are presented in Section 4.6.

4.1 Anglerfish abundance survey data

Since 2005, stratified bottom trawl surveys have been conducted annually to collect data for the anglerfish stock occupying the northern European shelf. The design for the annual abundance survey is *stratified random sampling* (see Thompson, 2002, p. 117), in which the population is partitioned into strata and the design within each stratum is simple random sampling. The definition of the strata is based on the expected anglerfish density and the distance from the surface to the sea bottom (referred to as the depth of the sea) is used as a proxy for the expected density, because anglerfish density has been found to vary with depth. For the shallow water (0 to 140 m) and very deep water (500 to 1000 m), the density of the anglerfish is expected to be low, while in the deep water there are two other density categories – high density and medium density. It is expected that the densities of these strata might differ greatly from each other. The survey is designed to give a decent number of hauls within each stratum (subject to constraints on available resources) to allow for abundance estimation within each stratum. Summing abundance estimates over strata provides an abundance estimate for the overall population of the survey area. However, due to the weather conditions and some unexpected technical problems in the field, it happened that there was only one haul successfully towed in one stratum in 2007 (the survey analysed in this chapter). This issue is addressed in Section 4.5.2 of this chapter.

As described in Section 1.2, the abundance survey in 2007 was extended into Irish waters. There is a total of $1,701$ fish captured by 158 hauls in the abundance survey. The location of each haul is given in Figure 4.1 with the radius of the circle being proportional to the logarithm of number of fish captured by each haul. There are 22 strata in total in the 2007 abundance survey and the number of hauls sampled in each stratum is given by Table 4.1. There are three different scenarios in terms of hauls in a stratum:

1. The stratum ‘North.L’ had only one empty haul (i.e., a haul that captured no fish at all) – a stratum with only one empty haul or empty hauls is referred to as an empty-haul stratum, and empty-haul strata have zero abundance estimates and associated variance estimates.

2. The stratum ‘Rockall.L2’ had only one haul which captured fish in the survey – a stratum with only one haul and this haul did capture fish is referred to as single-haul stratum.

3. The other 20 strata in the 2007 abundance survey had at least two hauls – a stratum with multiple hauls is referred to as a multi-haul stratum. The number of hauls in the multi-haul strata varies from 2 to 30 in the 2007 abundance survey data, with 8 hauls being the average number of hauls towed in one stratum.

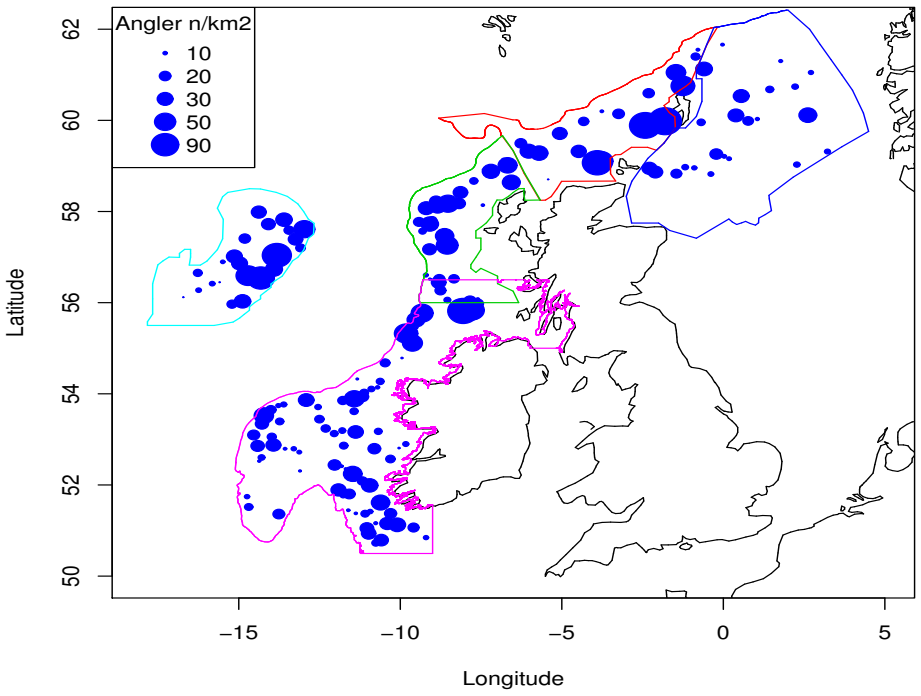


FIGURE 4.1. A plot of the survey area, showing the location of all the hauls in the 2007 survey with the Irish area, and the radius of each circle is proportional to the logarithm of the number of fish captured in each haul.

TABLE 4.1. The number of hauls towed in the 2007 anglerfish abundance survey.

Stratum	no. of hauls	Stratum	no. of hauls
East.M	25	Rockall.L1	6
East.L	4	Rockall.M	8
North.M1.E	5	Rockall.H	10
North.M2.E	3	Rockall.VH	6
North.H.E	7	Irish.L1.N	30
North.M1.W	5	Irish.H.N	3
North.M2.W	3	Irish.M.N	3
North.H.W	5	Irish.L2.N	2
West.L.56.5	5	North.L	1 [a]
West.M1.56.5	15	Rockall.L2	1 [b]
West.M2.56.5	3		
West.H.56.5	8		

[a] this one haul in stratum 'North.L' captured no fish at all and it is referred to as an empty haul;

[b] this one haul in stratum 'Rockall.L2' captured 20 fish in total.

The 2007 abundance survey data are used in this chapter to illustrate the method of abundance estimation that is developed below. This choice is due to the fact that the 2007 survey data is the most comprehensive data set among all annual abundance survey data. Specifically:

1. The 2007 abundance survey includes Irish waters, where the catch data have no age information (21.3% of the data has no age information). A method for imputing age for these missing-age data is addressed in Section 4.4.

2. The 2007 survey data have one single-haul stratum – variance estimation for this single-haul stratum causes problem in bootstrap variance estimation. How to 'borrow' information for the variance estimation of this single-haul stratum is addressed in Section 4.5.2.

In the context of the 2007 abundance survey data, the abundance estimation method is described in Section 4.3 with the corresponding variance estimation discussed in Section 4.5. Finally, the abundance estimation results are presented and discussed in Section 4.6. The abundance estimation using the 2007 survey data is implemented in the statistical software R.

The information of each individual catch includes the length and age of each captured fish, and the haul in which this fish was captured. The information of each haul includes the stratum in which the haul was towed, the size of the area swept by the whole gear (i.e., the swept area between the wings), and the size of the area between the wings and the doors (see Figure 1.2 for details).

To start with, the following gives a list of extended notation for abundance estimation in addition to the notation used for the capture probability estimation in Part II:

- s: a stratum in the overall survey area;
- i: a haul, and $\sum_{i \in s}$ denotes summing over all hauls in stratum s;
- l: the length of fish;
- a: the age of fish;
- N: the abundance in the overall survey area;
- n: the number of fish captured by hauls;
- ρ: the density of fish;
- A_s: the surface area of stratum s;
- v_s: the surface area swept by all hauls in stratum s, i.e., $v_s = \sum_{i \in s}(v_{1i} + v_{2i})$ (see Figure 1.2 for the definition of v_{1i} and v_{2i}); and
- nhaul$_s$: the total number of hauls towed in stratum s.

4.2 Abundance estimation using Horvitz-Thompson-like estimator

Abundance estimation from survey data is a key component of inference for many problems in statistical ecology, and the Horvitz-Thompson estimator (Horvitz &

Thompson, 1952) is a versatile and widely-used estimator of abundance (N). The estimator (which is abbreviated to 'HT estimator') has the general form

$$\widehat{N} \;=\; \sum_{i=1}^{n} \frac{s_i}{p_i}, \tag{4.1}$$

where n is the sample size, s_i is the number of individuals in the ith captured unit, and p_i is the probability that the ith unit is included in the sample, i.e., the 'inclusion probability'. However, for the anglerfish abundance survey data, p_i is unknown and involves estimated net retention probability $\hat{r}$ and herding factor $\hat{h}$. In the case that p_i is unknown and therefore (4.1) is used with an estimated inclusion probability, $\hat{p}_i$, it is referred to as a 'HT-like estimator', i.e.,

$$\widehat{N} \;=\; \sum_{i=1}^{n} \frac{s_i}{\hat{p}_i}. \tag{4.2}$$

The application of the HT-like estimator for the anglerfish abundance estimation starts with specifying the $\hat{p}_i$ in (4.2) for the anglerfish survey. Recall that anglerfish are sedentary fish, the net retention probability r is the probability that a fish is retained in the cod-end, given that it is in the area over which the net is dragged (i.e., the swept area). However in haul i, the anglerfish are not entirely sedentary, and there is some herding of anglerfish within the area swept by the net doors (the light grey area in Figure 1.2) into the net's path. A fraction h of the anglerfish between the doors and the wings (refer to Figure 1.2) are then herded into the path of the net. This herding factor h was estimated to be 0.017 by Allen (2006), and this is the value for $\hat{h}$ we used here.

We obtain an estimator conditional on the estimated retention probability $\hat{r}$ from the 2006-2007 experimental survey data, and the herding factor $\hat{h}$ based on simulations derived from video cameras observations of anglerfish behaviour (Reid *et al.*, 2007a; Allen, 2006). Then the inclusion probability of a fish with length l captured by haul i in stratum s is estimated by

$$\underbrace{\hat{r}}_{\text{(a)}} \times \underbrace{\frac{v_{1i} + \hat{h} v_{2i}}{v_{1i} + v_{2i}}}_{\text{(b)}} \times \underbrace{\frac{v_s}{A_s}}_{\text{(c)}}, \tag{4.3}$$

where:

- Part (a) is the estimated probability that a fish is retained in the cod-end of haul i, given that it contacts the fishing net. Note that $\hat{r}$ can be a function of length $\hat{r}(l)$ or a function of length and haul effect, $\hat{r}_i(l)$, both of which will be discussed in detail in Section 4.3.

- Part (b) gives the probability that a fish contacts the fishing net given that it was in the area swept by the haul i.

- Part (c) gives the probability that a fish is in the path of a haul towed in the stratum s.

The final inclusion probability that a fish with length l is included in the sample collected by haul i from stratum s is the product of (a), (b) and (c). Note that the product of (a) and (b) has been defined as the capture probability defined by (2.1) in Section 2.1, and the exact definition of this capture probability is the probability of catching a fish given that it was in the area covered by the sampling process. The inclusion probability is just the capture probability multiplied by proportion of the area covered by the sampling process.

Given that each captured fish has its length and age measured, let n_{ila} denote the number of fish with length l and age a captured by haul i. Therefore, with the inclusion probability given by (4.3), the abundance estimator for stratum s and fish with age a using the HT-like estimator is

$$\begin{aligned}\widehat{N}_{sa} &= \sum_{i\in s}\sum_{l}\frac{n_{ila}}{\hat{r}\left(\dfrac{v_{1i}+\hat{h}v_{2i}}{v_{1i}+v_{2i}}\right)\dfrac{v_s}{A_s}} \\ &= A_s\sum_{i\in s}\frac{(v_{1i}+v_{2i})}{v_s}\sum_{l}\frac{n_{ila}}{\hat{r}\,(v_{1i}+\hat{h}v_{2i})}. \qquad (4.4)\end{aligned}$$

The abundance estimator given by (4.4) is conditional on the estimated net retention probability, $\hat{r}$, whose exact form depends on which model to use in estimation. Section 4.3 gives the abundance estimators from assuming $r = 1$, $\hat{r}(l)$ is estimated by fixed-effects logistic regression, and $\hat{r}_i(l)$ is estimated by mixed-effects logistic regression incorporating haul i as a random effect.

In addition, given the abundance estimator (4.4), the density estimator can be easily obtained as

$$\hat{\rho}_{sa} = \frac{\widehat{N}_{sa}}{A_s} = \sum_{i \in s} \frac{(v_{1i} + v_{2i})}{v_s} \sum_{l} \frac{n_{ila}}{\hat{r}\,(v_{1i} + \hat{h} v_{2i})}, \tag{4.5}$$

and its variance estimator can be expressed as

$$\widehat{\text{Var}}\,[\hat{\rho}_{sa}] = \frac{1}{A_s^2} \widehat{\text{Var}}\,[\widehat{N}_{sa}].$$

Therefore, (4.5) gives the density estimator followed by the abundance estimator (4.4), which is derived from applying the Horvitz-Thompson method with the estimated inclusion probability. The estimator (4.5) is referred to as the HT-like density estimator.

Before giving the HT-like abundance estimators with different models for $\hat{r}$ in the next section, the following paragraphs are going to show that the HT-like density estimator is of the same form as the weighted average density estimator, which is commonly used in fishery research. This is of interest because Fernandes *et al.* (2007) used the weighted average density estimator in a previous study of the anglerfish abundance survey data.

To obtain the weighted average density estimator, since the annual abundance survey data are grouped by hauls resulting from the trawl catching process, the density estimation is carried out first for each haul within the given stratum, and then a weighted average over the density estimates for all hauls leads to the density estimate of this stratum. To start with, we consider the density estimator for each haul i given by Fernandes *et al.* (2007). Their density estimator for haul i involves the trawl efficiency Q, which is commonly used in fisheries literature. Somerton *et al.* (1999) defined Q, as applied to haul i, as follows:

$$\widehat{Q}_i = \hat{r} \left(1 + \hat{h} \frac{v_{2i}}{v_{1i}} \right); \tag{4.6}$$

see Appendix 4.A for its derivation.

For $\widehat{Q}_i$, the fishing effort is measured by the swept area between wings (i.e., v_{1i} in Figure 1.2). Then conditioning on the estimated net retention probability $\hat{r}$ and

herding factor $\hat{h}$, the abundance of anglerfish in v_{1i} can be estimated by

$$\widehat{N}^c_{ila} = \frac{n_{ila}}{\widehat{Q}_i} = \frac{n_{ila}}{\hat{r}\left(1 + \hat{h}\dfrac{v_{2i}}{v_{1i}}\right)},$$

and hence the corresponding density estimator is

$$\hat{\rho}_{ila} = \frac{n_{ila}}{v_{1i}} = \frac{n_{ila}}{\hat{r}\left(v_{1i} + \hat{h}v_{2i}\right)}. \tag{4.7}$$

In the case when information for estimating $r(l)$ is lacking, r is assumed to be unity, as in the abundance estimation provided by Fernandes *et al.* (2007).

In fisheries, age-specific density estimates are of more interest than length-specific density estimates, because age increases deterministically with time, whereas length increases stochastically with time and this makes age-based population dynamics models simpler than length-based models. Therefore, based on the density estimator $\hat{\rho}_{ila}$ given in (4.7), the density estimator for fish with age a and haul i can be obtained by summing over all the length classes for each age a, i.e.,

$$\hat{\rho}_{ia} \;=\; \sum_l \frac{n_{ila}}{\hat{r}(v_{1i} + \hat{h}v_{2i})}.$$

Taking the weighted average of $\hat{\rho}_{ia}$ for all i in stratum s and using the weights $w_i = (v_{1i}+v_{2i})/v_s$ gives an unbiased estimator of density if the individual $\hat{\rho}_{ia}$ are unbiased (see Appendix 4.B for the derivation of weight $w_i = (v_{1i} + v_{2i})/v_s$):

$$\hat{\rho}_{sa} \;=\; \sum_{i\in s} w_i\hat{\rho}_{ia} = \sum_{i\in s} \frac{(v_{1i} + v_{2i})}{v_s} \sum_l \frac{n_{ila}}{\hat{r}(v_{1i} + \hat{h}v_{2i})}. \tag{4.8}$$

Comparing (4.8) to (4.5), it can be seen that the density estimator (4.8) used by Fernandes *et al.* (2007), which is derived from the weighted average method using $\widehat{Q}_i$, has the same form as (4.5) derived from the HT-like estimator. Although $\widehat{Q}_i$ is widely used in fisheries literature, it is not a probability because if $\hat{r} = 1$ and $\hat{h} > 0$, then $\widehat{Q}_i$ given in (4.6) is larger than 1. It seems to us that the capture probability defined as the product of (a) and (b) in (4.3) is easier to understand, and more straightforward to be incorporated in the inclusion probability (4.3) for the HT-like abundance estimator. Therefore, the abundance estimators given in the next section are derived from the HT-like estimator.

The rest of this chapter focuses on the HT-like estimator with different models for the net retention probability. Conditional on the inclusion probability in which the net retention probability is estimated using fixed-effects models given in Section 3.2 and using mixed-effects models in Section 3.4, the HT-like estimators are presented in the next section with

1. capture probability with $r = 1$ (which is what has been assumed in previous assessments of this anglerfish stock),
2. capture probability with $\hat{r}(l)$, i.e., the net retention probability modelled as a function of fish length, and
3. capture probability with $\hat{r}_i(l)$, i.e. the length-based net retention probability with haul i as a random effect.

We consider four different forms of HT-like estimator with random effects in net retention probability, and these different forms of HT-like estimator with random effects will be studied in Part IV.

4.3 Anglerfish abundance estimators

Let r denote the net retention probability and p denote the inclusion probability. There are five different forms of the anglerfish abundance estimator presented in this section. Section 4.3.1 gives the abundance estimator under the assumption that $r = 1$. Section 4.3.2 gives the abundance estimator under the assumption that r is a function of fish length l, $\hat{r}(l)$, which is from the fixed-effects logistic regression studied in Section 3.2 using the anglerfish experimental survey data. Finally, Section 4.3.3 gives three different forms of the density estimator under the assumption that r is a function of fish length l with haul as a random effect, $\hat{r}_i(l)$, estimated by the mixed-effects logistic regression models in Section 3.4.

4.3.1 Estimators with perfect retention probability

If it is assumed that all the fish that contact the net are retained in the main cod-end, i.e., the net retention probability $r = 1$, then the inclusion probability defined in

(4.3) becomes

$$\left(\frac{v_{1i}+\hat{h}v_{2i}}{v_{1i}+v_{2i}}\right)\frac{v_s}{A_s},$$

which gives the abundance estimator as

$$\widehat{N}_{sa}=A_s\sum_{i\in s}\frac{(v_{1i}+v_{2i})}{v_s}\frac{\sum_l n_{ila}}{(v_{1i}+\hat{h}v_{2i})}=A_s\sum_{i\in s}\frac{n_{ia}}{(v_{1i}+\hat{h}v_{2i})}, \tag{4.9}$$

where $n_{ia}=\sum_l n_{ila}$.

Let $w_i=(v_{1i}+v_{2i})/v_s$. Then applying the Delta method described in Appendix 4.C.1, the variance of $\hat{\rho}_{sa}$ can be estimated as

$$\begin{aligned}&\widehat{\text{Var}}\,[\widehat{N}_{sa}]=A_s^2\,\widehat{\text{Var}}\,[\hat{\rho}_{sa}]\\ =\;& A_s^2\left\{\widehat{\text{Var}}\,[n_{ia}]\sum_{i\in s}\left[\frac{w_i}{v_{1i}+\hat{h}\,v_{2i}}\right]^2+\text{nhaul}_s\widehat{\text{Var}}\,[\hat{h}]\left[\sum_{i\in s}\frac{w_i\,n_{ia}\,v_{2i}}{(v_{1i}+\hat{h}\,v_{2i})^2}\right]^2\right\};\end{aligned} \tag{4.10}$$

see Appendix 4.C.3 for a detailed derivation of $\widehat{\text{Var}}\,[\hat{\rho}_{sa}]$.

Because of the escaping of fish beneath the footrope, it is not reasonable to assume that $r=1$, particularly for small fish. In order to obtain an estimate of r, experimental surveys were conducted in 2006 and 2007, and these data have been analyzed in Chapter 3. The next two sections describe the estimators with $\hat{r}$ estimated from the 2006-2007 experimental survey data.

4.3.2 Estimators with fixed-effects retention probability

When the net retention probability is estimated by the fixed-effects logistic regression models using the 2006-2007 experimental survey data, as described in Section 3.2, the net retention probability is predicted by the length of fish. This section gives the density estimator with $\hat{r}(l)$ and corresponding variance estimator derived by the Delta method. Let $\hat{r}(l)$ denote the estimated net retention probability by a fixed-effects logistic regression model. Then the inclusion probability for fish with length

l and age a captured by haul i in stratum s is

$$\hat{r}(l)\left(\frac{v_{1i}+\hat{h}v_{2i}}{v_{1i}+v_{2i}}\right)\frac{v_s}{A_s},$$

where $\hat{r}(l)$ can be either (3.36) obtained from a linear logistic regression or (3.43) from the extended logistic regression with non-unity asymptote (see Section 3.2.2 for full detail). Then the abundance estimator is

$$\widehat{N}_{sa}=A_s\sum_{i\in s}\frac{(v_{1i}+v_{2i})}{v_s}\sum_{l}\frac{n_{ila}}{\hat{r}(l)(v_{1i}+\hat{h}v_{2i})}. \tag{4.11}$$

In the case where $r(l)$ is estimated by a linear logistic regression model, i.e., $\hat{r}(l)$ is given by (3.36) in Section 3.2.1, we have

$$\begin{aligned}
&\widehat{\text{Var}}\left[\widehat{N}_{sa}\right]=A_s^2\widehat{\text{Var}}\left[\hat{\rho}_{sa}\right]\\
=\;& A_s^2\sum_{i\in s}w_i^2\left\{\sum_{l}\left[\frac{\widehat{\text{Var}}\left[n_{ila}\right]}{[\hat{r}(l)]^2(v_{1i}+\hat{h}v_{2i})^2}+\frac{\widehat{\text{Var}}\left[\hat{h}\right]n_{ila}^2v_{2i}^2}{[\hat{r}(l)]^2(v_{1i}+\hat{h}v_{2i})^4}\right.\right.\\
&\qquad+\frac{n_{ila}^2}{(v_{1i}+\hat{h}v_{2i})^2}\left(\widehat{\text{Var}}\left[\hat{\beta}_0\right]\exp(-2\hat{\beta}_0-2\hat{\beta}_1\,l)\right.\\
&\qquad\qquad+\widehat{\text{Var}}\left[\hat{\beta}_1\right]l^2\,\exp(-2\hat{\beta}_0-2\hat{\beta}_1\,l)\\
&\qquad\qquad\left.\left.+\;2\,\widehat{\text{Cov}}\left[\hat{\beta}_0,\hat{\beta}_1\right]l\,\exp(-2\hat{\beta}_0-2\hat{\beta}_1\,l)\right)\right]\right\};
\end{aligned} \tag{4.12}$$

see Appendix 4.C.4 for a detailed derivation of $\widehat{\text{Var}}\left[\hat{\rho}_{sa}\right]$ and see (4.62) for the variance estimator of $\hat{\rho}_{sa}$ with $r(l)$ estimated by asymptote-logistic regression, given by (3.43). The variance estimates presented later in Section 4.6 are obtained by bootstrapping, though in order to check the code to perform bootstrapping, the variance estimates with $r=1$ and $\hat{r}(l)$ estimated by a linear logistic regression are calculated based on the (4.10) and (4.12). The variance estimates obtained by using (4.12) are consistent with those obtained by bootstrapping.

4.3.3 Estimators with mixed-effects retention probability

In fisheries, it is very common to have a random haul effect in the capture probability model. Tschernij & Holst (1999), for example, showed that there were haul-level

effects on capture probability, and Fryer (1991) noted that neglecting the between-haul variation on such surveys can result in misleading inferences about capture probabilities. This section develops HT-like estimators with random effects in order to estimate fish abundance from a trawl survey. In addition to having length as the predictor in estimating the net retention probability, the haul effect was incorporated as a random effect by a mixed-effects logistic regression model in Section 3.4. Let $\hat{r}_i(l)$ denote the estimated net retention probability for fish with length l and captured by haul i. The variation in $\hat{r}_i(l)$ due to l was considered in two parts: the between- and within-haul parts. The between-haul part represents the haul effect, while the within-haul part represents the length effect of each fish in relation to its peers in the same haul. The between-haul effect was incorporated by including the average length of all fish captured by each haul ($\bar{l}_i$), referred to as group-mean length, and the within-haul effect via each fish's length centred on its group-mean length ($l - \bar{l}_i$).

The final model selected for $\hat{r}_i(l)$ is given by (3.72), which has random intercept (b_{0i}) only. However, abundance estimation with $\hat{r}_i(l)$ is considered for the more general case when $\hat{r}_i(l)$ has both random intercept b_{0i} and random slope b_{1i}. Let $\boldsymbol{b}_i$ denote $(b_{0i}, b_{1i})^T$ for haul i, and $\boldsymbol{b}_i \sim N\left(\mathbf{0}, \widehat{\Sigma}_b\right)$. This consideration is due to the incorporation of uncertainty about $\widehat{\Sigma}$ in bootstrapping for variance estimation. The rank of $\widehat{\Sigma}_b$ depends on the experimental survey data. Abundance estimation with haul effect considered below is for the most complicated case when $\hat{r}_i(l)$ has a full rank $\widehat{\Sigma}_b$, i.e., both random intercept and slope.

Given the random effects $\boldsymbol{b}_i$ and the group-mean length $\bar{l}_i$ for haul i , the net retention probability of a fish with length l and captured by haul i is estimated by

$$\hat{r}(l, \bar{l}_i|\boldsymbol{b}_i) \;=\; \text{logit}^{-1}[(\hat{\beta}_0 + b_{0i}) + \hat{\beta}_1^B\, \bar{l}_i + (\hat{\beta}_1^W + b_{1i})\,(l - \bar{l}_i)], \qquad (4.13)$$

where $\hat{\beta}_0$, $\hat{\beta}_1^B$ and $\hat{\beta}_1^W$ are the estimated fixed-effects coefficients. Then the inclusion probability of fish with length l captured by haul i in stratum s, p_{sil}, is

$$p_{sil}\,(l, \bar{l}_i \,|\, \boldsymbol{b}_i) \;=\; \hat{r}(l, \bar{l}_i|\boldsymbol{b}_i)\left(\frac{v_{1i} + \hat{h}\, v_{2i}}{v_{1i} + v_{2i}}\right)\frac{v_s}{A_s}. \qquad (4.14)$$

Note that the above inclusion probability is for a haul in the abundance survey and it is conditional on the random effects $\boldsymbol{b}_i$. However, the sample of hauls in the abundance survey is not the same as that in the experimental survey, though it is assumed

that they come from the same population (in the sense that the parameters estimated from the experimental survey apply to the abundance survey). In predicting the binary response for a new observation in a new group, using the mean of $\boldsymbol{b}_i$ is considered as a simpler alternative to integrating out $\boldsymbol{b}_i$ over its estimated distribution $N(\mathbf{0}, \widehat{\Sigma}_b)$ (Skrondal, 2009). Similarly, when predicting $\hat{r}_i(l)$ for hauls in the abundance survey, simply plugging in the mean $\mathbf{0}$ for $\boldsymbol{b}_i$ leads to $\widehat{N}^{(1)}$ considered below in (4.15). For the other two estimators, $\widehat{N}^{(2)}$ and $\widehat{N}^{3}$, prediction of the retention probability in the abundance survey uses the estimated distribution of $\boldsymbol{b}_i$, i.e., $N\left(\mathbf{0},\, \widehat{\Sigma}_b\right)$. Therefore, three different forms of HT-like abundance estimator with haul effect are considered in this section, depending on where the estimated random-effects distribution is incorporated in the HT-like estimator.

Let n_{ila} denote the number of fish of length class l and age a captured by haul i. Given the inclusion probability (4.14), there are three different HT-like estimators that can be applied to the anglerfish abundance survey data for the abundance estimation within stratum s:

$$\widehat{N}_{sa}^{(1)} = \sum_{i\in s;\, l} \frac{n_{ila}}{\hat{p}_{sil}\left(l, \bar{l}_i \,|\, \boldsymbol{b}_i = \mathbf{0}\right)}, \tag{4.15}$$

$$\widehat{N}_{sa}^{(2)} = \sum_{i\in s;\, l} \frac{n_{ila}}{\widehat{\mathbf{E}}_b[\hat{p}_{sil}\left(l, \bar{l}_i \,|\, \boldsymbol{b}_i\right)]}, \tag{4.16}$$

$$\widehat{N}_{sa}^{(3)} = \sum_{i\in s;\, l} \widehat{\mathbf{E}}_b \left[\frac{n_{ila}}{\hat{p}_{sil}\left(l, \bar{l}_i \,|\, \boldsymbol{b}_i\right)}\right]. \tag{4.17}$$

We now look at the estimators given by (4.15), (4.16) and (4.17) in more detail. There is no integration of $\boldsymbol{b}_i$ in $\widehat{N}^{(1)}$, and

$$\begin{aligned}\widehat{N}_{sa}^{(1)} &= \frac{A_s}{v_s} \sum_{i\in s;\, l} \left(\frac{v_{1i} + v_{2i}}{v_{1i} + \hat{h} v_{2i}}\right) \frac{n_{ila}}{\hat{r}(l, \bar{l}_i \,|\, \boldsymbol{b}_i = \mathbf{0})} \\ &= \frac{A_s}{v_s} \sum_{i\in s;\, l} \left(\frac{v_{1i} + v_{2i}}{v_{1i} + \hat{h} v_{2i}}\right) n_{ila} \left\{1 + \exp[-\hat{\beta}_0 - \hat{\beta}_1^B \bar{l}_i - \hat{\beta}_1^W \left(l - \bar{l}_i\right)]\right\}.\end{aligned} \tag{4.18}$$

However, for $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$, the abundance estimation involves integrating $\boldsymbol{b}_i$ with respect to its estimated distribution, which is a normal distribution with mean zero and variance $\hat{\sigma}_0^2$ (recall that $\hat{\sigma}_0 = 0.298$ and $\hat{\sigma}_1 = 0$ from the 2006-2007 experimental survey data). Analytical solution of the integrals in $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$ has been

attempted, but there is no solution for $\widehat{N}^{(2)}$, so numerical approximation is used for an approximation of the integral:

$$\begin{aligned}
\widehat{N}_{sa}^{(2)} &= \frac{A_s}{v_s}\sum_{i\in s;\, l}\left(\frac{v_{1i}+v_{2i}}{v_{1i}+\hat{h}v_{2i}}\right)\frac{n_{ila}}{\widehat{\mathbf{E}}_b[\hat{r}(l,\bar{l}_i\,|\,\boldsymbol{b}_i)]} \\
&= \frac{A_s}{v_s}\sum_{i\in s;\, l}\left(\frac{v_{1i}+v_{2i}}{v_{1i}+\hat{h}v_{2i}}\right)\frac{n_{ila}}{\int\int_{\mathbb{R}^2}\left\{1+\exp\left[-(\hat{\beta}_0+b_{0i})-\hat{\beta}_1^B\,\bar{l}_i-(\hat{\beta}_1^W+b_{1i})\,(l-\bar{l}_i)\right]\right\}^{-1}\hat{f}_b(\boldsymbol{b}_i)d\,\boldsymbol{b}_i} \\
&= \frac{A_s}{v_s}\sum_{i\in s;\, l}\left(\frac{v_{1i}+v_{2i}}{v_{1i}+\hat{h}v_{2i}}\right)\frac{n_{ila}}{\int_{\mathbb{R}}\left\{1+\exp\left[-\hat{\beta}_0-\hat{\beta}_1^B\,\bar{l}_i-\hat{\beta}_1^W(l-\bar{l}_i)-u_{il}\right]\right\}^{-1}f_{u_{il}}(u_{il})\,d\,u_{il}}
\end{aligned} \tag{4.19}$$

where $\hat{f}_b(\boldsymbol{b}_i)$ is a normal density function with mean $\mathbf{0}$ and variance-covariance matrix $\widehat{\Sigma}_b$. This expression shows that it is a two-dimensional integral and the function being integrated is well-behaved, so that numerical approximation will not be problematic. In addition, the two-dimensional integration with respect to $\boldsymbol{b}_i$ in (4.19) has been transformed into a one-dimensional integration by letting

$$u_{il} = b_{0i} + (l-\bar{l}_i)b_{1i}, \tag{4.20}$$

which is a linear transform of $\boldsymbol{b}_i$. Therefore, u_{il} is also normally distributed with mean 0 and variance $\hat{\sigma}_0^2 + (l-\bar{l}_i)^2\hat{\sigma}_1^2$. In this way, two-dimensional integration of $(b_{0i},\, b_{1i})^T$ becomes one-dimensional integration of u_{il}. This integral is obtained by a numerical approximation based on a 1000 equally spaced grid, which covers the interval within 5 standard errors from the mean of a linear combination of b_{0i} and b_{1i}, i.e., the u_{ij} in (4.20). Such a transformation does not only save computation time for $\widehat{N}^{(2)}$, but also makes the numerical approximation process the same for different cases of $\widehat{\Sigma}_b$: being full rank (both $\hat{\sigma}_0$ and $\hat{\sigma}_1$ are non-zero) or being singular (either $\hat{\sigma}_0$ or $\hat{\sigma}_1$ is zero). This provides great convenience in evaluating $\widehat{N}^{(2)}$ when incorporating uncertainty of $\widehat{\Sigma}_b$ in variance estimation.

For $\widehat{N}^{(3)}$, the analytical solution of the integral is given below:

$$\begin{aligned}
\widehat{N}_{sa}^{(3)} &= \frac{A_s}{v_s} \sum_{i \in s;\, l} \left(\frac{v_{1i} + v_{2i}}{v_{1i} + \hat{h} v_{2i}} \right) \widehat{\mathbf{E}}_b \left[\frac{n_{ila}}{\hat{r}(l, \bar{l}_i \,|\, \boldsymbol{b}_i)} \right] \\
&= \frac{A_s}{v_s} \sum_{i \in s;\, l} \left(\frac{v_{1i} + v_{2i}}{v_{1i} + \hat{h} v_{2i}} \right) n_{ila} \int \int_{\mathbb{R}^2} \left\{ 1 + \exp\left[-(\hat{\beta}_0 + b_{0i}) - \hat{\beta}_1^B \, \bar{l}_i - (\hat{\beta}_1^W + b_{1i}) \, (l - \bar{l}_i) \right] \right\} \hat{f}_b(\boldsymbol{b}_i) d\boldsymbol{b}_i \\
&= \frac{A_s}{v_s} \sum_{i \in s;\, l} \left(\frac{v_{1i} + v_{2i}}{v_{1i} + \hat{h} v_{2i}} \right) n_{ila} \left\{ 1 + \exp\left[-\hat{\beta}_0 - \hat{\beta}_1^B \, \bar{l}_i - \hat{\beta}_1^W \, (l - \bar{l}_i) + \frac{\hat{\sigma}_0^2 + (l - \bar{l}_i)^2 \hat{\sigma}_1^2}{2} \right] \right\}.
\end{aligned} \tag{4.21}$$

Note that the analytical solution of the integral in (4.21) is the bivariate case of (6.23), which is the general case discussed in in Section 6.3 of the HT-like estimator with capture probabilities estimated by a mixed-effects logistic regression model.

4.4 Age-imputation methods

The abundance estimators considered in Section 4.3 assume that the age and length of each captured fish has been measured. The length of fish can be easily measured at the scene. However, the age of each individual fish is decided by otoliths collected at the scene and brought back to the lab to allow the biologists to decide the specific age of each catch. In this whole process, some errors, such as contamination during transportation or handing mistakes at the scene that prevent the age identification, can occur. As a result, some fish do not have recorded ages in the annual abundance survey data. In addition, there is no age information for the fish captured in the Irish waters. The proportion of missing-age data for the 2007 abundance survey is 21.32%.

To illustrate how to deal with the missing age data, the 2007 anglerfish survey data are used, and a set of length classes with missing ages are given in Table 4.2. The last column 'Missing age' gives the number of captured fish that have no age information at each length class. Abundance estimation for each age group requires a proper way to deal with these missing age data. Section 4.4.1 and Section 4.4.2 describe two different ways to deal with these missing-age data.

TABLE 4.2. Part of age-length frequency table for the 2007 abundance survey.

