MARINE RESEARCH INSTITUTE TECHNICAL REPORT NR. 91

Statistical Analysis of Cod Catch Data from Icelandic Groundfish Surveys

Jenný Brynjarsdóttir

Statistical Analysis of Cod Catch Data from Icelandic Groundfish Surveys

Jenný Brynjarsdóttir

M.Sc. Thesis

June 2002
Department of Mechanical- and Industrial Engineering
University of Iceland

Abstract

Catch data for cod from the Icelandic groundfish surveys are analyzed in this thesis using mainly generalized linear models (GLM). The mail goal is to find environmental variables that affect the expected cod catch.

Two probability distributions are proposed for describing the data, the gamma distribution and the log-normal distribution. These two distributions have the same variance function, i.e. the variance is proportional to the mean squared. This relationship is investigated by a regression of the log sample variance on the log sample mean for small groups of data. This regression suggests a power relationship with an index slightly larger than 2.

The two proposed distribution are compared via a goodness of fit test. Such tests are complicated by the fact that the expected cod catch is variable and thus scaling of the data is needed to obtain a constant mean. The shape parameters are assumed to be constant but need to be estimated. This is done by a GLM where small survey areas and years are included as qualitative covariates. The hypothesis that the data, scaled with fitted values from the GLM, follow the $G(\widehat{r}, 1/\widehat{r})$ or the $LN(0, \widehat{\sigma}^2)$ distribution is tested. The shape parameters $\widehat{r}$ and $\widehat{\sigma}^2$ are estimated by the GLM. The test results are found to be better for the log-normal distribution. The models of the cod catch data investigated in this thesis therefore treat the log transformed data as the response and assume normally distributed errors.

The expected cod catch is assumed to be constant at a given place and time. The most straightforward model available is therefore a fully qualitative model with only spatial and time effects. This model is found to explain 63% of the variation in the cod catch data but that comes with a high cost of degrees of freedom.

A thorough examination of the environmental data collected in the groundfish surveys is given. Polynomials are proposed to describe the relationship between each environmental variable and the cod catch. The effects of these polynomials on the cod catch are then tested within the GLM framework. The most important environmental effects are found to be the bottom temperature, depth and surface temperature, although most of the tested effects are found to be significant.

Finally, an attempt is made to locate temperature fronts in the ocean by estimating the temperature gradient vector at each data point. The size of this gradient vector is then included in the qualitative model. This is a new approach to statistical groundfish analysis. The gradient term is, however, not found to be important (though significant) in the explanation of variation in the cod catch data.

Samantekt

Í þessari ritgerð eru rannsökuð aflagögn þorsks úr stofnmælingaleiðangrum Hafrannsóknarstofnunar. Þessir leiðangrar eru í daglegu tali nefndir togararall. Aðferðafræðin sem stuðst er við nefnist alhæfð línuleg líkön (e. generalized linear models, GLM). Megin markmiðið með þessari rannsókn er að finna umhverfisþætti sem hafa áhrif á þorskafla.

Skoðaðar eru tvær líkindadreifingar til að lýsa gögnunum, gamma dreifingin og log-normal dreifingin. Þær hafa sama dreifnifall, þ.e. dreifnin vex í hlutfalli við meðalgildið í öðru veldi. Þetta samband er rannsakað með aðhvarfi úrtaksdreifni á úrtaksmeðaltal lítilla gagnahópa. Þessi aðhvarfsgreining gefur til kynna veldisvísissamband með veldisvísi rúmlega tveir.

Þessar tvær líkindadreifingar eru bornar saman með mátunarprófum. Gert er ráð fyrir því að meðalgildið sé breytilegt og því þarf að staðla gögnin þannig að þau hafi sama væntigildi. Þetta er gert með GLM þar sem tiltölulega lítil leiðangurssvæði og ár eru notaðar sem skýribreytur. Þá er prófuð tilgátan um að gögnin, sköluð með metnum gildum frá GLM, lúti $G(\widehat{r}, 1/\widehat{r})$ eða $LN(0, \widehat{\sigma}^2)$ dreifingu. Formstikarnir $\widehat{r}$ og $\widehat{\sigma}^2$ eru einnig metnir með hjálp GLM líkansins. Niðurstöður prófanna eru betri í tilfelli log-normal dreifingar. Líkönin sem skoðuð eru í þessari ritgerð nota því logravörpuð gögn og gera ráð fyrir normal dreifðum skekkjum.

Gert er ráð fyrir því að þorskaflinn hafi sama væntigildi á litlum svæðum og svipuðum tíma. Þáttalíkan, sem inniheldur aðeins staðsetningu og tíma sem skýribreytur, er því er einfaldasta mögulega líkanið. Þetta líkan útskýrir 63% af heildar breytileikanum í gögnunum en notar til þess helst til margar frígráður.

Umhverfisgögnin sem safnað er í togararallinu eru rannsökuð ítarlega. Margliður af ýmsum stigum er notaðar til að lýsa sambandi hverrar umhverfisbreytu og þorskafla. Áhrif þessara margliða á þorskafla eru síðan prófuð með aðstoð GLM. Mikilvægustu umhverfisþættir eru botn hitastig dýpi og yfirborðs hitastig. Flest prófuð áhrif eru þó marktæk.

Að lokum er gerð tilraun til að staðsetja hitaskil í sjónum með því að meta hita stigulinn í hverjum gagna punkti. Lengd stigulsins er svo bætt inn sem skýribreytu í GLM líkanið. Þetta er ný nálgun í greiningu botnfisks gagna en það kemur í ljós að þessi stiguls liður er lítilvægur (en þó marktækur) í útskýringu á breytileika þorskafla.

Preface

This thesis has been prepared in the engineering department of the University of Iceland for the degree of M.Sc. in engineering.

This work was supported by the Marine Research Institute in Reykjavík, for which I am grateful.

I would like to thank my supervisors Gunnar Stefánsson, the mathematical faculty at the University of Iceland, Guðmundur R. Jónsson and Páll Jensson, the industrial engineering faculty at the University of Iceland for their help and guidelines during this work.

I also thank Lorna Taylor, statistician at the MRI, for sharing her insight and knowledge of the MRI databases and for her help with the S-plus program.

I thank my friend Ásta Herdís Hall for her very good comments and help with the English text.

Finally, I thank my husband Halldór Elías Guðmundsson for his never failing support and encouragement during this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Every March since 1985 the Marine Research Institute (MRI) in Iceland has conducted an annual groundfish survey in the Icelandic continental shelf. The purpose of this survey is to gather data which are used in the process of stock assessment for several demersal species. The data gathered contain numerous records of environmental measurements such as depth, temperature, weather etc. which give occasion for investigating the relation of such factors with catch rates. Such analysis for the Icelandic cod (*Gadus morhua*) is the subject of this thesis.

The methodology selected for this work is based on generalized linear models. As for any model type applied to biological data the models considered here will only be correct to a certain extent. An effort made to examine whether the data fulfill the model assumptions, with emphasis on which probability distribution function will be selected to describe the variability of the data.

The investigation of environmental effects and the probability distribution functions performed in this thesis can be utilized in several ways. For example, as an input into the stock assessment process of MRI (Anon. 2001) and the multi species model currently being developed by the MRI and other international marine institutes (Anon. 2002).

The main motivation for this work comes from a somewhat different vision. Many Icelandic trawler captains have since 1973 reported the estimated catch of each species in each tow along with a number of other parameters such as position, time of catch, depth etc. In 1990 these reports were made mandatory by law and since then every action of the entire Icelandic fleet has been recorded, providing a database containing very valuable information. These reports, along with several other data, are used by the MRI in the process of estimating stock abundance for several species of fish. The reports are confidential and only used for scientific purposes. However, the amount of data (over 3 million actions) awoke the idea of using the data for short term catch forecasts. Such forecasts can potentially be used by vessel captains to help selecting fishing grounds.

An examination of the current status of information systems used on board of fishing vessels, along with an analysis of the decisions made by vessel captains and the information that those decisions were based on was performed in the master's thesis of Bjarnason (1997). The choice of fishing grounds was found to be the most important decision and the most complicated one and it was shown that it could be crucial for the economy of the fishery. The captains currently make this decision based on experience and the latest information from other fishermen. Thus it is clear that short term catch forecasts could turn out to be a very helpful device.

The effects of environmental variable on catch rates found in this thesis will be very useful information if a forecast model is to be made using the trawler reports, and so will the investigation of probability distribution functions. Given that the survey data are collected more consistently than the trawler reports and contain much more complete environmental data, the present analysis can be used as an indicator of what effects should be considered when developing models for the commercial fleet.

1.1 Background

1.1.1 Groundfish survey data and generalized linear models

Classical linear models and ordinary least squares date back to the work of Gauss and Legendre who applied the methodology to astronomical data. The normal, or Gaussian, distribution was developed to describe measurement errors. Later in the nineteenth century the distribution was used to describe the variation between individuals in a biological population, leading to the numerous biological applications of linear models. (McCullagh & Nelder 1989, p.1).