	Age0-3	Age4	Age5	Age6	Age7	Age8-15	Missing age
55cm	0	1	48	20	0	0	3
56cm	0	0	33	33	0	0	2
57cm	0	0	19	37	0	0	4
58cm	0	0	14	34	0	0	1
59cm	0	0	11	33	1	0	3

4.4.1 Mode-based method

Given the length of a missing-age fish, the mode-based method picks up the most likely age based on the empirical distribution of the age of all the other fish with the same length. Taking all the fish with length 59 cm in Table 4.2 to illustrate this method, age 6 will be the most likely age. If we take all the fish with length 56 cm for another example, then age 5 and age 6 are equally likely. The question is how to decide a most likely age in this case. The existing method in Fernandes *et al.* (2007) is to choose the younger age, i.e., age 5 for length 56 cm missing-age fish. But there is no reason why age 6 cannot be assigned to those missing-age fish. This motivates the consideration in the next section of a method to distribute the missing-age fish to a set of ages in a probabilistic and rigorous way.

4.4.2 Probability-based method

TABLE 4.3. The observed age-distribution at given length 56 cm and 59 cm.

	Age0-4	Age5	Age6	Age7	Age8-15	Missing age
56cm	0	$2 \times \frac{33}{66}$	$2 \times \frac{33}{66}$	0	0	2
59cm	0	$3 \times \frac{11}{45}$	$3 \times \frac{33}{45}$	$3 \times \frac{1}{45}$	0	3

Instead of assigning the most likely age to a missing-age catch in the mode-based method, the probability-based method spreads out the missing-age catch into related age groups. How the spread is done is based on the empirical frequency distribution of age given the length. The implementation of the probability-based method is

illustrated by the catch data of length 56 cm and 59 cm in Table 4.3, and the proportions in the table are calculated from the observed frequencies of catches at each age group in Table 4.2.

Comparing the mode-based method described in Section 4.4.1 to the above probability-based method, it is apparent that the mode-based method is easier to implement in abundance estimation. However, the probability-based method is more rigorous in a way that it incorporates all the available information about the connection between age and length from the catch data. Therefore, the probability-based method is used to deal with missing-data in the anglerfish abundance survey data and all the results presented in Section 4.6 are obtained from the estimation using the probability-based method.

To formulate the probability-based method for abundance estimation using the HT-like estimator (4.1), the age is treated as an extra layer of random effect (or random variable) for those fish whose ages are unidentified, and the distribution of age as a random variable is estimated by the empirical distribution from the observed catch data (see Figure 4.2 for the empirical distribution of age at each length class using the 2007 abundance survey data). Let $f_{a|l}(a|l)$ denote the probability mass function of age given length for captured fish. $\hat{f}_{a|l}(a|l)$ can be expressed as a step function estimated from the observed frequencies in Figure 4.2 for each age classes from 0 to 15.

Let $\widehat{N}_s(a)$ be a proposed HT-like estimator for the missing-age data (i.e. a is a random variable here), where $\widehat{N}$ can be any of the abundance estimators presented in Section 4.3: (4.9), (4.11), (4.15), (4.16) and (4.17). For fish of known age, those estimators can be applied directly, but for fish of unknown age but known length, the following two estimators are applied

$$\widehat{N}_s(a) = \widehat{N}_s(\widehat{\mathbf{E}}_{a|l}[a|l]), \tag{4.22}$$

$$\widehat{N}_s(a) = \widehat{\mathbf{E}}_{a|l}[\widehat{N}_s(a|l)], \tag{4.23}$$

where $\widehat{\mathbf{E}}_{a|l}[\ |l]$ is the estimated conditional expectation with respect to age, given length, according to the estimated age distribution given length, $\hat{f}_{a|l}(a|l)$.

In a previous analyses of the Scottish anglerfish abundance survey data, Fernandes *et al.* (2007) used an abundance estimator similar to that of (4.22) given above, but with the mode of the observed ages of fish of length class l in place of $\widehat{\mathbf{E}}_{a|l}[a|l]$, which

is the mode-based method given in Section 4.4.1. The probability-based method uses the observed age distribution of fish of length class l as an estimator of $f_{a|l}(a|l)$ in the estimator given by (4.23).

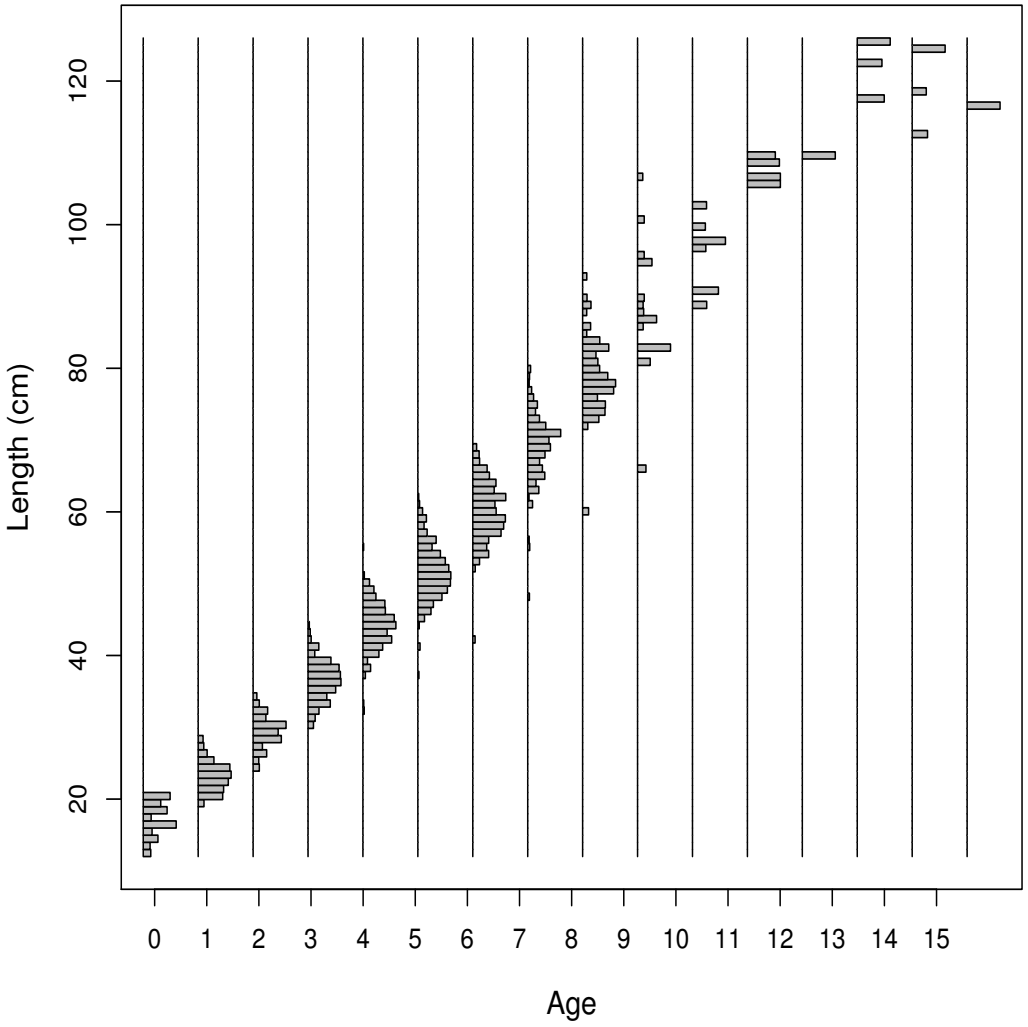


FIGURE 4.2. Plot of the empirical age distribution for each length class for the 2007 anglerfish abundance survey data.

4.5 Variance and interval estimation

Given the stratified-random sampling design for the anglerfish abundance survey, the samples in different strata are collected independently, and hence the variances

of the abundance estimator for the whole population can be obtained by adding the variance estimates for each individual stratum together. For the abundance estimator of fish of a given age a in the survey area, its variance estimator is obtained by

$$\widehat{\text{Var}}[\widehat{N}_a] = \sum_s \widehat{\text{Var}}[\widehat{N}_{sa}]. \tag{4.24}$$

Then, under the assumption of independence among different age groups, for the overall total abundance estimates, it follows that

$$\widehat{\text{Var}}[\widehat{N}_{\text{total}}] = \sum_{a,\,s} \widehat{\text{Var}}[\widehat{N}_{sa}].$$

As described by (4.24), the estimation of $\text{Var}[\widehat{N}_a]$ needs the variance estimate for each stratum in the survey area, i.e., $\widehat{\text{Var}}[\widehat{N}_{sa}]$.

As shown in Section 4.3, analytical expressions of the variance estimator are available only for the simple case of net retention probability, i.e., $r = 1$, and fixed effects logistic regression models. But in this case, the uncertainty from missing ages cannot be included in the variance estimation. This will not be a serious problem if there are few missing ages in the abundance survey data. However, there is no age information available for the catch data in the Irish waters. As a result, age-specific abundance estimation for Irish waters needs to borrow the age-length information from Scottish waters, and it is not analytically straightforward to work out the uncertainty caused by missing ages. Therefore, to include all the sources of uncertainty, bootstrapping is used for variance estimation of the abundance estimators within each stratum and then the percentile method is used to estimate their confidence intervals (Efron & Gong, 1983).

As the anglerfish survey data are grouped by hauls, the 2007 abundance survey data are sampled by hauls with replacement within each stratum, to include the variation caused by the sampling process. However, bootstrapping becomes problematic in the case when there is only one haul in the stratum. Recall that the number of hauls in each stratum is given in Table 4.1. The stratum 'Rockall.L2' has only one haul, which captured 20 fish in total. Variance estimation for this single-haul stratum is different from that for strata with multiple hauls (the multi-haul stratum defined in Section 4.1). Therefore, this section starts with the variance estimation for multi-haul

stratum in Section 4.5.1, and then comes to the variance estimation for single-haul stratum in Section 4.5.2.

4.5.1 Bootstrap variance estimation

There are two components in variance estimation: the variation in the number of fish in the sampling process and the variation in the inclusion probability. For the inclusion probability, based on the inclusion probability estimated by (4.3), there are two sources of uncertainty: the estimated net retention probability $\hat{r}$ and the estimated herding factor $\hat{h}$. Therefore, there are three sources of uncertainty in the anglerfish abundance estimation:

1. the catch data from the annual abundance survey,
2. the net retention probability estimated from the 2006-2007 experimental survey data, and
3. the herding factor estimated from a sister project (Allen, 2006).

Both parametric and non-parametric bootstrap methods are used to include all three sources of uncertainty listed above. To include the uncertainty from the annual abundance survey, a non-parametric bootstrap method is applied to accommodate uncertainty due to the catch data with missing ages and the sampling process: within each multi-haul stratum, hauls are sampled with replacement to obtained a set of sampled fish. Then for each set of sampled fish, the empirical age distribution at each length class, like that plotted in Figure 4.2, is then used in the probability-based method described in Section 4.4.2 to impute age to all the missing-age fish in the re-sampled catch data.

To include the uncertainty of $\hat{r}$ in the variance estimation of the abundance estimator, parametric and non-parametric bootstrap methods are used, depending on which model is used to estimate r.

- When using fixed-effects logistic regression models to estimate r (see Section 3.2 for full detail), parametric bootstrapping is used to include the uncertainty of $\hat{r}$ in abundance estimator (4.11). The regression coefficients are simulated from a multivariate normal distribution, which is assumed for the

asymptotic distribution of the estimated regression coefficients. For linear logistic regression, the mean and variance-covariance matrix of its asymptotic normal distribution is given by (3.33), (3.34), and (3.35). For extended logistic regression with non-unity asymptote, the mean and variance-covariance matrix of its asymptotic normal distribution are given by (3.38), (3.39), (3.40) and (3.41).

- When using the mixed-effects logistic regression model described in Section 3.4 to estimate r, non-parametric bootstrap method, unlike the fixed-effect logistic regression, is implemented to include all the uncertainty in r in abundance estimator (4.15), (4.16) and (4.17). The reason for this is that the distributions of the random-effects variance estimates are highly skewed and are not readily approximated by any standard parametric distribution. Therefore, the standard errors of $\hat{\sigma}_0$ and $\hat{\sigma}_1$ are not listed in the model-fitting output of the package `lme4` (Bates, D. and Maechler. M and Bolker, B., 2011), which was used for estimation of r. Therefore, the inclusion of all the uncertainty in a mixed-effects $\hat{r}_i(l)$ is implemented by non-parametric bootstrapping of the 2006-2007 experimental survey data: first, re-sample the experimental survey data with replacement in hauls, and then fit a two-level mixed-effects logistic regression model of the form (4.13) to the re-sampled data set.

Figure 4.3 gives the histograms of all parameter estimates for all 999 re-sampled experimental survey data (with replacement in hauls). The first three plots are for the histograms of the fixed-effects parameter estimates, i.e., $\hat{\beta}_0$, $\hat{\beta}_1^B$ and $\hat{\beta}_1^W$ in (4.13), which shows that it is still reasonable to assume that they are normally distributed. However, high skewness in the distribution of $\hat{\sigma}_0$ and $\hat{\sigma}_1$ is indicated by the last two histograms in Figure 4.3. A very high proportion of zeros in the histograms of $\hat{\sigma}_0$ and $\hat{\sigma}_1$ indicates small haul effect, but there is some evidence for the random intercept. In addition, the histogram of $\hat{\beta}^B - \hat{\beta}^W$ shows that its percentile CI includes 0, which suggests no difference between the with-haul effect and between-haul effect of length.

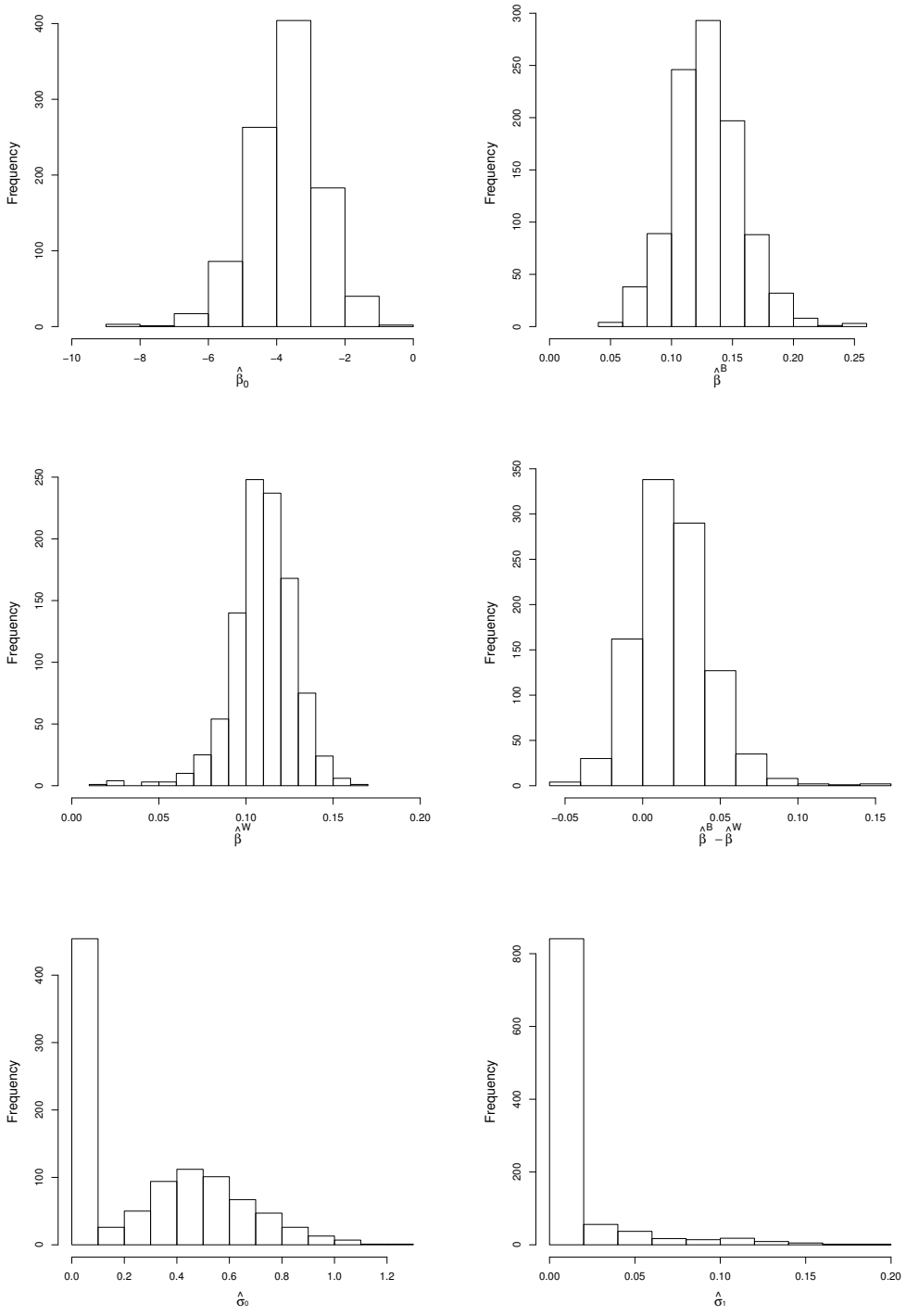


FIGURE 4.3. Histograms of parameter estimates for the net retention probability based on 999 re-sampled experimental survey data (with replacement of hauls): a two-level mixed–effects logistic regression model of the form (4.13) is fitted to each simulated experimental survey data, which allows both random intercept and random slope.

Finally, to include the uncertainty of $\hat{h}$, a non-parametric bootstrap method is used based on results from a sister project (see Allen, 2006, for details). This project set up a simulation based on the information about the anglerfish behaviour summarized from the video camera data. This simulation generated 100 values for herding factor, whose histogram is given in Figure 4.4.

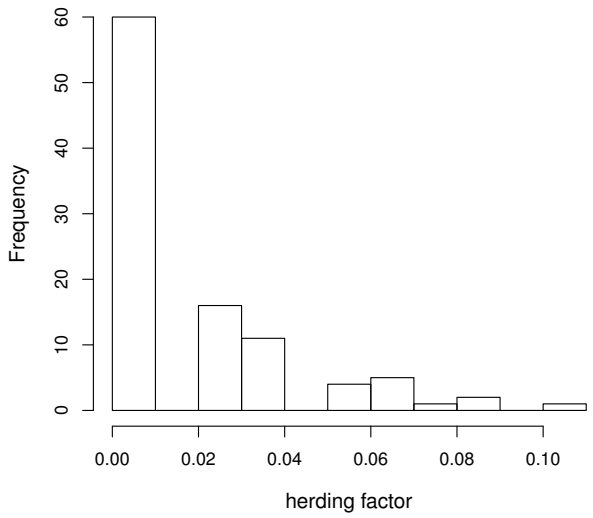


FIGURE 4.4. Histogram of simulation results of h from Allen (2006).

Therefore, for the multi-haul strata in the abundance survey, all the potential sources of uncertainty in abundance estimation are included in the bootstrap variance estimation described above. The complication of the bootstrap method depends on which model to use for the estimation of r. However, no matter which model is chosen to estimate r, re-sampling the annual abundance survey data within each stratum is an essential part in the variance estimation. Sampling hauls with replacement from the abundance survey data becomes problematic when there is only one haul in that stratum. It happens that several strata have only one haul due to some unforeseen reasons, such as bad weather or technical failure. In this case, variance estimation for the single-haul stratum is obtained by borrowing the information of uncertainty in the abundance estimates from a similar stratum. The next section will address the problem of how to estimate variance for a single-haul stratum.

4.5.2 Variance estimation for single-haul strata

Recall that $\widehat{N}_{sa}$ denotes the abundance estimate for stratum s and age a, and its variance estimator is given by (4.25) if every stratum in the abundance survey has at least two hauls. As this is not the case in the 2007 abundance survey, strata with at least one non-empty haul are divided into two groups: strata with only one haul (single-haul strata) and strata with at least two hauls (multi-haul strata). Then the variance estimation of the abundance estimator for fish with age a over all strata becomes

$$\widehat{\text{Var}}[\widehat{N}_a] = \underbrace{\sum_{s:\,(\text{m})} \widehat{\text{Var}}[\widehat{N}_{sa}^{(\text{m})}]}_{\text{from bootstrapping}} + \underbrace{\sum_{s:\,(\text{s})} \widehat{\text{Var}}[\widehat{N}_{sa}^{(\text{s})}]}_{\text{'borrowed'}} \tag{4.25}$$

where the superscript (m) denotes the multi-haul strata, and (s) denotes the single-haul strata.

Section 4.5.1 has described how to use bootstrap methods to include all the potential sources of uncertainty in $\widehat{N}_{sa}^{(\text{m})}$ for a multi-haul stratum, i.e., the first summation in (4.25). This section focuses on the variance estimation for a single-haul stratum, i.e., the second summation in (4.25). The idea is to 'borrow' information about the uncertainty in the abundance estimation for a single-haul stratum from a 'similar' multi-haul stratum, and then the question arises as to which stratum to borrow from and which uncertainty measure is borrowed.

Let s' denote the single-haul stratum for which we need to borrow uncertainty information for its abundance estimation. The uncertainty of $\widehat{N}_{s'a}$ is borrowed from a multi-haul stratum, which is denoted by s^*. The stratum s^* has the most similar density estimate over all age groups to that of the single-haul stratum s', i.e., the least difference between $\hat{\rho}_{s^*}$ and $\hat{\rho}_{s'}$.

In order to illustrate how to take into account the differences in the sampling effort between the two strata, s^* and s', we temporarily ignore the age structure in the abundance estimation. For the single-haul stratum s', $\widehat{N}_{s'} = \widehat{N}_i$ as there is only one haul in this stratum. However, for the multi-haul stratum s^*, if we assume that $\widehat{N}_i$ are independent and identically distributed random variables, then $\widehat{N}_{s^*}$ can be thought of as an average of $\widehat{N}_i$ over all hauls in stratum s^* (recall the derivation based on weighted average method given by (4.8)). Then it follows that

$$\widehat{\text{Var}}\,[\widehat{N}_{s^*}] = \widehat{\text{Var}}\left[\frac{\sum_{i\in s^*}\widehat{N}_i}{\text{nhaul}_{s^*}}\right] = \widehat{\text{Var}}\,[\widehat{N}_i]/\text{nhaul}_{s^*},$$

and equivalently,

$$\widehat{\text{sd}}[\widehat{N}_{s^*}] = \widehat{\text{sd}}[\widehat{N}_i]/\sqrt{\text{nhaul}_{s^*}}, \tag{4.26}$$

where nhaul_{s^*} denotes the number of hauls towed in the multi-haul stratum s^*.

When borrowing uncertainty estimates from a multi-haul stratum s^*, it is important to note that the borrowed uncertainty is for the abundance estimator based on the data collected by this single haul in stratum s'. Therefore, the sampling effort in the multi-haul stratum s^* needs to be taken into account. According to (4.26), we use $\sqrt{\text{nhaul}_{s^*}}$ as a measure of sampling effort when borrowing uncertainty estimates for stratum s'

For the measurement of uncertainty, it is the coefficient of variation (CV) that is borrowed for the stratum s'. The CV can be thought of as a measure of the relative uncertainty. It is more reasonable to borrow the information about the relative uncertainty as it is very likely that these two strata have great differences in their abundance estimates, and then the CV is preferred as it measures the standard error of the abundance estimator relative to its mean.

The variance of the abundance estimator for the single-haul stratum s' is estimated as

$$\begin{aligned}\widehat{\text{Var}}\,[\widehat{N}_{s'a}] &= \left[\widehat{\text{CV}}[\widehat{N}_{s'a}]\,\widehat{N}_{s'a}\right]^2 &(4.27)\\ &= \left[\left(\widehat{\text{CV}}[\widehat{N}_{s^*a}]\sqrt{\text{nhaul}_{s^*}}\right)\widehat{N}_{s'a}\right]^2, &(4.28)\end{aligned}$$

where the red part in (4.27) cannot be obtained in bootstrapping as there is no replication of hauls in stratum s'. The red part in (4.28) is then borrowed for s' from stratum s^*, with $\widehat{\text{CV}}[\widehat{N}_{s^*a}]$ obtained from the bootstrap method described in Section 4.5.1 and $\sqrt{\text{nhaul}_{s^*}}$ as a measure of sampling effort according to (4.26).

Because $\widehat{N}_{s'a}$ is a positive random variable and its sampling distribution is right-skewed, a log-normal distribution is assumed for $\widehat{N}_{s'a}$. Given the CV 'borrowed' in

(4.28), according to (4.66) derived in Appendix 4.D, the variance of $\log(\widehat{N}_{s'a})$ can be estimated as

$$\widehat{\text{Var}}\left[\log\left(\widehat{N}_{s'a}\right)\right] = \log\left\{1 + \left(\widehat{\text{CV}}\left[\widehat{N}_{s'a}\right]\right)^2\right\}. \tag{4.29}$$

Under the log-normal assumption for the distribution of $\widehat{N}_{s'a}$, given the significance level α, an approximate $100(1-\alpha)\%$ confidence interval is

$$(\widehat{N}_{s'a}/\widehat{C}_{s'a}, \widehat{N}_{s'a} \times \widehat{C}_{s'a}),$$

where $\widehat{N}_{s'a}$ is the point estimate given the sampled data in the abundance survey and

$$\widehat{C}_{s'a} = \exp\left[z_{(1-\alpha/2)} \times \sqrt{\text{Var}\left[\log(\widehat{N}_{s'a})\right]}\right],$$

and $z_{(1-\alpha/2)}$ is the $1-\alpha/2$ quantile of a standard normal distribution.

As the abundance estimates of fish at each age class over the whole survey area, $\widehat{N}_a$, are often of interest in stock assessment. It is important to show how uncertainty these estimates are when present the abundance estimation results, such as the results shown in Figures 4.5 and 4.6 in the next section. The confidence intervals presented in these two figures are also based on the log-normal assumption. This assumption is made because bootstrap CIs are only available if the abundance survey has only multi-haul strata. However, this is not the case in the 2007 abundance survey as the stratum 'North.L' has one single haul. Therefore the process of borrowing uncertainty estimates described above is carried out for stratum 'North.L'. The variance of $\widehat{N}_a$ is then estimated according to (4.25) given the variance estimates of stratum 'North.L', which is borrowed as (4.28). The 95% CIs for $\widehat{N}_a$ are calculated the same as the process desribed above for obtaining the CI of $\widehat{N}_{s'a}$.

4.6 Results and discussion

This section presents the abundance estimation results using all the estimators presented in Section 4.3 for the 2007 anglerfish abundance survey: (4.9), (4.11), (4.15), (4.16) and (4.17), for all of which the missing-age data are dealt with using the

probability-based method described in Section 4.4.2. The confidence intervals presented here are obtained by the bootstrap method described in Section 4.5. However, in the case of the estimator (4.9) with $r = 1$, its analytical variance estimator given by (4.10) can be easily calculated, which is consistent with the bootstrap variance estimates.

Even though the estimated standard deviation of the random slope b_0 is zero in (4.13) fitted using 2006-2007 experimental survey data, both random intercept and slope are allowed in the bootstrap variance estimation when fitting mixed-effects logistic regression models to the re-sampled experimental survey data. Given the re-sampled experimental survey data, the estimated variance-covariance matrix $\widehat{\Sigma}_b$ can be either singular or of full rank. A singular $\widehat{\Sigma}_b$ means that the fitted model only has random intercept or random slope, while a full rank $\widehat{\Sigma}_b$ means that the fitted model has both random intercept and slope. However, no matter whether $\widehat{\Sigma}_b$ is singular or full rank, the corresponding abundance estimators are of the same form given by (4.19) and (4.21) for $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$, respectively. Note that $\widehat{N}^{(1)}$ does not involve $\widehat{\Sigma}_b$ since it assumes $\boldsymbol{b}_i = \mathbf{0}$ for any haul i. $\widehat{N}^{(2)}$ is obtained by a numerical approximation to (4.19), and this approximation is done the same way as described in Section 4.3.3.

Table 4.4 presents the estimates of the overall abundance, i.e., the abundance estimates over all ages and all strata, with 95% CIs estimated by (4.25) based on the 999 samples from bootstrapping with haul as the sampling unit. It can been seen that

TABLE 4.4. Anglerfish abundance estimates in millions of fish with 95% CIs: for the 'Fixed-effects $\hat{r}(l)$' part, '$\widehat{N}$ (linear logistic)' means the abundance estimator (4.11) with $\hat{r}(l)$ estimated by a linear logistic regression model (see (3.36) in Section 3.2.1 for full detail), and '$\widehat{N}$ (asymptote-logistic)' means the abundance estimator (4.11) with $\hat{r}(l)$ estimated by an asymptote-logistic regression model (see (3.43) in Section 3.2.1 for full detail). For the 'Mixed–effects $\hat{r}_i(l)$' part, $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$, and $\widehat{N}^{(3)}$ are given by (4.15), (4.16) and (4.17), respectively.

	Point estimate	95% CI	
Estimator	$\widehat{N}(\times 10^6)$	Lower	Upper
$\widehat{N}$ with $r = 1$	21.42	17.85	25.69
Fixed-effects $\hat{r}(l)$:			
$\widehat{N}$ (linear logistic)	28.19	23.02	34.53
$\widehat{N}$ (asymptote-logistic)	28.96	23.25	36.05
Mixed-effects $\hat{r}_i(l)$: [a]			
$\widehat{N}^{(1)}$	27.79	22.47	34.36
$\widehat{N}^{(2)}$	27.82	22.63	34.38
$\widehat{N}^{(3)}$	28.06	1.51×10^{-7}	5.21×10^{9}

[a] the confidence interval estimates incorporate se$(\hat{\sigma}_0)$ and se$(\hat{\sigma}_1)$ by bootstrapping the experimental survey data, and for each re-sampled data set, fitting a two-level mixed-effects logistic regression model given by (4.13).

- $\widehat{N}$ with $r = 1$ has the most narrow confidence interval among all estimators presented in Table 4.4, which is expected because this $\widehat{N}$ with $r = 1$ ignores the uncertainty in the net retention probability.

- There is no big difference among the other estimators except that the CI of $\widehat{N}^{(3)}$ is much wider than the other estimators with $\hat{r}_i(l)$. This can be explained by comparing (4.18) with (4.21). Let $\hat{r}_{il}^{(1)}$ denote the estimated net retention probability conditional on $\boldsymbol{b}_i = \mathbf{0}$, i.e.,

$$\hat{r}_{il}^{(1)} = \hat{r}(l, \bar{l}_i | \boldsymbol{b}_i = \mathbf{0}) = \left(1 + \exp\left[-\hat{\beta}_0 - \hat{\beta}_1^B \bar{l}_i - \hat{\beta}_1^W (l - \bar{l}_i)\right]\right)^{-1}.$$

It follows that $\left(\hat{r}_{il}^{(1)}\right)^{-1}$ is the expression inside the curly brackets of (4.18), and $\hat{r}_{il}^{(1)} \in (0,\, 1)$. Then for the fish of length l and age a captured by haul i, the ratio of $\widehat{N}_{ila}^{(3)}$ to $\widehat{N}_{ila}^{(1)}$ is

$$\begin{aligned} & \frac{1+\exp\left[-\hat{\beta}_0-\hat{\beta}_1^B\,\bar{l}_i-\hat{\beta}_1^W\,(l-\bar{l}_i)+\frac{\hat{\sigma}_0^2+(l-\bar{l}_i)^2\hat{\sigma}_1^2}{2}\right]}{\left(\hat{r}_{il}^{(1)}\right)^{-1}} \\ = \;& \frac{1+\left(\hat{r}_{il}^{(1)}\right)^{-1}\exp\left[\frac{\hat{\sigma}_0^2+(l-\bar{l}_i)^2\hat{\sigma}_1^2}{2}\right]-\exp\left[\frac{\hat{\sigma}_0^2+(l-\bar{l}_i)^2\hat{\sigma}_1^2}{2}\right]}{\left(\hat{r}_{il}^{(1)}\right)^{-1}} \\ = \;& \hat{r}_{il}^{(1)}+\left(1-\hat{r}_{il}^{(1)}\right)\exp\left[\frac{\hat{\sigma}_0^2+(l-\bar{l}_i)^2\hat{\sigma}_1^2}{2}\right]\geq 1\,. \end{aligned} \tag{4.30}$$

Because $\exp\left[\frac{\hat{\sigma}_0^2+(l-\bar{l}_i)^2\hat{\sigma}_1^2}{2}\right] \geq 1$, (4.30) is no less than 1. When $\hat{\sigma}_1$ is non-zero and $\hat{r}_{il}^{(1)}$ is small, this ratio can be quite large. Hence, we could see a very large upper bound of the percentile CI for $\widehat{N}^{(3)}$, when comparing it with $\widehat{N}^{(1)}$ for the multi-haul strata. However, the CI of $\widehat{N}^{(3)}$ in Table 4.4 is extremely wide with a very small lower bound, and this suggests that the log-normal assumption that is assumed for all three estimators holds for $\widehat{N}^{(1)}$ and $\widehat{N}^{(2)}$, but not for $\widehat{N}^{(3)}$.

- We also check the effect of conditioning on $\widehat{\Sigma}_b$ in variance estimation of $\widehat{N}$, as is often done with this kind of model. For the anglerfish application, if conditioning on the $\widehat{\Sigma}_b$ from the 2006-2007 experimental survey data, the bootstrap variance estimation has $\hat{\sigma}_0 = 0.298$ and $\hat{\sigma}_1 = 0$. In this case, the CIs of $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$ are very different from those obtained by including uncertainty of $\widehat{\Sigma}_b$. The comparison of these two sets of CIs is given by Table 4.5. The column 'Conditional on $\widehat{\Sigma}_b$' is obtained by parametric bootstrapping $\hat{\beta}_0$, $\hat{\beta}_1^B$ and $\hat{\beta}_1^w$ from their estimated asymptotic distribution, with $\hat{\sigma}_0 = 0.289$ and $\hat{\sigma}_1 = 0$. When variance estimation is conditioned on $\hat{\sigma}_0 = 0.289$ and $\hat{\sigma}_1 = 0$, the only difference between (4.18) and (4.21) for fish of length l captured by haul i is $\exp\left[\frac{\hat{\sigma}_0^2+(l-\bar{l}_i)^2\hat{\sigma}_1^2}{2}\right]$, and it remains equal to 1.043. This explains the small difference in CIs between $\widehat{N}^{(1)}$ and $\widehat{N}^{(3)}$ in Table 4.5 when variance estimation is conditioned on $\widehat{\Sigma}_b$.

TABLE 4.5. Abundance estimators with random effects in millions of fish with 95% CIs: variance estimation conditional on $\widehat{\Sigma}_b$ vs including se$(\widehat{\Sigma}_b)$. $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$, and $\widehat{N}^{(3)}$ are given by (4.15), (4.16) and (4.17), respectively.

Estimator	Conditional on $\widehat{\Sigma}_b$	Including se$(\widehat{\Sigma}_b)$
$\widehat{N}^{(1)}$	(22.39, 34.50)	(22.47, 34.36)
$\widehat{N}^{(2)}$	(22.43, 34.51)	(22.63, 34.38)
$\widehat{N}^{(3)}$	(22.56, 34.90)	$(1.51\times 10^{-7}, 5.21 \times 10^{9})$

Therefore, conditioning on $\widehat{\Sigma}_b$ estimated from the experimental survey data can lead to misleading CIs for the abundance estimators with haul effect, especially $\widehat{N}^{(3)}$. In the rest of this section, we only discuss the variance estimation for $\widehat{N}^{(m)}$, $m = 1, 2, 3$, when the uncertainty of $\widehat{\Sigma}_b$ is incorporated in variance estimation by fitting a two-level mixed-effects model to the simulated experimental survey data.

For a better understanding of the difference between the estimators with haul effect, we plot the abundance estimates for each age group in Figure 4.5. This figure shows the difference between the abundance estimators with haul effect: there is almost no difference in point estimates among $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$, except that $\widehat{N}^{(3)}$ is slightly larger than the other two estimators from age 0 to age 4 (as the green bars are slightly higher); for the estimated CIs, only at age 0 and age 1, $\widehat{N}^{(1)}$ has slightly wider CIs than $\widehat{N}^{(2)}$. Note that the CIs for $\widehat{N}^{(3)}$ are not given for any age group in Figure 4.5. This is because of the extremely wider CI for $\widehat{N}^{(3)}$.

Given the little difference among the abundance estimator with haul effect shown in Figure 4.5, Figure 4.6 plots the estimation results for $\widehat{N}^{(2)}$ and all the other abundance estimators without haul effect, i.e., $\widehat{N}$ with assuming $r = 1$ and $\widehat{N}$ with $\hat{r}(l)$ estimated by fixed-effects regression models. It can been seen that

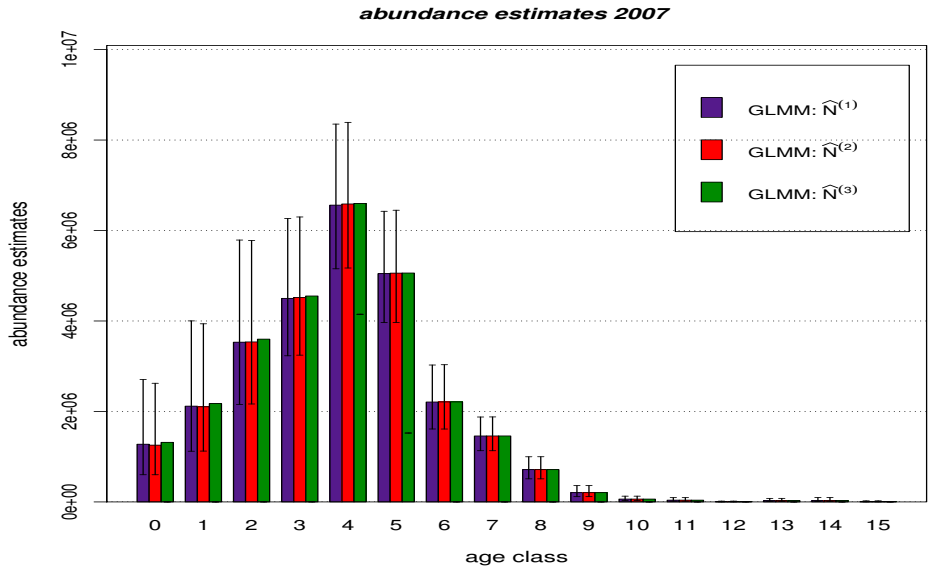


FIGURE 4.5. The 2007 abundance estimation results with r estimated by a two-level mixed–effects logistic regression model, i.e., the estimators $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$, together with their 95% CI obtained from bootstrapping both the abundance survey and experimental survey data. $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$ are given by (4.15), (4.16) and (4.17), respectively. The net retention probability used in these estimators, $\hat{r}_i(l)$, is given by (3.72); see Section 3.4 for full detail.

- The greatest difference among these four estimators occurs at the very small age groups (i.e., age 0 and age 1) and for age groups beyond age 8, there is no difference in the estimates and their CIs among these four estimators.

- $\widehat{N}$ with $r = 1$ has the lowest estimates with the most narrow CIs for fish up to age 6, which is caused by assuming $r = 1$. Such an assumption for r ignores the length-based net retention probability, and fish of difference sizes have the same inclusion probability. This leads to underestimation of the abundance estimates together with their variances, particularly for small fish.

- $\widehat{N}$ with asymptote-logistic $\hat{r}(l)$ (the 'asymptote' in Figure 4.6) gives higher abundance estimates than the other estimators and it also has the widest CIs. This difference can be explained by the asymptote-logistic curve in Figure 3.3, which shows that the most difference of the asymptote-logistic curve from the linear logistic curve ('linear logistic' in Figure 4.6) occurs at small length classes. The lower bound of its CI for $\hat{r}(l)$ reaches almost zero at small length classes (see Figure 3.3 for details), which explains why the upper bound of

the 95% CI for asymptote-logistic is much higher than that of $\widehat{N}$ with linear logistic $\hat{r}(l)$ at age 0.

About the potential extrapolation problem mentioned in Section 3.2.2 when using the estimated asymptote-logistic regression model to predicting $r(l)$ for the abundance survey, though there are 10% of fish larger than 70 cm in the abundance survey (maximum length of fish captured in the experimental survey is 70 cm), the effect of extraplotion for larger fish is neglectible when using the HT-like estimators. This is because there is little uncertainty in the HT-like estimator when the capture probability is almost certain for the fish larger than 60 cm.

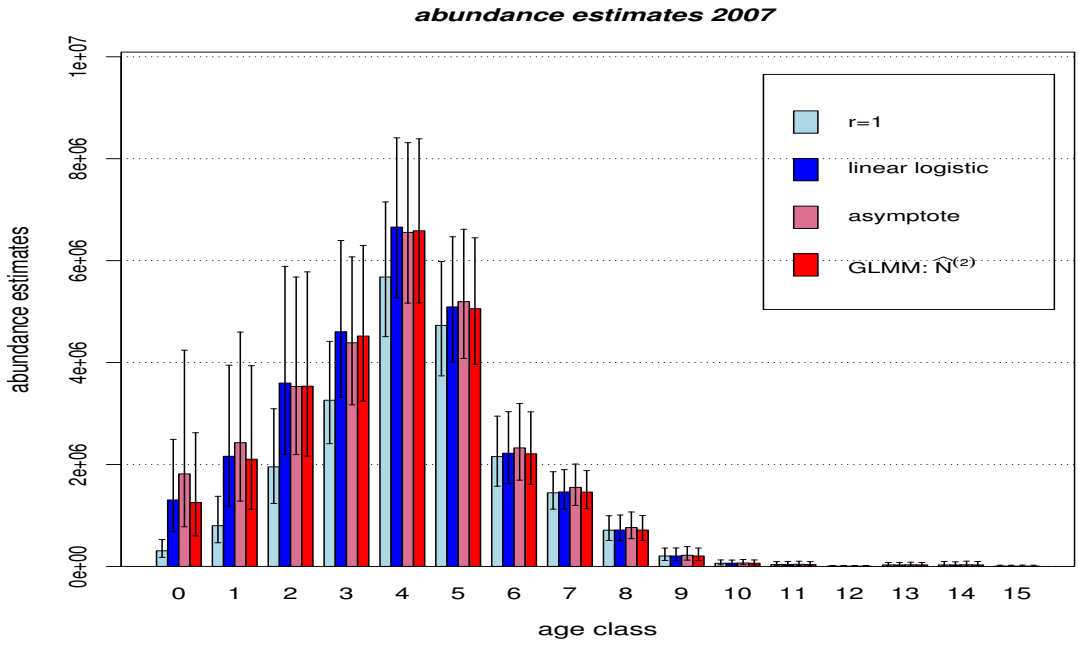


FIGURE 4.6. The 2007 abundance estimation results using the abundance estimators (4.9) with $r = 1$, (4.16) with $\hat{r}(l)$ estimated by a linear logistic regression model, (4.11) with $\hat{r}(l)$ estimated by fixed-effects logistic regression, and (4.16) with $\hat{r}_i(i)$ estimated by a fixed-effects logistic regression model. The 'linear logistic' and 'asymptote' stand for the selected fixed-effects models given by (3.36) and (3.43), respectively. The 'GLMM: $\widehat{N}^{(2)}$' stands for the estimator $\widehat{N}^{(2)}$ given by (4.16) with $\hat{r}_i(l)$ of the form (4.13).

The fish captured in the Irish waters make up about 20% of the 2007 anglerfish abundance survey data. The data from the Irish waters have no age information, and therefore age imputation is required when including the Irish data in the anglerfish

abundance estimation. The probability-based method described in Section 4.4.2 is implemented to obtain the abundance estimates for each age class. To check the effect of the missing-age data collected in the Irish waters, we apply the abundance estimation method, which is the same as the one used to obtain the results in Figures 4.5 and 4.6, to the data from the Scottish waters only (i.e., excluding the missing-age data from the Irish waters). The obtained results for the Scottish waters are plotted in Figures 4.7 and 4.8.

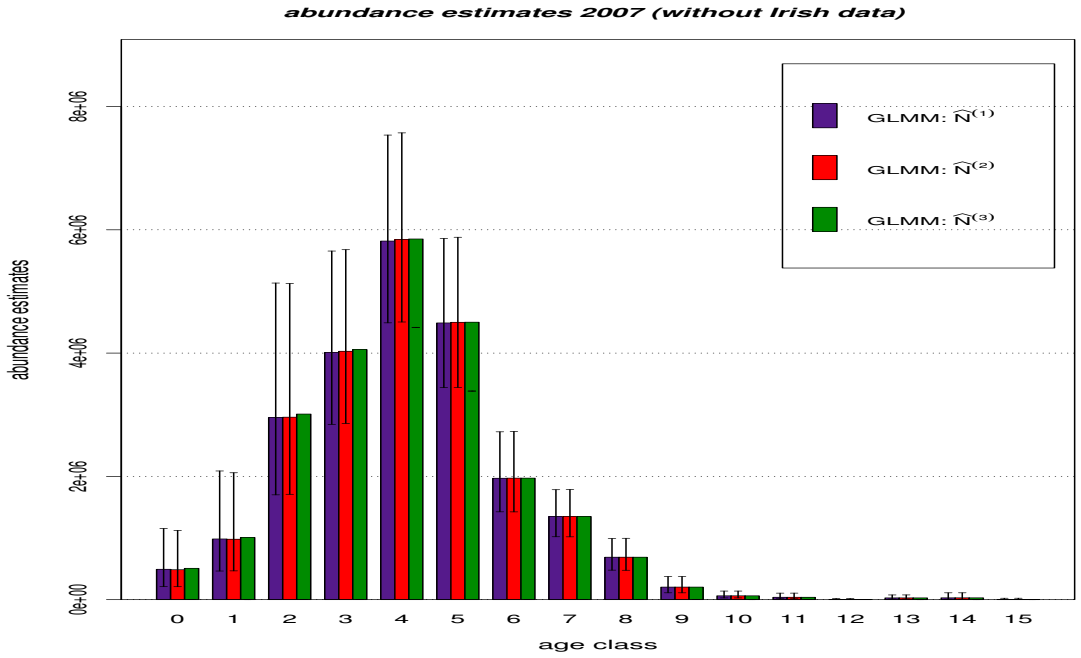


FIGURE 4.7. The 2007 abundance estimation results with r estimated by a two-level mixed–effects logistic regression model, i.e., the estimators $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$, together with their 95% CI obtained from bootstrapping both the abundance survey and experimental survey data. $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$ are given by (4.15), (4.16) and (4.17), respectively. The net retention probability used in these estimators, $\hat{r}_i(l)$, is given by (3.72); see Section 3.4 for full detail. The abundance estimation in this plot uses the 2007 abundance survey data without the fish captured in the Irish waters, and the estimation process (including the bootstrap variance estimation) is the same as the process used for obtaining the results shown in Figure 4.5.

Comparing Figure 4.5 to Figure 4.7 and Figures 4.6 to Figure 4.8, the results without the Irish data are very similar to those with the Irish waters: the same pattern over all estimators at each age class except the lower point estimates and narrower CIs, which is expected to happen after excluding the missing-age data from the Irish waters. Therefore, the probability-based method for the missing-age data does not

affect the performance of the abundance estimators discussed here. As long as the sample size of the survey data with age information is sufficient for estimating the age distributions at each length class, the probability-based method can be used and it does not affect the performance of the abundance estimators except more uncertainty (e.g., wider CIs) contributed by the missing age data.

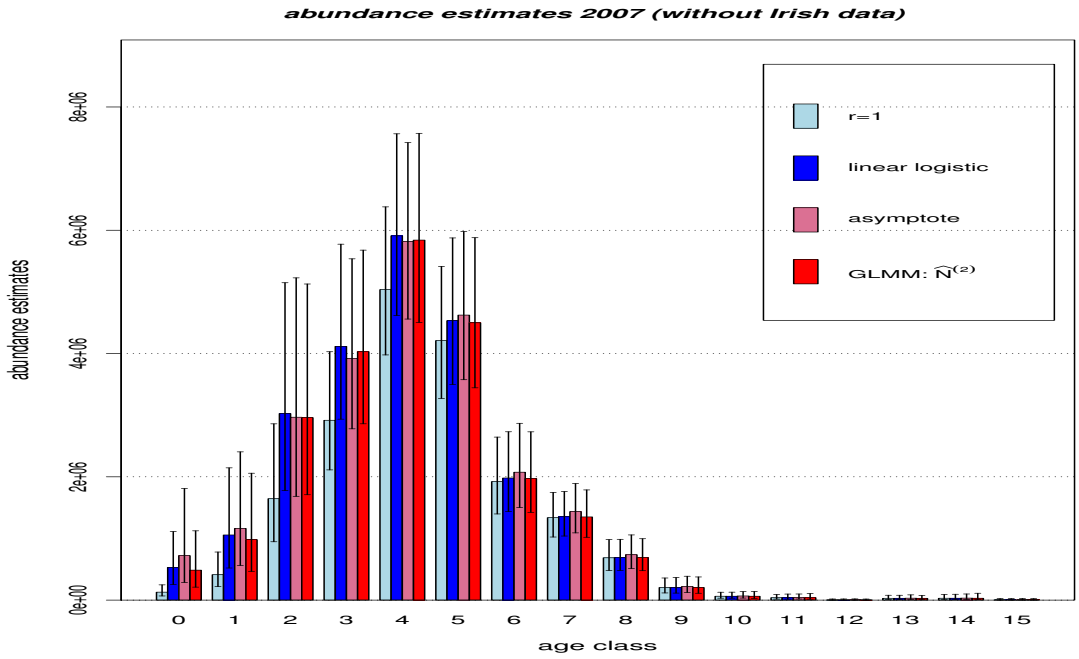


FIGURE 4.8. The 2007 abundance estimation results using the abundance estimators (4.9) with $r = 1$, (4.16) with $\hat{r}(l)$ estimated by a linear logistic regression model, (4.11) with $\hat{r}(l)$ estimated by fixed-effects logistic regression, and (4.16) with $\hat{r}_i(i)$ estimated by a fixed-effects logistic regression model. The 'linear logistic' and 'asymptote' stand for the selected fixed-effects models given by (3.36) and (3.43), respectively. The 'GLMM: $\widehat{N}^{(2)}$' stands for the estimator $\widehat{N}^{(2)}$ given by (4.16) with $\hat{r}_i(l)$ of the form (4.13). The abundance estimation in this plot uses the 2007 abundance survey data without the fish captured in the Irish waters, and the estimation process (including the bootstrap variance estimation) is the same as the process used for obtaining the results shown in Figure 4.6.

The density estimates are also of great interest in making fishery management decisions. Identifying the change in density for a particular area is important in setting up regional fishing regulations to achieve a long-term sustainability of the targeted stock. Figure 4.9 shows the density estimates and their bootstrap CIs for the multi-haul strata in the 2007 abundance survey. Again, there are no very large differences in the density estimates for different estimators: the density estimates with $r = 1$ are lower than that with $\hat{r}(l)$ and $\hat{r}_i(l)$, especially for those strata with more small

fish, such as 'North.M1.E' and 'Rockall.L1'. The highest densities (the number of fish per square kilometre) occurred in the stratum 'North.M1.E', which is 140-200 metres deep and in the survey area 'North of Scotland' east of -4° longitude. The lowest density estimates occurred in the stratum 'North.M2.W'– a 200-500 metres deep water stratum east in the survey area 'North of Scotland' west of -4° longitude. Note that there was one haul towed in the very deep (500-1000 metres) stratum 'North.L' in the survey area 'North of Scotland', but it captured no fish.

To sum up, the estimates with and without haul effect, i.e., $\hat{r}(l)$ and $\hat{r}_i(l)$, produce almost the same density or abundance estimation results among all age groups, except that for the point estimates up to age 5, $\widehat{N}^{(2)}$ gives lower estimates than $\widehat{N}$ with linear logistic, and the CIs of $\widehat{N}^{(2)}$ are slightly narrower than those of $\widehat{N}$ with linear logistic $\hat{r}(l)$ for age 1 to age 4. The effect of haul on the HT-like estimators $\widehat{N}^{(m)}$, $m = 1, 2, 3$, is not clear, except the CIs are much wider for $\widehat{N}^{(3)}$ than the other two. The haul effect makes little difference in the anglerfish abundance estimation when using $\widehat{N}^{(m)}$, $m = 1, 2, 3$. This might be a result of the small haul effect ($\hat{\sigma}_0 = 0.289$ relative to $\hat{\beta}_0 = -3.606$), or the fact that capture is almost certain for a fish with length beyond 50 cm based on the estimation results for both the fixed-effects models in Figure 3.3 and the random-effects model in Figure 3.9. However, the $\widehat{N}^{(1)}$ is much easier to calculate and it produces almost the same results as $\widehat{N}^{(2)}$ in Figure 4.5. Then the question becomes which estimator to use in abundance estimation, and what would happen if the haul effect is ignored. To answer these questions, it requires a better understanding of the performance of HT-like estimators with random effects, i.e., $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$, and $\widehat{N}^{(3)}$ as given by (4.15) – (4.17). The next chapter performs a simulation study to examine the performance of these estimators, and the haul effect in abundance estimation.

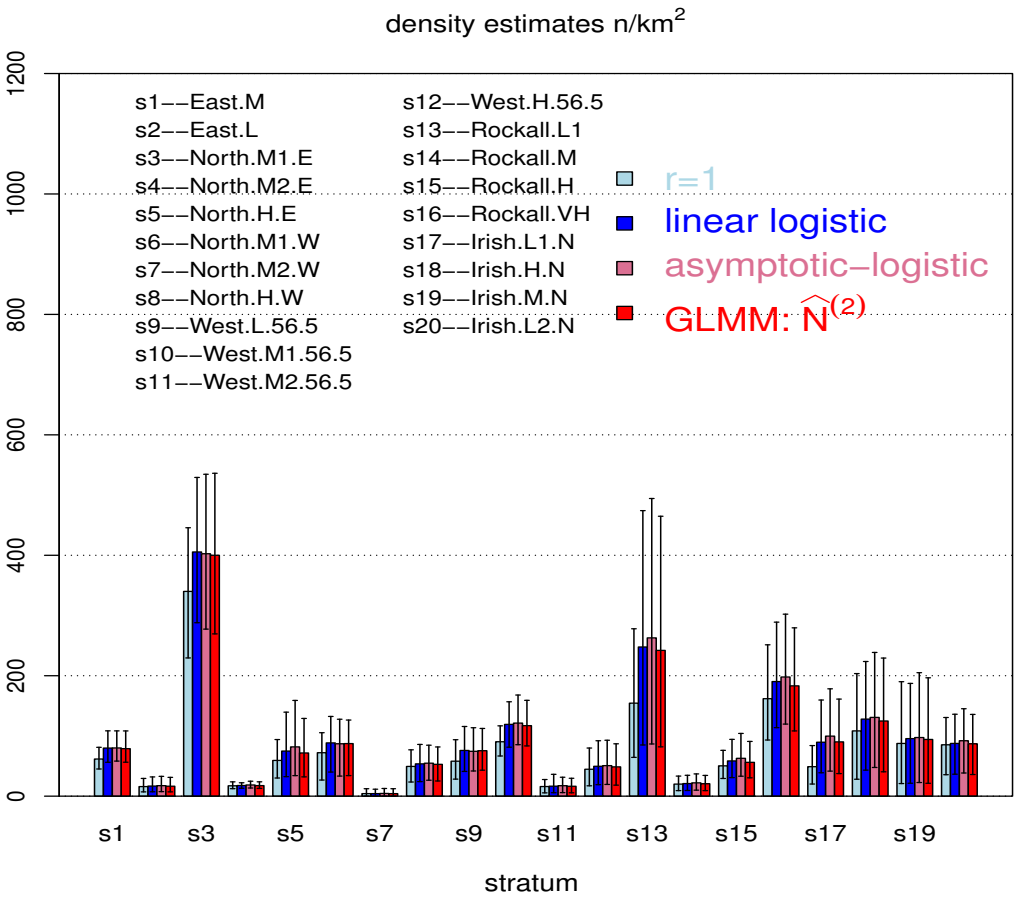


FIGURE 4.9. Plot of the density estimates for the multi-haul strata in the 2007 abundance survey with their 95% CIs for the 2007 abundance survey. The density estimates are obtained with net retention probability ('$r = 1$'), estimated by a linear logistic regression model ('linear logistic'), and estiimated by a two-level mixed-effects logistic regression model with abundance estimator $\widehat{N}^{(2)}$ given by (4.19).

Appendices – **Chapter 4**

4.A Survey trawl efficiency

The survey trawl *efficiency* is widely used in fishery research for density estimation. Somerton *et al.* (1999) defined the efficiency of a survey trawl (denoted by Q) as the proportion of the fish removed by one unit of fishing effort (usually measured by the swept area of the trawl). Somerton *et al.* (1999) also pointed out that in terms of using experimental surveys to estimate Q, three components of the catching process should be considered in the survey: vertical herding, horizontal herding and escapement. As described in Chapter 2, only horizontal herding and escapement beneath the footrope are the sources of catchability for the anglerfish survey. They are denoted by h and r respectively. Then Q for anglerfish survey is

$$Q = r + r\,h\,\frac{v_2}{v_1}, \tag{4.31}$$

where v_1 is the area swept by wings, and v_2 is the area between the wing ends and the doors, both of which have been defined in Figure 1.2. Note that it is not necessary for Q to be less than 1, though Somerton *et al.* (1999) defined it as a proportion. When assuming $r = 1$, the Q given in (4.31) is larger than 1 if $h > 0$. Therefore, the trawl efficiency is not a probability.

Given the trawl efficiency defined by (4.31), the question then arises of how to incorporate Q in the stock assessment process, such as abundance, density or biomass estimation of the targeted species. The rest of this appendix derives the density estimator from two different perspectives, both of which lead to the same estimator. This resulting estimator provides the basis for the density estimator (4.7) for a given haul, age and length class.

Without loss of generality, only one haul is considered here, and both r and h are assumed to be unknown constants in the derivation and the following gives a list of the notation used later in this appendix,

- N_1: number of fish in area v_1 before trawling,
- N_2: number of fish in area v_2 before trawling,

- n_1: number of fish captured out of the population N_1,
- n_2: number of fish captured out of the population N_2,
- n: total number of catch given the trawl, and
- ρ: the density of the fish in the area covered by this haul.

4.A.1 Derivation of (4.31)

It is assumed that fish are uniformly distributed in the survey area, and hence the ratio of the population size in area v_2 to that in area v_1 is the ratio of the two swept areas, v_2/v_1. Therefore, given N_1 and the two swept areas, $N_2 = v_2/v_1 N_1$. It is also assumed that $n_1 \sim \text{Binomial}(N_1, r)$ and $n_2 \sim \text{Binomial}(N_2, r\,h)$, where h gives the probability of a fish being herded from v_2 into v_1 after the trawling process. There are two sources, N_1 and N_2, of the captured n fish, therefore,

$$\begin{aligned}
\mathbf{E}[n] &= \mathbf{E}[n_1 \,|\, N_1] \,+\, \mathbf{E}[n_2 \,|\, N_2] \\
&= r\,N_1 \,+\, r\,h\,\mathbf{E}[n_2 \,|\, N_2] \\
&= r\,N_1 \,+\, r\,h\,\frac{v_2}{v_1} N_1 \\
&= \left(r \,+\, r\,h\,\frac{v_2}{v_1}\right) N_1,
\end{aligned}$$

and it follows that the estimator of N_1 is

$$\widehat{N}_1 = \frac{n}{\left(r \,+\, r\,h\,\dfrac{v_2}{v_1}\right)}, \tag{4.32}$$

of which the denominator is the trawl efficiency defined in (4.31). Hence the density estimator is

$$\begin{aligned}
\hat{\rho} &= \frac{\widehat{N}_1}{v_1} \\
&= \frac{n}{r\,v_1 \,+\, r\,h\,v_2}.
\end{aligned} \tag{4.33}$$

This estimator was introduced by Somerton (1996).

4.A.2 Derivation of (4.33) using the Horvitz-Thompson method

The definition of the trawl efficiency in (4.31) can also be illustrated in the context of the Horvitz-Thompson method. The central part of the Horvitz-Thompson estimator is the probability that the ith captured fish is included in the sample, i.e., the inclusion probability. In the setting of the above definition for swept areas v_1 and v_2, the inclusion probability is therefore the probability that a fish is captured given it was in the swept area $v_1 + v_2$ before the trawling process. Let $c = 1$ indicate that a fish is captured by the trawl, and $c = 0$ otherwise. Then the probability that a fish is captured, $\mathbf{P}\{c = 1\}$, can be expressed as

$$\begin{aligned}
& \mathbf{P}\{c = 1|\text{it comes from } v_1 + v_2\} \\
= \; & \mathbf{P}\{c = 1|\text{it comes from } v_1\}\, \mathbf{P}\{\text{it comes from } v_1\} \\
& + \mathbf{P}\{c = 1|\text{it comes from } v_2\}\, \mathbf{P}\{\text{it comes from } v_2\} \\
= \; & \mathbf{P}\{c = 1|\text{it comes from } v_1\}\, \frac{v_1}{v_1 \,+\, v_2} \\
& + \mathbf{P}\{c = 1|\text{it comes from } v_1\} \times \mathbf{P}\{\text{herded into } \; v_1|\text{it comes from } v_2\} \\
= \; & r\, \frac{v_1}{v_1 \,+\, v_2} \,+\, r\, h\, \frac{v_2}{v_1 \,+\, v_2}.
\end{aligned}$$

Given the above capture probability, applying the Horvitz-Thompson method, the abundance of the fish in swept areas v_1 and v_2 can be estimated by

$$\begin{aligned}
\widehat{N}_1 \,+\, \widehat{N}_2 \;&=\; \frac{n}{\mathbf{P}(c = 1|\text{it comes from } v_1 + v_2)} \\
&=\; \frac{n\,(v_1 \,+\, v_2)}{r\, v_1 \,+\, r\, h\, v_2}.
\end{aligned} \tag{4.34}$$

Therefore the density can be estimated by

$$\begin{aligned}
\hat{\rho} \;&=\; \frac{\widehat{N}_1 \,+\, \widehat{N}_2}{v_1 \,+\, v_2} \\
&=\; \frac{n}{r\, v_1 \,+\, r\, h\, v_2},
\end{aligned} \tag{4.35}$$

which is the same as the density estimator (4.33) given by Somerton (1996), which is commonly used in fisheries research.

4.B New weights for an unbiased density estimator

In a previous analysis of the anglerfish annual abundance survey data, Fernandes *et al.* (2007) proposed a density estimator by taking the average of the density estimates by hauls within each stratum with weight the reciprocal of the total number of hauls towed in each stratum, i.e.,

$$\hat{\rho}_{sa} = \frac{1}{\text{nhaul}_s} \sum_{i \in s} \hat{\rho}_{ia} \tag{4.36}$$

where nhaul_s denotes the total number of hauls in stratum s. Let $v_i = v_{1i} + v_{2i}$ denote the area swept by the trawl i (both the wings and doors, see Figure 1.2 for details), and v_s denote the swept area by all the hauls towed in stratum s. We propose the weight $w_i = v_i / v_s$ for the ith stratum.

This appendix shows why these new weights are proposed for an unbiased density estimator (if all the $\hat{\rho}_{ia}$ in (4.36) are unbiased). Instead of estimating abundance over all the survey area, the discussion here focuses on the abundance in the covered area within each stratum, i.e., the abundance of fish from the swept area by all the hauls in each stratum, denoted by N^c_{sa} with superscript c meaning the abundance in the covered area to distinguish it from the N_{sa}, the abundance in stratum s. Therefore, the weighted-average density estimator with the new proposed weights is

$$\hat{\rho}_{sa} = \sum_{i \in s} w_i \hat{\rho}_{ia} = \sum_{i \in s} \frac{v_i}{v_s} \hat{\rho}_{ia}, \tag{4.37}$$

where $\hat{\rho}_{ia} = \widehat{N}^c_{ia} / v_i$ and $\widehat{N}^c_{ia}$ is obtained by using the HT-like estimator (see Section 4.2 for details). If the HT-like estimator $\widehat{N}^c_{ia}$ is unbiased, then the question becomes how to choose the averaging weights so that $\hat{\rho}_{sa}$ is an unbiased estimator. To address this question, the expectation of both sides of the following equation is examined

$$\hat{\rho}_{sa} = \sum_{i \in s} w_i \hat{\rho}_{ia}.$$

For the left side, if $\hat{\rho}_{sa}$ is unbiased, then

$$\mathbf{E}[\hat{\rho}_{sa}] = \rho_{sa} = \frac{N^c_{sa}}{v_s} = \frac{\sum_{i \in s} N^c_{ia}}{v_s}. \tag{4.38}$$

On the right side,

$$\mathbf{E}\left[\sum_{i\in s} w_i \hat{\rho}_{ia}\right] = \sum_{i\in s} w_i \mathbf{E}\left[\frac{\widehat{N}_{ia}^c}{v_s}\right]. \tag{4.39}$$

On equating (4.38) and (4.39), it follows that

$$\frac{\sum_{i\in s} N_{ia}^c}{v_s} = \sum_{i\in s} w_i \frac{\mathbf{E}[\widehat{N}_{ia}^c]}{v_i}. \tag{4.40}$$

Therefore, given that $\widehat{N}_{ia}^c$ is unbiased for each haul in stratum s, i.e., $\mathbf{E}[\widehat{N}_{ia}^c] = N_{ia}^c$ for all i on the right side of (4.40), it follows that the weight $w_i = v_i/v_s$ will lead to an unbiased weighted average density estimator within each stratum s.

4.C Delta method and analytical variance estimation for anglerfish survey

For a nonlinear function of random variables, its variance can be estimated by the Delta method. This appendix introduces the Delta method and shows how it can be used to approximate the variance of a function of random variables.

4.C.1 Delta method

Let X_i be a random variable with mean μ_i, where $i = 1, 2, \ldots, n$, and $\boldsymbol{X}$ and $\boldsymbol{\mu}$ be the corresponding vectors, i.e., $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$. The aim is to find the mean, variance and covariance of a function of $\boldsymbol{X}$, denoted by $g(\boldsymbol{X})$. To start with, the Taylor series of $g(\boldsymbol{X})$ is expanded at $\boldsymbol{\mu}$ with the terms with order higher than two ignored. Therefore, $g(\boldsymbol{\mu})$ can be approximated as

$$\begin{aligned} g(\boldsymbol{X}) \;\approx\; & g(\boldsymbol{\mu}) + \sum_{i=1}^{n} (X_i - \mu_i) \frac{\partial g}{\partial X_i}\bigg|_{\boldsymbol{\mu}} \\ & \underbrace{+ \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{(X_i - \mu_i)(X_j - \mu_j)}{2!} \frac{\partial^2 g}{\partial X_i \partial X_j}\bigg|_{\boldsymbol{\mu}}}_{\text{quadratic term}}, \end{aligned} \tag{4.41}$$

where all the partial derivatives are evaluated at $\boldsymbol{\mu}$.

1. Mean of $g(\boldsymbol{X})$:

   Taking expectation of both sides of (4.41),

$$\begin{aligned}\mathbf{E}[g(\boldsymbol{X})] &\approx g(\boldsymbol{\mu}) + \sum_{i=1}^{n} \underbrace{\mathbf{E}[X_i - \mu_i]}_{0} \frac{\partial g}{\partial X_i}\bigg|_{\boldsymbol{\mu}} \\ &\qquad + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{2}\mathbf{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right] \frac{\partial^2 g}{\partial X_i \partial X_j}\bigg|_{\boldsymbol{\mu}}}_{w} \\ &= g(\boldsymbol{\mu}) + w,\end{aligned}$$

   and

$$\begin{aligned} w &= \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{2}\widehat{\text{Cov}}\left[X_i,\, X_j\right] \frac{\partial^2 g}{\partial X_i \partial X_j}\bigg|_{\boldsymbol{\mu}} \\ &= \sum_{i=1}^{n} \frac{1}{2}\text{Var}\left[X_i\right] \frac{\partial^2 g}{\partial X_i^2}\bigg|_{\boldsymbol{\mu}} + \sum\sum_{i<j} \text{Cov}\left[X_i,\, X_j\right] \frac{\partial^2 g}{\partial X_i \partial X_j}\bigg|_{\boldsymbol{\mu}}, \end{aligned}$$

   which means that if X_i and X_j are independent for any $i \neq j$, then the expectation of $g(\boldsymbol{X})$ can be approximated as

$$\mathbf{E}[g(\boldsymbol{X})] \approx g(\boldsymbol{\mu}) + \sum_{i=1}^{n} \frac{1}{2}\text{Var}\left[X_i\right] \frac{\partial^2 g}{\partial X_i^2}\bigg|_{\boldsymbol{\mu}}.$$

2. Variance of $g(\boldsymbol{X})$:

$$\text{Var}\left[g(\boldsymbol{X})\right] = \mathbf{E}\left[\{g(\boldsymbol{X}) - \mathbf{E}[g(\boldsymbol{X})]\}^2\right]. \tag{4.42}$$

   If the quadratic term in (4.41) is ignored, then

$$g(\boldsymbol{X}) \approx g(\boldsymbol{\mu}) + \sum_{i=1}^{n} (X_i - \mu_i) \frac{\partial g}{\partial X_i}\bigg|_{\boldsymbol{\mu}}, \tag{4.43}$$

   and

$$\mathbf{E}[g(\boldsymbol{X})] \approx g(\boldsymbol{\mu}). \tag{4.44}$$

Together with the expression (4.43), the variance of $g(\boldsymbol{X})$ given in (4.42) can be approximated as

$$\begin{aligned}
& \text{Var}\,[g(\boldsymbol{X})] \\
\approx\ & \mathbf{E}\left[\left\{\sum_{i=1}^{n}(X_i-\mu_i)\frac{\partial g}{\partial X_i}\right\}^2\right] \\
=\ & \sum_{i=1}^{n}\text{Var}\,[X_i]\left(\frac{\partial g}{\partial X_i}\right)^2\Big|_{\boldsymbol{\mu}} + 2\sum\sum_{i<j}\text{Cov}\,[X_i,\,X_j]\,\frac{\partial g}{\partial X_i}\frac{\partial g}{\partial X_j}\Big|_{\boldsymbol{\mu}} \\
=\ & \nabla g(\boldsymbol{\mu})^T\,\text{Var}\,[\boldsymbol{X}]\,\nabla g(\boldsymbol{\mu}), \qquad (4.45)
\end{aligned}$$

where Var $[\boldsymbol{X}]$ denotes the variance-covariance matrix of $\boldsymbol{X}$.