Groundfish trawl surveys are commonly conducted for the purpose of obtaining an average catch per tow, to be used as an indicator of stock abundance used in a stock assessment process. The single most common method for estimating means or total abundances and associated variances is probably the standard formulas for stratified random design. (Stefánsson 1996). In such design-based analysis, all inferences about the estimates of population quantities are based on treating the measured characteristics (such as number of fish) as fixed. It is the method of choosing the sample units as specified by the survey design which provides the source of randomness. (Smith 1990).

In general, abundance estimates from trawl surveys have large variance estimates and often the mean and variance are directly related, i.e. larger means have larger variances associated with them. A number of statistical models have been suggested for the estimation of mean catch per tow in an attempt to account for this variability. In such model-based inference the data are assumed to be generated by some underlying stochastic process whose structure is entirely known, apart from a fixed number of parameters. (Smith 1990). Many of these methods use skewed probability distributions, for example the delta-distribution (Pennington 1983). Here the probability of obtaining an empty tow is modeled as a Bernoulli trial but the distribution conditioned on positive values is assumed to be the log-normal or the gamma distribution. Smith (1990) provides an interesting comparison of these two approaches, design-based analysis and model-based analysis.

Another possibility of model-based analysis is to use a generalized linear model (GLM) where the survey strata can simply be treated as covariates. Myers & Pepin (1986) were the first to use the linear regression model as a method for estimating population size in groundfish surveys and to implement the method on the American plaice. The GLM has the advantage over the stratified analysis that the underlying spatial pattern of the fish density can be explicitly modeled, an aspect ignored by the stratified analysis. Also, data from all years of the survey can be analyzed at once and data from uncompleted surveys can be included. Furthermore, the use of environmental variables as explanatory variables separates the inter-annual variation in population size from the variation due to differences in the environmental data. The difficulty of the model is that the parameter estimates are sensitive to violations of the model assumptions. In addition, the abundance indices obtained so far in the literature have not performed better than simple arithmetic or weighted averages when compared to results of virtual population analysis (VPA) (Anon. 1992).

Survey data often include a large proportion of zeros which causes problems when the log transformation of data is used or when the log link is used in a GLM. Stefánsson (1996) combines the delta-distribution and the GLM framework and develops a maximum likelihood method where an explicit formula is obtained for the probability distribution of the catch at each station. The resulting model allows formal testing of which factors influence survey catches, as well as the computation of abundance indices.

The GLM framework has been used to obtain stock abundance indices from Icelandic marine data, but is not currently used in the stock assessment process. An example of this can be found in Stefánsson (1988) where cod catch data from Icelandic trawler reports are analyzed using a GLM. The total catch of a trawler in one month in a single statistical rectangle is assumed to be dependent on the total towing time of that trawler in that month and area. Other effects in the model include vessel effects, spatial effect (statistical rectangle) month and year effects. The year effects give annual cpue (catch per unit of effort) indices. These new indices were found to give at least as good correlations with stock size as indices computed with methods such as VPA.

Another example of the use of the GLM can be found in Stefánsson (1996). The main purpose of that paper is to establish a method combining the delta and GLM approaches but it also provides an example of using a GLM for haddock catch data from the Icelandic groundfish surveys. Only spatial and depth factors were considered in addition to the year effect and the data consisted of number of haddock per age and by station. The paper concludes that the smallest survey areas are needed (in this case statistical rectangles) and yet the model only explains about 40% of the variation of non-empty tows. The estimated abundance indices obtained in this model are found to perform well compared to other indices and VPA runs.

Unlike the usual analysis of survey data and any previous analysis of Icelandic marine data the emphasis in this thesis is on the variability of the survey catch. The purpose is not to obtain an abundance index, keeping a possible catch forecast in mind. Numerous environmental measurements from the survey are tested for significance in order to cast light on why the fish is caught at one place rather than another. This thesis provides a more thorough examination of the environmental factors measured in the survey than has been done before.

Only the tows where the catch included a positive number of cod (non-empty tows) are included in this work. This will not undermine the usefulness of this analysis for forecasting purposes since captains focusing on cod will be concerned with areas where the probability of getting tows containing no cod is very low. Furthermore a Bernoulli model of empty/non-empty tows can be combined with the analysis afterwards, following the delta approach of Stefánsson (1996).

1.1.2 Gamma distribution versus the log-normal distribution

The cod catch data from the Icelandic groundfish surveys are highly skewed. Several authors have suggested the use of the gamma distribution to describe skewed marine data. This includes Stefánsson (1996) who used the gamma distribution for age-disaggregated haddock catch data from the Icelandic groundfish surveys, also Goñi, Alvarez & Adlerstein (1999) for Western Mediterranean fisheries and Ye, Al-Husaini & Al-Baz (2001) for the Kuwait drift net fishery. A log transformation has frequently been used as a normalizing and/or variance stabilizing method and the log-normal distribution has therefore also been used to describe skewed marine data. Examples of this are Myers & Pepin (1986) for the American plaice, Stefánsson (1988) for cod cpue data from the Icelandic trawler reports, Lo, Jacobson & Squire (1992) for northern anchovy data collected by aircrafts and Pennington (1996) for several marine survey data. Furthermore, Steinarsson & Stefánsson (1986) fitted several probability distributions to the cod catch data from the Icelandic groundfish surveys 1985-1986 and found that among tested distributions, the gamma, log-normal and negative binomial distribution had the best fit. In light of this literature, the adequacy of both the gamma distribution and the log-normal distribution for the cod catch data will be investigated in this thesis. The negative binomial distribution will not be considered here.

The gamma and log-normal distributions share some characteristics which often makes it difficult to choose between them. Both distribution have a positive probability mass only for positive values and can describe data with the majority of the probability mass on low values but a heavy tail to the right. They also share the same relationship between the mean and variance:

$$\text{var}(Y) = \sigma^2 E(Y)^2 \tag{1.1}$$

This relationship differs from other distributions like the normal, poisson and the negative binomial distribution and can therefore be used to distinguish these two distributions from others. A common approach to check this relationship is to examine a plot of log(sample variance) versus log(sample mean) for homogeneous groups of data, see for example McCullagh & Nelder (1989, p.306), Stefánsson & Pálsson (1997) and Goñi et al. (1999). If the points form a straight line with the slope close to 2, the gamma and log-normal distributions can not be rejected as the true underlying distribution. This check will be performed in this thesis. That has not been done before for the cod catch data from the Icelandic groundfish surveys. Such investigation can not, however, distinguish between these two distributions.

A GLM using gamma distributed errors usually assumes a constant coefficient of variation (CV). Since the CV is defined as

$$CV = \frac{\sqrt{\text{var}(Y)}}{|E(Y)|},$$

the σ in (1.1) is simply the CV. When the CV is small the variance stabilizing transformation $\log(Y)$ has approximate moments

$$E(log(Y)) \approx \log(E(Y)) - \sigma^2/2 \quad \text{and} \quad \text{var}(\log(Y)) \approx \sigma^2$$

where σ denotes the CV. Hence, assuming constant CV for Y is approximately equivalent to assuming a constant variance for $\log(Y)$ for low values of σ. Furthermore, if the model assumes that the independent variables have multiplicative effects on the response on the original scale the effects are additive on the log scale. Then, with the exception of the intercept or constant term in the linear model, consistent estimates of the parameters and of their precision may be obtained by transforming the data to the log scale and applying ordinary least squares. The intercept is then biased by approximately $-\sigma^2/2$. (McCullagh & Nelder 1989, p.285). For small σ^2 a GLM analysis assuming gamma distributed errors (equivalent to assuming a constant CV) and a GLM analysis on log transformed data assuming normally distributed errors will therefore usually produce the same conclusions. Atkinson (1982) concludes that these two methods of analysis should provide similar results for σ^2 as large as 0.6. When σ^2 is small, the choice of distribution function should then depend on the purpose of the investigation. For a dataset where the exact distribution of Y is unknown, the first method is convenient and desirable if the analysis is exploratory or if only a graphical presentation is needed. But if Y is a variable with a physical dimension and we want to present conclusions on the original scale of measurement, it is preferable to retain that scale and not to transform the response. (McCullagh & Nelder 1989, p.286). Furthermore, if the distribution is unknown, Firth (1988) compares the efficiencies of parameter estimates using the gamma model when the errors are in fact log-normally distributed with using the log-normal model when the errors are really gamma distributed. He concludes that the gamma model performs slightly better.

When σ^2 is large, however, these two kinds of analysis can give different results (see for example Wiens (1999)). Therefore, a comparison of how well these two distributions fit the data is desirable. In Stefánsson (1988) and Stefánsson & Pálsson (1997) this is done by scaling the observations with the fitted values from a GLM. Then a Kolmogorov-Smirnov test on the scaled data can be used to distinguish between the gamma and log-normal distributions. This comparison will be conducted in this thesis. That has not been done for the data from Icelandic groundfish surveys before.