3. Covariance between $g(\boldsymbol{X})$ and $f(\boldsymbol{Y})$:
   The Delta method can also be used for approximating the covariance between two functions of random vectors, denoted by $g(\boldsymbol{X})$ and $f(\boldsymbol{Y})$ respectively, where $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ and $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_m)^T$. Similarly for $f(\boldsymbol{Y})$, i.e., applying (4.44) and (4.45) for the expectation and variance estimation of $f(\boldsymbol{Y})$, the covariance between function $g(\boldsymbol{X})$ and $f(\boldsymbol{Y})$ can be approximated as

$$\begin{aligned}
& \text{Cov}\,[g(\boldsymbol{X}), f(\boldsymbol{Y})] \\
=\ & \mathbf{E}\,[\{g(\boldsymbol{X})-\mathbf{E}[g(\boldsymbol{X})]\}\{f(\boldsymbol{Y})-\mathbf{E}[f(\boldsymbol{Y})]\}] \\
\approx\ & \mathbf{E}\left[\left\{\sum_s (X_s-\mathbf{E}[X_s])\frac{\partial g}{\partial X_s}\Big|_{\mathbf{E}[\boldsymbol{X}]}\right\}\left\{\sum_t (Y_t-\mathbf{E}[Y_t])\frac{\partial f}{\partial Y_t}\Big|_{\mathbf{E}[\boldsymbol{Y}]}\right\}\right] \\
=\ & \sum_s\sum_t \text{Cov}\,[X_s,\,Y_t]\frac{\partial g}{\partial X_s}\Big|_{\mathbf{E}[\boldsymbol{X}]}\frac{\partial f}{\partial Y_t}\Big|_{\mathbf{E}[\boldsymbol{Y}]} \\
=\ & \nabla g\,(\mathbf{E}[\boldsymbol{X}])^T\ \text{Cov}\ [\boldsymbol{X},\,\boldsymbol{Y}]\ \nabla f\,(\mathbf{E}[\boldsymbol{Y}])\,, \qquad (4.46)
\end{aligned}$$

where Cov $[\boldsymbol{X},\,\boldsymbol{Y}]$ denotes the $n \times m$ covariance matrix between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

4.C.2 Coefficient of variation

Quite a few statistical problems can be simplified into the product or ratio of two random variables, and in practice the coefficient of variation is often calculated to consider the standard deviation relative to the mean of a random variable. Let CV stand for the coefficient of variation. For a random variable X, it is defined as

$\text{CV}[X] = \sigma_X/\mu_X$, where σ_X and μ_X denote the standard deviation and mean of X, respectively. Let Y denote another random variable independent of X. Let μ_X and μ_Y denote $\mathbf{E}[X]$ and $\mathbf{E}[Y]$ respectively.

- Approximation: the following equations are mostly used to approximate $\text{CV}[XY]$ and $\text{CV}[X/Y]$

$$\begin{aligned} \text{CV}^2[XY] &\approx \text{CV}^2[X] + \text{CV}^2[Y], && (4.47)\\ \text{CV}^2[X/Y] &\approx \text{CV}^2[X] + \text{CV}^2[Y]. && (4.48) \end{aligned}$$

Based on the Delta method, the derivation of (4.47) is given as follows,

$$\text{Var}\,[XY] \approx \text{Var}\,[X] \left(\frac{\partial(XY)}{\partial X}\right)^2\Bigg|_{(\mu_X,\mu_Y)} + \text{Var}\,[Y] \left(\frac{\partial(XY)}{\partial Y}\right)^2\Bigg|_{(\mu_X,\mu_Y)},$$

where all the derivatives are evaluated at the mean of X and Y, thus

$$\begin{aligned} \text{Var}\,[XY] &\approx \text{Var}\,[X](\mu_Y)^2 + \text{Var}\,[Y](\mu_X)^2 \\ &= \text{CV}^2[X](\mu_X\mu_Y)^2 + \text{CV}^2[Y](\mu_X\mu_Y)^2. && (4.49) \end{aligned}$$

Since X and Y are independent from each other, it follows that $\mathbf{E}[XY] = \mu_X\mu_Y$. Then dividing both sides of equation (4.49) by $(\mu_X\mu_Y)^2$, it follows that

$$\text{CV}^2[XY] \approx \text{CV}^2[X] + \text{CV}^2[Y]. \quad (4.50)$$

Similarly, for $\text{CV}^2[X/Y]$,

$$\begin{aligned} \text{Var}\,[X/Y] &\approx \text{Var}\,[X] \left(\frac{\partial(X/Y)}{\partial X}\right)^2\Bigg|_{(\mu_X,\mu_Y)} + \text{Var}\,[Y] \left(\frac{\partial(X/Y)}{\partial Y}\right)^2\Bigg|_{(\mu_X,\mu_Y)} \\ &= \text{Var}\,[X] \left(\frac{1}{Y^2}\right)\Bigg|_{(\mu_X,\mu_Y)} + \text{Var}\,[Y] \left(\frac{X^2}{Y^4}\right)\Bigg|_{(\mu_X,\mu_Y)} \\ &= \text{Var}\,[X]\frac{1}{(\mu_Y)^2} + \text{Var}\,[Y]\frac{(\mu_X)^2}{(\mu_Y)^4} \\ &= \text{CV}^2[X] \left(\frac{\mu_X}{\mu_Y}\right)^2 + \text{CV}^2[Y] \left(\frac{\mu_X}{\mu_Y}\right)^2 \\ &= \text{CV}^2[X] \left(\mathbf{E}[X/Y]\right)^2 + \text{CV}^2[Y] \left(\mathbf{E}[X/Y]\right)^2, \end{aligned}$$

therefore,

$$\mathrm{CV}^2[X/Y] \approx \mathrm{CV}^2[X] + \mathrm{CV}^2[Y]. \tag{4.51}$$

The above gives the approximation of $\mathrm{CV}[XY]$ and $\mathrm{CV}[X/Y]$ in (4.50) and (4.51), respectively. This approximation is obtained by applying the Delta method to XY and X/Y as functions of X and Y. However, in the case of $\mathrm{CV}[XY]$, its exact expression is available, which is derived in the following paragraph.

- Exact expression : start with the trick that leads to the exact expression of $\mathrm{CV}[XY]$, writing XY as

$$XY = \mu_X\mu_Y + (X-\mu_X)\mu_Y + (Y-\mu_Y)\mu_X + (X-\mu_X)(Y-\mu_Y).$$

Then, $\mathbf{E}[XY]$ and $\mathbf{E}[(XY)^2]$ can be calculated as

$$\begin{aligned}
& \mathbf{E}[XY] \\
= \; & \mathbf{E}\left\{\mu_X\mu_Y + (X-\mu_X)\mu_Y + (Y-\mu_Y)\mu_X + (X-\mu_X)(Y-\mu_Y)\right\} \\
= \; & \mu_X\mu_Y + \mu_Y\mathbf{E}[X-\mu_X] + \mu_X\mathbf{E}[Y-\mu_Y] + \mathbf{E}[(X-\mu_X)(Y-\mu_Y)] \\
= \; & \mu_X\mu_Y + \mathrm{Cov}\,[X,\,Y],
\end{aligned}$$

and

$$\begin{aligned}
& \mathbf{E}[(XY)^2] \\
= \; & \mathbf{E}\left[\{\mu_X\mu_Y + (X-\mu_X)\mu_Y + (Y-\mu_Y)\mu_X \right. \\
& \qquad \left. +(X-\mu_X)(Y-\mu_Y)\}^2\right] \\
= \; & \mathbf{E}\left[\mu_X^2\mu_Y^2 + (X-\mu_X)^2\mu_Y^2 + (Y-\mu_Y)^2\mu_X^2 \right. \\
& \qquad +(X-\mu_X)^2(Y-\mu_Y)^2 \\
& \qquad +2\mu_X\mu_Y^2(X-\mu_X)^2 + 2\mu_X^2\mu_Y(Y-\mu_Y)^2 \\
& \qquad +4\mu_X\mu_Y(X-\mu_X)(Y-\mu_Y) \\
& \qquad \left. +2\mu_Y(X-\mu_X)^2(Y-\mu_Y) + 2\mu_X(X-\mu_X)(Y-\mu_Y)^2\right] \\
= \; & \mu_X^2\mu_Y^2 + \mu_Y^2\mathrm{Var}\,[X] + \mu_X^2\mathrm{Var}\,[Y] + \mathbf{E}[(X-\mu_X)^2(Y-\mu_Y)^2] \\
& +2\mu_Y\mathbf{E}[(X-\mu_X)^2(Y-\mu_Y)] + 2\mu_X\mathbf{E}[(X-\mu_X)(Y-\mu_Y)^2] \\
& +4\mu_X\mu_Y\mathrm{Cov}\,[X,\,Y].
\end{aligned}$$

Based on the expressions of $\mathbf{E}[XY]$ and $\mathbf{E}[(XY)^2]$ given above, Var $[XY]$ can be evaluated exactly as

$$\begin{aligned}
&\text{Var}\,[XY] \\
=\;& \mathbf{E}[(XY)^2] - (\mathbf{E}[XY])^2 \\
=\;& \Big(\mu_X^2\mu_Y^2 + \mu_Y^2\text{Var}\,[X] + \mu_X^2\text{Var}\,[Y] + \mathbf{E}[(X-\mu_X)^2(Y-\mu_Y)^2] \\
& +2\mu_Y\mathbf{E}[(X-\mu_X)^2(Y-\mu_Y)] + 2\mu_X\mathbf{E}[(X-\mu_X)(Y-\mu_Y)^2] \\
& +4\mu_X\mu_Y\text{Cov}\,[X,\,Y]\Big) - \Big(\mu_X\mu_Y + \text{Cov}\,[X,\,Y]\Big)^2 \\
=\;& \mu_Y^2\text{Var}\,[X] + \mu_X^2\text{Var}\,[Y] + 2\mu_X\mu_Y\text{Cov}\,[X,\,Y] - (\text{Cov}\,[X,\,Y])^2 \\
& +\mathbf{E}[(X-\mu_X)^2(Y-\mu_Y)^2] + 2\mu_Y\mathbf{E}[(X-\mu_X)^2(Y-\mu_Y)] \\
& +2\mu_X\mathbf{E}[(X-\mu_X)(Y-\mu_Y)^2]. \qquad (4.52)
\end{aligned}$$

In the case that X and Y are independent from each other, it follows that

$$\begin{aligned}
\mathbf{E}[(X-\mu_X)^2(Y-\mu_Y)^2] &= \text{Var}\,[X]\text{Var}\,[Y], \\
\mathbf{E}[(X-\mu_X)^2(Y-\mu_Y)] &= \mathbf{E}[(X-\mu_X)^2]\mathbf{E}[(Y-\mu_Y)] = 0, \\
\mathbf{E}[(X-\mu_X)(Y-\mu_Y)^2] &= \mathbf{E}[(X-\mu_X)]\mathbf{E}[(Y-\mu_Y)^2] = 0,
\end{aligned}$$

Substituting these in (4.52) gives

$$\text{Var}\,[XY] = \mu_Y^2\text{Var}\,[X] + \mu_X^2\text{Var}\,[Y] + \text{Var}\,[X]\text{Var}\,[Y].$$

Finally, on dividing both sides by $\mu_X^2\mu_Y^2$ gives,

$$\text{CV}^2[XY] = \text{CV}^2[X] + \text{CV}^2[Y] + \text{CV}^2[Y]\text{CV}^2[Y]. \qquad (4.53)$$

4.C.3 Variance estimation for $\hat{\rho}_{sa}$ with $r = 1$

Under the assumption of perfect net retention, i.e., $r = 1$, the density estimator is expressed as (4.10) in Section 4.3.1, from which it follows that the variance of $\hat{\rho}_{sa}$ comes from n_{ia} $(i = 1,\, 2,\, \ldots, m_s)$ and $\hat{h}$, where the averaging weight for haul i is $w_i = v_i/v_s$. It is assumed that $\hat{h}$ and n_{ai} are independent for all i. In addition, the estimator given by (4.10) can be thought of as a ratio of two random variables, in which case the exact expression of CV$[X/Y]$ is not available. Therefore, the Delta

method described in Appendix 4.C.1 is applied to obtain the CV of the estimation given by (4.10).

With $r = 1$, the density estimator for stratum s and age class a is

$$\hat{\rho}_{sa} = \sum_{i \in s} w_i \left(\frac{n_{ia}}{v_{1i} + \hat{h} v_{2i}} \right),$$

where $w_i = v_i / \sum_{i \in s} v_i$.

Applying the Delta method given by expression (4.45) to the above density estimator, the variance of $\hat{\rho}_{sa}$ can then be approximated. To start with, the partial derivatives of $\hat{\rho}_{sa}$ with respect to n_{ia} and h are worked out for use later in applying the Delta method:

$$\begin{aligned} \frac{\partial \hat{\rho}_{sa}}{\partial n_{ia}} &= w_i \left[\frac{1}{v_{1i} + \hat{h}\, v_{2i}} \right], \\ \frac{\partial \hat{\rho}_{sa}}{\partial \hat{h}} &= \sum_{i \in s} \frac{w_i\, n_{ia}\, v_{2i}}{(v_{1i} + \hat{h}\, v_{2i})^2}. \end{aligned}$$

Given the above and applying the Delta method given by (4.45), the variance of $\hat{\rho}_{sa}$ can be estimated as

$$\begin{aligned} \widehat{\text{Var}}[\hat{\rho}_{sa}] &= \sum_{i \in s} \left\{ \widehat{\text{Var}}[n_{ia}] \left[\frac{\partial \hat{\rho}_{sa}}{\partial n_{ia}} \right]^2 + \widehat{\text{Var}}[\hat{h}] \left[\frac{\partial \hat{\rho}_{sa}}{\partial \hat{h}} \right]^2 \right\} \\ &= \widehat{\text{Var}}[n_{ia}] \sum_{i \in s} \left[\frac{w_i}{v_{1i} + \hat{h}\, v_{2i}} \right]^2 + \text{nhaul}_s \widehat{\text{Var}}[\hat{h}] \left[\sum_{i \in s} \frac{w_i\, n_{ia}\, v_{2i}}{(v_{1i} + \hat{h}\, v_{2i})^2} \right]^2, \end{aligned}$$

where nhaul_s denotes the total number of hauls in stratum s.

4.C.4 Variance estimation for $\hat{\rho}_{sa}$ with $\hat{r}(l)$

This appendix focuses on analytical variance estimation for the density estimator with the net retention probability r estimated by the fixed-effects logistic regression models in Section 4.3.2. Let $\hat{r}(l)$ denote the estimate net retention probability as a function of length l of fish. Recall the density estimator given by (4.11) for fish with

age a in stratum s:

$$\hat{\rho}_{sa} = \sum_{i \in s} \underbrace{\frac{(v_{1i} + v_{2i})}{v_s}}_{w_i} \underbrace{\sum_l \frac{n_{ila}}{\hat{r}(l)(v_{1i} + \hat{h} v_{2i})}}_{\hat{\rho}_{ia}} \tag{4.54}$$

$$= \sum_{i \in s} w_i \hat{\rho}_{ia}, \tag{4.55}$$

Further $\hat{\rho}_{ia}$ can be written as

$$\hat{\rho}_{ia} = \sum_l \hat{\rho}_{ila}, \tag{4.56}$$

where

$$\hat{\rho}_{ila} = \frac{n_{ila}}{\hat{r}(l)(v_{1i} + \hat{h} v_{2i})}.$$

Then assuming the independence of $\hat{\rho}_{ila}$ for different lengths l, based on (4.56), it follows that

$$\text{Var}\left[\hat{\rho}_{ia}\right] = \sum_l \text{Var}\left[\hat{\rho}_{ila}\right].$$

Further, assuming the independence of $\hat{\rho}_{ia}$ for different i, then based on (4.55), the variance of $\hat{\rho}_{sa}$ can be calculated as

$$\begin{aligned} \text{Var}\left[\hat{\rho}_{sa}\right] &= \sum_{i \in s} w_i^2 \, \text{Var}\left[\hat{\rho}_{ia}\right] \\ &= \sum_{i \in s} w_i^2 \left(\sum_l \text{Var}\left[\hat{\rho}_{ila}\right] \right). \end{aligned} \tag{4.57}$$

Equation (4.57) shows that the calculation of $\text{Var}\left[\hat{\rho}_{sa}\right]$ reduces to that of $\text{Var}\left[\hat{\rho}_{ia}\right]$ for each haul i in stratum s. The Delta method is then applied to $\hat{\rho}_{ila}$ for $\widehat{\text{Var}}\left[\hat{\rho}_{ila}\right]$, which are summed over all lengths with the same age a.

Therefore, starting from the $\text{Var}\,[\hat{\rho}_{ila}]$ in (4.57) by applying the Delta method given in (4.45), the first-order partial derivatives of $\hat{\rho}_{ila}$ with respect to n_{ila}, $\hat{h}$ and $\hat{r}(l)$ are

$$\begin{aligned}
\frac{\partial \hat{\rho}_{ila}}{\partial n_{ila}} &= \frac{1}{\hat{r}(l)(v_{1i} + \hat{h} v_{2i})}, \\
\frac{\partial \hat{\rho}_{ila}}{\partial \hat{h}} &= \frac{-n_{ila} v_{2i}}{\hat{r}(l)(v_{1i} + \hat{h} v_{2i})^2}, \\
\frac{\partial \hat{\rho}_{ila}}{\partial \hat{r}} &= \frac{-n_{ila}}{[\hat{r}(l)]^2 (v_{1i} + \hat{h} v_{2i})}.
\end{aligned}$$

Using the Delta method, the variance of $\hat{\rho}_{ila}$ can be approximated by

$$\widehat{\text{Var}}\,[\hat{\rho}_{ila}] = \widehat{\text{Var}}\,[n_{ila}] \left(\frac{\partial \hat{\rho}_{ila}}{\partial n_{ila}}\right)^2 + \widehat{\text{Var}}\,[\hat{h}] \left(\frac{\partial \hat{\rho}_{ila}}{\partial \hat{h}}\right)^2 + \widehat{\text{Var}}\,[\hat{r}(l)] \left(\frac{\partial \hat{\rho}_{ila}}{\partial \hat{r}}\right)^2. \tag{4.58}$$

Note that $\hat{r}(l)$ is a function of l which is estimated from the experimental survey data by using the fixed-effects logistic regression models, for the asymptote-logistic regression model formulated as (2.7),

$$\hat{r}_3(l) = \frac{\hat{\gamma}}{1 + \exp(-\hat{\beta}_0 - \hat{\beta}_1 l)}. \tag{4.59}$$

Thus estimation of $\text{Var}\,[\hat{r}(l)]$ requires applying the Delta method as well. Note that the subscript 3 in $\hat{r}_3(l)$ denotes the number of unknown parameters in $r(l)$, in order to distinguish it from the linear logistic regression model without parameter γ, or in other words, assuming $\gamma = 1$.

To start with,

$$\begin{aligned}
\frac{\partial \hat{r}_3(l)}{\partial \hat{\beta}_0} &= \frac{\hat{\gamma} \exp(-\hat{\beta}_0 - \hat{\beta}_1 l)}{[1 + \exp(-\hat{\beta}_0 - \hat{\beta}_1 l)]^2}, \\
\frac{\partial \hat{r}_3(l)}{\partial \hat{\beta}_1} &= \frac{\hat{\gamma}\, l\, \exp(-\hat{\beta}_0 - \hat{\beta}_1 l)}{[1 + \exp(-\hat{\beta}_0 - \hat{\beta}_1 l)]^2}, \\
\frac{\partial \hat{r}_3(l)}{\partial \hat{\gamma}} &= \frac{1}{1 + \exp(-\hat{\beta}_0 - \hat{\beta}_1 l)}.
\end{aligned}$$

Then based on the Delta method given by (4.45), $\text{Var}\,[\hat{r}_3(l)]$ can be approximated as

$$\begin{aligned}\widehat{\text{Var}}\,[\hat{r}_3(l)] &= \widehat{\text{Var}}\,[\hat{\beta}_0]\left(\frac{\partial \hat{r}_3(l)}{\partial \hat{\beta}_0}\right)^2 + \widehat{\text{Var}}\,[\hat{\beta}_1]\left(\frac{\partial \hat{r}_3(l)}{\partial \hat{\beta}_1}\right)^2 + \widehat{\text{Var}}\,[\hat{\gamma}]\left(\frac{\partial \hat{r}_3(l)}{\partial \hat{\gamma}}\right)^2 \\ &\quad +2\,\widehat{\text{Cov}}\,[\hat{\beta}_0,\hat{\beta}_1]\,\frac{\partial \hat{r}_3(l)}{\partial \hat{\beta}_0}\,\frac{\partial \hat{r}_3(l)}{\partial \hat{\beta}_1} + 2\,\widehat{\text{Cov}}\,[\hat{\beta}_0,\hat{\gamma}]\,\frac{\partial \hat{r}_3(l)}{\partial \hat{\beta}_0}\,\frac{\partial \hat{r}_3(l)}{\partial \hat{\gamma}} \\ &\quad +2\,\widehat{\text{Cov}}\,[\hat{\beta}_1,\hat{\gamma}]\,\frac{\partial \hat{r}_3(l)}{\partial \hat{\beta}_1}\,\frac{\partial \hat{r}_3(l)}{\partial \hat{\gamma}}, \end{aligned} \tag{4.60}$$

where the variances of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\gamma}$ and corresponding covariances can be obtained from the inverse of observed Fisher information matrix (see Section 3.1.4 for details).

For a linear logistic regression, $\hat{r}(l)$ is of the form similar to (4.59) but with the assumption that $\gamma = 1$. Therefore, based on the above process for $\widehat{\text{Var}}\,[\hat{r}_3(l)]$ with $\hat{r}_3(l)$ given by (4.59), thinking of $\hat{\gamma}$ as constant ($\hat{\gamma} = 1$), we have

$$\begin{aligned}\frac{\partial \hat{r}(l)}{\partial \hat{\beta}_0} &= \frac{\exp(-\hat{\beta}_0 - \hat{\beta}_1\, l)}{[1+\exp(-\hat{\beta}_0 - \hat{\beta}_1\, l)]^2}, \\ \frac{\partial \hat{r}(l)}{\partial \hat{\beta}_1} &= \frac{l\,\exp(-\hat{\beta}_0 - \hat{\beta}_1\, l)}{[1+\exp(-\hat{\beta}_0 - \hat{\beta}_1\, l)]^2}. \end{aligned}$$

Using the Delta method, the variance of

$$\hat{r}(l) = \frac{1}{1+\exp(-\hat{\beta}_0 - \hat{\beta}_1\, l)}$$

can be approximated as

$$\widehat{\text{Var}}\,[\hat{r}(l)] = \widehat{\text{Var}}\,[\hat{\beta}_0]\left(\frac{\partial \hat{r}(l)}{\partial \hat{\beta}_0}\right)^2 + \widehat{\text{Var}}\,[\hat{\beta}_1]\left(\frac{\partial \hat{r}(l)}{\partial \hat{\beta}_1}\right)^2 + 2\,\widehat{\text{Cov}}\,[\hat{\beta}_0,\hat{\beta}_1]\,\frac{\partial \hat{r}(l)}{\partial \hat{\beta}_0}\,\frac{\partial \hat{r}(l)}{\partial \hat{\beta}_1}. \tag{4.61}$$

Finally, combining (4.57), (4.58) and (4.60), the variance of $\hat{\rho}_{sa}$ with $r_3(l)$ being estimated by asymptote-logistic regression can be approximated by

$$\begin{aligned}
&\widehat{\text{Var}}\,[\hat{\rho}_{sa}] \\
&= \sum_{i\in s} w_i^2 \left\{ \sum_l \left[\frac{\widehat{\text{Var}}\,[n_{ila}]}{[\hat{r}_3(l)]^2(v_{1i}+\hat{h}v_{2i})^2} + \frac{\widehat{\text{Var}}\,[\hat{h}]\, n_{ila}^2\, v_{2i}^2}{[\hat{r}_3(l)]^2(v_{1i}+\hat{h}v_{2i})^4} \right.\right. \\
&\qquad + \frac{n_{ila}^2}{(v_{1i}+\hat{h}v_{2i})^2} \left(\widehat{\text{Var}}\,[\hat{\beta}_0] \frac{\exp(-2\hat{\beta}_0 - 2\hat{\beta}_1\, l)}{\hat{\gamma}^2} \right. \\
&\qquad\qquad + \widehat{\text{Var}}\,[\hat{\beta}_1] \frac{l^2\, \exp(-2\hat{\beta}_0 - 2\hat{\beta}_1\, l)}{\hat{\gamma}^2} \\
&\qquad\qquad + \frac{\widehat{\text{Var}}\,[\hat{\gamma}]}{[\hat{r}_3(l)]^2\, \hat{\gamma}^2} \\
&\qquad\qquad + 2\, \widehat{\text{Cov}}\,[\hat{\beta}_0, \hat{\beta}_1] \frac{l\, \exp(-2\hat{\beta}_0 - 2\hat{\beta}_1\, l)}{\hat{\gamma}^2} \\
&\qquad\qquad + 2\, \widehat{\text{Cov}}\,[\hat{\beta}_0, \hat{\gamma}] \frac{\exp(-\hat{\beta}_0 - \hat{\beta}_1\, l)}{\hat{r}_3(l)\hat{\gamma}^2} \\
&\qquad\qquad \left.\left. + \; 2\, \widehat{\text{Cov}}\,[\hat{\beta}_1, \hat{\gamma}] \frac{l\, \exp(-\hat{\beta}_0 - \hat{\beta}_1\, l)}{\hat{r}_3(l)\hat{\gamma}^2} \right) \right] \right\}.
\end{aligned} \tag{4.62}$$

Similarly, combining (4.57), (4.58) and (4.61), the variance of $\hat{\rho}_{sa}$ with $r(l)$ being estimated by linear logistic regression can be approximated by

$$\begin{aligned}
&\widehat{\text{Var}}\,[\hat{\rho}_{sa}] \\
&= \sum_{i\in s} w_i^2 \left\{ \sum_l \left[\frac{\widehat{\text{Var}}\,[n_{ila}]}{[\hat{r}(l)]^2(v_{1i}+\hat{h}v_{2i})^2} + \frac{\widehat{\text{Var}}\,[\hat{h}]\, n_{ila}^2\, v_{2i}^2}{[\hat{r}(l)]^2(v_{1i}+\hat{h}v_{2i})^4} \right.\right. \\
&\qquad + \frac{n_{ila}^2}{(v_{1i}+\hat{h}v_{2i})^2} \left(\widehat{\text{Var}}\,[\hat{\beta}_0] \exp(-2\hat{\beta}_0 - 2\hat{\beta}_1\, l) \right. \\
&\qquad\qquad + \widehat{\text{Var}}\,[\hat{\beta}_1]\, l^2\, \exp(-2\hat{\beta}_0 - 2\hat{\beta}_1\, l) \\
&\qquad\qquad \left.\left. + \; 2\, \widehat{\text{Cov}}\,[\hat{\beta}_0, \hat{\beta}_1]\, l\, \exp(-2\hat{\beta}_0 - 2\hat{\beta}_1\, l) \right) \right] \right\}.
\end{aligned} \tag{4.63}$$

From (4.62) and (4.63), it can be seen that the analytical evaluation for approximating $\text{Var}\,[\hat{\rho}_{sa}]$ is not straightforward. Therefore, bootstrapping is preferred for a complicated problem like this. In addition, it is easier to include all the sources of

uncertainty in bootstrapping such as the uncertainty from missing-age data, which is not considered in the analytical evaluation. However, the analytical approximation to $\widehat{\text{Var}}\,[\hat{\rho}_{sa}]$ can be compared to the bootstrap variance for debugging purposes.

4.D Log-normal distribution

In practical applications, there are many cases in which the random variable can take only positive values, such as animal abundance, survival time of a certain bacterium and family incomes. In these cases, instead of the usual normality assumption, a log-normal distribution (see Figure 4.10 for an illustration) is often assumed for further statistical inferences, such as confidence intervals. A random variable X is log-normally distributed if its logarithm $\log X$, denoted by Y, is normally distributed. A log-normal distribution is usually denoted by $X \sim LN(\mu, \sigma^2)$. It should be noted that here μ and σ are not the mean and standard deviation of X, but they are the mean and standard deviation of Y, i.e., $Y = \log(X) \sim N(\mu, \sigma^2)$. Given that Y is normally distributed with mean μ and standard deviation σ, its probability density function is

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

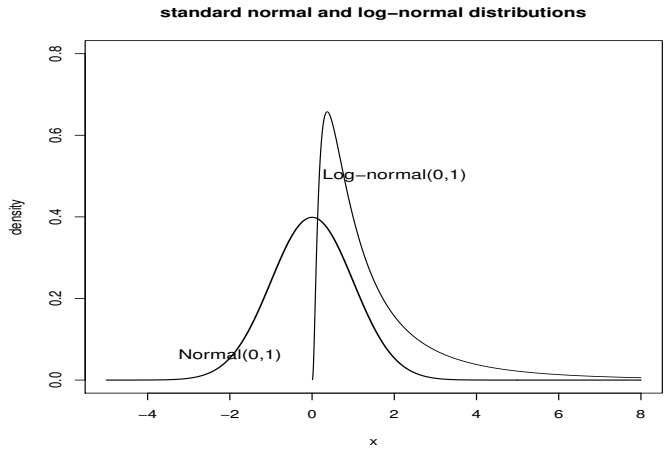


FIGURE 4.10. Density plots of standard normal, $N(0,1)$, and log-normal $LN(0,1)$, distributions.

Then the probability density function of X can be obtained via the following equation

$$|f_X(x)dx| = |f_Y(y)dy|,$$

each side of which gives the probability contained in a differential area for X and $Y = \log(X)$, and this probability is invariant under change of random variables. Therefore,

$$f_X(x) = f_Y(y(x)) \left|\frac{dy}{dx}\right| = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left\{-\frac{[\log(x)-\mu]^2}{2\sigma^2}\right\}.$$

Therefore, the mean of the log-normally distributed random variable X can be derived as

$$\mathbf{E}[X] = \int_0^{+\infty} x f_X(x)\, dx = \int_0^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{[\log(x)-\mu]^2}{2\sigma^2}\right\} dx\,.$$

Replacing $\log(x)$ by y, the above expression can be re-written as

$$\begin{aligned}
\mathbf{E}[X] &= \int_0^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] d\, \exp(y) \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \exp(y)\, dy \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2} + y\right] dy \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{[y-(\mu+\sigma^2)]^2}{2\sigma^2} + \frac{\sigma^4+2\mu\sigma^2}{2\sigma^2}\right\} dy \\
&= \exp\left(\frac{\sigma^4+2\mu\sigma^2}{2\sigma^2}\right) \underbrace{\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{[y-(\mu+\sigma^2)]^2}{2\sigma^2}\right\} dy}_{=1} \\
&= \exp(\mu + \sigma^2/2)\,. \qquad (4.64)
\end{aligned}$$

Similarly, it can be shown that

$$\mathbf{E}[X^2] = \exp(\sigma^2)\exp(2\mu+\sigma^2).$$

Therefore, the variance of a log-normally distributed random variable X is

$$\text{Var}\,[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \exp(2\mu+\sigma^2)[\exp(\sigma^2)-1]. \qquad (4.65)$$

Then we take the ratio of (4.65) over (4.64) and then move the constant -1 to the other side of the equation. The variance of the corresponding normally distributed random variable X can be worked out as

$$\text{Var}\left[Y\right] = \sigma^2 = \log\left\{1 + \left(\text{CV}[X]\right)^2\right\}. \tag{4.66}$$

Chapter 5

Properties of anglerfish abundance estimators with haul effect

Key Idea: perform simulation studies to examine the properties of anglerfish abundance estimators with random haul effects

In order to examine the performance of the HT-like estimators with haul effect which are used in the previous chapter for the anglerfish abundance estimation, simulation studies are carried out for both experimental and abundance surveys. To start with, Section 5.1 describes how to set up the simulation study aiming at being representative of the 2007 abundance survey and the 2006-2007 experimental survey, and then Section 5.2 presents the results obtained from the simulation study and discusses the performance of the estimators in relation to the haul effect.

5.1 Simulation of anglerfish surveys

In order to replicate the abundance estimation scenario in Chapter 4, simulations are set up for both experimental and abundance surveys. The data obtained from simulating the experimental survey are used to estimate the capture probability, which is then used in abundance estimation with the catch data from simulating the abun-

dance survey. In order to simplify the simulation study but without loss of any key component in the abundance estimation performed in Chapter 4, only the net retention probability is considered for the capture probability in the abundance survey. This means that the herding factor is not considered here and the capture probability is the net retention probability estimated from the simulated experimental survey data. In addition, it is assumed that there is no difference in the swept area between the hauls in the abundance survey, and abundance estimation is considered only for the area sampled by hauls. This means that the capture probability is the inclusion probability in the HT-like estimator and there is no need to consider swept area, as it is assumed to be a constant over all hauls and therefore has no effect on the performance of the abundance estimators.

The process of simulating the anglerfish survey data and the way in which its components are combined in obtaining the abundance estimates are summarized by a tree diagram in Figure 5.1. The components in the tree diagram are given in four different colours: red is for the model components that are based on the estimation results using $\widehat{N}^{(2)}$ in Chapter 4, and these assumptions are the same for simulation of both experimental and abundance surveys; blue is for the simulation of the experimental survey; green is for the simulation of the abundance survey; and black is for the abundance estimation which is based on the simulated experimental and abundance survey data.

First, we explain the red part in Figure 5.1. The true model assumed for the capture probability (also the inclusion probability) in the simulation study is

$$r_i(l) = \text{logit}^{-1}(-3.606 + b_{0i} + 0.125\bar{l}_i + 0.110(l - \bar{l}_i)), \tag{5.1}$$

where the mean and standard deviation of b_0 are assumed to be 0 and 0.289 respectively. The $r_i(l)$ assumed here is of the same form as that obtained in applying a two-level mixed-effects model to estimate r using the 2006-2007 experimental survey data (see Section 3.4 for full details). Therefore, in the simulation study, a realized value of random intercept, b_{0i}, is simulated from $N(0, 0.289^2)$ in calculating the true capture probability for haul i.

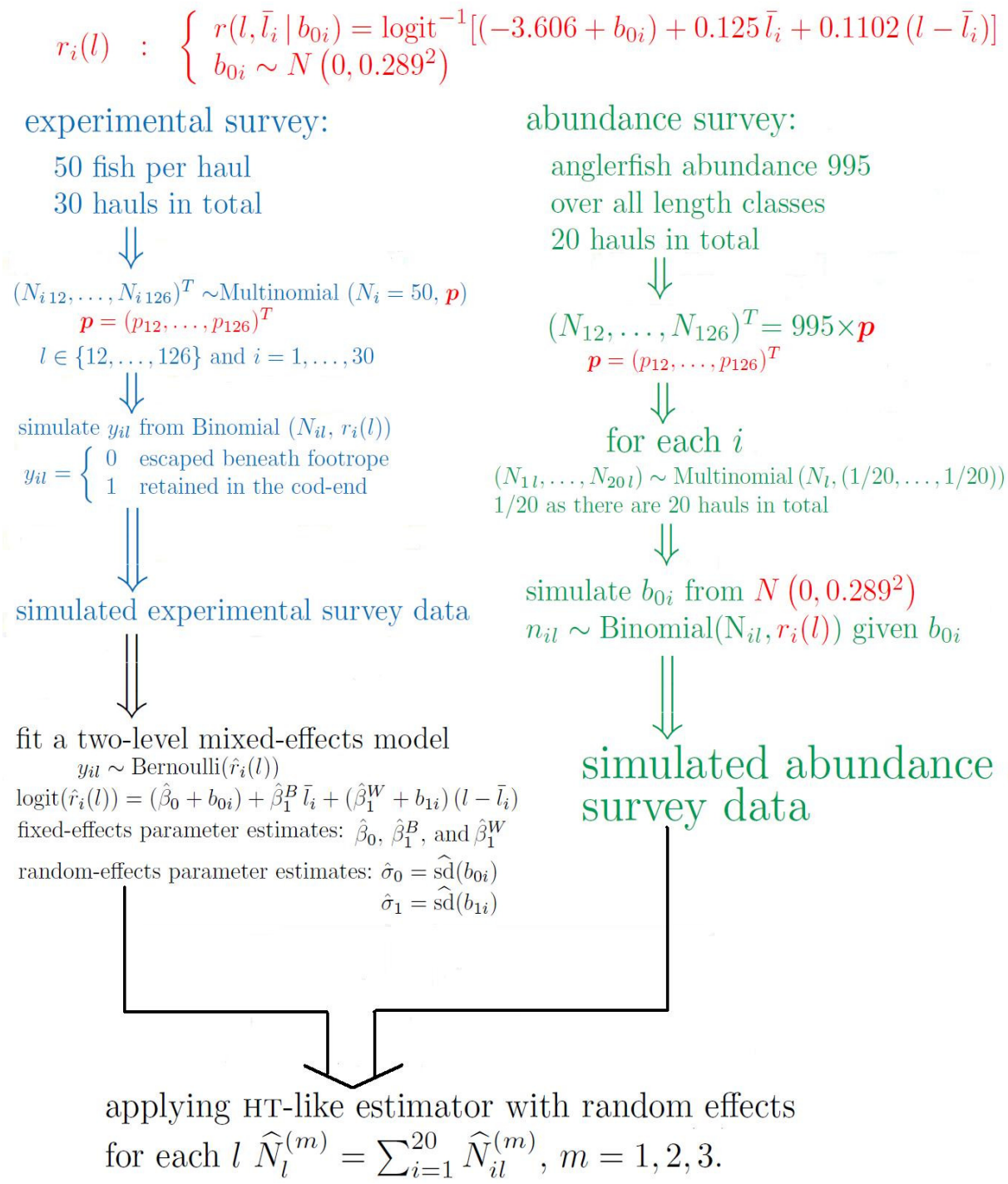


FIGURE 5.1. Tree diagram of the simulation process for anglerfish simulation study. Objects in red give details of the models from which data were simulated. The simulation process for the experimental survey is coloured in blue, while the process for the abundance survey is coloured in green. The objects in black are involved in the abundance estimation for the simulated anglerfish survey data.

The length classes considered in the simulation are from 12 cm to 126 cm, and this range is the same as that in the 2007 anglerfish abundance survey data. In order to assign a length class to each fish assumed in the simulation study, a distribution of length is estimated based on the estimation results using $\widehat{N}^{(2)}$ given by (4.16). The abundance estimates at each length class and age group are plotted in Figure 5.2, and then some probability distributions are fitted to the marginal estimates at each length class. Based on AIC, a gamma distribution (the smooth line on right side in Figure 5.2) is finally chosen to estimate the marginal distribution of length for the anglerfish population in the simulation study. The gamma distribution is not a great fit but it is considered adequate for the purposes of this simulation study.

Let p_l denote the probability that a fish has length l, and $\boldsymbol{p}$ denote the probability vector $(p_{12}, p_{13}, \ldots, p_{126})^T$ for all potential length classes in the population. Given the assumed population size $N = 995$ for the abundance survey, the population size at all the length classes, $(N_{12}, N_{13}, \ldots, N_{126})^T$, follows a multinomial distribution with probabilities $\boldsymbol{p}$, i.e., Multinomial$(N, \boldsymbol{p})$. In order to examine the performance of the estimators, we fix the population size at each length as this simplifies the interpretation and calculation of bias and MSE. Therefore, instead of simulating $(N_{12}, N_{13}, \ldots, N_{126})^T$ from Multinomial$(N, \boldsymbol{p})$, N_l is fixed at the expectation of Multinomial$(N, \boldsymbol{p})$, i.e., $N_l = N\, p_l$ for each length class $l \in \{12, 13, \ldots, 126\}$.

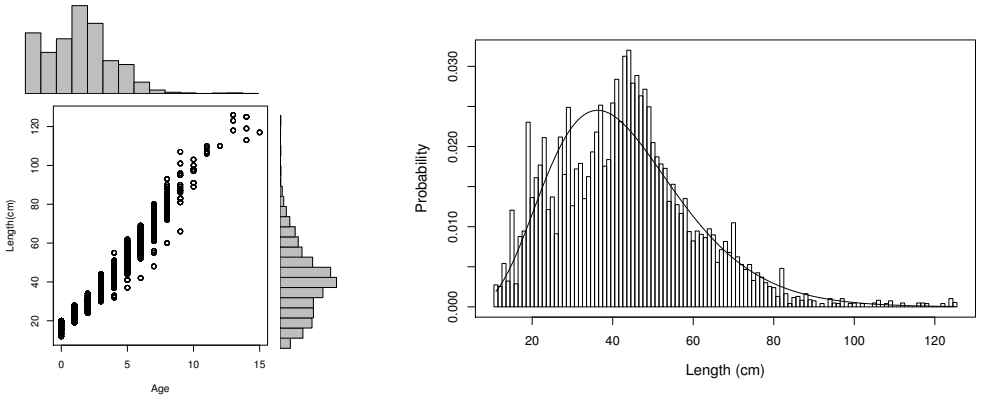


FIGURE 5.2. Plot of length vs age (left) and of the estimated marginal distribution of length (right) based on the abundance estimates using estimator (4.16). The smooth line is the fitted gamma distribution.

The green component in Figure 5.1 shows the simulation process for the abundance survey in one stratum with 20 hauls and population size assumed to be 995. This pro-

cess is considered in 3 steps, which are represented by the green downward arrows in Figure 5.1:

- Step (1): Set up N_l for each length class by $(N_{12}, N_{13}, \ldots, N_{126})^T = 995 \times \boldsymbol{p}$. Note that N_l is fixed for any l in the 999 replications of the simulation process, as plotted in Figure 5.3.

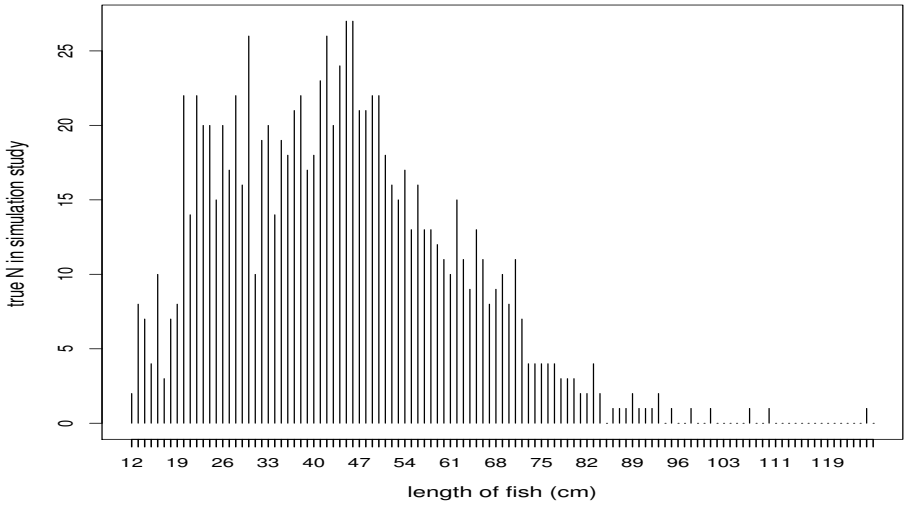


FIGURE 5.3. Plot of the true abundance at each length class over 20 hauls in the simulation study of the abundance survey

- Step (2): The swept areas for the 20 hauls are assumed to be the same. Then, $(N_{1\,l}, \ldots, N_{20\,l}) \sim$ Multinomial $(N_l, (1/20, \ldots, 1/20))$, where N_{il} denotes the population size for fish of length l in the area sampled by haul i. One realization of this step is given in Table 5.1, where the row 'Total' coloured in red is from step (1).

- Step (3): First, a random intercept b_{0i} is simulated from $N(0, 0.0289^2)$ for haul i and the corresponding true capture probability $r_i(l)$ given by (5.1) is then calculated conditional on the simulated b_{0i} and mean length of fish $\bar{l}_i$. Second, given $r_i(l)$, the number of fish, n_{il}, of length l that are captured by haul i (i.e., being retained in the cod-end) is simulated from Binomial$(N_{il}, r_i(l))$. Then n_{il}, $i = 1, 2, \ldots, 20$ and $l = 12, 13, \ldots, 126$ cm, are the simulated catch data from the abundance survey.

TABLE 5.1. One re-sampled population for the abundance survey with 20 hauls in total and length classes from 12 cm to 125 cm.

Haul \ Length(cm)	12	13	14	...	110	125	Total
Haul 1	0	1	0	...	0	0	51
Haul 2	0	0	0	...	0	0	53
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Haul 19	0	1	0	...	1	0	46
Haul 20	0	0	0	...	0	0	49
Total [a]	2	8	7	...	1	1	995

[a] this row gives the assumed population size at each length class over all 20 hauls for the simulation of the abundance survey

For the blue part in Figure 5.1, the simulation of the experimental survey is carried out in three steps, which are represented by the blue downward arrows in Figure 5.1:

- Step (1): Note that Snijders & Bosker (1999) concluded that a sample size less than 20 at level-two usually gives limited information about the random effect distribution. Therefore, the simulation of the experimental survey assumes 30 hauls in total with a population size of 50 fish in each haul.
- Step (2): Given that $N_i = 50$ for $i = 1, 2, \ldots, 30$ assumed in step (1), let N_{il} denote the population size at length class l for haul i, then it follows that $N_{il} \sim$ Multinomial$(N_i = 50, \boldsymbol{p})$.
- Step (3): For each haul i, simulate a value for b_{0i} from the normal distribution with mean 0 and variance 0.0289^2, then calculate the assumed capture probability $r_i(l)$ according to (5.1). For each individual fish out of N_{il} simulated in step (2), whether or not it is retained in the main cod-end is denoted by a binary response y_{il}, where $y_{il} = 1$ means being retained and $y_{il} = 0$ otherwise, and $y_{il} \sim$ Bernoulli$(r_i(l))$. The binary response data y_{il}, the group mean $\bar{l}_i$ and the length of each individual fish assumed for the experimental survey population constitute the simulated experimental survey data.

The black components in Figure 5.1 concern the abundance estimation given both simulated experimental and abundance survey data, i.e., the simulated data sets from the blue and green parts in Figure 5.1. First, capture probability estimation is conducted: a two-level mixed-effects logistic regression model of the form (4.13) is fitted using simulated experimental survey data. Note that this model allows both random intercept and slope, and the fitted $\hat{r}(l)$ may have only either random intercept b_{0i} or random slope b_{1i}, or both. This depends on the simulated experimental survey data.

Given the fitted $\hat{r}_i(l)$ from simulated experimental survey data, abundance estimation is then carried out for the simulated abundance survey data using HT-like estimators. The calculations for $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$ are the same as that given by (4.18), (4.19) and (4.21), except that here it is for only one stratum. Note that the numerical approximation of the integral in (4.19) for $\widehat{N}^{(2)}$ is based on a grid of 1000 equally spaced cells over 5 standard errors away from the mean of a linear combination of b_{0i} and b_{1i} given by (4.20). Such a grid has been tested for convergence in abundance estimation.

In addition to the HT-like estimators with random effects, the HT-like estimator with fixed-effects $\hat{r}(l)$ is applied here, in order to understand the effect of ignoring the haul effect in abundance estimation. Therefore, a fixed-effects linear logistic regression of the form (3.32) is fitted to the same simulated experimental survey data from step (3) in the green part of Figure 5.1. This leads to an estimated capture probability $\hat{r}(l)$ as a function of length. Then the abundance estimator given by (4.11) is applied to the simulated abundance survey data with the resulting $\hat{r}(l)$.

5.2 Results and discussion

The simulation process described in Section 5.1 is repeated 999 times to examine the performance of the HT-like estimators with haul effect. These estimators $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$ are given by (4.15), (4.16) and (4.17), respectively. Performance of each estimator is then measured in terms of bias and MSE, scaled as percentages of

true (simulated) abundance (N_l) given length class l and its square (N_l^2), as follows:

$$\%\text{bias}(\widehat{N}_l^{(m)}) = \text{bias}(\widehat{N}_l^{(m)})/N \times 100 \quad (5.2)$$

$$\%\text{MSE}(\widehat{N}_l^{(m)}) = \text{MSE}(\widehat{N}_l^{(m)})/N_l^2 \times 100, \quad (5.3)$$

where $m = 0, 1, 2, 3$.

Note that in order to examine the effect of ignoring the haul effect, the %bias and %MSE given above are also calculated for the HT-like estimator without random effects, i.e., with only fixed-effects length, given by (4.11), where the capture probability is estimated by a fixed-effects linear logistic regression model. In the following discussion on the simulation results, $\widehat{N}$ denotes the HT-like estimator without haul effect, which is given by (4.11).

The simulation results for $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$ are plotted in Figure 5.4. The plot is truncated at length 70 cm, since there is little difference among these three estimators at length classes beyond 70 cm.

- $\widehat{N}^{(2)}$ always performs the best – lowest %bias and %MSE among all three estimators with haul effect.

- The influence of the haul effect on the performance of these three estimators depends on the capture probability. The distribution of the capture probability assumed in the simulation study is plotted in Figure 5.5. Note that this capture probability is also the inclusion probability used in HT-like estimators. The distribution of $r_i(l)$ is given for a haul with group-mean length equal to 48.8 cm and the density plots are for fish of length 10 cm, 30 cm, 40 cm and 60 cm. These density plots show that capture is almost certain at 60 cm and the estimated capture probability has little variance for fish larger than 60 cm. This explains why the estimators $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$ all have little bias and small MSE in Figure 5.4. If capture is almost certain, there is little difference in the estimation results obtained by different estimators (see Section 4.6 for details), and the haul effect makes no difference to the performance of different forms of estimators.

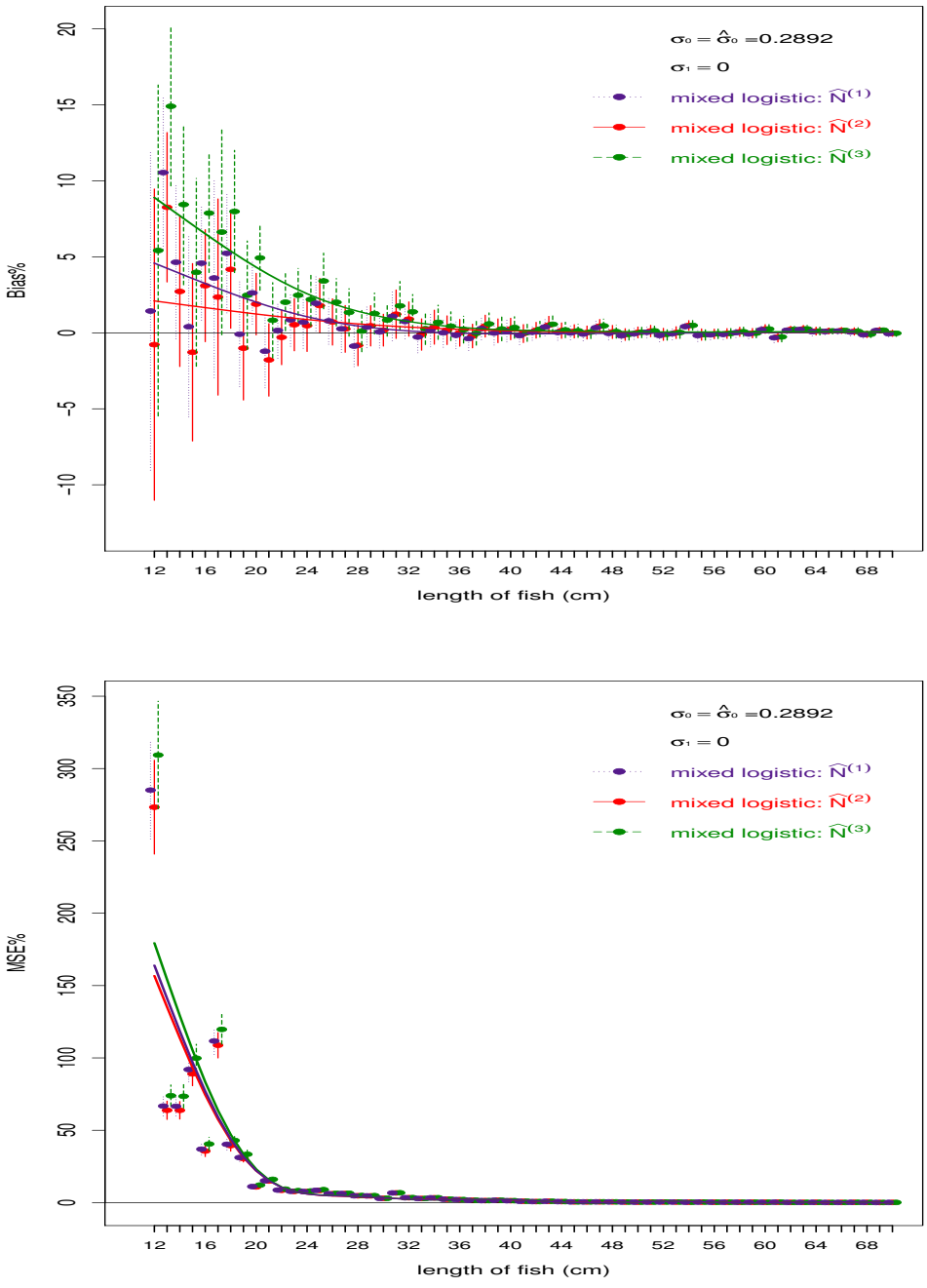


FIGURE 5.4. Simulation results for the anglerfish survey up to length 70 cm (for length beyond 70 cm, the %bias and %MSE are almost zero). The top plot shows %bias of $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$, which is scaled as a percentage of true, i.e. simulated, abundance (N) for each length class, together with its empirical CI. The bottom plot shows the %MSE of these three estimators, which is scaled as a percentage of squared true abundance (N^2) for each length class, together with its empirical CI. Each solid line is a generalized additive model fitted to the dots for each colour using the default generalized cross validation criterion (Wood, 2006).

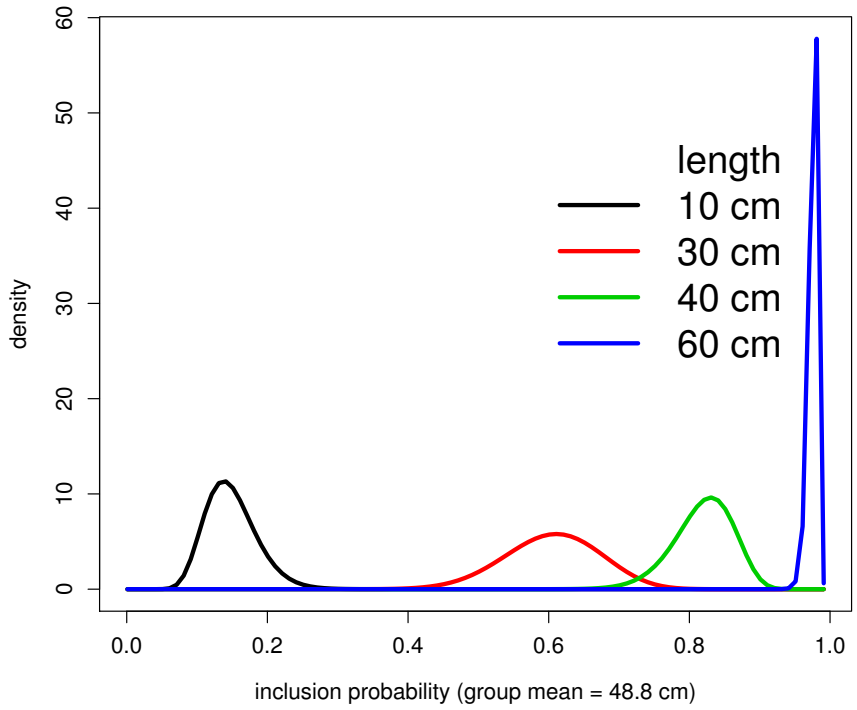


FIGURE 5.5. Density plot of the capture probability assumed in simulation study for the HT-like estimator with random effects. The density plot of the capture probability given by (5.1) with $\text{sd}(b_{0i}) = 0.0289$. The group mean of length ($\bar{l}_i$) used in the plot is 48.8 cm, and the density plots are for fish with length 10 cm, 30 cm, 40 cm and 60 cm respectively.

- The importance of the capture probability is underlined further by a simulation study with a larger haul effect, which is 5 times the standard deviation of b_{0i} assumed in Figure 5.4. The results of the simulation study with the larger haul effect are given in Figure 5.6. All the results are still truncated at 70 cm, as %bias and %MSE are both zero at lengths classes beyond 70 cm even with the increased haul effect. $\widehat{N}^{(2)}$ still performs the best, whereas $\widehat{N}^{(1)}$ is more positively biased than $\widehat{N}^{(2)}$ and this difference is more obvious with the increased haul effect. $\widehat{N}^{(3)}$ still performs the worst, much worse than $\widehat{N}^{(1)}$ and $\widehat{N}^{(2)}$ with the increased haul effect. However, when capture is almost certain, i.e., fish length larger than 60 cm, even when the haul effect is 5 times larger than its original size, there is still no difference in %bias and %MSE among these three estimators.

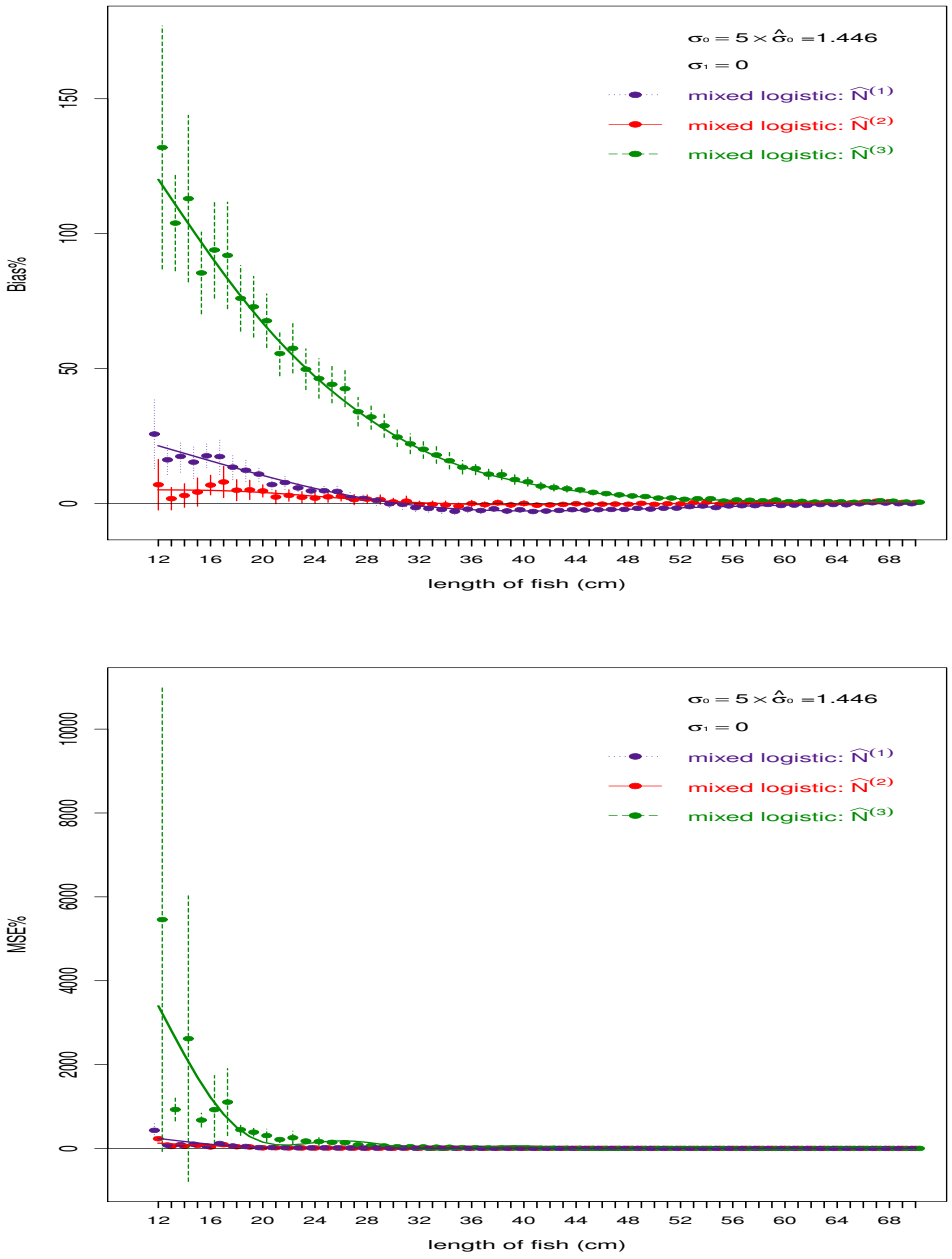


FIGURE 5.6. Simulation results for an artificial anglerfish survey with larger haul effect – 5 times the MLE of sd(b_{0i}) from experimental survey data. This means the standard deviation of b_{0i} to obtain results shown in this figure is 5 times that in Figure 5.4. The results shown here are based on the same simulation process as summarized in Figure 5.1 except that the standard deviation of the random intercept is 5 times $\hat{\sigma}_0$, i.e., 5×0.289. %bias and %MSE are calcuated as (5.2) and (5.3), respectively. Their empirical CIs are presented by vertical lines. The results for length beyond 70 cm are truncated as %bias and %MSE are almost zero for length class beyond 70 cm. Each solid line is a smooth through the dots for each colour.

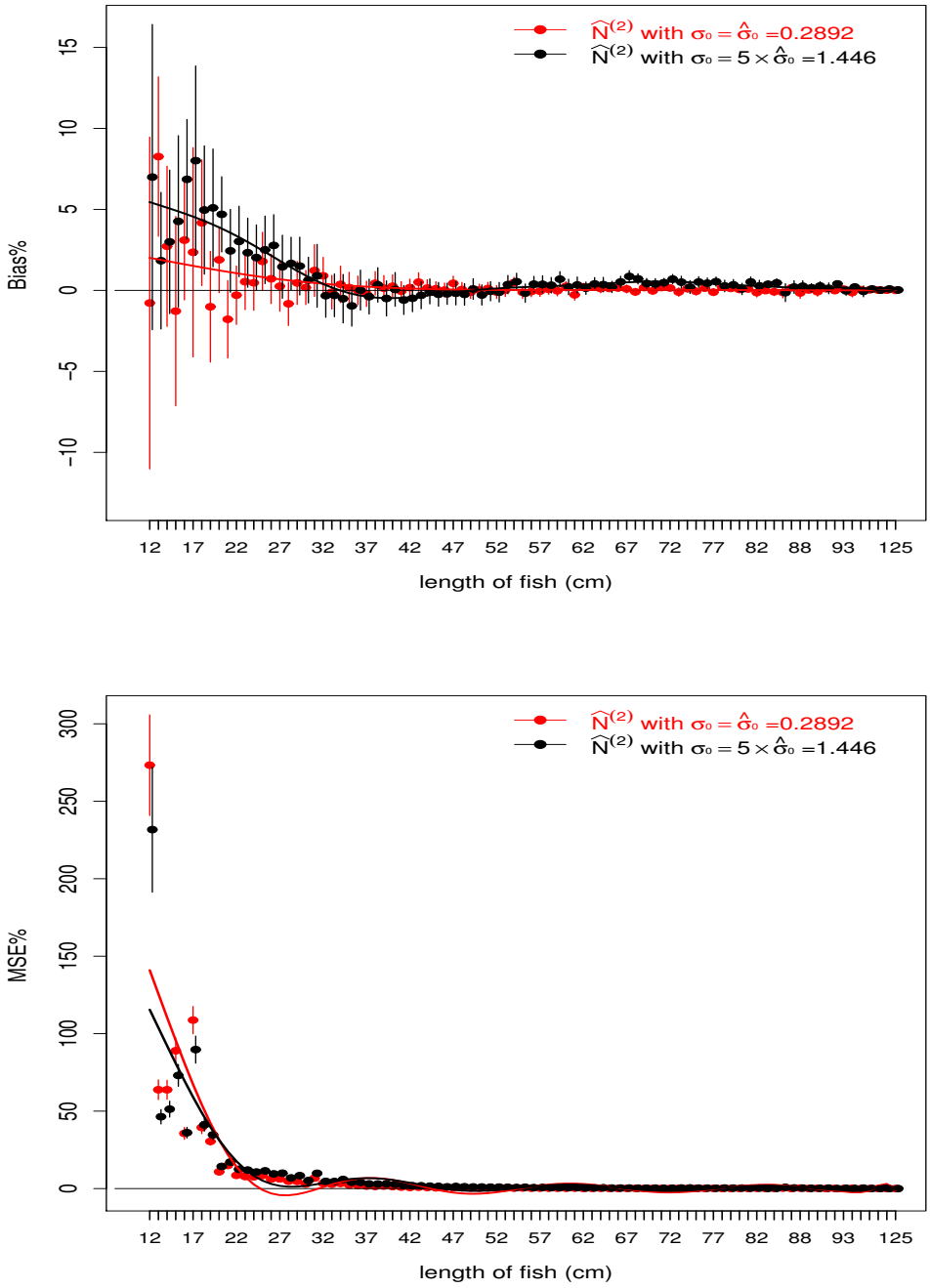


FIGURE 5.7. Plot of %bias (top) and %MSE (bottom) for $\widehat{N}^{(2)}$ only. The colour red is for the simulation study with the standard deviation of b_{0i} assumed to be 0.289, and the colour black is for the simulation study with the standard deviation of b_{0i} assumed to be 5×0.289. Other than the different sd(b_{0i}) assumed in simulation, everything else is the same as the process summarized in Figure 5.1. The %bias and %MSE are calculated according to (5.2) and (5.3) using the simulation results. The assumed abundance at each class is given in Figure 5.3. Each solid line is a smooth through the dots for each colour.

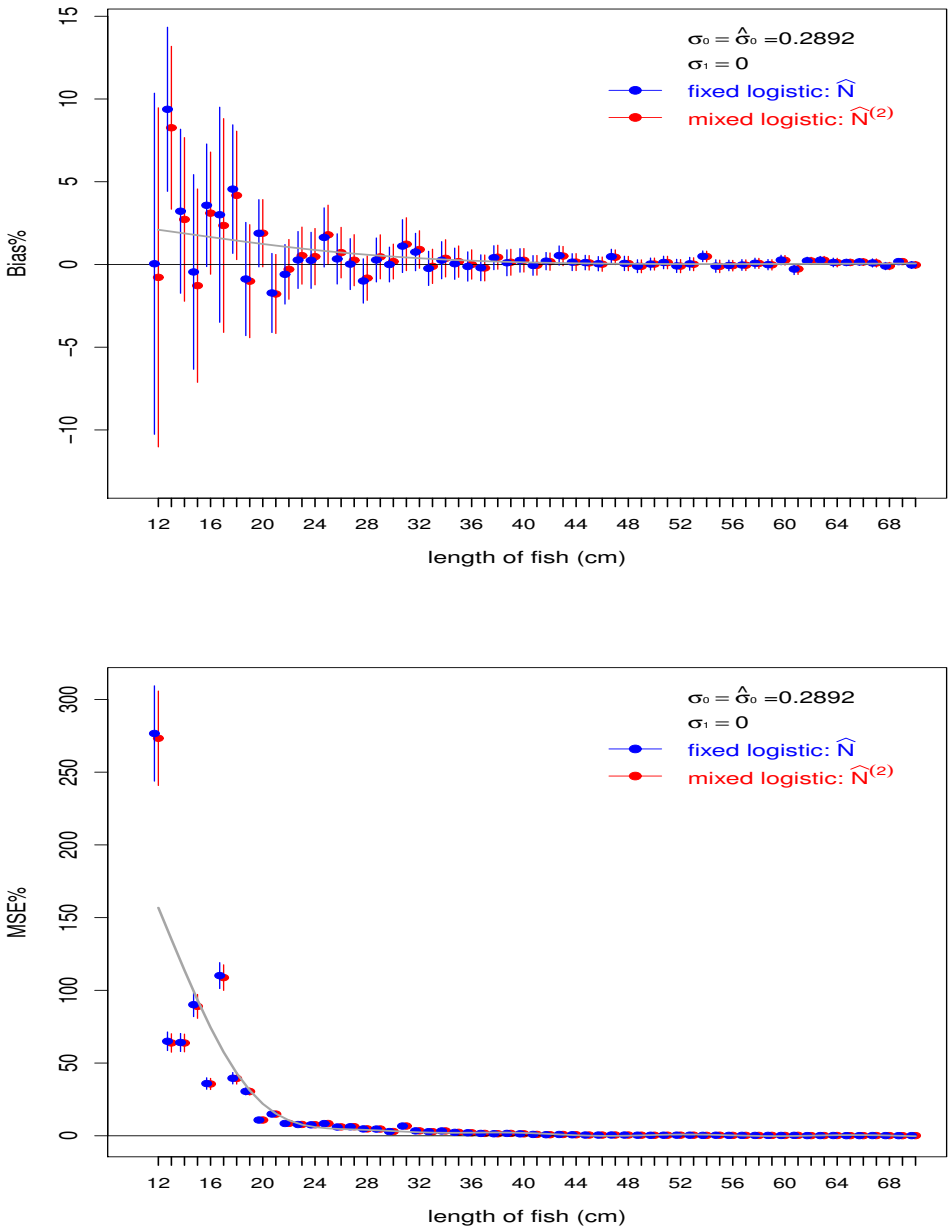


FIGURE 5.8. Plot of the simulation results to check the effect of ignoring haul effect in abundance estimation. The %bias and %MSE are calculated as (5.2) and (5.3) using the simulation results. The calculated %bias and %MSE are plotted up to length 70 cm. For length classes beyond 70 cm, both %bias and %MSE are almost zero. The top plot shows %bias of the $\widehat{N}^{(2)}$ given by (4.16) and $\widehat{N}$ without haul effect given by (4.11). The bottom plot shows the %MSE of these two estimators. Their empirical CIs are given by vertical lines. The assumed abundance in each class is given in Figure 5.3. The solid line is a smooth through the red dots for $\widehat{N}^{(2)}$.

- The performance of $\widehat{N}^{(2)}$ stays about the same in spite of a larger haul effect. Focusing on the performance of $\widehat{N}^{(2)}$ in Figure 5.4 and 5.6, Figure 5.7 checks the performance of $\widehat{N}^{(2)}$ after increasing the haul effect. This figure shows that there is no change in %MSE when the haul effect is increased, and the %bias increases only by about 5% compared with %bias with the original haul effect.

- Given that $\widehat{N}^{(2)}$ performs the best for abundance estimation with haul effect, then the question arises of what is the effect of ignoring the haul effect. To answer this question, a fixed-effects linear logistic model of the form (3.32) is fitted to the same simulated experimental survey data, which are used to obtain the results plotted in Figure 5.4. In other words, both fixed-effects and mixed-effects models are fitted to the same simulated experimental survey data to obtain $\hat{r}(l)$ and $\hat{r}_i(l)$. Then $\widehat{N}^{(2)}$ with $\hat{r}_i(l)$ and $\widehat{N}$ with $\hat{r}(l)$ are applied to the same simulated abundance survey data for abundance estimation. Comparing the %bias and %MSE between the two estimators in Figure 5.8, we can see that ignoring the haul effect in abundance estimation results only in some bias for small fish.

To sum up, $\widehat{N}^{(2)}$ should be the estimator used to incorporate the haul effect in the anglerfish abundance estimation, as it has the lowest bias and MSE in comparison with the alternative HT-like abundance estimators with the haul effect, i.e., $\widehat{N}^{(1)}$ and $\widehat{N}^{(3)}$. It is much easier to obtain the HT-like estimator without the haul effect, $\widehat{N}$, than $\widehat{N}^{(2)}$. The computational time for $\widehat{N}$ is about 6% of that for $\widehat{N}^{(2)}$ when using the same computer. However, if $\widehat{N}$ is used, then the haul effect in the survey data must be ignored because prediction of the haul effect in the abundance survey is problematic when treating haul as a fixed effect (see Section 3.5 for more discussions). In the case of the anglerfish abundance survey, the abundance estimation does not show great difference between $\widehat{N}$ and $\widehat{N}^{(2)}$. However, it is very likely that the difference between these estimators will be greater if the random effect is greater, but this is not the case here due to the limited information contained in the 2006-2007 experimental survey data.