1.1.3 Temperature fronts

It has been suggested that the food for cod, such as capelin, may aggregate in frontal regions where cold sea meets with warmer sea, i.e. where there is a sudden change in temperature (Vilhjálmsson 1994). Therefore it could be expected that the cod would pursue such temperature fronts.

In light of this it would be interesting to test the effect of temperature fronts on the catch rates of cod. In order to do that, estimates of the size of the temperature gradient in each station is needed. A temperature surface over the survey area is estimated in this thesis, using locally weighted regression on bottom temperature which is measured in each station in the survey. This surface is then used to get an estimate of the size of the temperature gradient at each station, which is then tested for significance in a GLM.

Incorporation of temperature fronts in the analysis of groundfish catch data has not been done before[1]. The analysis in this thesis is limited and it would be interesting to develop this approach further.

[1] In Sakuma & Ralston (1995) spatial distributions of some late larval groundfish species off central California where found to be dependent on a temperature front, but no literature was found where temperature fronts are linked to groundfish catch data.

1.2 Outline of the thesis

In this introduction the motivation of the project was described. An introduction of the background of this work was given and a discussion was provided on in what sense this thesis is different from what has previously been done in the field. In chapter 2, the data used in the analysis of this thesis is described. In chapter 3, an overview of the methodology is given. This includes mainly subjects related to the generalized linear model, but also a short introduction to locally weighted regression and additive models. Chapter 4 contains some analysis related to the choice of a probability function (gamma or log-normal) that will be used to describe the cod catch data. In chapter 5, three linear models of the cod catch data are considered. Each model is described and the results are examined. Finally, chapter 6 contains a discussion of the conclusions of this work and some ideas for future work.

Chapter 2

The Icelandic groundfish surveys

The data analyzed in this thesis are cod catch data from the Icelandic groundfish surveys. These surveys have been conducted in March every year since 1985 by the Marine Research Institute (MRI) in Iceland. The survey design is outlined in section 2.1 and in section 2.2 an overview is given of the data.

2.1 Groundfish survey design

The groundfish survey area contains the Icelandic continental shelf inside the 500 meters depth contour (see figure 2.1), which covers the fishing grounds for the most important commercial species of demersal fish in Icelandic waters.

The stratification scheme is based on cod density patterns pre-estimated from historical commercial, as well as research vessel, catch data. The whole survey area is divided into statistical rectangles of half degree latitude and one degree longitude, on which the stratification scheme is based. Each statistical rectangle can then be divided into four sub-rectangles. Based on biological and hydrographic considerations, the survey area is divided into two main areas, northern and southern areas, and ten sub areas (strata). Stations are allocated among strata in direct proportion to the area of each stratum and its estimated cod density. Finally, the trawl stations of a stratum are allocated to each statistical rectangle within the stratum in direct proportion to the area of the rectangle. The main areas, sub areas and the statistical rectangles are shown on the map in figure 2.1.

Stations within each statistical rectangle were initially divided equally between fishermen, and project members from MRI. MRI selected random places as the middle of a tow and fishermen decided on the direction. Fishermen were asked to fix their stations in each rectangle in accordance with their knowledge and experience of fishing and fishing grounds. The stratification is therefore semi-random, but the survey design is systematic since the same stations are repeated every year. The (middle) locations of the stations which gave positive cod catches in the survey years 1985-2001 are shown in figure 2.1.

In order to visit all stations in the time limit of 2-3 weeks, five commercial fishing vessels are leased every year (not necessarily always the same vessels). The bottom trawl was standardized

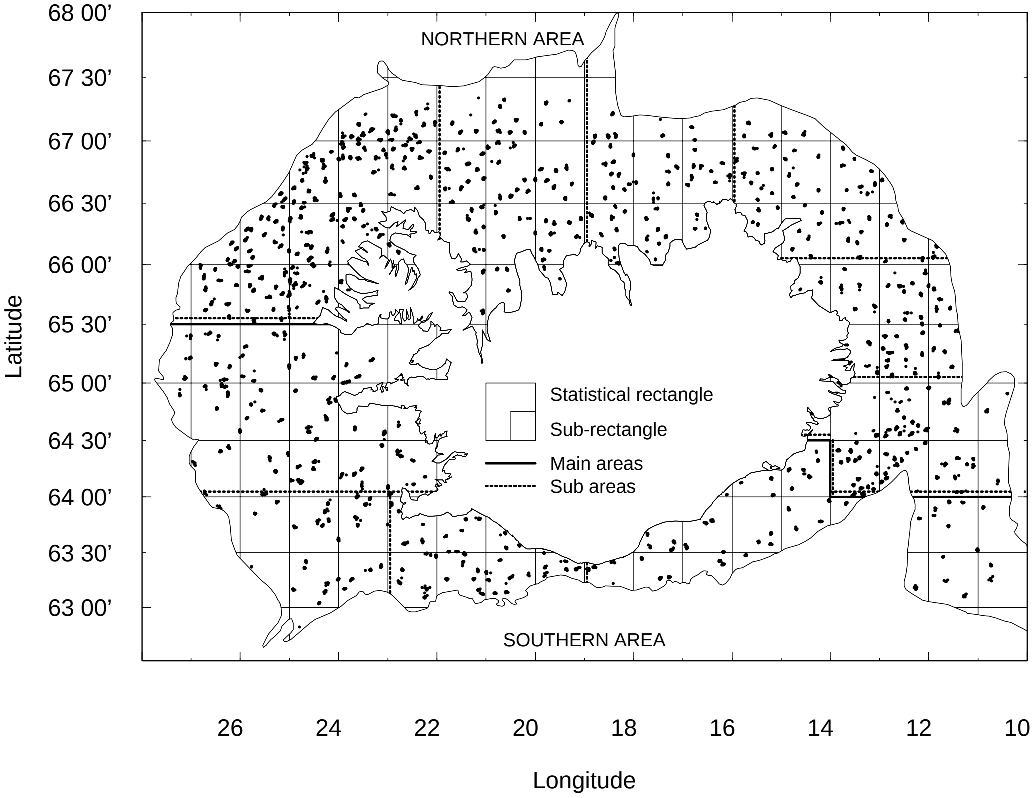

Figure 2.1: A map of the survey area showing the 500 meters contour line, main areas, sub areas and statistical rectangles. The points denote the (middle) locations of stations for all survey years were positive cod catches were obtained.

in cooperation between the scientists of MRI and the trawler captains. The main consideration was that the trawl would effectively sample groundfish species, especially cod, of different lengths ensuring adequate data for the catchable stock as well as the recruiting year-classes.

Fishing methods are standardized as far as possible. Standard towing speed is 3.8 knots over the bottom and the towing distance is 4.0 nautical miles. Trawling is done uniformly over the 24 hours in the day. The vessels are supposed to stop trawling when the wind force exceeds 8 on the Beaufort scale.

For a more detailed description of the survey design see Pálsson, Jónsson, Schopka, Stefánsson & Steinarsson (1989).

2.2 Collected data

The collected data can be categorized into trawl station data, trawl catch data and environmental observations.

year	1985	1986	1987	1988	1989	1990	1991	1992	1993
no. stations	579	576	553	531	556	556	555	555	575

year	1994	1995	1996	1997	1998	1999	2000	2001	Total
no. stations	579	578	535	525	498	521	515	516	9303

Table 2.1: Number of stations visited each year of the survey.

The trawl station data recorded are position, time, direction and depth of the tow, distance towed and trawling speed. In addition, information on trawl performance was recorded.

The trawl catch data include length measurements, age determination from otolith samples and sex determination. The sampling of otoliths is performed in a length stratified procedure for each fish species in each of the 10 sub areas. The total catch per species is not weighted but can be obtained by length-weight relationships.

In the environmental category the following meteorological and hydrographic data are recorded for each trawl station: wind force (Beaufort scale) and direction, air-, surface- and near-bottom temperature, weather conditions, cloud coverage, wave height, ice conditions and barometric pressure.

The survey handbook (Einarsson, Jónsson, Björnsson, Pálsson, Schopka & Bogason 2002) provides detailed descriptions of the data collection.

A total of 600 stations were initially considered a reasonable number for the objectives of the survey, such as precision in stock abundance estimates. Over the years the total number of visited stations has been variable due to a number of reasons, but it has always exceeded 500. In this analysis, only stations were the cod catch is strictly positive are considered (as explained in section 1.1.1). Table 2.1 displays the number of such stations per survey year. The total number of available observations are 9303.

Chapter 3

Methods

In this chapter an overview of the methodology used in this thesis is given. In section 3.1, subjects related to the generalized linear model are described and in section 3.2, a short introduction to locally weighted regression and additive models is given.

3.1 Generalized linear models

The term ‘generalized linear model’ (GLM) was introduced by Nelder & Wedderburn (1972). This model class unifies many statistical techniques such as analysis of variance, regression analysis, analysis of covariance, probit analysis and contingency tables. The GLM can be viewed as a method that aims to describe the mean structure in a set of observations as a linear function of several explanatory variables or factors. The main goal of the method is often to understand what affects the behavior of the phenomenon of interest.

A GLM consists of a systematic part and a random part. For the random part, a vector of observations (the phenomenon of interest) $\mathbf{y} \in \mathbb{R}^N$, also called the *response*, is assumed to be a realization of a random variable, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$. The components of $\mathbf{Y}$ are assumed to be statistically independent and belong to the same exponential family of probability distributions with a constant dispersion parameter (see section 3.1.1). The means, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)$, are however not required to be equal. The systematic part specifies a transformation of the means as a linear function of unknown parameters, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)^T$, and known explanatory variables, $\mathbf{x}_1, \ldots, \mathbf{x}_m$, called *covariates.* This can be written in matrix form:

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

where x_{ij} is the value of the jth covariate corresponding to observation i, $j = 1, 2, \ldots, m$ and $i = 1, 2, \ldots, N$. The matrix $\mathbf{X}$ is called the *model matrix* and the transformation $g(\cdot)$ is called the *link function* as it links together the linear terms and the mean vector. The model matrix is assumed to be regular. The systematic part of GLMs can also be written in a parametric form:

$$g(\mu_i) = \eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im}$$

(McCullagh & Nelder 1989, p.26-27).

Both qualitative and quantitative covariates can be included in a GLM. A quantitative covariate occupies one column of the model matrix where the ith entry in that column is simply the value of the covariate corresponding to the ith observation. A qualitative covariate, also called a *factor*, occupies one column for each level. Then the value of x_{ij} is 1 if observation i belongs to level j but 0 otherwise. This means that more parameters are required in the model for a factor (one for each level) than for a quantitative covariate. As an example of this, a model with one quantitative covariate $\mathbf{x}$ and one factor that has l levels, can be written in parametric form as follows:

$$\eta_i = \gamma x_i + \sum_{j=1}^{l} \beta_j \delta_j(i), \quad \text{where } \delta_j(i) = \begin{cases} 1 & \text{if observation } i \text{ belongs to level } j, \\ 0 & \text{otherwise.} \end{cases}$$

and in matrix form:

$$\boldsymbol{\eta} = \begin{bmatrix} 1 & & & & x_1 \\ 1 & & & & x_2 \\ \vdots & & & & \vdots \\ 1 & & & & x_i \\ & 1 & & & x_{i+1} \\ & \vdots & & & \vdots \\ & 1 & & & \\ & & \ddots & & \vdots \\ & & & 1 & \\ & & & \vdots & \vdots \\ & & & 1 & x_N \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_l \\ \gamma \end{bmatrix}$$

3.1.1 The exponential family

A random variable Y is said to follow a probability distribution from the *exponential family* if it has a probability density function of the form

$$f(y; \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are functions of the *canonical parameter* θ and the *dispersion parameter* ϕ. The dispersion parameter is sometimes denoted with σ^2. The function $a(\phi)$ is commonly of the form $a(\phi) = \phi/w$ where w is a known prior weight that varies from observation to observation. (McCullagh & Nelder 1989, p.28).

The definition of the exponential family is merely for the convenience of being able to talk about many different distributions that have common properties. In addition, the theory of GLMs is set forth for an arbitrary exponential family. Examples of probability distributions in the exponential family are the normal-, poisson-, binomial- and gamma distributions.

The variance and mean of Y can be written in terms of the functions $a(\cdot)$ and $b(\cdot)$:

$$E(Y) = \mu = b'(\theta) \quad \text{and} \quad \text{var}(Y) = b''(\theta)a(\phi)$$

Let $\tau(\theta) := b'(\theta) = \mu$. It can be shown that τ is a bijective mapping from the interior of the domain of τ onto the range of τ so there is a one-to-one correspondence between the canonical parameter and the mean, $\theta = \tau^{-1}(\mu)$. The function $b''(\theta)$ can therefore be considered as a function of μ

$$b''(\theta) = b''(\tau^{-1}(\mu)) =: V(\mu)$$

and is called the *variance function.* This function is characteristic for an exponential family, i.e. it defines the family uniquely. (McCullagh & Nelder (1989, p.28-29) and Thyregod (1998*b*, p.31-35)).

The probability distributions considered in this thesis are the gamma distribution and the log-normal distribution. In what follows, some characteristics of these distributions will be summarized.

The gamma distribution

A random variable Y is said to be gamma distributed, $Y \sim G(r, \mu/r)$, if it has the probability function

$$f(y) = \frac{y^{r-1}e^{-yr/\mu}}{(\mu/r)^r\Gamma(r)}$$

where $\Gamma(\cdot)$ is the Gamma function,

$$\Gamma(r) = \int_0^\infty x^{r-1}e^{-x}dx$$

The gamma distribution is an exponential distribution with mean, variance, variance function and dispersion parameter as follows:

$$E(Y) = \mu, \quad \text{var}(Y) = \frac{\mu^2}{r}, \quad V(\mu) = \mu^2, \quad \phi = \frac{1}{r}$$

(Thyregod 1998*b*, p.149-153). A useful fact about the gamma distribution is that if $Y \sim G(r, \mu/r)$ and $c \in \mathbb{R}$ then $cY \sim G(r, c\mu/r)$ (Conradsen 1999, p.120). Furthermore, if $c = 1/\mu$ then

$$\frac{Y}{\mu} \sim G\left(r, \frac{1}{r}\right)$$

The log-normal distribution

A random variable Y is said to be log-normally distributed, $Y \sim LN(a, b^2)$, if $Z = \log(Y)$ is normally distributed with mean a and variance b^2, $Z \sim N(a, b^2)$. The variance and mean of Y are

$$E(Y) = e^{a+\frac{1}{2}b^2} \quad \text{and} \quad \text{var}(Y) = e^{2a+b^2}(e^{b^2} - 1) = E(Y)^2(e^{b^2} - 1) \tag{3.1}$$

(Conradsen 1999, p.134-135). The log-normal distribution is not a member of the exponential family but can be used in the GLM framework by log transforming the data and assume normally distributed errors.

Although a variance function is not defined for the log-normal distribution it can be seen from (3.1) that the relationship between $E(Y)$ and $\text{var}(Y)$ is the same as for a gamma distributed variable, i.e.

$$\text{var}(Y) = \phi \cdot E(Y)^2$$

where $\phi = e^{b^2} - 1$.

A useful fact about the normal distribution is that if $Z \sim N(a, b^2)$ and $c \in \mathbb{R}$ then $Z - c \sim N(a - c, b^2)$ (Conradsen 1999, p.125). Furthermore, if $c = a$ then

$$Z - a \sim N(0, b^2) \quad \text{and} \quad e^{Z-a} \sim LN(0, b^2)$$

3.1.2 Estimation of parameters in GLMs

The parameters in a GLM are estimated with the maximum likelihood method. The log-likelihood function for an exponential family is

$$\begin{aligned} l(\theta, \phi; y) &= \log f(y; \theta, \phi) \\ &= \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \\ &= \frac{y\tau^{-1}(\mu) - b(\tau^{-1}(\mu))}{a(\phi)} + c(y, \phi) \\ &= l(\mu, \phi; y) \end{aligned} \tag{3.2}$$

This function measures the likelihood of arbitrary values of the parameters μ and ϕ for a given observation y and for a given distribution family. Assuming that $\boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$ the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is the one that maximizes the log-likelihood function (3.2). This is found as the solution of the score function:

$$\left(\text{diag}\left\{\frac{1}{g'(g^{-1}(\mathbf{x}_i\boldsymbol{\beta}))}\right\} \mathbf{X}\right)^T \text{diag}\left\{\frac{w_i}{V(g^{-1}(\mathbf{x}_i\boldsymbol{\beta}))}\right\} (\mathbf{y} - g^{-1}(\mathbf{X}\boldsymbol{\beta})) = 0 \tag{3.3}$$

which is the log-likelihood function differentiated with respect to $\boldsymbol{\beta}$. (Thyregod 1998*a*, p.179-180)

For a GLM assuming gamma distributed errors, with a logarithmic link function and $w_i = 1 \; \forall i$ (3.3) becomes:

$$X^T \text{diag}\left\{\frac{1}{e^{\mathbf{x}_i\boldsymbol{\beta}}}\right\} \left(\mathbf{y} - e^{\mathbf{X}\boldsymbol{\beta}}\right) = 0$$

For a GLM assuming normally distributed errors, with an identity link function and $w_i = 1 \; \forall i$ (3.3) becomes:

$$\mathbf{X}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \quad \Rightarrow \quad \widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{y}$$

Usually, equation (3.3) has to be solved iteratively with the Newton-Raphson method or the Fishers scorings method for example. See Thyregod (1998*a*, p.191) for details.