Although it was expected that the inclusion of haul as a random effect would be important in the anglerfish abundance estimation, it turns out that except in the case of small fish, the random effects are so small that their inclusion in the logistic regression has little effect. In addition, small fish constitute a tiny proportion in the catch data from the abundance survey – only 3% of the captured fish are smaller than 20

cm. As a result, the haul effect is further attenuated by a very small sample of small fish. For the majority of the fish captured in the abundance survey, the capture probability is quite high–about 0.6 for fish larger than 30 cm, which constitutes 87% of all the captured fish. However, such a high capture probability is not generally the case in wildlife surveys, and from the HT-like estimators with random effects developed in Section 4.3.3 in order to estimate anglerfish abundance with haul effect from the trawl survey data, it then became apparent that such estimators could be useful with a range of other abundance estimation methods, including distance sampling and mark-recapture. Therefore, in Part IV, Chapter 6 develops the HT-like method to deal with random effects in a more general context, and Chapter 7 applies the estimators to different types of survey data and investigates estimator properties by simulation studies.

Part IV

Horvitz-Thompson-like estimators with random effects

Chapter 6

Horvitz-Thompson-like estimator with random effects

Key Idea: derive the properties (bias and MSE) of the HT estimator with random effects

Chapter 4 used different forms of the HT-like estimator with random effects to estimate the anglerfish abundance, and then in Chapter 5, the performance of these estimators was examined by a simulation study. It is noted that random-effects models for capture or detection probability are frequently used in the analysis of mark-recapture surveys and are sometimes used for other surveys in ecology. Though fixed-effects models are simpler and sometimes sufficient for modelling the capture/detection probability in terms of certain explanatory variables, mixed-effects models are more appropriate when capture/detection probability depends on unobserved variables, and when it depends on observed variables but inferences are to be made about a broader population than that from which the observed variable values are drawn. This chapter develops the general forms of the HT-like estimators in the context of wildlife surveys. The estimators are applicable to any survey in which estimated inclusion probabilities have random effects.

When the individual inclusion probabilities are known, the properties of the HT estimator are well-known. However, the properties of HT-like estimators (those in which inclusion probabilities are estimated) are not in general known. The properties depend on the distribution of the estimated inclusion probabilities. In wildlife surveys, the inclusion probabilities are frequently unknown and must be estimated from the survey data or supplementary data, such as the anglerfish surveys, so that estimation of inclusion probabilities is a key component of inference.

HT-like estimators have been developed for all widely-used animal abundance estimation methods (see Borchers *et al.*, 2002), including distance sampling methods (Borchers & Burnham, 2004), mark-recapture methods (Huggins, 1989; Alho, 1990a), and removal methods (Huggins & Yip, 1997). They have also been developed and applied in studies of human populations. Alho (1990b), for example, used a HT-like estimator to estimate sample moments under non-ignorable nonresponse, using logistic regression to estimate inclusion probabilities. And Volz & Heckathorn (2008) developed a HT-like estimator to estimate the population mean of any quantity measured on the sampled individuals, using "respondent-driven" sampling to estimate inclusion probabilities (see Gile & Handcock, 2010, for a review of estimators within respondent-driven sampling).

Having developed HT-like estimators with random effects modelled by a mixed-effects logistic regression for the anglerfish abundance estimation in Section 4.3.3, it became apparent that these estimators can be useful for a range of other applications with capture probability estimated by a mixed-effects logistic regression. The utility of these estimators is not only in ecology, but also in the social sciences, when surveying human populations. For example, in estimating substance abuse in a population of students, Ibiwoye & Adeleke (2011) postulate "the existence of an unobserved continuous variable, Z, which can be thought of as the propensity toward abuse". Although estimators like ours are not used by these authors, random effects are a natural way of modelling such unobservable variables.

HT-like estimators accommodate covariates in models for inclusion probabilities, but they have not to date included random effects. Chapter 4 addressed the problem of abundance estimation for a demersal (i.e. bottom-dwelling) fish, anglerfish, using trawl survey data with a capture probability than includes haul as a random effect. This motivates the consideration in this chapter of a unified framework for abundance estimation using HT-like estimators in which the inclusion probability

includes random effects, in particular when a separate experimental survey is designed to collect data for estimation of capture or detection probability.

6.1 The HT estimators with random effects

Consider a finite population of unknown size, N, in which each member of the population has associated with it a vector $\boldsymbol{r}_i$ ($i = 1, \ldots, N$). We assume also that $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N$ are independent draws from a probability density function $f_{\boldsymbol{r}}(\boldsymbol{r})$. The vector $\boldsymbol{r}$ may include covariates and random coefficients, as will be illustrated below. A sample of size n is taken from the population in such a way that individual i is included with probability $p(\boldsymbol{r}_i)$. Let δ_i be a binary random variable that is 1 if individual i is in the sample, and 0 otherwise. Denote the expected values of $\boldsymbol{r}$, $p(\boldsymbol{r})$ and $1/p(\boldsymbol{r})$ as $\mu_{\boldsymbol{r}}$, μ_p, $\mu_{1/p}$, respectively.

Although the properties of estimators of N are of greatest interest when the function $p(\boldsymbol{r})$ and the pdf $f_{\boldsymbol{r}}(\boldsymbol{r})$ are estimated, analytic investigation of the properties in this case is difficult. Therefore, to begin with, Section 6.2 considers the properties of estimators in the case where the function $p(\boldsymbol{r})$ and the probability density $f_{\boldsymbol{r}}(\boldsymbol{r})$ are known. In addition to $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$ considered for anglerfish abundance estimation, there is one more estimator available, when the random effects are observable. In distance sampling, for example, distance can be considered an observable random effect, with a known distribution. Let $\widehat{N}^{(0)}$ denote the HT-like estimator with observed random effects. Here are the four HT-like estimators that will be considered in Section 6.2:

$$\widehat{N}^{(0)} = \sum_{i=1}^{N} \frac{\delta_i}{p(\boldsymbol{r}_i)}, \tag{6.1}$$

$$\widehat{N}^{(1)} = \sum_{i=1}^{N} \frac{\delta_i}{p(\mu_{\boldsymbol{r}})}, \tag{6.2}$$

$$\widehat{N}^{(2)} = \sum_{i=1}^{N} \frac{\delta_i}{\mu_p}, \tag{6.3}$$

$$\widehat{N}^{(3)} = \sum_{i=1}^{N} \delta_i \, \mu_{1/p}. \tag{6.4}$$

Section 6.2 will examine the properties of estimators (6.1)–(6.4) with known inclusion probability.

6.2 Estimators properties

Armed now with these four HT-like estimators with random effects given by (6.1)–(6.4), we investigate their properties analytically as far as we are able to. To start with, the laws of total expectation and total variance are applied to work out the $\mathbf{E}[\delta_i]$ and $\text{Var}[\delta_i]$, where δ_i is the binary random variable for an animal being captured or not, and $\boldsymbol{r}_i$ is a random vector attached to individual i. For $i = 1, 2, \ldots, N$:

$$\begin{aligned} \mathbf{E}_{\delta_i}[\delta_i] &= \mathbf{E}_{\boldsymbol{r}_i}\{\mathbf{E}_{\delta_i|\boldsymbol{r}_i}[\delta_i \mid \boldsymbol{r}_i]\} = \mathbf{E}_{\boldsymbol{r}_i}[p(\boldsymbol{r}_i)] = \mu_p, \qquad (6.5) \\ \text{Var}_{\delta_i}[\delta_i] &= \text{Var}_{\boldsymbol{r}_i}\{\mathbf{E}_{\delta_i|\boldsymbol{r}_i}[\delta_i \mid \boldsymbol{r}_i]\} + \mathbf{E}_{\boldsymbol{r}_i}\{\text{Var}_{\delta_i|\boldsymbol{r}_i}[\delta_i \mid \boldsymbol{r}_i]\} \\ &= \mu_p - \mu_p^2. \qquad (6.6) \end{aligned}$$

Using the above and treating $\boldsymbol{r}_i$ $(i = 1, \ldots, N)$ as independent and identically distributed random vectors, the following gives the expectations and variances of the four estimators $\widehat{N}^{(m)}$ $(m = 0, 1, 2, 3)$ given by (6.1)–(6.4):

$$\mathbf{E}[\widehat{N}^{(0)}] = \sum_{i=1}^{N} \mathbf{E}_{\boldsymbol{r}_i}\left\{\mathbf{E}_{\delta_i|\boldsymbol{r}_i}\left[\frac{\delta_i}{p(\boldsymbol{r}_i)}\,\middle|\,\boldsymbol{r}_i\right]\right\} = \sum_{i=1}^{N} \mathbf{E}_{\boldsymbol{r}_i}\left[\frac{p(\boldsymbol{r}_i)}{p(\boldsymbol{r}_i)}\right] = N, \quad (6.7)$$

$$\mathbf{E}[\widehat{N}^{(1)}] = \frac{1}{p(\mu_{\boldsymbol{r}})}\sum_{i=1}^{N} \mathbf{E}_{\delta_i}[\delta_i] = \frac{1}{p(\mu_{\boldsymbol{r}})}\sum_{i=1}^{N} \mu_p = N\,\frac{\mu_p}{p(\mu_{\boldsymbol{r}})}, \quad (6.8)$$

$$\mathbf{E}[\widehat{N}^{(2)}] = \frac{1}{\mu_p}\sum_{i=1}^{N} \mathbf{E}_{\delta_i}[\delta_i] = \frac{1}{\mu_p}\sum_{i=1}^{N} \mu_p = N, \quad (6.9)$$

$$\mathbf{E}[\widehat{N}^{(3)}] = \mathbf{E}\left[\frac{1}{p(\boldsymbol{r})}\right]\sum_{i=1}^{N} \mathbf{E}_{\delta_i}[\delta_i] = N\,\mu_p\,\mu_{1/p}\,, \quad (6.10)$$

$$\text{Var}[\widehat{N}^{(0)}] = 0 + \sum_{i=1}^{N} \mathbf{E}_{\boldsymbol{r}_i}\left[\frac{1}{p(\boldsymbol{r_i})} - 1\right] = N(\mu_{1/p} - 1)\,, \quad (6.11)$$

$$\text{Var}[\widehat{N}^{(1)}] = \frac{1}{p(\mu_{\boldsymbol{r}})^2}\sum_{i=1}^{N} \text{Var}_{\delta_i}[\delta_i] = [\mu_p - \mu_p^2]\frac{N}{p(\mu_{\boldsymbol{r}})^2}\,, \quad (6.12)$$

$$\text{Var}[\widehat{N}^{(2)}] = \frac{1}{\mu_p^2}\sum_{i=1}^{N} \text{Var}_{\delta_i}[\delta_i] = [\mu_p - \mu_p^2]\frac{N}{\mu_p^2} = N\,\left[\frac{1}{\mu_p} - 1\right], \quad (6.13)$$

$$\text{Var}[\widehat{N}^{(3)}] = \mu_{1/p}^2 \sum_{i=1}^{N} \text{Var}_{\delta_i}[\delta_i] = N\,\mu_{1/p}^2\,[\mu_p - \mu_p^2]. \tag{6.14}$$

Based on (6.7)–(6.14), the MSEs of these four estimators are then compared with each other, using the fact that $\text{MSE}(\widehat{N}) = \text{Var}[\widehat{N}] + \Big(\text{bias}[\widehat{N}]\Big)^2$.

1. $1/p(\boldsymbol{r})$ is a convex function of $p(\boldsymbol{r})$ and by Jensen's inequality [1], it follows that $\mu_{1/p} \geq 1/\mu_p$ for any p. Given this and (6.11) and (6.13), it follows that $\text{Var}[\widehat{N}^{(0)}] \geq \text{Var}[\widehat{N}^{(2)}]$. Because both estimators are unbiased, $\text{MSE}[\widehat{N}^{(0)}] \geq \text{MSE}[\widehat{N}^{(2)}]$.

2. Given (6.8), $\widehat{N}^{(1)}$ is unbiased when $p(\boldsymbol{r})$ is a linear function of $\boldsymbol{r}$, in which case $p(\mu_{\boldsymbol{r}}) = \mu_p$. Otherwise, by Jensen's equality, $\mu_p \geq p(\mu_{\boldsymbol{r}})$ when $p(\boldsymbol{r})$ is convex; $\mu_p \leq p(\mu_{\boldsymbol{r}})$ when $p(\boldsymbol{r})$ is concave. In other words, $\widehat{N}^{(1)}$ changes its direction of bias at the inflexion point of $p(\boldsymbol{r})$.

   Furthermore, from (6.12) and (6.13), noting that $\mu_p \geq p(\mu_{\boldsymbol{r}})$ for convex $p(\boldsymbol{r})$, it follows that $\text{Var}[\widehat{N}^{(1)}] \geq \text{Var}[\widehat{N}^{(2)}]$ with a convex $p(\boldsymbol{r})$. Since $\widehat{N}^{(2)}$ is unbiased while $\widehat{N}^{(1)}$ is not, it follows that $\text{MSE}[\widehat{N}^{(1)}] \geq \text{MSE}[\widehat{N}^{(2)}]$ when $p(\boldsymbol{r})$ is convex.

   When $p(\boldsymbol{r})$ is concave, $p(\mu_{\boldsymbol{r}}) \geq \mu_p$. Given this inequality, together with the variance estimators of $\widehat{N}^{(1)}$ and $\widehat{N}^{(2)}$ given in (6.12) and (6.13), it is clear that $\text{Var}[\widehat{N}^{(1)}] \leq \text{Var}[\widehat{N}^{(2)}]$. Further, the MSE's of $\widehat{N}^{(1)}$ and $\widehat{N}^{(2)}$ are,

$$\begin{aligned} \text{MSE}[\widehat{N}^{(1)}] &= \text{Var}[\widehat{N}^{(1)}] + (\text{bias}[\widehat{N}^{(1)}])^2, \\ \text{MSE}[\widehat{N}^{(2)}] &= \text{Var}[\widehat{N}^{(2)}] + 0. \end{aligned}$$

   Therefore, the relationship between $\text{MSE}[\widehat{N}^{(1)}]$ and $\text{MSE}[\widehat{N}^{(2)}]$ with concave $p(\boldsymbol{r})$ is not straightforward. For a better understanding of this relationship, we now consider conditions under which the inequality

$$\text{MSE}[\widehat{N}^{(2)}] \geq \text{MSE}[\widehat{N}^{(1)}] \tag{6.15}$$

[1] In the context of probability theory, the Jensen's inequality is stated as: if X is a random variable and f is a convex function, then $f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)]$.

holds. In terms of variances and squared bias, inequality (6.15) between MSE's can be written as

$$\text{Var}[\widehat{N}^{(2)}] - \text{Var}[\widehat{N}^{(1)}] - \left(\text{bias}[\widehat{N}^{(1)}]\right)^2 \geq 0.$$

The question then becomes when the squared bias of $\widehat{N}^{(1)}$ is large enough to compensate for the difference between the variances of the two estimators, $\widehat{N}^{(2)}$ and $\widehat{N}^{(1)}$. To answer this question, consider the following inequalities:

$$\begin{aligned}
\frac{\text{Var}[\widehat{N}^{(2)}] - \text{Var}[\widehat{N}^{(1)}]}{(\text{bias}[\widehat{N}^{(1)}])^2} &= \frac{\left[\mu_p - \mu_p^2\right]\; N\; \left(\dfrac{1}{\mu_p^2} - \dfrac{1}{p(\mu_{\boldsymbol{r}})^2}\right)}{N^2\left[\dfrac{\mu_p}{p(\mu_{\boldsymbol{r}})} - 1\right]^2} \\
&= \frac{\mu_p^2\left[\dfrac{1}{\mu_p} - 1\right]\; N\; \left[\dfrac{p(\mu_{\boldsymbol{r}})^2 - \mu_p^2}{\mu_p^2\, p(\mu_{\boldsymbol{r}})^2}\right]}{N^2\,\dfrac{[\mu_p - p(\mu_{\boldsymbol{r}})]^2}{p(\mu_{\boldsymbol{r}})^2}} \\
&= \frac{1}{N}\left[\frac{1}{\mu_p} - 1\right]\left[\frac{p(\mu_{\boldsymbol{r}}) + \mu_p}{p(\mu_{\boldsymbol{r}}) - \mu_p}\right] \qquad (6.16) \\
&\geq \frac{1}{N}\left[\frac{1}{\mu_p} - 1\right]. \qquad (6.17)
\end{aligned}$$

If $\widehat{N}^{(2)}$ has larger MSE than $\widehat{N}^{(1)}$, then (6.16) must be larger than one. To guarantee this, it can be seen from (6.17) that μ_p needs to be no more than $1/(N+1)$. Therefore, when the mean detection probability, μ_p, is no more than $1/(N+1)$, $\widehat{N}^{(2)}$ has higher MSE than $\widehat{N}^{(1)}$.

The above shows that if $\mu_p \leq 1/(N+1)$, then (6.17)≥ 1, hence $\text{MSE}[\widehat{N}^{(2)}] \geq \text{MSE}[\widehat{N}^{(1)}]$. However, in most applications μ_p is unlikely to be this small, and in this case it is not apparent which of the two estimators has lower MSE.

3. Given (6.10), $\widehat{N}^{(3)}$ is always positively biased because $\mu_{1/p}\,\mu_p \geq 1$ for any $p(\boldsymbol{r})$. This follows from $\mu_{1/p} \geq 1/\mu_p$ for any $p(\boldsymbol{r})$ given in item (1) above. Using this result together with (6.13) and (6.14) implies that $\text{Var}[\widehat{N}^{(3)}] \geq \text{Var}[\widehat{N}^{(2)}]$ for any $p(\boldsymbol{r})$. Furthermore, because $\widehat{N}^{(2)}$ is unbiased and $\widehat{N}^{(3)}$ is always positively biased, it follows that $\text{MSE}[\widehat{N}^{(3)}] \geq \text{MSE}[\widehat{N}^{(2)}]$ for any $p(\boldsymbol{r})$.

In summary, this section shows that the estimators $\widehat{N}^{(0)}$ and $\widehat{N}^{(2)}$ are unbiased, while the expectations of $\widehat{N}^{(1)}$ and $\widehat{N}^{(3)}$ are $N\mu_p/p(\mu_{\boldsymbol{r}})$ and $N\mu_p\,\mu_{1/p}$ respectively. In addition, it also shows that $\widehat{N}^{(2)}$ has the lowest mean square error (MSE) among the four estimators when $p(\boldsymbol{r})$ is convex, and that $\widehat{N}^{(1)}$ has the lowest MSE when $p(\boldsymbol{r})$ is concave and $\mu_p \leq 1/(N+1)$. When $p(\boldsymbol{r})$ is concave and $\mu_p > 1/(N+1)$, the analytical results shown above cannot be used to determine which estimator out of $\widehat{N}^{(1)}$ and $\widehat{N}^{(2)}$ has the lower MSE.

While these results are not conclusive, they suggest that the estimators of the form of $\widehat{N}^{(2)}$ may perform well relative to the others when the inclusion probabilities with random effects are estimated, not known. This suggestion is consistent with the simulation results for anglerfish in Chapter 5. To investigate the estimator properties in more general contexts than the anglerfish survey, the performances of different forms of the HT-like estimators are investigated in the context of another three different kinds of survey in Chapter 7. These surveys include a trapping point transect survey, a line transect survey and a mark-recapture survey. A simulation study is conducted for these estimators in context of each of these three surveys in Chapter 7, and the results from the simulation study are consistent with the analytical results shown above.

6.3 HT-like estimators with random effects modelled by a mixed-effects logistic regression model

For cases in which a separate experimental survey is used to estimate detection or capture probabilities with random effects, a mixed-effects logistic regression is normally used to model the detection probability (see Section 3.3 for more detail about the mixed-effects logistic regression model). This section develops the HT-like estimators, $\widehat{N}^{(1)}$, $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$ specifically for the capture probability that is estimated by a mixed-effects logistic regression model. It also gives an analytic way of integrating out the random effects in $\widehat{N}^{(3)}$ and a method for reducing the multi-dimensional integration in $\widehat{N}^{(2)}$ to a one-dimensional integration.

To begin with, recall that mixed-effects modelling is also known as hierarchical or multilevel modelling. Taking a two-level mixed-effects model for illustration, the term 'level-one' is used for the observations within each cluster, group, or subject,

and the term 'level-two' for the clusters, groups or subjects, in which the level-one observations are clustered. The random effects arise in the level-two unit, which is the sampling unit in both the experimental and abundance surveys.

The following notation is used for the HT-like abundance estimators with a two-level mixed-effects logistic detection function:

- i: the ith unit at level-two, and $i = 1, \ldots, n$, where n is the total number of distinct units sampled at level-two;
- $\boldsymbol{\beta}$: the $q \times 1$ fixed-effects parameter vector for the capture probability, where $q - 1$ is the number of explanatory variables;
- $\boldsymbol{x}_{ij}$: the $q \times 1$ fixed-effects model vector of the jth observation within the ith level-two unit, with the first element being constant 1;
- $\boldsymbol{b}_i$: the $m \times 1$ random-effects coefficient vector of the ith level-two unit, where $m = \dim(\boldsymbol{b}_i)$. Unlike $\boldsymbol{\beta}$, $\boldsymbol{b}_i$ is not a parameter vector; it is a random draw from the assumed distribution for $\boldsymbol{b}_i$, i.e., $\boldsymbol{b}_i \sim N(\mathbf{0}\,, \Sigma_b)$, and its pdf is denoted by $f_b(\boldsymbol{b})$. The random-effects parameter is Σ_b;
- $\boldsymbol{z}_{ij}$: the $m \times 1$ random effects model vector of the jth observation within the ith level-two unit;
- A_i^c: the area covered by the ith level-two sampling unit;
- A_s: the area of the survey region.

Note that we consider the case in which $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are observed in the abundance survey. If there is any unobserved predictor in $\boldsymbol{x}_{ij}$, then this predictor can be considered as an additional random effect. The estimators considered below are complicated by adding another layer of random effect, which is the case for the application that will be described in Section 7.1.

Let $\hat{g}_{ij}$ denote the estimated detection probability of the jth level-one observation within the area sampled by the ith level-two unit. Given the random effects $\boldsymbol{b}_i$, the estimated detection probability is $\text{logit}^{-1}(\boldsymbol{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}}+\boldsymbol{z}_{ij}^{\mathrm{T}}\boldsymbol{b}_i)$, which is estimated by a mixed-effects logistic regression model (see Section 3.3 for details) using the experimental survey data. Conditional on the $\hat{g}_{ij}$, the inclusion probability in the abundance survey of the jth level-one observation within the ith level-two unit is $\hat{p}_{ij} = \hat{g}_{ij}\,\mathrm{A}_i^c/\mathrm{A}_s$.

Therefore, the HT-like estimator and its alternative forms with a mixed-effects logistic $\hat{g}_{ij}$ are:

$$\widehat{N}^{(0)} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{1}{\hat{p}_{ij}(\boldsymbol{x}_{ij},\boldsymbol{z}_{ij};\boldsymbol{b}_i)}, \tag{6.18}$$

$$\widehat{N}^{(1)} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{1}{\hat{p}_{ij}(\boldsymbol{x}_{ij},\boldsymbol{z}_{ij};\boldsymbol{b}=\boldsymbol{0})}, \tag{6.19}$$

$$\widehat{N}^{(2)} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{1}{\widehat{\mathbf{E}}_b\left[\hat{p}_{ij}(\boldsymbol{x}_{ij},\boldsymbol{z}_{ij};\boldsymbol{b})\right]}, \tag{6.20}$$

$$\widehat{N}^{(3)} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\widehat{\mathbf{E}}_b\left[\frac{1}{\hat{p}_{ij}(\boldsymbol{x}_{ij},\boldsymbol{z}_{ij};\boldsymbol{b})}\right]. \tag{6.21}$$

In practice, the random effects vector $\boldsymbol{b}_i$ is not observable for sampling unit i, and $\widehat{N}^{(0)}$ is therefore not available. The estimator $\widehat{N}^{(1)}$ takes $\boldsymbol{b}_i$ to be equal to its expectation, $\boldsymbol{0}$ in the detection probability $\hat{p}_{ij}$. It then follows that there is no integration of $\boldsymbol{b}$ in $\widehat{N}^{(1)}$. The following gives the analytic solution to integrating out $\boldsymbol{b}$ in $\widehat{N}^{(3)}$, and shows how to reduce the dimension of integration of $\boldsymbol{b}$ in $\widehat{N}^{(2)}$ from m, giving a substantial saving in computing time. The analytic solution for integrating out the $\boldsymbol{b}$ in a mixed-effects logistic $\hat{g}_{ij}$ in $\widehat{N}^{(3)}$ is:

$$\begin{aligned}
&\widehat{\mathbf{E}}_b\left[1/\hat{p}_{ij}(\boldsymbol{x}_{ij},\boldsymbol{z}_{ij};\,\boldsymbol{b})\right]\\
=\;&\widehat{\mathbf{E}}_b[\mathrm{A}_s/(\mathrm{A}_i^c\,\hat{g}_{ij})]\\
=\;&\frac{\mathrm{A}_s}{\mathrm{A}_i^c}\int_{\mathbb{R}^m}\cdots\int\left\{1+\exp[-\boldsymbol{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}}-\boldsymbol{z}_{ij}^{\mathrm{T}}\boldsymbol{b}]\right\}\,\hat{f}_b(\boldsymbol{b})\,d\,\boldsymbol{b}\\
=\;&\frac{\mathrm{A}_s}{\mathrm{A}_i^c}\int_{\mathbb{R}^m}\cdots\int\left\{1+\exp[-\boldsymbol{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}}-\boldsymbol{z}_{ij}^{\mathrm{T}}\boldsymbol{b}]\right\}\,\frac{1}{(2\pi)^{m/2}|\widehat{\Sigma}_b|^{1/2}}\,\exp\left(-\frac{\boldsymbol{b}^{\mathrm{T}}\,\widehat{\Sigma}_b^{-1}\,\boldsymbol{b}}{2}\right)\,d\,\boldsymbol{b}\\
=\;&\frac{\mathrm{A}_s}{\mathrm{A}_i^c}\left\{1+\exp(-\boldsymbol{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}})\underbrace{\int_{\mathbb{R}^m}\cdots\int\frac{1}{(2\pi)^{m/2}|\widehat{\Sigma}_b|^{1/2}}\exp\left[-\frac{\boldsymbol{b}^{\mathrm{T}}\,\widehat{\Sigma}_b^{-1}\,\boldsymbol{b}}{2}-\boldsymbol{z}_{ij}^{\mathrm{T}}\boldsymbol{b}\right]\,d\,\boldsymbol{b}}\right\}
\end{aligned} \tag{6.22}$$

$$= \frac{\mathrm{A}_s}{\mathrm{A}_i^c}\left[1+\exp\left(-\boldsymbol{x}_{ij}^{\mathrm{T}}\,\widehat{\boldsymbol{\beta}}+\frac{\boldsymbol{z}_{ij}^{\mathrm{T}}\,\widehat{\Sigma}_b\,\boldsymbol{z}_{ij}}{2}\right)\right], \tag{6.23}$$

where in (6.22) the part with $\underbrace{\quad}$ is equal to

$$\begin{aligned} & \int_{\mathbb{R}^m} \cdots \int \frac{1}{(2\pi)^{m/2}|\widehat{\Sigma}_b|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{b}+\widehat{\Sigma}_b \boldsymbol{z}_{ij})^{\mathrm{T}}\, \widehat{\Sigma}_b^{-1}(\boldsymbol{b}+\widehat{\Sigma}_b \boldsymbol{z}_{ij}) + \frac{1}{2}\boldsymbol{z}_{ij}^{\mathrm{T}}\widehat{\Sigma}_b \boldsymbol{z}_{ij}\right] d\,\boldsymbol{b} \\ = \quad & \exp\left[\frac{1}{2}\boldsymbol{z}_{ij}^{\mathrm{T}}\widehat{\Sigma}_b \boldsymbol{z}_{ij}\right] \underbrace{\int_{\mathbb{R}^m} \cdots \int \frac{1}{(2\pi)^{m/2}|\widehat{\Sigma}_b|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{b}+\widehat{\Sigma}_b \boldsymbol{z}_{ij})^{\mathrm{T}}\, \widehat{\Sigma}_b^{-1}(\boldsymbol{b}+\widehat{\Sigma}_b \boldsymbol{z}_{ij})\right] d\,\boldsymbol{b}}_{=1} \end{aligned}$$

.

The vector $\boldsymbol{b}$ in $\widehat{N}^{(2)}$ cannot be integrated out analytically in closed form. However, the integration of $\boldsymbol{b}$ in $\widehat{N}^{(2)}$ can be reduced from m dimensions to one dimension:

$$\begin{aligned} & \widehat{\mathbf{E}}_b[\,\hat{p}_{ij}(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij};\, \boldsymbol{b})] \\ = \quad & \widehat{\mathbf{E}}_b[\,\hat{g}_{ij}\, \mathrm{A}_i^c/\mathrm{A}_s] \\ = \quad & \frac{\mathrm{A}_i^c}{\mathrm{A}_s} \int_{\mathbb{R}^m} \cdots \int \left\{1+\exp[-\boldsymbol{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \boldsymbol{z}_{ij}^{\mathrm{T}}\boldsymbol{b}]\right\}^{-1}\, \hat{f}_b(\boldsymbol{b})\, d\,\boldsymbol{b} \\ = \quad & \frac{\mathrm{A}_i^c}{\mathrm{A}_s} \int_{-\infty}^{\infty} \left\{1+\exp[-\boldsymbol{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - u_{ij}]\right\}^{-1}\, \hat{f}_{u_{ij}}(u_{ij})\, d\,u_{ij}, \end{aligned} \tag{6.24}$$

where $u_{ij} = \boldsymbol{z}_{ij}^{\mathrm{T}}\, \boldsymbol{b}$, a linear combination of $\boldsymbol{b}_i$. It follows that

$$u_{ij} \sim N(0,\, \boldsymbol{z}_{ij}^{\mathrm{T}}\, \widehat{\Sigma}_b\, \boldsymbol{z}_{ij}),$$

which means the probability density function of μ_{ij} is

$$\hat{f}_{u_{ij}}(u_{ij}) = (2\pi\, \boldsymbol{z}_{ij}^{\mathrm{T}}\, \widehat{\Sigma}_b\, \boldsymbol{z}_{ij})^{-1/2} \exp[-u_{ij}^2/(2\, \boldsymbol{z}_{ij}^{\mathrm{T}}\, \widehat{\Sigma}_b\, \boldsymbol{z}_{ij})].$$

Further, it follows that

$$\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}} + u_{ij} \;\sim\; N(\boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}},\, \boldsymbol{z}_{ij}^{\mathrm{T}}\, \widehat{\Sigma}_b\, \boldsymbol{z}_{ij}),$$

and so, the pdf of $\hat{g}_{ij}$ conditional on $\boldsymbol{x}_{ij}$, $\boldsymbol{z}_{ij}$, $\hat{\boldsymbol{\beta}}$ and $\widehat{\Sigma}_b$ is

$$\frac{1}{\sqrt{2\pi \boldsymbol{z}_{ij}^{\mathrm{T}}\, \widehat{\Sigma}_b\, \boldsymbol{z}_{ij}}} \frac{1}{\hat{g}_{ij}(1-\hat{g}_{ij})} \exp\left\{-\frac{[\mathrm{logit}(\hat{g}_{ij}) - \boldsymbol{x}_{ij}\hat{\boldsymbol{\beta}}]^2}{2\boldsymbol{z}_{ij}^T\widehat{\Sigma}_b\boldsymbol{z}_{ij}}\right\}.$$

Chapter 7

Applications of HT-like estimators in ecology

Key Idea: study the properties of the HT-like estimators with random effects in the context of three different types of survey data with different forms of random effects

The previous chapter explored analytically, as far as was possible, the performance of four different forms of HT-like estimator with random effects, given by (6.1)–(6.4). The analytic results suggested that $\widehat{N}^{(2)}$ given by (6.3) is unbiased with known inclusion probability, and $\widehat{N}^{(2)}$ has the lowest MSE when the inclusion probability is a concave function of random effects. These analytical results are obtained under a condition that everything related to the inclusion probability is known. However, this condition is seldom met for the wildlife surveys. As was the case with the studies for the anglerfish survey, neither the capture probability nor the distribution of the random effect is known. A simulation study is therefore performed to examine the properties of the HT-like estimators.

Motivated by the anglerfish application in Part III, together with the analytic results obtained above, this chapter investigates the performance of HT-like estimators with estimated inclusion probabilities and random effects, in the context of four different

kinds of surveys. These are the anglerfish survey, a trapping point transect survey (woodrat), a mark-recapture survey (wood mouse), and a line transect survey (Dall's Porpoise). See Figure 7.1 for an image of each species. Simulation studies are conducted in the context of each survey, and then the performance of each estimator is measured in terms of bias and MSE, both of which are scaled as percentages according to (5.2) and (5.3). The results are discussed in Section 7.4.

FIGURE 7.1. Images of the species involved in the four applications: anglerfish (top left), woodrat (top right), Dall's porpose (bottom left) and wood mouse (bottom right).

7.1 Trapping point transect survey: woodrats

Point transect sampling with traps (Buckland *et al.*, 2006) can be thought of as a kind of distance sampling (point transect) survey in which the key explanatory variable distance, x, is not observed. Because x is not observed, distance sampling methods cannot be used to estimate the detection function $g(x)$; instead a separate experiment is conducted for this purpose. As in the anglerfish experimental survey, the survey data collected from the experimental survey data have binary response with multiple trials on different individual woodrats. Therefore, in order to incorporate the individual effect in the detection probability, in Section 7.1.1 we fit a mixed-effect logistic regression model for detectability estimation using the experimental survey data with individual woodrat as the random effect.

A trapping point transect survey was conducted to estimate the abundance of the Key Largo woodrat (*Neotoma floridana smalli*) in 2008 and 2009 (Potts, 2011). The Key Largo woodrat is an endangered rodent restricted to 850 hectares of remnant vegeta-

tion on northern Key Largo island, Florida, USA. Obtaining reliable estimates of the abundance of Key Largo woodrats is of great importance for making management decisions to maintain the ecosystem inhabited by this subspecies. In Section 7.1.2 we conduct the abundance estimation for the 2008 abundance survey, which used a randomly-placed systematic grid of 137 trapping points with 250 meters between traps. Data at each trapping point were collected on three consecutive nights, and 19 woodrats were detected by traps.

7.1.1 Detectability estimation

Detection probability depends on the distance of the woodrats' nests from the traps. Because these distances were not observed in the abundance survey, a separate survey (referred to as the experimental survey) was conducted to estimate detection probability as a function of distance. This survey was conducted near in time to the abundance survey, and in the same region. It contained multiple trials on each of a number of radio-collared woodrats, with each trial involving placing traps at a given distance from a location near a woodrat's nest. Let y_{ij} denote the response for the jth trial on the ith individual woodrat. Then $y_{ij} = 1$ means the woodrat was detected and $y_{ij} = 0$ means that it was not. Individual variation in response probability is expected, with some individuals being more susceptible to capture than others.

Analogous to modelling the capture probability in the anglerfish survey as a logistic function of fish length with haul as the random effect in Section 3.4, the probability of capture was modelled as a logistic function of distance with individual woodrat random effects. However, the individual woodrat random effects are expected to be much greater than the haul effect in the anglerfish survey. A two-level random-effects model of the form given by (3.71) is applied to estimate the detectability of woodrats, with individual woodrats comprising the level-two effects for this analysis, while separate trials on the same individual animal comprise the level-one effects.

Due to the concern about the sample size requirement for estimating the distribution of random effects, the 2008-2009 experimental survey data are pooled together to give a total of 512 trials on 28 woodrats. Release distances varied from 1 meter to 319 meters, by which distance detection is believed to be extremely unlikely. The trials with the largest distances were conducted to facilitate accurate estimation of the tail of the detection function.