3.1.3 The deviance

Instead of using the log-likelihood function $l(\boldsymbol{\mu}, \phi; \mathbf{y})$ directly as a criterion for comparing the goodness-of-fit of different models, a linear function called the *deviance* is used:

$$\begin{aligned} D(\mathbf{y}; \boldsymbol{\mu}) &:= 2\phi(l(\mathbf{y}, \phi; \mathbf{y}) - l(\boldsymbol{\mu}, \phi; \mathbf{y})) = 2\phi \sum_{i=1}^{N} (l(y_i, \phi; y_i) - l(\mu_i, \phi; y_i)) \\ &= 2\sum_{i=1}^{N} w_i\{y_i(\tau^{-1}(y_i) - \tau^{-1}(\mu_i)) - b(\tau^{-1}(y_i)) + b(\tau^{-1}(\mu_i))\} \end{aligned}$$

In other words, the deviance measures the difference in log-likelihood between using $\mathbf{y}$ and $\boldsymbol{\mu}$ as the means. Note that $l(\mathbf{y}, \phi; \mathbf{y})$ is the maximum likelihood achievable for an exact fit in which the fitted values are equal to the observed data. As $l(\mathbf{y}, \phi; \mathbf{y})$ does not depend on $\boldsymbol{\mu}$, maximizing $l(\boldsymbol{\mu}, \phi; \mathbf{y})$ is equivalent to minimizing $D(\mathbf{y}, \boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$. (McCullagh & Nelder 1989, p.24, 33-34).

If $Y_i \sim G(r, \mu_i/r)$ and $w_i = 1\ \forall i$, the deviance becomes

$$D(\mathbf{y}; \boldsymbol{\mu}) = \sum_{i=1}^{N} \left\{ \frac{y_i}{\mu_i} - \log\left(\frac{y_i}{\mu_i}\right) - 1 \right\}$$

If $Z_i \sim N(a_i, b^2)$ and $w_i = 1\ \forall i$, the deviance simply becomes the sum of squares

$$D(\mathbf{z}; \boldsymbol{\mu}) = \sum_{i=1}^{N} (z_i - a_i)^2$$

If the dispersion parameter is known, the *scaled deviance*

$$D^*(\mathbf{y}, \boldsymbol{\mu}) = 2(l(\mathbf{y}, \phi; \mathbf{y}) - l(\boldsymbol{\mu}, \phi; \mathbf{y})) = \frac{D(\mathbf{y}, \boldsymbol{\mu})}{\phi}$$

will approximately follow a $\chi^2(N - m)$ distribution and can therefore be used as a measure of the goodness-of-fit of the model. If the dispersion parameter is not known (as in the analysis of this thesis) it can be estimated by

$$\widehat{\phi} = \frac{D(\mathbf{y}, \boldsymbol{\mu})}{N - m} \tag{3.4}$$

(Thyregod 1998*a*, p.219-220). It is obvious that in this case, the scaled deviance can not be used as a measure of the goodness-of-fit of the model. Instead the deviance is used to measure the difference in the goodness-of-fit of sub models relative to a given model.

3.1.4 Analysis of deviance

In order to investigate the parameters in a GLM the following fact can be used. If the hypothesis $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ is true then

$$\frac{\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sqrt{\phi}} \overset{\text{as}}{\sim} N_m(\mathbf{0}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\Sigma} = \left[\mathbf{X}^T \mathrm{diag} \left\{ \frac{w_i}{(g'(g^{-1}(\mathbf{x}_i\boldsymbol{\beta})))^2 V(g^{-1}(\mathbf{x}_i\boldsymbol{\beta}))} \right\} \mathbf{X} \right]^{-1}$$

The $\boldsymbol{\Sigma}$ matrix can be estimated by replacing $\boldsymbol{\beta}$ with $\widehat{\boldsymbol{\beta}}$ (Thyregod 1998*a*, 187). For the hypothesis $H_0:\ \beta_j = \beta_0$ the two sided test statistic

$$\left| \frac{\widehat{\beta}_j - \beta_0}{\sqrt{\phi \widehat{\sigma}_{jj}}} \right| \tag{3.5}$$

should be compared to the $(1 - \alpha/2)$ quantiles of $N(1, 0)$. If the dispersion parameter ϕ is unknown it can be estimated with (3.4) and the Student's t-distribution with $(N - m)$ degrees of freedom should be used instead. (Thyregod 1998*a*, p.284).

Usually, one is more concerned with the hypothesis that all parameters of a factor or all parameters for an n-degree polynomial can be assumed to be 0 at the same time. In other words, if a whole term can be left out from the model. Consider two hypotheses:

$$H_i : \ g(\boldsymbol{\mu}) = \mathbf{X}^i \boldsymbol{\beta}^i, \ \boldsymbol{\beta}^i \in \mathbb{R}^{m_i} \quad \text{and} \quad H_{i+1} : \ g(\boldsymbol{\mu}) = \mathbf{X}^{i+1} \boldsymbol{\beta}^{i+1}, \ \boldsymbol{\beta}^{i+1} \in \mathbb{R}^{m_{i+1}}$$

where $m_i > m_{i+1}$ and the columns of the matrix $\mathbf{X}^{i+1}$ are all contained in $\mathbf{X}^i$. Then the model H_{i+1} is said to be contained in model H_i, written $H_{i+1} \subset H_i$. Let $\boldsymbol{\mu}_i = g^{-1}(\mathbf{X}^i \widehat{\boldsymbol{\beta}}^i)$ and $\boldsymbol{\mu}_{i+1} = g^{-1}(\mathbf{X}^{i+1} \widehat{\boldsymbol{\beta}}^{i+1})$ where $\widehat{\boldsymbol{\beta}}^i$ and $\widehat{\boldsymbol{\beta}}^{i+1}$ are estimated maximum likelihood parameters from model H_i and H_{i+1} respectively. Tests for goodness-of-fit of these models relative to the full model are

$$G^2(H_i) = D^*(\mathbf{y}, \boldsymbol{\mu}_i) \overset{\text{approx}}{\sim} \chi^2(N - m_i) \quad \text{and} \quad G^2(H_{i+1}) = D^*(\mathbf{y}, \boldsymbol{\mu}_{i+1}) \overset{\text{approx}}{\sim} \chi^2(N - m_{i+1})$$

(Thyregod 1998*a*, p.288-290). In general, a hierarchically organized chain of models

$$H_{min} \subset H_k \subset \cdots \subset H_2 \subset H_1 \subset H_0 \subset H_{full} \tag{3.6}$$

can be considered. The minimal model, H_{min}, is the smallest possible model usually containing only an intercept $\boldsymbol{\eta} = \beta_0$, i.e. assuming that all observations have the same expected value. The full model, H_{full}, is the largest possible model containing one parameter for each observation, assuming that every observation have different expected values. A quotient test for the hypothesis of H_i given the smaller model H_{i+1} is

$$G^2(H_i | H_{i+1}) = G^2(H_i) - G^2(H_{i+1}) = D^*(\boldsymbol{\mu}_{i+1}, \boldsymbol{\mu}_i) \overset{\text{as.}}{\sim} \chi^2(m_i - m_{i+1}) \tag{3.7}$$

The chain in (3.6) usually denotes a sequential addition of terms. The test in (3.7) is then for the hypothesis that all coefficients of one term are zero (the term can be dismissed) given that all the terms already included have been accepted. (Thyregod 1998*a*, p.309).

When the dispersion parameter is unknown, the largest available model, H_0, will have to be assumed to be correct and successive testings of models are then done relative to that model. A quotient test for H_i given the smaller model H_{i+1} and relative to H_0 is then

$$\frac{G^2(H_i | H_{i+1})/(m_i - m_{i+1})}{G^2(H_0)/(N - m_0)} = \frac{D(\boldsymbol{\mu}_{i+1}, \boldsymbol{\mu}_i)/(m_i - m_{i+1})}{D(\mathbf{y}, \boldsymbol{\mu}_0)/(N - m_0)} \overset{\text{approx}}{\sim} F(m_i - m_{i+1}, N - m_0) \tag{3.8}$$

since the dispersion parameter is canceled out. (Thyregod 1998*a*, p.298-300). A general analysis of deviance table that gives these test statistics is shown in table 3.1.

3.1.5 Residuals and model control

Residuals are used to measure the difference between the observations and the fitted values given by the model. Let $\mathbf{y}$ be the observations and $\widehat{\boldsymbol{\mu}}$ the fitted values. The response residual is defined as:

$$r_i := y_i - \widehat{\mu}_i, \quad i = 1, 2, \ldots, N$$

and the deviance residual is defined as:

$$r_i^D := \text{sign}(y_i - \widehat{\mu}_i) \sqrt{w_i d(y_i, \widehat{\mu}_i)}, \quad i = 1, 2, \ldots, N$$

Source of variation	Deviance	Df	Deviance/Df	F-Test
From H_{min} to H_k	$D(\boldsymbol{\mu}_{min}, \boldsymbol{\mu}_k)$	$m_k - m_{min}$	$\frac{D(\boldsymbol{\mu}_{min},\boldsymbol{\mu}_k)}{m_k - m_{min}}$	$\frac{D(\boldsymbol{\mu}_{min},\boldsymbol{\mu}_k)/(m_k - m_{min})}{D(\mathbf{y},\boldsymbol{\mu}_0)/(N-m_0)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
From H_2 to H_1	$D(\boldsymbol{\mu}_2, \boldsymbol{\mu}_1)$	$m_1 - m_2$	$\frac{D(\boldsymbol{\mu}_2,\boldsymbol{\mu}_1)}{m_1 - m_2}$	$\frac{D(\boldsymbol{\mu}_2,\boldsymbol{\mu}_1)/(m_1 - m_2)}{D(\mathbf{y},\boldsymbol{\mu}_0)/(N-m_0)}$
From H_1 to H_0	$D(\boldsymbol{\mu}_1, \boldsymbol{\mu}_0)$	$m_0 - m_1$	$\frac{D(\boldsymbol{\mu}_1,\boldsymbol{\mu}_0)}{m_0 - m_1}$	$\frac{D(\boldsymbol{\mu}_1,\boldsymbol{\mu}_0)/(m_0 - m_1)}{D(\mathbf{y},\boldsymbol{\mu}_0)/(N-m_0)}$
Residuals	$D(\mathbf{y}, \boldsymbol{\mu}_0)$	$N - m_0$	$\frac{D(\mathbf{y},\boldsymbol{\mu}_0)}{(N-m_0)}$	
Total	$D(\mathbf{y}, \boldsymbol{\mu}_{min})$	$N - m_{min}$		

Table 3.1: A general analysis of deviance table when the dispersion parameter is unknown. The largest model, H_0, is assumed to be valid and successive testings of models are made relative to H_0. The terms are added sequentially and each model H_i is tested given that H_{i+1} is valid.

where $d(y_i, \widehat{\mu}_i) = 2\phi(l(y_i, \phi; y_i) - l(\widehat{\mu}_i, \phi; y_i))$. The deviance residual therefore measures the difference in log-likelihood. If Y is assumed to be normally distributed, these residuals become equal since $d(y_i, \widehat{\mu}_i) = (y_i - \widehat{\mu}_i)^2$. The standardized deviance residual is defined as

$$r_i^{Ds} := \frac{r_i^D}{\sqrt{\widehat{\phi}(1 - h_i)}}, \quad i = 1, 2, \ldots, N$$

Here h_i denotes the ith diagonal element of the 'hat' matrix

$$H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2}, \quad \text{where } W = \text{diag}\left\{ \frac{1}{(g'(\mu_i))^2 V(\mu_i)} \right\} \text{ is the}$$

matrix that maps $\mathbf{y}$ into $\widehat{\boldsymbol{\mu}}$. The standardized deviance residuals have constant variance. (Thyregod 1998*a*, p.202-215).

Informal checks of systematic departures from the model can be made by plotting the standardized deviance residuals versus the fitted values or a covariate. If there are no systematic departures from the model, the residuals on these plots have mean 0 and constant range. The plots can be used to spot systematic departures in the model like curvature in the mean and systematic change of range with fitted values or covariate values. Curvature of the mean may arise from several causes, including the wrong choice of the link function, wrong choice of scale of one or more covariates, omission of a quadratic term in a covariate or simply that the chosen covariates do not succeed to model the mean structure of the data.

A plot of absolute standardized deviance residuals against fitted values gives an informal check on the adequacy of the assumed variance function. An ill-chosen variance function will result in a trend in the mean. A positive trend indicates that the current variance function is increasing too slowly with the mean and a negative trend indicates the reverse effect. (McCullagh & Nelder 1989, p.398-400).

3.2 Locally weighted regression

Locally weighted regression is a part of a model class called ‘smoothers’. A smoother is a tool for summarizing the trend of a response measurement Y as a function of one or more predictor measurements, $X_1, \ldots, X_m$, similar to the classical linear models. But unlike the classical linear models, a smoother is nonparametric in its nature since it does not assume a rigid form for the dependence of Y on $X_1, \ldots, X_m$.

Let y_i, $i = 1, \ldots, N$, be measurements of the dependent variable and let $x_{i1}, \ldots, x_{im}$ be N measurements of m independent variables. Suppose that the data has the form

$$y_i = g(\mathbf{x}_i) + \epsilon_i$$

where the ϵ_i are independent normal variables with mean 0 and constant variance σ^2. In the GLM framework the function $g(\cdot)$ has to be a parametric function, but here there are no parameters and the only assumption of $g(\cdot)$ is that it is a smooth function in the independent variables (i.e. infinitely differentiable). However, the effective number of parameters or degrees of freedom is defined, see Hastie & Tibshirani (1990, p.52-55).

In the case of a locally weighted regression smoother, ‘loess’, an estimated surface is provided by fitting a function of the independent variables locally and in a moving fashion, analogous to how a moving average is computed for a time series. Let q be an integer, where $1 \leq q \leq N$. A neighborhood of $\mathbf{x}_0$ consists of the q observations whose $\mathbf{x}_i$ values are closest to $\mathbf{x}_0$. Each point in the neighborhood is weighted according to its distance from $\mathbf{x}_0$; close points have greater weight than points further away. A linear or quadratic function of the independent variables is fitted to the dependent variable using weighted least squares and the estimate $\widehat{g}(\mathbf{x}_0)$ is taken to be the value of this fitted function in $\mathbf{x}_0$. This procedure is repeated for all observations in the data.

There are of course several possible distance functions $\rho(\cdot, \cdot)$ that can be used to obtain the neighborhoods. In the case where the independent variables measures position in physical space (like in section 5.3) the Euclidean distance is a natural choice.

The weight function used with the loess smoothers is the tricube function:

$$W(u) = \begin{cases} (1-u^3)^3 & \text{if } 0 \leq u \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let $D(\mathbf{x}_0)$ denote the neighborhood of $\mathbf{x}_0$. The weight for observation y_i becomes:

$$w_i(\mathbf{x}_0) = W\left(\frac{\rho(\mathbf{x}_0, \mathbf{x}_i)}{\max\limits_{\mathbf{x}_i \in D(\mathbf{x}_0)} \{\rho(\mathbf{x}_0, \mathbf{x}_i)\}} \right) \tag{3.9}$$

Thus $w_i(\mathbf{x}_0)$, as a function of i, has a maximum for $\mathbf{x}_i$ close to $\mathbf{x}_0$. It decreases as the distance increases and becomes 0 outside the neighborhood.

The size of the neighborhoods is denoted by $f = q/N$ and is called the *scope*. As f increases, $\widehat{g}(\mathbf{x})$ becomes smoother. If locally linear fitting is used, the fitting variables are just the independent variables. If locally quadratic fitting is used, the fitting variables are the independent variables, their squares and their cross-products. Locally quadratic fitting tends to perform better in situations where the regression surface has substantial curvature, such as local maxima and minima.

The statistical properties of the loess smoother will not be discussed here. For more details of the loess smoother see Cleveland & Devlin (1988) and about smoothers in general see Hastie & Tibshirani (1990).

An additive model is defined by

$$Y_i = \alpha + \sum_{j=1}^{m} f_j(\mathbf{X}_j) + \epsilon_i$$

where the errors ϵ_i are independent of the X_is with mean 0 and constant variance. The f_j are smoothers, arbitrary functions of one or more predictors. This model is an extension of the classical linear model. It allows an arbitrary function of X instead of one that is linear in some parameters. It does not achieve an arbitrary regression surface as the terms are assumed to be additive. This restriction, however, makes statistical inferences about each individual term $f_j(\mathbf{X}_j)$ possible, since the variance of the fitted response holding all but one predictor fixed does not depend on the values of the other predictors. This means that the estimated functions $\widehat{f_j}$ can be plotted separately in order to examine the roles of the predictors in modeling the response. (Hastie & Tibshirani 1990, 82-87).

An additive model can be fitted by a back-fitting algorithm. It fits the smoothers f_j one at a time by taking the residuals

$$Y - \sum_{k \neq j}^{m} f_k(\mathbf{X}_k)$$

and smoothing them against X_j. The process is repeated until it converges. (Venables & Ripely 1997, p.327).

The generalized additive models extends the GLM in the same manner as the additive model extends the linear model. For detailed treatment of this subject see Hastie & Tibshirani (1990).

Chapter 4

Selecting the probability distribution function

This chapter deals with deciding which distribution, gamma or log-normal, is more appropriate for the cod catch data from the Icelandic groundfish surveys 1985-2001. The data are highly skewed as can be seen on the histograms in figure 4.1. Furthermore, the mean catch in these surveys is 153 cod but the median is only 53 cod. The histogram to the left in figure 4.1 contains all the positive cod catch data and because the data contains a few very large catches[1] almost all the data appear in one bar. The histogram on the right contains only tows that contained less than 400 cod and gives a better view of the distribution. A histogram of log transformed data is shown on figure 4.2. This histogram indicates that the normal distribution can possibly be used to describe the log transformed data, except that the lower tail is probably too thick. On the other hand, this skewness could be explained by a mean structure in the data.