A small simulation study was conducted to check whether or not the woodrats experimental survey data are sufficient to estimate the distribution of the random effects. This simulation study also considered the choice of the distances set up in the experimental survey, with the same amount of effort (i.e., the number of traps set up in the survey) and different choices of distances from traps. The choice of distances from traps is made for more precise estimates in the detection function. For each individual woodrat in the experimental survey, 21 traps are set up with different distances in simulation, since a sample size less than 20 at level-two usually gives limited information about the random effect distribution (Snijders & Bosker, 1999). The distances of these 21 traps have three different protocols

1. distances that are the same as those in 2008-2009 experimental survey data;

2. distances that lead to equally spaced traps from 0 to 319 m; and

3. distances with more traps closer to a collared woodrat and fewer traps further away (i.e., more effort is spent on setting up experiments for estimating high detectability). The experimental distances per woodrat are: 1, 1, 1, 6, 6, 6, 11, 11, 11, 21, 21, 41, 41, 51, 61, 71, 81, 141, 201, 261, 320 meters.

The results show that protocol 3 works the best among the three in terms of precision of parameter estimates, and therefore the third protocol is the one used in the simulation study described in the next section. The simulation study also shows that even for a sample of 50 woodrats and a balanced survey design, i.e., 21 trials for each individual woodrat, the correlation between random intercept b_0 and random slope b_1 is very poorly estimated. This leads to some implausible detection functions that do not decay to zero over the range of the experimental survey distances; see the top plot in Figure 7.2. This in turn causes negative bias in abundance estimation.

The assumed detection function in Figure 7.2 is

$$\text{logit}[g(x, \boldsymbol{b}_i)] = (0.233 + b_{0i}) + (-0.076 + b_{1i})\, x, \tag{7.1}$$

where the standard deviation of b_{0i}, $\sigma_0 = 0.651$, the standard deviation of b_{1i}, $\sigma_1 = 0.026$ and the correlation between b_{0i} and b_{1i}, $\text{corr}(b_{0i}, b_{1i}) = 0.277$. This is a non-centred one-level mixed-effects logistic regression model and it is the model fitted to the woodrats experimental survey data. This assumed detection function, i.e., the

true detection function in the simulation study, is given by a thick black line in both plots of Figure 7.2.

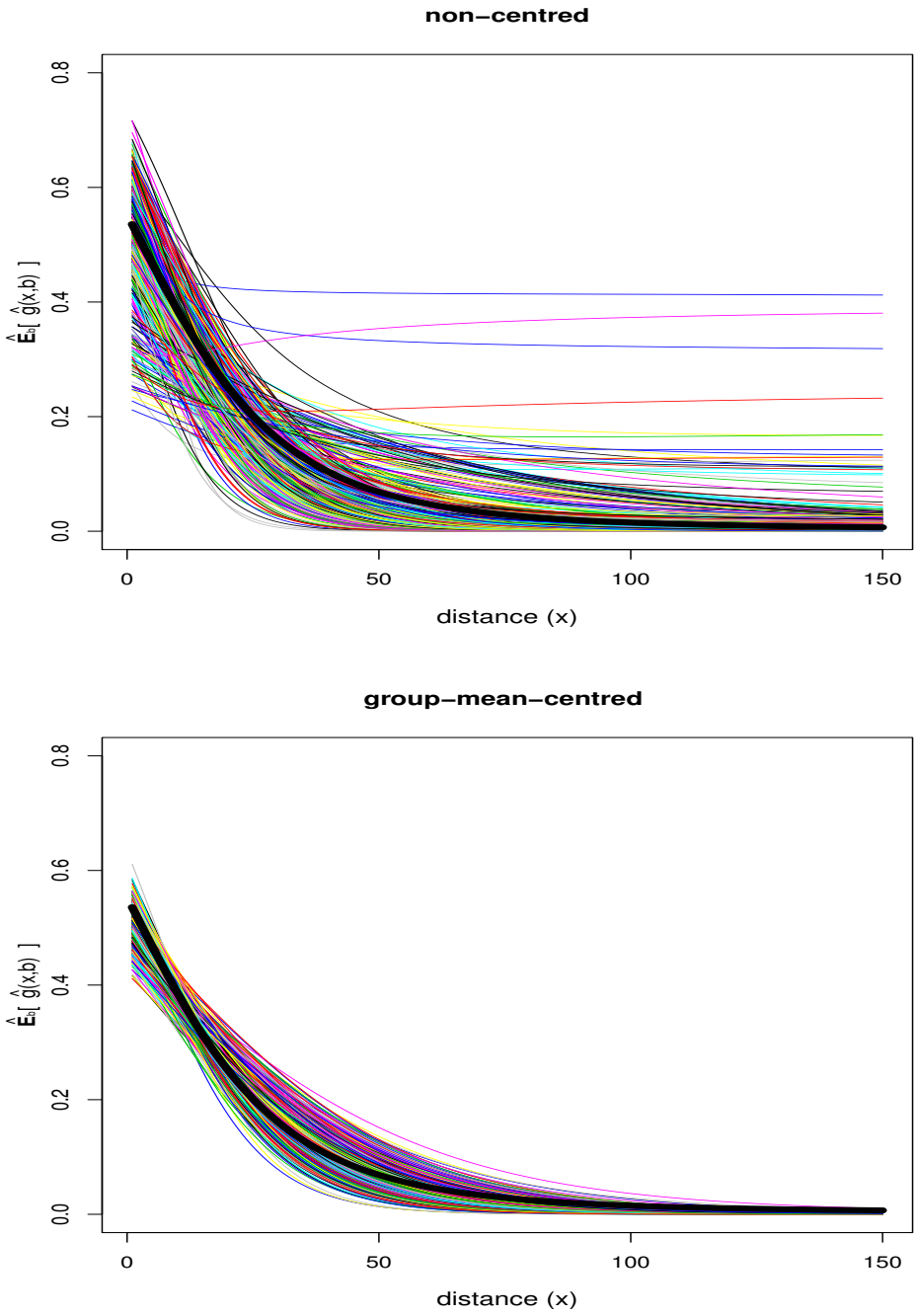


FIGURE 7.2. The plots of the estimated non-centred mixed-effects logistic regression and group-mean centred mixed-effects logistic regression model for a simulation study of the woodrats experimental survey. Each estimated model is given in a different colour and the assumed true detection function (7.1) is given by a thick black line in each plot.

The simulation study shown in Figure 7.2 is to examine the effect of centring x in estimating detectability using the woodrats experimental survey data. The data are simulated with the detection function given by (7.1) and the third protocol for the distances of the traps set up for each individual woodrat. Then both a non-centred model

$$\text{logit}[g(x, \boldsymbol{b}_i)] = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\, x, \tag{7.2}$$

and a group-mean centred model

$$\text{logit}[g(x, \boldsymbol{b}_i)] = (\beta_0 + b_{0i}) + \beta_0^* \bar{x}_i + (\beta_1^* + b_{1i})\, (x - \bar{x}_i), \tag{7.3}$$

are fitted to the same simulated experimental survey data, which are plotted with colour in Figure 7.2. It is important to note that when fitting the non-centred model given by (7.2), correlation between b_{0i} and b_{1i} is allowed; while when fitting the group-centred model given by (7.3), independence between b_{0i} and b_{1i} is assumed as the group-mean centring approach eliminates the correlation between b_{0i} and b_{1i} (Kreft *et al.*, 1995).

For plausible models of detection probability the detection function has decayed to zero by 150 m. This means that the non-centred mixed-effects model, whose estimation results are given in the top plot of Figure 7.2, cannot be used to estimate detectability for woodrats. The non-zero asymptote at large distances in the fitted non-centred model results from the correlation between b_0 and b_1. However, as the group-mean centring strategy eliminates the correlation between b_0 and b_1, when estimating the detectability for woodrats, the distances in the experimental survey data are centred about the average of all distances within woodrats for a plausible detection function.

A group-mean centred two-level random-effects model of the form (7.3) is then fitted to estimate the woodrats' detectability using the pooled experimental survey data, with individual woodrat as the random effect. Model selection is conducted according to the steps given in Section 3.3.4, and given the random effects $\boldsymbol{b}_i$ for the ith individual woodrat, the final selected model for estimating detectability is

$$\hat{g}(x,\, \bar{x}_i |\, \boldsymbol{b}_i) = \text{logit}^{-1}[b_{0i} - 0.068\bar{x}_i + (-0.053 + b_{1i})(x - \bar{x}_i)], \tag{7.4}$$

where b_{0i} and b_{1i} are independent and normally distributed with mean 0 and standard deviation 0.568 and 0.012 respectively.

7.1.2 Abundance estimation

Conditional on the estimated detection function given by (7.4), the inclusion probability in the abundance survey for a woodrat at distance x from the trapping point in the abundance survey is estimated by:

$$\hat{p}\,(x, \bar{x}_{\text{trial}} \,|\, b_{0i}, b_{1i}) \;\; = \;\; \hat{g}(x, \bar{x}_{\text{trial}} \,|\, b_{0i}, b_{1i})\, A_c / A_s, \tag{7.5}$$

where $A_c \;=\; 3\,K\,2\,\pi\,\int_0^w r\,q(r)\,dr$ is the area of the survey region falling within circles of diameter $w \;=\; 0.15$ km about K trapping points (each point was used three times), $A_s = 8.94$ km^2 is the area of the survey region, and $\bar{x}_{\text{trial}}$ is the average distance of the experimental survey trials.

Noting that neither distance x nor the random effects b_{0i} and b_{1i} are observed, the HT-like estimators for the woodrat survey are

$$\widehat{N}^{(1)} \;\; = \;\; \frac{n}{\hat{p}\,(\mathbf{E}[x],\, \bar{x}_{\text{trial}} \,|\, b_{0i} = 0, b_{1i} = 0)}, \tag{7.6}$$

$$\widehat{N}^{(2)} \;\; = \;\; \frac{n}{\mathbf{E}_x \left\{ \widehat{\mathbf{E}}_b \left[\hat{p}\,(x,\, \bar{x}_{\text{trial}} \,|\, b_{0i}, b_{1i}) \right] \right\}}, \tag{7.7}$$

$$\widehat{N}^{(3)} \;\; = \;\; \mathbf{E}_x \left\{ \widehat{\mathbf{E}}_b \left[\frac{1}{\hat{p}\,(x,\, \bar{x}_{\text{trial}} \,|\, b_{0i}, b_{1i})} \right] \right\}. \tag{7.8}$$

Expectations above are with respect to radial distance x and the random effects (b_{0i}, b_{1i}). In common with ordinary point transect surveys, the pdf of radial distances x of detected and undetected animals from points is assumed to be triangular, i.e., $\pi(x) = 2x/w^2$. The transect points, i.e., the trapping locations in the woodrats abundance survey, are randomly placed in the sampled area. In addition, because some parts of the edges of circles about trapping points fell outside the survey region (hence in habitat in which woodrats do not occur), it is important to exclude these parts from consideration. This exclusion is implemented by using a so-called 'edge effect function', $q(x)$, which gives the average proportion of the circumference of a circle of radius x centred on trapping points that lies in the survey region. This

function is specified as

$$q(x) = \text{logit}^{-1}(2.984 - 0.009\,x);$$

see Potts (2011) p. 155 for details. The specification of $q(x)$ leads to the following expression for the probability density function of x when combined with edge effect in the woodrats abundance survey:

$$\pi(x) = x\,q(x) \Big/ \int_0^w x\,q(x)\,dx. \tag{7.9}$$

Given the pdf of x with edge effect in (7.9) and $w = 150$ m, it can be shown that $\mathbf{E}[x] = 98.8$ m in (7.6). Distances and random-effects parameters are assumed to be independent. Integration with respect to x with pdf given in (7.9) is numerical, while details of integration with respect to (b_{0i}, b_{1i}) for $\widehat{N}^{(2)}$ and $\widehat{N}^{(3)}$ are based on (6.24) and (6.23) respectively. The numerical integration with respect to x is implemented with a 1000 equally spaced grid from 0 to w in estimators (7.7) and (7.8). The numerical integration of (b_{0i}, b_{1i}) in $\widehat{N}^{(2)}$ is based on a 1000 grid of equally spaced points which cover the interval within 5 standard errors from the mean of a linear combination of b_{0i} and b_{1i}, i.e., the u_{ij} in (6.24).

7.1.3 Results and discussion

Point estimates of abundance are given in Table 7.1, together with 95% CIs, which were estimated by bootstrapping with trapping point as the sampling unit and using 999 re-samples. A simulation study was also conducted with 999 simulations of both the experimental and abundance survey to examine the performance of (7.6)–(7.8). The experimental survey was simulated with 50 woodrats and a balanced survey design with distances of each trial given by the third protocol in Section 7.1.1. The model assumed for the detection probability is the two-level group-mean-centered logistic model fitted to the 2008-2009 experimental survey data, which was given by (7.4). Figure 7.3 plots the estimated detection functions for the first 28 individual woodrats assumed in the simulation study of the experimental survey. These plots show that the maximum detection probability is no larger than 0.5 and there is substaintial heterogeneity in detectability among different individual woodrats.

For each simulation of the abundance survey, the same w and A_c as for the 2008 abundance survey were used, and the population size was assumed to be $N = 411$. The results are given in Table 7.2. Unlike the results of the anglerfish abundance estimation given in Section 5.2, there are clear differences in the performances of the estimators. $\widehat{N}^{(2)}$ performs best, much better than both $\widehat{N}^{(1)}$ and $\widehat{N}^{(3)}$ (Table 7.1). $\widehat{N}^{(1)}$ is much higher than $\widehat{N}^{(2)}$ and has very high positive bias and MSE (Table 7.2). $\widehat{N}^{(3)}$ is extremely biased and performs worst.

TABLE 7.1. Woodrat abundance estimates with 95% CIs using the estimators given by (7.6), (7.7) and (7.8). The distribution of distance is given by (7.9) and the inclusion probability given by (7.5) conditional on the estimated detection function given by (7.4).

	Point estimate	95% CI	
Estimator	$\widehat{N}$	Lower	Upper
$\widehat{N}^{(1)}$	$2,947$	373	$81,077$
$\widehat{N}^{(2)}$	411	115	973
$\widehat{N}^{(3)}$	$17,010$	982	3.75×10^6

TABLE 7.2. Simulation results for woodrat survey: %bias and %MSE for the estimators (7.6), (7.7) and (7.8) are calculated according to (5.2) and (5.3) respectively, and their empirical standard deviations are given in brackets.

Estimator	%bias (sd)	%MSE (sd)
$\widehat{N}^{(1)}$	$1,137\,(61)$	$50,890\,(11,067)$
$\widehat{N}^{(2)}$	$5.01\,(1.77)$	$31.54\,(1.89)$
$\widehat{N}^{(3)}$	$9,835\,(960)$	$10.1\,(3.56) \times 10^6$

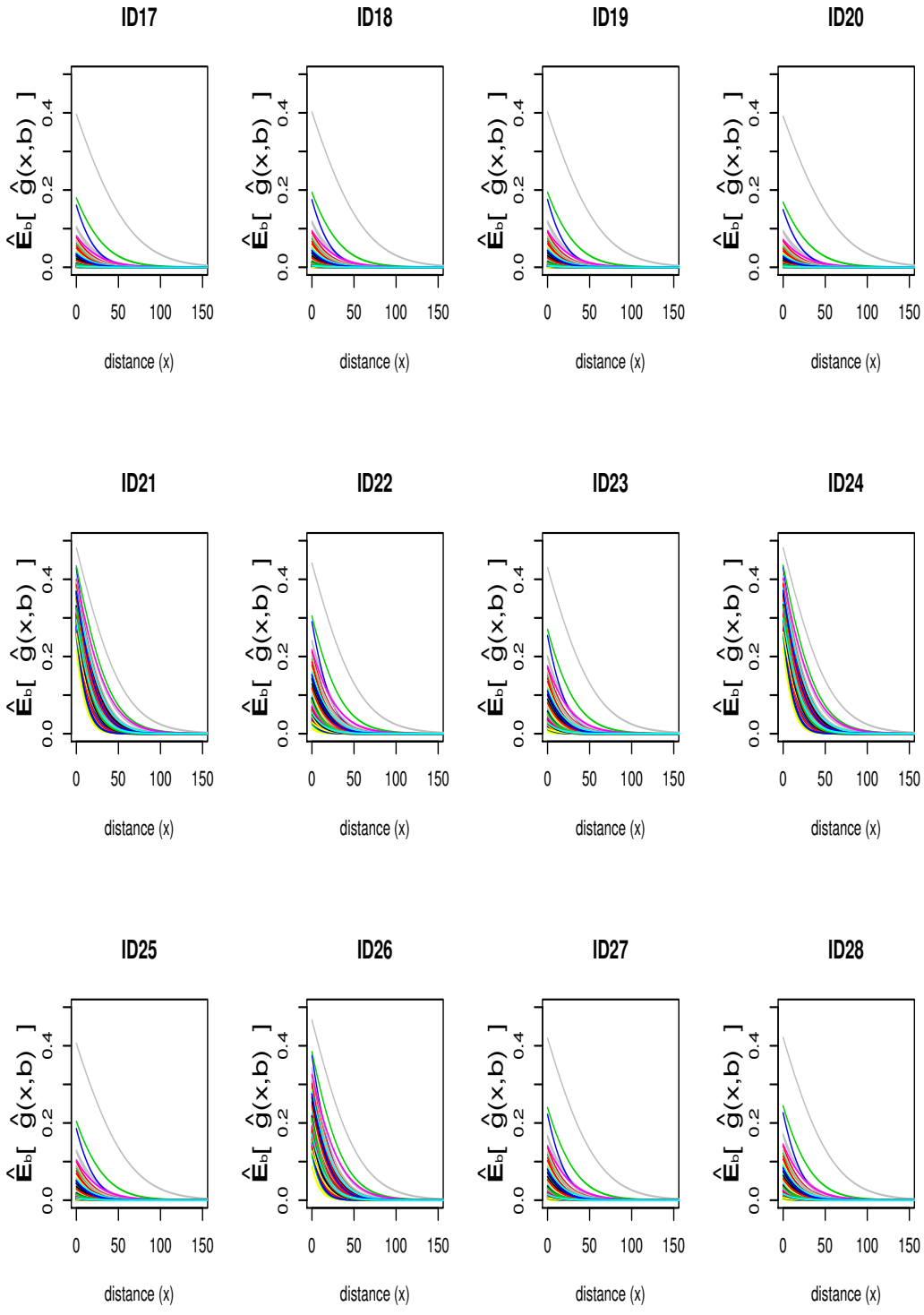


FIGURE 7.3. Plots of the first 100 simulated detection functions for 28 individual woodrats in the simulation study. Each curve in the plots has the random intercept and slope integrated out with respect to their estimated distribution. The assumed detection function in the simulation study is given by (7.4). The 'ID' on top of each plot provides the identity of each individual woodrat in the simulation study.

7.2 Line transect: Dall's porpoise

In the woodrats application described in the previous section, the distance x from the trap was treated as another layer of random effects in addition to the random effect for individual heterogeneity. Note that $\widehat{N}^{(2)}$ in (7.7) has the form of the conventional line transect estimator of abundance in the survey region, given by Buckland *et al.* (2001, p. 37), but with the addition of random effects for individual heterogeneity. The random effects $\boldsymbol{b}_i$ for individual heterogeneity are not observable. However, unlike the woodrats application on line transect and point transect surveys, the random effect x can be observed.

The line transect and point transect surveys are the most widely used of a group of methods for abundance estimation collectively known as distance sampling. A feature of these methods is that although the main explanatory variable, distance x, is observed for all detected animals, it is treated as a random variable with known probability distribution. This leads to the consideration of using the HT-like estimators for a conventional distance sampling survey. The distance, x, is treated as a random effect and it is observed for each animal. In this case, $\widehat{N}^{(0)}$ given by (6.1) is available but with unknown p. As detection probability depends on distance, randomness in x generates randomness in detection probability and any of the estimators $\widehat{N}^{(0)}$ to $\widehat{N}^{(3)}$ is applicable, although $\widehat{N}^{(2)}$ is the one conventionally used.

Line transect surveys involve traversing lines and searching for animals within a strip of half-width w about the lines. Let n be the number of animals detected within the strips, L be the total length of all lines, and P_a be the probability of detecting an animal in the searched strips. Providing lines are randomly located with respect to animals, it can be assumed that animals are uniformly distributed with respect to distance from the line, i.e. the pdf of x is $\pi(x) = 1/w$. Note that x is the distance from the line, irrespective of which side of the line it is, giving a folded distribution on $(0, w)$.

The conventional line transect estimator of abundance in the survey region is $\widehat{N} = A_s\, n/(2wL\widehat{P}_a)$, given by Buckland *et al.* (2001, p. 37), where A_s is the area of the survey region, $\widehat{P}_a = \mathbf{E}_x[\hat{g}(x)] = \int_0^w \hat{g}(x)\pi(x)dx$, and $g(x)$ is the detection function, i.e., the probability of detecting an animal at distance x. The detection function, and hence P_a, is usually estimated by maximizing the conditional likelihood function, i.e., the likelihood that is conditional on the captured animals.

Given n detected distances and a half-normal detection function

$$g(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right),$$

the conditional likelihood function is

$$L_c(\sigma \mid x_1, \ldots, x_n) = \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}\right) \Big/ \left\{\int_0^w g(x)dx\right\}^n, \tag{7.10}$$

and $\hat{\sigma}$ is the MLE of σ obtained by maximizing the log of (7.10). The detection function estimated from the Dall's porpoise line transect data (described below) is plotted in Figure 7.4.

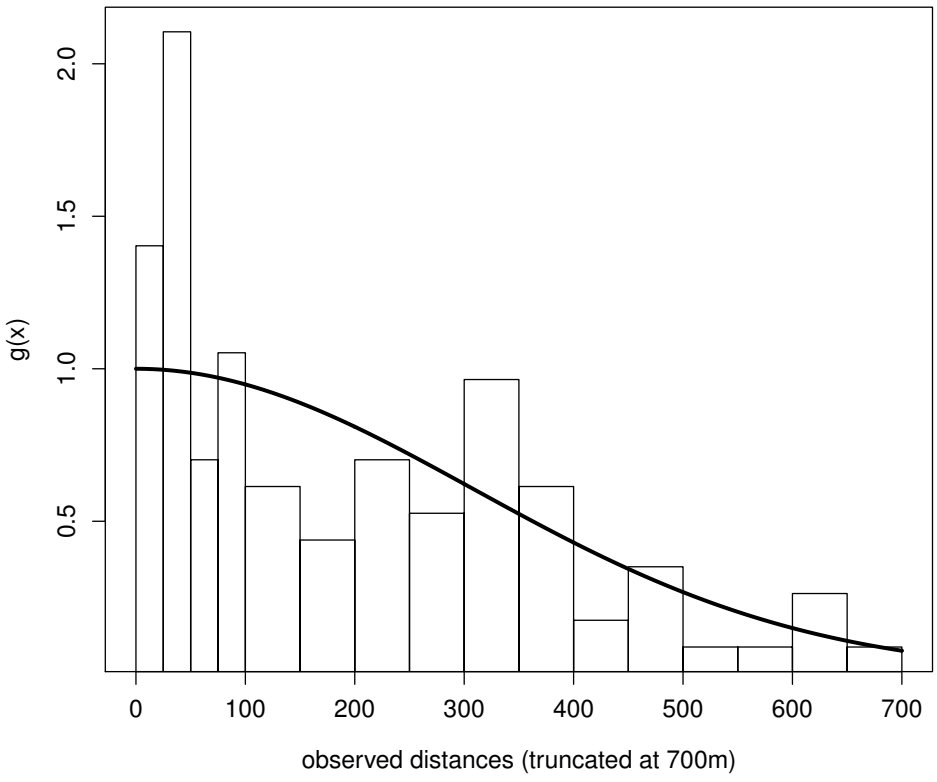


FIGURE 7.4. Plot of estimated detection function $\hat{g}(x)$ for the Dall's Porpoise data in stratum 1 (2004) with the histogram of observed distances.

Our four estimators are as follows in the case of a line transect survey:

$$\widehat{N}^{(0)} = \sum_{i=1}^{n} \frac{1}{\hat{p}(x_i)} = \frac{n\,A_s}{2wL} \sum_{i=1}^{n} \frac{1}{\hat{g}(x_i)}, \tag{7.11}$$

$$\widehat{N}^{(1)} = \sum_{i=1}^{n} \frac{1}{\hat{p}(\mu_x)} = \frac{n\,A_s}{2wL\,\hat{g}(\mu_x)}, \tag{7.12}$$

$$\widehat{N}^{(2)} = \sum_{i=1}^{n} \frac{1}{\mathbf{E}_x[\hat{p}(x)]} = \frac{n\,A_s}{2wL\,\mathbf{E}_x[\hat{g}(x)]}, \tag{7.13}$$

$$\widehat{N}^{(3)} = \sum_{i=1}^{n} \mathbf{E}_x\left[\frac{1}{\hat{p}(x)}\right] = \frac{n\,A_s}{2wL}\,\mathbf{E}_x\left[\frac{1}{\hat{g}(x)}\right]. \tag{7.14}$$

These estimators are applied to Dall's porpoise (*Phocoenoides dalli*) data from a line transect survey in British Columbia that is reported in Williams & Thomas (2007). The data used here are from stratum 1 of the 2004 survey, which has 35 transect lines placed according to a systematic zigzag design with random start, and a strip of 700 meters on each side of a transect line. Dall's porpoise occur in groups and estimation is performed for group abundance here. A total of $n = 57$ groups was detected on transects of total length $L = 963$ nautical miles in a survey region of $A_s = 18,360$ square nautical miles. A half-normal detection function form is chosen for $g(x)$, after selecting models on the basis AIC; see Figure 7.4 for the plot of the estimated detection function. Group abundance estimates using the estimators (7.11)–(7.14) are presented in Table 7.3, together with 95% CIs, estimated by bootstrapping non-parametrically with transect as the sampling unit, with 999 re-samples.

A simulation study was also conducted with 999 simulations of the survey, using the same w and L as in the 2004 survey, the detection function is assumed to be the one obtained by maximizing the conditional likelihood function plotted in Figure 7.4. The number of groups in the population N is assumed to be $2,386$. The results are given in Table 7.4. The results show that $\widehat{N}^{(2)}$ performs best, with $\widehat{N}^{(0)}$ a surprisingly close second. We found its performance surprising, because it does not use the information that $\pi(x) = 1/w$, and in our experience this can lead to substantially worse performance when some $\hat{g}(x_i)$, $i = 1, \ldots, n$, are small. In this survey, the smallest $\hat{g}(x_i)$ is equal to 0.078, which is not very small and this may explain the reasonable performance of $\widehat{N}^{(0)}$. The bias of $\widehat{N}^{(1)}$ is in the opposite direction to that in the case of the woodrats data. As was the case with the woodrats data, $\widehat{N}^{(3)}$ performs worst by some way.

TABLE 7.3. Dall's porpoise abundance estimates with 95% CIs using the estimators given by (7.11)–(7.14).

Estimator	Point estimate $\widehat{N}$	95% CI Lower	Upper
$\widehat{N}^{(0)}$	$2,039$	961	$3,749$
$\widehat{N}^{(1)}$	$2,452$	$1,122$	$4,473$
$\widehat{N}^{(2)}$	$2,386$	$1,136$	$4,151$
$\widehat{N}^{(3)}$	$4,212$	$1,727$	$10,893$

TABLE 7.4. Simulation results for Dall's porpoise survey: %bias and %MSE for the estimators (7.11)–(7.14) are calculated according to (5.2) and (5.3) respectively, and their empirical standard deviations are given in brackets.

Estimator	%bias (sd)	%MSE (sd)
$\widehat{N}^{(0)}$	$2.30\,(0.61)$	$3.71\,(0.38)$
$\widehat{N}^{(1)}$	$9.43\,(0.82)$	$7.67\,(0.61)$
$\widehat{N}^{(2)}$	$2.23\,(0.51)$	$2.69\,(0.12)$
$\widehat{N}^{(3)}$	$136\,(6.28)$	$579\,(113)$

7.3 Mark-recapture: wood mouse data

In this section we conclude the investigation of HT-like estimators with random effects in ecology by considering mark-recapture methods – where models with random effects are quite common. The mark-recapture literature includes many contributions on what are commonly called 'M_h' models ('h' is for heterogeneity), in which detection probability is treated as a random variable. Dorazio & Royle (2003) contains an up-to-date overview of such models. A key difference between most of

the likelihood-based estimators in the mark-recapture literature and those considered here is that, like the estimators of Huggins (1989) and Alho (1990b), the estimators using the HT-like method are based on a conditional likelihood model, conditioning on capture. M_h estimators in the mark-recapture literature that model randomness in detection probability due to unobserved random variables are based on unconditional likelihood functions.

An advantage that the estimators proposed here have over estimators based on unconditional likelihood functions is that the probability density functions of observed covariates that affect detection probability do not need to be modelled. When there are few such covariates and one has reasonable *a priori* models for the distribution of these covariates, this is much less of an advantage than when there are many covariates and one has little or no *a priori* knowledge of the form of the distribution of these covariates. On the other hand, the bias of conditional-likelihood estimators is bigger than that of the unconditional-likelihood estimators (Fewster & Jupp, 2009). Maximum likelihood estimators based on the conditional and unconditional likelihood functions are compared in Fewster & Jupp (2009), although they do not consider estimators with random coefficients.

The wood mouse (*Apodemus sylvaticus*) dataset of Tanton (1965) is used here to illustrate our estimators in a mark-recapture context. The data, which do not contain any covariates, consist of 334 individuals captured at least once over 21 capture occasions. Morgan & Ridout (2008) re-analysed the wood mouse data, and investigated a number of estimators based on unconditional likelihoods. They concluded that a beta-binomial mixture model for the distribution of detection probabilities was best among the models considered. This likelihood is given by

$$\begin{aligned} & L(N,\alpha,\beta \mid n_1, n_2, \ldots, n_K) \\ = \; & \frac{N!}{\prod_{k=0}^{K} n_k!} \prod_{k=0}^{K} \left\{ \binom{K}{k} \frac{B(k+\alpha, K-k+\beta)}{B(\alpha,\beta)} \right\}^{n_k} \\ = \; & \frac{N!}{(N-n)!} \left\{ \frac{B(\alpha, K-\beta)}{B(\alpha,\beta)} \right\}^{N-n} \prod_{k=1}^{K} \frac{\left[\binom{K}{k} \frac{B(k+\alpha,K-k+\beta)}{B(\alpha,\beta)} \right]^{n_k}}{n_k!}. \end{aligned} \tag{7.15}$$

Here n_k is the number of individual animals that were captured exactly k times out of the $K = 21$ occasions, the total number of captured individual animals is $n = n_1 + \ldots + n_K$, and $B(\alpha,\, \beta)$ denotes the beta function with parameters α and β. Let

ϕ denote the individual-specific capture probability on a single sampling occasion, which is treated as random effect here. The likelihood is then obtained by modelling ϕ as independent Beta(α, β) variables.

The probability that an animal is captured k times given that it has been captured at least once, $p_{k|c}(\alpha, \beta)$, can be shown to be

$$p_{k|c}(\alpha, \beta) = \binom{K}{k} \frac{B(\alpha + k, \beta + K - k)}{[B(\alpha, \beta) - B(\alpha, \beta + K)]}; \tag{7.16}$$

see (7.28) in Appendix 7.A.

It follows that the conditional likelihood for (α, β), given capture, is

$$\begin{aligned} L_c(\alpha, \beta \mid K, n_1, n_2, \ldots, n_K) &= \frac{n!}{\prod_{k=1}^{K} n_k!} \prod_{k=1}^{K} \{p_{k|c}(\alpha, \beta)\}^{n_k} \\ = \quad \frac{n!}{\prod_{k=1}^{K} n_k!} &\left\{ \binom{K}{k} \frac{B(\alpha + k, \beta + K - k)}{[B(\alpha, \beta) - B(\alpha, \beta + K)]} \right\}^{n_k}. \end{aligned} \tag{7.17}$$

Conditional maximum likelihood estimates, $\hat{\alpha} = 0.525$ and $\hat{\beta} = 3.010$, are obtained by maximizing the likelihood given by (7.17). For any given ϕ, the inclusion probability used in the HT-like estimators is the probability that an individual animal is captured at least once on K occasions, which is denoted by $p_{\cdot}(\phi)$, and $p_{\cdot}(\phi) = 1 - (1 - \phi)^K$.

Because ϕ is not observed, estimator $\widehat{N}^{(0)}$ cannot be used, but given the conditional maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$, the other estimators are as follows:

$$\widehat{N}^{(1)} = \frac{n}{1 - \left(1 - \frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}}\right)^K}, \tag{7.18}$$

$$\widehat{N}^{(2)} = \frac{n}{1 - \frac{B(\hat{\alpha}, \hat{\beta}+K)}{B(\hat{\alpha}, \hat{\beta})}}, \tag{7.19}$$

$$\widehat{N}^{(3)} = \frac{n}{B(\hat{\alpha},\, \hat{\beta})} \int_0^1 \frac{\phi^{\hat{\alpha}-1}(1-\phi)^{\hat{\beta}-1}}{[1 - (1-\phi)^K]}\, d\,\phi\,; \tag{7.20}$$

a more detailed derivation is given in Appendix 7.A.

Interval estimates of N were obtained using a non-parametric bootstrap with 999 re-samples, re-sampling capture histories (Buckland & Garthwaite, 1991). The point

estimates for each estimator are shown in Table 7.5, together with 95% CIs. The point estimate for $\widehat{N}^{(2)}$ agrees well with the estimator $\widehat{N}^{(full)}$, obtained by maximizing the unconditional likelihood with respect to N, α and β, whereas those for $\widehat{N}^{(1)}$ and $\widehat{N}^{(3)}$ do not. The confidence interval for N for estimator $\widehat{N}^{(2)}$ is substantially wider than that for $\widehat{N}^{(full)}$. This is not unexpected as it is consistent with the findings of Fewster & Jupp (2009), who also showed that log $\widehat{N}^{(2)}$ and log $\widehat{N}^{(full)}$ have the same distributions to order $O(N^{-1/2})$.

To investigate the properties of the estimators for a scenario similar to that of the wood mice, a simulation study is also conducted with 999 simulations of the survey. The simulation assumes that the true population size is 500, there are 21 sampling occasions, and detection probability has a Beta($\hat{\alpha}, \hat{\beta}$) distribution, where $\hat{\alpha}$ and $\hat{\beta}$ are the conditional maximum likelihood estimates from the wood mice analysis. The survey was simulated 999 times. The percentage bias (%bias) and percentage MSE (%MSE) are shown in Table 7.6. $\widehat{N}^{(2)}$ performs best. Its bias and MSE are very similar to those of $\widehat{N}^{(full)}$.

TABLE 7.5. Wood mouse abundance estimates with 95% CIs using the HT-like estimators given by (7.18), (7.19) and (7.20), together with the $\widehat{N}^{(full)}$ obtained by maximizing the full likelihood given by (7.15).

	Point estimate	95% CI	
Estimator	$\widehat{N}$	Lower	Upper
$\widehat{N}^{(1)}$	346	289	360
$\widehat{N}^{(2)}$	494	387	765
$\widehat{N}^{(3)}$	$2,757$	121	$18,525$
$\widehat{N}^{(full)}$	489	462	521

TABLE 7.6. Simulated bias and MSE for wood mouse survey: %bias and %MSE for the estimators (7.18), (7.19), (7.20), and $\widehat{N}^{(full)}$ are calculated according to (5.2) and (5.3) respectively, and their empirical standard deviations are given in brackets.

Estimator	%bias (sd)	%MSE(sd)
$\widehat{N}^{(1)}$	$-29.80\,(0.08)$	$8.95\,(0.05)$
$\widehat{N}^{(2)}$	$1.35\,(0.39)$	$1.51\,(0.26)$
$\widehat{N}^{(3)}$	$222.7\,(18.6)$	$3,971\,(2,059)$
$\widehat{N}^{(full)}$	$0.16\,(0.37)$	$1.34\,(0.21)$

7.4 Discussion

For simplicity, the anglerfish survey data is referred to as case a1, the woodrats trapping point survey data as case a2, the Dall's porpoise line transect data as case b, and the wood mouse mark-recapture data as case c.