The idea here is to divide the data into small homogeneous groups where the expected number of cod in tow can be assumed to be constant. If there are reasonably many observations in each group the mean and variance can be estimated for each group. These estimates can be used to investigate the variance function (section 4.1), the assumption of constant coefficient of variation (CV) (section 4.2) and the assumption of constant variance of log transformed data (section 4.3). Furthermore, estimated parameters from a generalized linear model (GLM) where these groups are treated as qualitative covariates (factors) can be used to test the goodness-of-fit of the two distributions (section 4.4).

The expected number of caught fish is assumed to be dependent on the environmental conditions in the sea and annual fluctuations of the size and catchability of the cod stock. Homogeneous groups of data can therefore be obtained by dividing the survey data into small areas. The sub-rectangles are well suited for these purposes since they are small (quarter of degree latitude times a half degree longitude) and the density of fish can therefore be assumed to be constant throughout that area. The drawback of using the sub-rectangles as homogeneous groups is that there are few observations in each group, the highest number is 7 observations in one year. Therefore, the statistical rectangles with up to 16 observations in one rectangle are also considered. But since they are four times bigger than the sub-rectangles, the assumption of homogeneity is not as reliable.

[1] The by far largest tow contained 14918 cod, performed in Skagafjörður 1985.

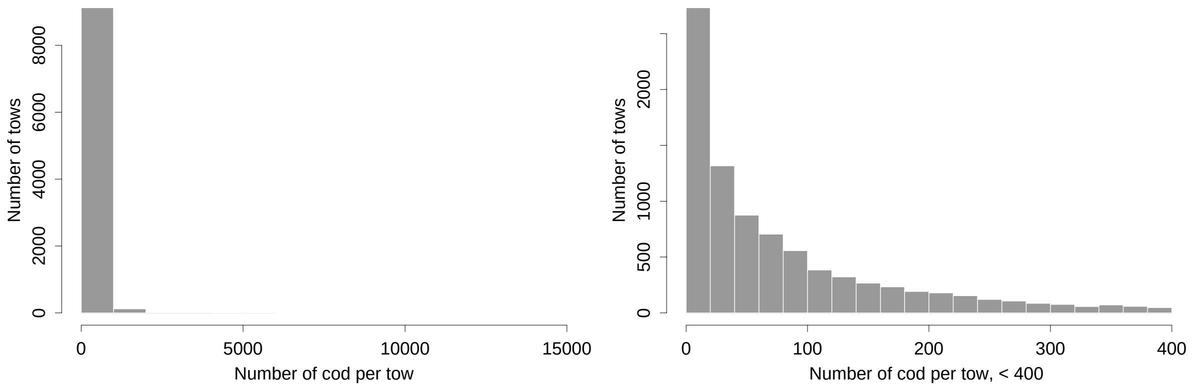

Figure 4.1: Histograms of number of cod caught in each tow. Each bar shows the number of tows that contain the amount of cod shown on the x-axis. On the left all data are included, a total of 9303 tows. On the right only tows with less than 400 cod caught are shown, a total of 8548 tows.

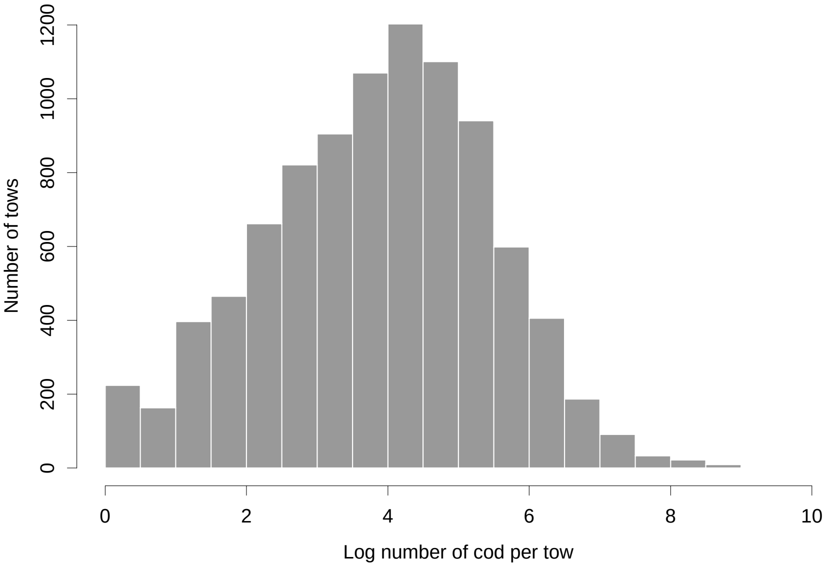

Figure 4.2: Histogram of log number of cod caught in each tow. All data are included.

4.1 The variance function

Both the gamma distribution and the log-normal distribution have the following relationship between the mean and the variance

$$\mathrm{var}(Y) = \phi \cdot E(Y)^2 \tag{4.1}$$

where ϕ is a constant. Furthermore, ϕ is the squared coefficient of variation (CV). The relationship of (4.1) can be written as:

$$\log(\mathrm{var}(Y)) = 2\log(E(Y)) + \log\phi$$

One way to determine whether the data can be assumed to fulfill the relationship of (4.1) is to plot the log sample variance, $\log(s_j^2)$, versus the log sample mean, $\log(\overline{y}_j)$, for sub-rectangles

$j = 1, 2, \dots, J$ and see if the points form a straight line with slope 2. Note that

$$\overline{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ji} \quad \text{and} \quad s_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(y_{ji} - \overline{y}_j)^2 \quad j = 1, 2, \dots, J \tag{4.2}$$

where n_j is the number of observations in sub-rectangle j and J is the total number of sub-rectangles. If a regression line $\log(s_j^2) = \beta_1 + \beta_2 \log(\overline{y}_j)$, using $1/n_j$ as weights and assuming Normal errors, is fitted to the data, the estimated slope $\widehat{\beta_2}$ will approximately follow a Normal distribution:

$$\widehat{\beta_2} \overset{\text{as}}{\sim} N\left(\beta_2, \frac{\sigma^2}{\sum_{j=1}^{J} \frac{\log(\overline{y}_j)^2}{n_j}}\right) =: N(\beta_2, \sigma^2_{\beta_2}) \tag{4.3}$$

(see equation (3.5)).

Figure 4.3 displays a scatter plot of $\log(s_j^2)$ versus $\log(\overline{y}_j)$ for sub-rectangles $j = 1, 2, \dots, J$. For every 17 years of the survey only sub-rectangles which contain 5 or more observations are included in this plot ($n_j \geq 5$), a total of $J = 241$ points. A weighted linear regression of the log variance on the log mean gave a slope of 2.23 ± 0.05 (mean ± standard error) and an intercept of -1.6 ± 0.2. The CV can be estimated from the intercept, $\widehat{\text{CV}} = \sqrt{e^{-1.6}} = 0.45$.

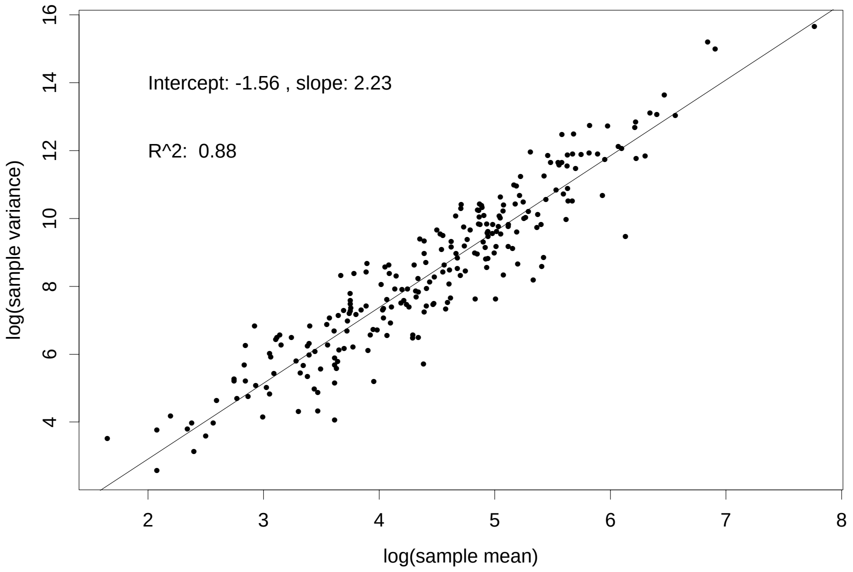

Figure 4.3: Scatter plot of $\log(s_j^2)$ versus $\log(\overline{y}_j)$ for every sub-rectangle j that has 5 or more observations. The regression line $\log(s^2) = 2.23\log(\overline{y}) - 1.6$ is also included.

Calculating the sample variance of only 5 observations does not give a very precise estimate of the variance. Therefore, the same analysis was done for statistical rectangles, see figure 4.4. Only rectangles which contained 10 or more observations were included in this plot ($n_j \geq 10$), a total of $J = 293$ points. A weighted linear regression of the log variance on the log mean gave the same slope as for the sub-rectangles, 2.23 ± 0.05, but a different intercept, -1.1 ± 0.3 indicating a slightly different CV, $\widehat{\text{CV}} = \sqrt{e^{-1.1}} = 0.58$.