In Sections 3.4, 7.1.1, 7.2 and 7.3, three different methods were used to model randomness in detection probability:

- method 1: model coefficients of the detection probability function as random with unknown distribution parameters,
- method 2: model detection probability as a function of predictor variables that are treated as random with a known distribution,
- method 3: model detection probability itself as a random effect with unknown distribution parameters.

The inclusion probability in $\widehat{N}^{(m)}$ $(m = 0, 1, 2, 3)$ can be modelled by any one of the above three methods, or a combination of them. In all cases it is found that $\widehat{N}^{(2)}$ performed best, with the least bias and lowest MSE, while $\widehat{N}^{(3)}$ was found to

be consistently the worst estimator. Then the performance of $\widehat{N}^{(2)}$ is considered in relation to the distribution of inclusion probabilities for all four cases.

Let p denote the inclusion probability in the sampled area with mean μ_p and standard deviation σ_p. We give the pdfs of p for all cases in (7.21)–(7.23); the corresponding μ_p and σ_p are given in Table 7.7. (7.21) gives the conditional pdf of p given the model matrices for the case when a logistic mixed-effect regression model is used to estimate the capture probability, such as the anglerfish and woodrats applications. This case is then referred to as case a. Expressions for the pdfs of p for cases b and c are given by (7.22) and (7.23) respectively:

$$f^{(logit)}_{p|(\boldsymbol{x},\boldsymbol{z})}(p|\boldsymbol{x},\boldsymbol{z}) = \frac{1}{\sqrt{2\pi \boldsymbol{z}^T \Sigma_b \boldsymbol{z}}} \frac{1}{p(1-p)} \exp\left\{-\frac{[\text{logit}(p) - \boldsymbol{x}^T\boldsymbol{\beta}]^2}{2\boldsymbol{z}^T \Sigma_b \boldsymbol{z}}\right\}, \tag{7.21}$$

where $p(\boldsymbol{b}|\boldsymbol{x},\boldsymbol{z}) = \text{logit}^{-1}(\boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{z}^T\boldsymbol{b})$ and $\boldsymbol{b} \sim N(\mathbf{0},\, \Sigma_b)$.

$$f^{(b)}_p(p) = \frac{\sigma}{w\, p\sqrt{-2\log(p)}}, \tag{7.22}$$

where $p(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right)$ and $f_x(x) = 1/w$.

$$f^{(a)}_p(p) = \frac{\left[1-(1-p)^{1/K}\right]^{\alpha-1}(1-p)^{\beta/K-1}}{K\, B(\alpha,\beta)}, \tag{7.23}$$

where $p = 1-(1-\phi)^K$ and $\phi \sim \text{Beta}(\alpha,\beta)$.

The pdfs are shown graphically in Figure 7.5. Salient features of the pdfs are given in Table 7.7, together with the corresponding %bias and %MSE of $\widehat{N}^{(m)}$, $m = 1, 2, 3$, from the simulation studies. The pdfs of cases a2, b and c have a similar shape to a beta distribution, while the pdfs in case a1 are similar to normal distributions. For case a1, the plot shows that when the length of fish gets bigger, μ_p increases and σ_p decreases, giving smaller cv_p at greater lengths.

It is apparent from Table 7.7 that poor estimator performance is associated with (i) low mean detection probability μ_p and (ii) high cv_p. Note that in case a1 the model is fitted to all lengths, although abundance estimates are shown only for a few selected lengths.

TABLE 7.7. Features of the inclusion probability pdfs assumed in simulation study, together with %bias and %MSE of $\widehat{N}^{(2)}$ for cases a1, a2, b and c.

	a1: anglerfish[*1] , given length (l)				a2: woodrats[*2]	b: DP[*3]	c: wood mouse
	12 cm	30 cm	40 cm	60 cm			
μ_p	0.179	0.604	0.819	0.976	0.015	0.551	0.682
σ_p	0.034	0.068	0.043	0.007	0.042	0.293	0.354
$cv_p(\sigma_p/\mu_p)$	0.237	0.112	0.052	0.007	2.8	0.532	0.519
N[*4]	2	26	18	11	411	2386	500
%bias of $\widehat{N}^{(1)}$	2.56	-0.570	-0.094	-0.145	1,137	9.43	-29.80
(sd)	(5.69)	(0.608)	(0.405)	(0.176)	(61)	(0.82)	(0.08)
%MSE of $\widehat{N}^{(1)}$	323.9	3.696	1.641	0.311	50,890	7.67	8.95
(sd)	(19.29)	(0.179)	(0.077)	(0.017)	(11,067)	(0.61)	(0.05)
%bias of $\widehat{N}^{(2)}$	2.10	-0.551	-0.029	-0.124	5.01	2.23	1.35
(sd)	(5.67)	(0.608)	(0.405)	(0.176)	(1.77)	(0.51)	(0.39)
%MSE of $\widehat{N}^{(2)}$	321.1	3.696	1.642	0.311	31.54	2.69	1.51
(sd)	(19.16)	(0.179)	(0.077)	(0.017)	(1.89)	(0.12)	(0.26)
%bias of $\widehat{N}^{(3)}$	3.27	-0.55	0.028	-0.122	9,835	136	222.7
(sd)	(5.73)	(0.61)	(0.406)	(0.176)	(960)	(6.28)	(18.6)
%MSE of $\widehat{N}^{(3)}$	328.3	3.70	1.643	0.311	10.1×10^6	579	3,971
(sd)	(19.51)	(0.18)	(0.076)	(0.017)	(3.56×10^6)	(113)	(2,509)

[*1]For case a1, μ_p, σ_p and cv_p are calculated with the average length equal to the average length in the experimental survey data (49 m). Because the average length varied across simulations, the %bias and %MSE do not correspond exactly to that from a pdf with μ_p and σ_p given above.

[*2]For case a2, μ_p, σ_p and cv_p are based on the marginal density of p, integrating out the unobserved distance x numerically with respect to its triangular distribution.

[*3]DP is short for Dall's porpoise.

[*4]N is the simulated abundance.

In all cases of Table 7.7, we find that $\widehat{N}^{(2)}$ has the smallest %MSE and small bias even in the presence of very small mean inclusion probabilities and high cv_p. $\widehat{N}^{(3)}$ performs very poorly for cases a2, b and c, except when μ_p is large and cv_p very small in case a1, in which case little difference between the estimators' performance can be seen (see Figure 5.4 for all length classes). But in this case it may not be necessary to use random effects models for detection probability at all. $\widehat{N}^{(1)}$ performs worse as cv_p increases and as the proportion of small detection probabilities in the population increases; compare cases b and c with reference to Figure 7.5 and Table 7.7. We therefore recommend the use of $\widehat{N}^{(2)}$ when using a HT-like estimator in conjunction with a conditional likelihood approach in the presence of estimated inclusion probabilities with random effects.

Alternative approaches to that used above include full-likelihood approaches with maximum likelihood or Bayesian estimation of abundance. These require models for all explanatory variables included in the detection function, and the main reason for choosing a conditional-likelihood approach and HT-like estimator over a full-likelihood approach for abundance estimation is to avoid having to model explanatory variable probability density functions in the presence of limited knowledge of their form, a problem that is exacerbated if there are many explanatory variables.

Finally, as noted in Chapter 6, the HT-like estimators developed for incorporating random effects have utility not only in ecology, but in social science applications too. Similar estimators have been developed independently in these two fields in the past. A recent example is the estimator of the population mean developed by Volz & Heckathorn (2008), (see equation (4) of Gile & Handcock, 2010), which can be viewed as a special case of the estimator in equations (16) and (20) of Borchers *et al.* (1998). It would no doubt benefit both ecology and social science if there was rapid transfer of developments in methodology between them.

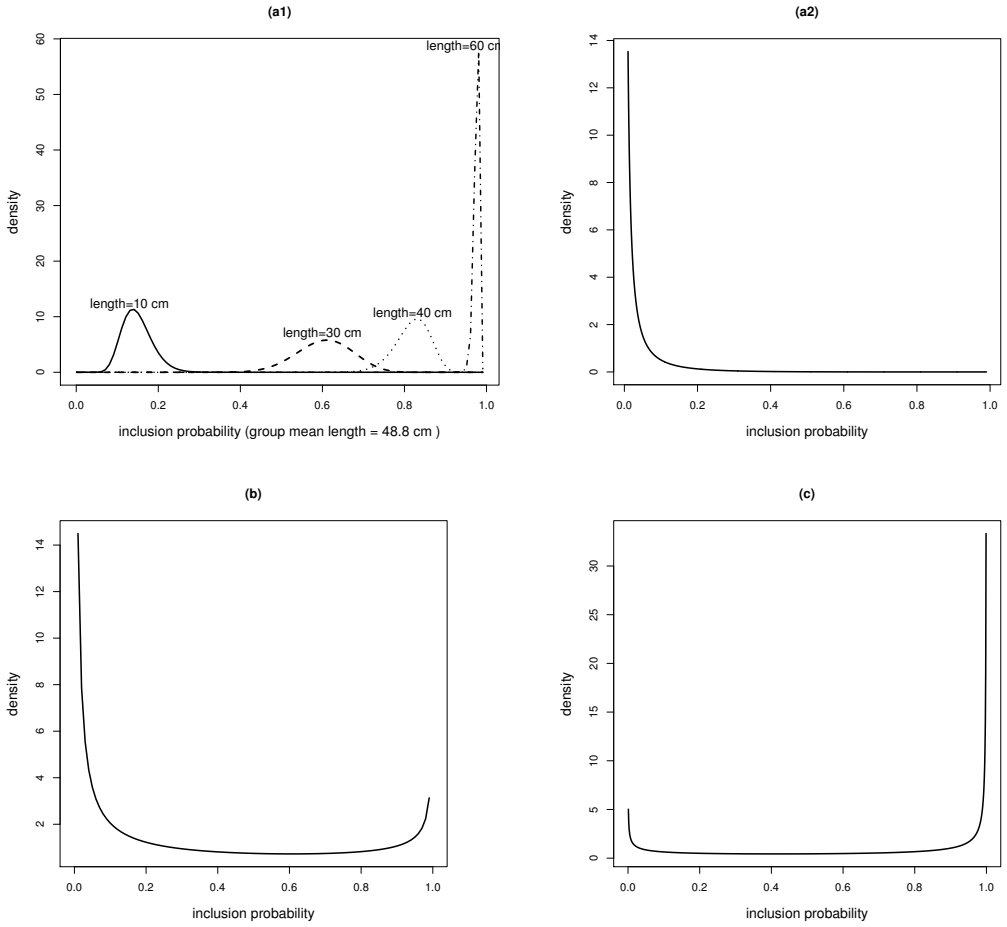


FIGURE 7.5. Density plots of detection probability for all four datasets: (a1) the density plot of the detection probability in anglerfish data set, in which the estimated detection function is $\text{logit}^{-1}(-3.606+b_0+0.125\bar{l}+0.110\times l)$ with $\bar{l}=48.8$ cm, which is the grand mean length of all fish captured in the anglerfish survey, length (l) equal to 10, 30, 40 and 60 centimeters respectively, and random intercept (b_0) is normally distributed with mean 0 and standard deviation 0.289. (a2) the density plot of the detection probability in the woodrat dataset, in which the estimated detection function is $\text{logit}^{-1}[b_0 - 0.068\bar{x} + (-0.053 + b_1)(x - \bar{x})]$, assuming x has a triangular distribution, the grand mean distance in the experimental survey $\bar{x} = 53.05$ m, and both b_0 and b_1 are normally distributed with mean 0 and standard deviation estimated to be 0.568 and 0.012 respectively. (b) the density plot of detection probability in the Dall's porpoise dataset, which is modeled by a half-normal with $\hat{\sigma} = 308$ and $w = 700$ meters. (c) the detection probability in the plot is the probability that an individual wood mouse is captured at least once over 21 sampling occasions, and the individual heterogeneity is modeled by Beta(0.525, 3.010).

Appendices – **Chapter 7**

7.A Mark-recapture conditional likelihood with a beta distribution for capture probability

Let ϕ be the individual-specific capture probability and $C^{(K)}$ denote a random variable corresponding to the capture frequency for an individual over all K sampling occasions. The probability that an individual is captured k times out of K sampling occasions is

$$\Pr\{C^{(K)} = k \,|\, \phi\} = \binom{K}{k} \phi^k \, (1-\phi)^{K-k}. \tag{7.24}$$

Therefore, the inclusion probability is the probability that an individual is captured at least once:

$$\Pr\{C^{(K)} \geq 1 \,|\, \phi\} = 1 - P\{C^{(K)} = 0 \,|\, \phi\} = 1 - (1-\phi)^K. \tag{7.25}$$

In the beta-binomial full-likelihood and the HT-like estimators examined in the paper, we assumed a beta distribution with parameters α and β for ϕ. Let B denote the beta function. The probability density function of ϕ is

$$f_\phi(\phi \,|\, \alpha, \beta) = \phi^{\alpha-1} \, (1-\phi)^{\beta-1} / B(\alpha, \beta).$$

Based on (7.24) and (7.25), we have

$$\begin{aligned} \Pr\{C^{(K)} = k \,|\, \alpha,\, \beta\} &= \int_0^1 \Pr\{C^{(K)} = k \,|\, \phi\} \, f(\phi \,|\, \alpha, \beta) \, d\phi \\ &= \int_0^1 \binom{K}{k} \phi^k \, (1-\phi)^{K-k} \, \frac{\phi^{\alpha-1} \, (1-\phi)^{\beta-1}}{B(\alpha, \beta)} \, d\phi \\ &= \binom{K}{k} \frac{B(\alpha + k,\, \beta + K - k)}{B(\alpha, \beta)}. \end{aligned} \tag{7.26}$$

It follows that the inclusion probability given α and β is

$$\Pr\{C^{(K)} \geq 1 \,|\, \alpha,\, \beta\} = 1 - \Pr\{C^{(K)} = 0 \,|\, \alpha,\, \beta\} = 1 - \frac{B(\alpha,\, \beta + K)}{B(\alpha, \beta)}. \tag{7.27}$$

Therefore, the probability that an animal is captured k times given that it has been captured at least once, $p_{k|c}(\alpha, \beta)$, can be shown to be

$$p_{k|c}(\alpha,\, \beta) = \frac{\Pr\{C^{(K)} = k \,|\, \alpha,\, \beta\}}{\Pr\{C^{(K)} \geq 1 \,|\, \alpha,\, \beta\}} = \binom{K}{k} \frac{B(\alpha + k,\, \beta + K - k)}{[B(\alpha, \beta) - B(\alpha,\, \beta + K)]}. \tag{7.28}$$

Based on the inclusion probability conditional on the random effect ϕ given in (7.25), the three HT-like abundance estimators with $\phi \sim \text{Beta}(\hat{\alpha}, \hat{\beta})$ are as follows, where $\hat{\alpha}$ and $\hat{\beta}$ are the MLE of the conditional likelihood.

$$\begin{aligned}
\widehat{N}^{(1)} &= \frac{n}{\Pr\{C^{(K)} \geq 1 \,|\, \phi = \hat{\mu}_\phi\}} = \frac{n}{1 - \left(1 - \dfrac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}}\right)^K}, \\
\widehat{N}^{(2)} &= \frac{n}{\widehat{E}_\phi\left[\Pr\{C^{(K)} \geq 1 \,|\, \phi\}\right]} = \frac{n}{\int_0^1 \Pr\{C^{(K)} \geq 1 \,|\, \phi\}\, f_\phi(\phi \,|\, \hat{\alpha}, \hat{\beta})\, d\phi} \\
&= \frac{n}{\Pr\{C^{(K)} \geq 1 \,|\, \hat{\alpha}, \hat{\beta}\}} \quad \text{using (7.27)}, \\
&= \frac{n}{1 - \dfrac{B(\hat{\alpha}, \hat{\beta} + K)}{B(\hat{\alpha}, \hat{\beta})}}, \\
\widehat{N}^{(3)} &= \widehat{E}_\phi\left[\frac{n}{\Pr\{C^{(K)} \geq 1 \,|\, \phi\}}\right] = n \int_0^1 \frac{f(\phi \,|\, \hat{\alpha}, \hat{\beta})}{1 - (1 - \phi)^K}\, d\phi \\
&= \frac{n}{B(\hat{\alpha},\, \hat{\beta})} \int_0^1 \frac{\phi^{\hat{\alpha}-1}(1 - \phi)^{\hat{\beta}-1}}{1 - (1 - \phi)^K}\, d\,\phi.
\end{aligned}$$

Part V

General discussion

Chapter 8

General discussion

8.1 Discussion

Estimating abundance using HT-like estimators is not a completely likelihood-based method. HT-like estimators of abundance are not obtained as an optimum of some objective function; they are obtained by adapting ideas from design-based inference. In estimating the detection function, we maximize the conditional likelihood, i.e., the likelihood conditional on the captured individuals. The parameter of primary interest (abundance) does not appear in this likelihood. Our results show that among HT-like estimators that include random effects, the HT-like estimator $\widehat{N}^{(2)}$ performs the best, although it is sometimes slightly positively biased. The main difference between these HT-like estimators and conditional likelihood estimators like those of Huggins (1989) and Alho (1990a) is that the former includes random effects, whereas the latter deal only with fixed effects. There are two scenarios in which a HT-like estimator with random effects would seem to be better than either a conditional likelihood method without random effects or a full likelihood method with random effects:

- They are better when a separate experimental survey is necessary to estimate the detection probability and there are individual or spatial random effects. In this case it is clearly better to use a method that allows random effects to be

included in the model, but a full likelihood method cannot be used to analyse the abundance survey data because its data are inadequate for estimation of detection probability.

- They may also be preferable when the main aim of inference is to estimate density of abundance in some area from samples in the area and detection probability includes random effects. Two possible approaches in this case are (a) to conduct inference on the basis of the full likelihood with random effects and (b) to use the HT-like estimator with random effects, as described in this thesis. The advantage of the latter is that, conditional on the estimated inclusion probabilities, inference about abundance or density is design-based and so there is no need to model the spatial distribution of the animals in the area of interest, whereas a full-likelihood approach requires this to be modelled. Not modelling it is advantageous because spatial distributions can be complex both in their systematic structure (or spatial trend) and in their correlation structure and this can be difficult to model adequately. In addition, inference based on an inadequate spatial model may be biased. The HT-like estimator approach produces density and abundance estimates without the need to model complex spatial processes. Moreover, the results of this thesis suggest that the resulting estimates are quite reliable, and may be asymptotically unbiased.

  Asymptotic unbiasedness has not been proven, although Huggins (1989) and Alho (1990a) showed HT-like estimators without random effects are asymptotically unbiased, although they may be biased with finite samples.

8.2 Future research

In the case of the anglerfish and woodrats applications, a full-likelihood method is not as easy as the HT-like estimator to incorporate the haul effect in detection/capture probabilities. The full likelihood must incorporate both the experimental and abundance survey data, which is not easy to implement in a frequentist framework. However, under a Bayesian framework, the posterior distribution of the estimated net retention probability can be easily incorporated in another Bayesian analysis for the abundance survey data.

For the anglerfish survey, haul random effects can be thought of as a spatial random effect, and if we use the full likelihood method, we have to assume a certain model for the distribution of density at each haul location and for each length class. One important feature of the abundance survey data is that there is a very large proportion of zeros, especially for small and larger length classes. There are two sources of excess zeros: false zeros and true zeros (Lambert (1992)). The so-called false zeros are referred to as the structural zeros, which come from areas from which it is not possible to make non-zero observations, in other words, the population density at the sampled location is zero, because, for example, the location is unsuitable habitat for the species. The problem caused by structural zeros can be solved by including a latent variable for distinguishing the structural and sampling zeros. This type of model is referred as zero-inflated models in the literature, such as the zero-inflated Poisson (ZIP) model when the distribution of the count data is assumed to be Poisson, or alternative the zero-inflated Negative Binomial model (Ridout *et al.*, 1998; Hall, 2000).

For abundance estimation using a full-likelihood method, one could fit a spatial model (e.g. Cressie, 1993) to estimate a spatial trend of the anglerfish density over the survey area, and then obtain the abundance estimates by integrating the density surface over the space. The spatial process can be independent between different locations, or have an auto-correlation within a defined neighbourhood area. If there is no auto-correlation in the spatial process, then a ZIP model fitted under a Bayesian framework can incorporate both the experimental and abundance survey data. On the other hand, if there is auto-correlation structure, this leads to fitting a ZIP auto-regressive spatial model under the Bayesian framework. In this type of model, the distribution of the spatial random effects is usually assumed to be normal with a non-diagonal variance-covariance matrix, which is defined by the conditional distribution of one location given its neighbours.

Notation

A list of notation is given below, which is standard across all the thesis. The notation is listed according to the sequence of the chapters and detailed information will be found in the related chapter

N : population size or abundance of the fish stock.

ρ : population density of the fish stock.

n : the number of fish captured during the survey.

$\hat{}$: a "hat" is used to denote an estimate or estimator. For example, $\hat{N}$ denotes an estimator or estimate of N.

a : the age of the captured fish.

l : the length of the captured fish.

s : the stratum specified by the survey design.

$\rho_{(.)}$: the density of the population with sub-notation for age/length and stratum.

A_s : the surface area of the stratum s.

A_i^c: the covered area by the ith sampling unit.

$r(l)$: the retention curve as a function of fish length l.

$r_i(l)$: the retention curve as a function of fish length l and captured by haul i.

$r(a)$: the retention curve as a function of fish age a.

h : the herding factor.

v_{1i} : area swept by the wings of the trawl i.

v_{2i} : area swept by the doors minus that swept by the wings of the trawl i.

$\sum_{i \in s}$: sum over all the trawls in stratum s.

y: response variable.

δ: a binary response variable.

q: the number of fixed-effects parameter and $q - 1$ is the number of fixed-effects explanatory variables.

$\boldsymbol{x}_{ij}$: the $q \times 1$ fixed-effects model vector for the jth observation within ith group or cluster, with the first element being 1.

$\boldsymbol{\beta}$: the q-dimensional fixed-effects parameter vector.

$\boldsymbol{z}_{ij}$: the $m \times 1$ random-effects model vector.

$\boldsymbol{b}_i$: the m-dimensional random-effects parameter vector.

Σ_b: the variance-covariance matrix of $\boldsymbol{b}_i$.

b_{0i}: the random intercept for the ith group or cluster.

b_{1i}: the random slope for the ith group or cluster.

σ_0: the standard deviation of b_{0i}.

σ_1: the standard deviation of b_{1i}.

^{B}: a superscript for between-group effect.

^{W}: a superscript for within-group effect.

References

Agresti, A., Caffo, B., & Ohman-Strickland, P. 2004. Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis*, **47**, 639–653.

Aitkin, M., & Longford, N. T. 1986. Statistical modelling issues in school effectiveness studies (with discussion). *Journal of the Royal Statistical Society: Series A*, **149**, 1–43.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. *Pages 267–281 of:* Petrov, B. N., & Csáki, F. (eds), *International Symposium on Information Theory*. Budapest, Hungary: Akadèmiai Kiadi.

Alho, J. M. 1990a. Adjusting for nonresponse bias using logistic regression. *Biometrika*, **77(3)**, 617–24.

Alho, J. M. 1990b. Logistic regression in capture-recapture models. *Biometrics*, **46**, 623–635.

Allen, V. J. 2006. *Using an Individual Based Model to Investigate the Influence of Gear Parameters and Fish Behaviour on the Efficiency of an Anglerfish* (Lophius spp.) *Survey Trawl.* MSc thesis, King's College, London.

Anon. 2005. A Sustainable Framework for Scottish Sea Fisheries. *Scottish Executive, Edinburgh*, 46.

Aranda-Ordaz, F. J. 1981. On two families of transformations to additivity for binary response data. *Biometrika*, **68**, 357–363.

Bates, D., & DebRoy, S. 2004. Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, **91**, 1–17.

Bates, D. and Maechler. M and Bolker, B. 2011. *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-39.

Belcher, C. N., & Jennings, C. A. 2009. Use of a fishery-independent trawl survey to evaluate distribution patterns of subadult sharks in Georgia. *Marine and Coastal Fisheries: Dynamics, Management, and Ecosystem*, **Science 1**, 218–229.

Bonate, P. L. 1999. The effect of collinearity on parameter estimates in nonlinear mixed effect models. *Pharmaceutical Research*, **16**, 709–717.

Borchers, D. L., & Burnham, K. P. 2004. General formulation for Distance Sampling. *In:* Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., & Thomas, L. J. (eds), *Advanced Distance Sampling.* Oxford University Press.

Borchers, D. L., Buckland, S. T., Goedhart, P. W., Clarke, E. D., & Hedley, S. L. 1998. Horvitz-Thompson estimators for double-platform line transect surveys. *Biometrics*, **54**, 1221–1237.

Borchers, D. L., Buckland, S. T., & Zucchini, W. 2002. *Estimating Animal Abundance: Closed Populations.* Springer.

Box, G. E. P. 1979. Robustness in the strategy of scientific model building. *In:* Launer, R. L., & Wilkinson, G. N. (eds), *Robustness in Statistics*. New York: Academic Press.

Breslow, N. E., & Clayton, D. G. 1993. Approximate Inference in generalized linear mixed models. *Journal of Computational and Graphical Statistics*, **88**, 9–25.

Brooks, S. P. 1998. Markov chain Monte Carlo method and its application. *The Statistician*, **47**, 69–100.

Buckland, S. T., & Garthwaite, P. H. 1991. Quantifying precision of mark-recapture estimates using bootstrap and related methods. *Biometrics*, **47**, 255–268.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., & Thomas, L. 2001. *Introduction to Distance Sampling*. Oxford: Oxford University Press.

Buckland, S. T., Summers, R. W., Borchers, D. L., & Thomas, L. 2006. Point transect sampling with traps or lures. *Journal of Applied Ecology*, **43**, 377–384.

Burstein, L. 1980. The analysis of multilevel data in educational research and evaluation. *Review of Reserach in Education*, **8**, 158–233.

Burton, R. 1976. *The Mating Game*. Crown Publishers.

Chen, J., Thompson, M. E., & Wu, C. 2004. Estimation of fish abundance indices based on scientific research trawl surveys. *Biometrics*, **60**, 116–123.

Cressie, N. A. C. 1993. *Statistics for Spatial Data.* Revised Edition. New York: Wiley.

Dorazio, R, M., & Royle, J. A. 2003. Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, **59**, 351–364.

Efron, B., & Gong, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-Validation. *The American Statistician*, **37**.

Enders, C. K., & Tofighi, D. 2007. Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods*, **12**, 121–138.

Fernandes, P. G., Armstrong, F., Burns, F., Copland, P., Davis, C., Graham, N., Harlay, X., O'Cuaig, M., Penny, I., Pout, A. C., & Clarke, E. D. 2007. Progress in estimating the absolute abundance of anglerfish on the European northern shelf from a trawl survey. *ICES CM 2007*, **K:12**, 16.

Fewster, R. M., & Jupp, P. E. 2009. Inference on population size in binomial detectability models. *Biometrika*, **96**, 805–820.

Florin, V., & Blanchard, S. 2005. Conditional Akaike information for mixed-effects models. *Biometrika*, **92**, 351–370.

Fryer, R. J. 1991. A model of between-haul variation in selectivity. *ICES Journal of Marine Science*, **48**, 281–290.

Gile, K. J., & Handcock, M. S. 2010. Respondent-driven sampling: an assessment of current methodology. *Socialogical Methodology*, **40**, 285–327.

Godø, O. R., & Totland, A. 1996. A stationary acoustic system for monitoring undisturbed and vessel affected fish behaviour. *ICES CM 1996*, **B:12**, 1–11.

Greven, S., & Kneib, T. 2010. On the behaviour of marginal and conditional Akaike information criteria in linear mixed models. *Biometrika*, **97**, 773–789.

Hall, D. B. 2000. Zero-inflated Poisson and binomial regression with random effects: a case-study. *Biometrics*, **56**, 1030–1039.

Hodges, J. S., & Sargent, D. J. 2001. Counting degress of freedom in hierarchical and other richly parameterized models. *Biometrika*, **88**, 367–379.

Hofmann, D. A., & Gavin, M. B. 1998. Centering decisions in hierarchical linear models: implications for research in organizations. *Journal of Management*, **24**, 623–641.

Horvitz, D. G., & Thompson, D. J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Hosmer, D., & Lemeshow, S. 1980. A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, **A10**, 1043–1069.

Hosmer, D. W., & Lemeshow, S. 2000. *Applied Logistic Regression*. Second Edition. Wiley-Interscience Publication.

Hosmer, D. W., Hosmer, T., le Cessie, S., & Lemeshow, S. 1997. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, **16**, 965–980.

Huggins, R. M. 1989. On the statistical analysis of capture experiments. *Biometrika*, **76**, 133–140.

Huggins, R. M., & Yip, P. S. F. 1997. Statistical analysis of removal experiments with the use of auxiliary variables. *Statistica Sinica*, **7**, 705–712.

Ibiwoye, A., & Adeleke, I. A. 2011. Estimating the proportion of undergraduates in Lagos at risk of substance abuse. *European Journal of Scientific Research*, **48**, 505–515.

Komárek, A., & Lesaffre, E. 2008. Generalized linear mixed models with a penalized Gaussian mixture as a random effects distribution. *Computational Statistics and Data Analysis*, **52**, 3441–3458.

Kreft, I. G. G, de Leeuw, J., & Aiken, L. S. 1995. The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, **30**, 1–21.

Kreft, I.G. G., & de Leeuw, J. 1998. *Introducing Multilevel Modelling*. London: Sage.

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

Lesaffre, E., & Spiessens, B. 2001. On the effect of the number of quadrature points in a logistic random-effects model: an example. *Applied Statistics*, **50**, 325–335.

Litière, S., Alonso, A., & Molenberghs, G. 2008. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in Medicine*, **27**, 3125–3144.

Liu, Q., & Pierce, D. A. 1994. A note on Gauss-Hermite quadrature. *Biometrika*, **81**, 624–629.

Longford, N. T. 1989. To center or not to center? *Multilevel Modelling Newsletter*, **1**, 7.

Maas, C. J. M., & Hox, J. J. 2005. Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, **1**, 85–91.

Madsen, N., Moth-Poulsen, T., Holst, R., & Wileman, D. 1999. Selectivity experiments with escape windows in the North Sea Nephrops (*nephrops norvegicus*) trawl fishery. *Fisheries Research*, **42**, 167–181.

McCullagh, P., & Nelder, J. A. 1989. *Generalized Linear Models*. Second Edition. Chapman and Hall.

Millar, R. B., & Fryer, R. J. 1999. Estimating the size-selection curves of towed gear, traps, nets and hooks. *Reviews in Fish Biology and Fisheries*, **9**, 89–116.

Moineddin, R., Matheson, F. I., & Glazier, R. H. 2007. A simulation study of sample size for multilevel logistic regression models. *Medical Research Methodology*, 7–34.

Molenberghs, G., & Verbeke, G. 2007. Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician*, **61**, 22–27.

Morgan, B. J. T., & Ridout, M. S. 2008. A new mixture model for capture heterogenity. *Journal of the Royal Statistical Society: Series C*, **57**, 433–446.

Munro, P. T., & Somerton, D. A. 2001. Maximum likelihood and non-parametric methods for estimating trawl footrope selectivity. *ICES Journal of Marine Science*, **58**, 220–229.

Paccagnella, O. 2006. Centering or not centering in multilevel models? – The role of the group mean and the assessment of group effects. *Evaluation review*, **30**, 66–85.

Pinheiro, J. C., & Bates, D. M. 1995. Approximations to the log-likelihood function in the non-linear mixed effects model. *Journal of Computational and Graphical Statistics*, **4**, 12–35.

Pinheiro, J. C., & Bates, D. M. 2000. *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

Pinheiro, J. C., & Chao, E. C. 2006. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational Statistics and Graphical Statistics*, **15**, 58–51.

Plewis, I. 1989. Comment on "centering" predictors in multilevel analysis. *Multilevel Modelling Newsletter*, **1**, 10–12.

Potts, J. M. 2011. *Estimating Abundance of Rare, Small Mammals: a Case Study of the Key Largo Woodrat* (Neotoma floridana smalli). Ph.D. thesis, University of St Andrews.

Rago, P. J. 2005. Fishery independent sampling: survey techniques and data analyses. *Fisheries Technical Paper*, **474**, 201–215.

Raudenbush, S. W., & Bryk, A. S. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Second Edition. Thousand Oaks, CA: Sage.

Raudenbush, S. W., Yang, M. L., & Yosef, M. 2000. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, **9(1)**, 141–157.

Reid, D. G., Allen, V. J., Bova, D. J., Jones, E. G., Kynoch, R. J., Peach, K. J., Fernandes, P. G., & Turrell, W. R. 2007a. Anglerfish catchability for swept area abundance estimates in a new survey trawl. *ICES Journal of Marine Science*, **64**, 1503–1511.

Reid, D. G., Kynoch, R. J., Penny, I., & Peach, K. 2007b. Estimation of catch efficiency in a new anglerfish survey trawl. *ICES CM 2007*, **Q:22**.

Richards, F. J. 1959. A flexible growth function for empirical use. *Journal of Experimental Botany*, **10**, 290–301.

Ridout, M., Demétrio, C. G. B., & Hinde, J. 1998. Models for count data with many zeros. *Invited paper presented at the Nineteenth International Biometric Conference, Capetown, South Africa*.

Rodriguez, G., & Goldman, N. 1995. An assemessment of estimation procedure for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A*, **158**, 73–90.

Rue, H., Martino, S., & Chopin, N. 2009. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society: Series B*, **71**, 319–392.

Self, S. G., & Liang, K. Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

Skrondal, A. 2009. Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Series A*, **172**, 659–687.

Snijders, T. A. B., & Bosker, R. J. 1993. Standard errors and sample sizes for 2-Level research. *Journal of Educational Statistics*, **18**, 237–259.

Snijders, T. A. B., & Bosker, R. J. 1999. *Mutilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Somerton, D. 1996. Estimating trawl catchability. *Alaska Fisheries Science Center Quarterly Report*, **April-May-June**, 1–3.

Somerton, D., Ianelli, J., Walsh, S., Smith, S., Godø, O., & Ramm, D. 1999. Incorporating experimental derived estimates of survey trawl efficiency into the stock assessment process: a discussion. *ICES Journal of Marine Science*, **56**, 299–302.

Tanner, M. A., & Wong, W. H. 1987. The calculation of posterior distributions by data augmentation. *Journal of the A*, **82**, 528–540.

Tanton, M. T. 1965. Problems of live-trapping and population estimation for the woodmouse, *Apodemus sylvaticus* (L.). *Journal of Animal Ecology*, **34**, 1–22.

Thomas, Y. 1976. *Abyss: The Deep Sea and the Creatures That Live in It*. Crowell Company.

Thompson, S. K. 2002. *Sampling*. Wiley.

Tierney, L., & Kadane, J. B. 1986. Accurate approximation for posterior moments and marginal densities. *Journal of American Statistical Association*, **81**, 82–86.

Tschernij, V., & Holst, R. 1999. Evidence of factors at vessel-level affecting codend selectivity in Baltic cod demersal trawl fishery. *ICES CM 1999*, **R:02**.

Venables, W. N., & Ripley, B. D. 2002. *Modern Applied Statistics with S*. Fourth Edition. Springer.

Volz, E., & Heckathorn, D. D. 2008. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, **24**, 79–97.

Williams, R., & Thomas, L. 2007. Distribution and abundance of marine mammals in the coastal waters of British Columbia, Canada. *Journal of Cetacean Research and Management*, **9(1)**, 15–28.

Wood, S. N. 2006. *Generalized Additive Models*. Chapman & Hall.

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. 2009. *Mixed Effects Models and Extensions in Ecology with R*. First Edition. Springer.