It should be noted here that the weighting does not make any difference in this regression since the size of the groups does not vary much, 5-7 for sub-rectangles and 10-16 for statistical rectangles. Figure A.1 shows the regression without the weighting.

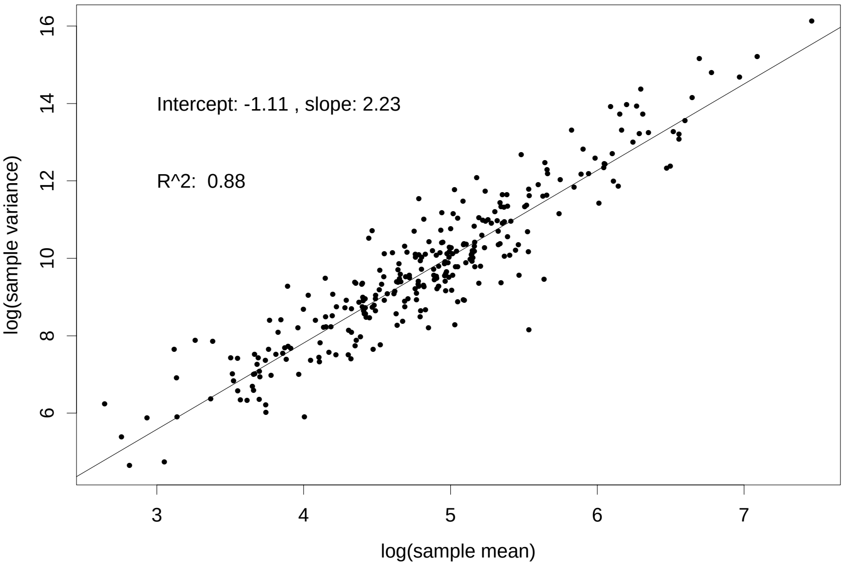

Figure 4.4: Scatter plot of $\log(s_j^2)$ versus $\log(\overline{y}_j)$ for every statistical rectangle j that has 10 or more observations. The regression line $\log(s^2) = 2.23 \log(\overline{y}) - 1.1$ is also included.

In both figures 4.3 and 4.4 the points are close to form a straight line and the regression has a R^2 value of 0.88. The two-sided test statistic (4.3) for the hypothesis $H_0 : \beta_2 = \beta_0 = 2.00$ is

$$\frac{\widehat{\beta_2} - \beta_0}{\widehat{\sigma_{\beta_2}}} = \frac{2.23 - 2.00}{0.05} = 4.60$$

Since the 0.975 percentile of $N(0, 1)$ is 1.96 this hypothesis is rejected on 5% significance level and the conclusion is that the slope is significantly higher than 2.

In the following analysis it will nevertheless be assumed that the relationship of (4.1) is valid for the data at hand and the gamma and log-normal distributions are proposed for describing the data. This assumption will simplify the analysis. A variance function of $V(\mu) = \mu^{2.23}$ does not define any known exponential distribution that could be used within the GLM framework. On the other hand, it is possible to estimate the parameters of the GLM by constructing a quasi-likelihood function with the variance function $V(\mu) = \mu^{2.23}$. This can be considered an avenue of future research. See McCullagh & Nelder (1989, chapter 9) for further discussion of quasi-likelihoods.

4.2 The coefficient of variation

For a generalized linear model where the response variable Y_i is gamma distributed it is assumed that the coefficient of variation

$$\mathrm{CV}_i = \frac{\sqrt{\mathrm{var}(Y_i)}}{E(Y_i)}$$

is constant for all i. An informal check of this assumption can be made with help of simulation.

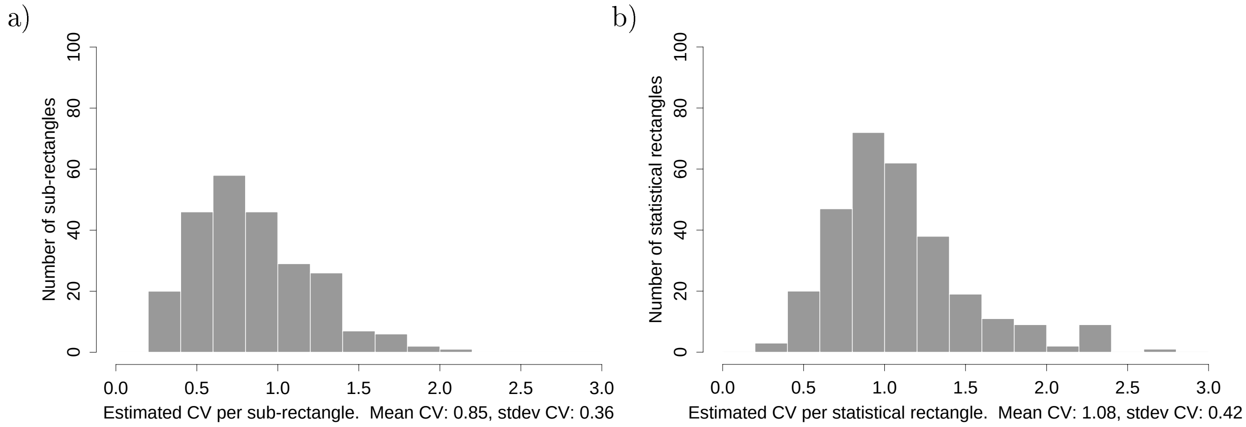

Figure 4.5: a) Histogram of estimated CVs, $s_j/\overline{y}_j$, for sub-rectangles that have 5 or more observations. The mean estimated CV is 0.85 and the standard error is 0.36. b) Histogram of estimated CVs, $s_j/\overline{y}_j$, for statistical rectangles that have 10 or more observations. The mean estimated CV is 1.08 and the standard error is 0.42.

Histograms of estimated CVs show that the CVs vary a lot. Figure 4.5 shows estimated CVs, $s_j/\overline{y}_j$, for sub-rectangles that have 5 or more observations (on the left) and for statistical rectangles that have 10 or more observations (on the right). It is difficult to tell whether CV is constant between sub-rectangles or statistical rectangles. In addition, the standard error is 40% of the mean and the difference between the highest and the lowest value is tenfold. Note that even though s^2 and $\overline{y}$ are central estimates of the variance and mean, the ratio $s/\overline{y}$ is not necessarily central estimator of the CV. The mean estimated CV obtained here, 0.85 and 1.08, are therefore not in contrast with the estimated values of 0.45 and 0.58 obtained in section 4.1.

Simulation was used to investigate whether the CVs estimated from the data vary more than could be expected of a gamma distributed variable. The idea is to simulate equally large groups as in figure 4.5 of gamma distributed random variables with mean 1 but the same CV. Since $s_j/\overline{y}_j$ is a biased estimator of the CV, a gamma distribution with CV equal to 0.85 and 1.08 will not produce comparable plots to those in figure 4.5. Therefore, the parameter r in the Gamma distribution $G(r, 1/r)$ was tuned until the mean estimated CV was close to 0.85 and 1.08. Figure 4.6 shows estimated CVs for 250 groups containing 5 observations from $G(1.05, 1/1.05)$ which can be compared to the CVs of sub-rectangle data plotted on the left of figure 4.5. Figure 4.7 shows estimated CVs for 300 groups containing 10 observations from $G(0.70, 1/0.70)$ which can be compared to the right hand side of figure 4.5.

In the case of sub-rectangles, the observed CVs do not seem to vary more than can be expected from a Gamma distributed variable. The simulations gave standard errors of 0.27-0.30 compared to observed value of 0.36. The simulations gave a highest value of 1.90 and a lowest of 0.22 but the highest observed value of the CV is 2.09 and the lowest is 0.20, which is within the bounds determined by the simulations. Furthermore, a Kolmogorov-Smirnov two sample test did not reject the hypothesis that the estimated CVs from data and the estimated CVs from the simulations come from the same probability distribution. The p-values of these tests are given in table 4.1.

In the case of statistical rectangles, the observed CVs can on the other hand not be assumed to be constant. The simulations gave standard errors 0.26-0.28 compared to observed value of 0.42. The highest observed value of the CV is 2.62 and the lowest is 0.22 which is a wider range than

Simulation no.	1	2	3	4
Sub-rectangles	0.1038	0.1686	0.2758	0.1080
Statistical rectangles	0.0129	0.0001	0.0155	0.0076

Table 4.1: p-values for Kolmogorov-Smirnov two sample test, comparing the estimated CVs from the data and estimated CVs from four simulations.

was found by the simulations, where the highest value is 2.25 and the lowest is 0.40. Furthermore, a Kolmogorov-Smirnov two sample test rejected the hypothesis that the estimated CVs from data and the estimated CVs from the simulations come from the same probability distribution. The p-values of these tests are given in table 4.1. This indicates that sub-rectangles are better suited as homogeneous areas than statistical rectangles.

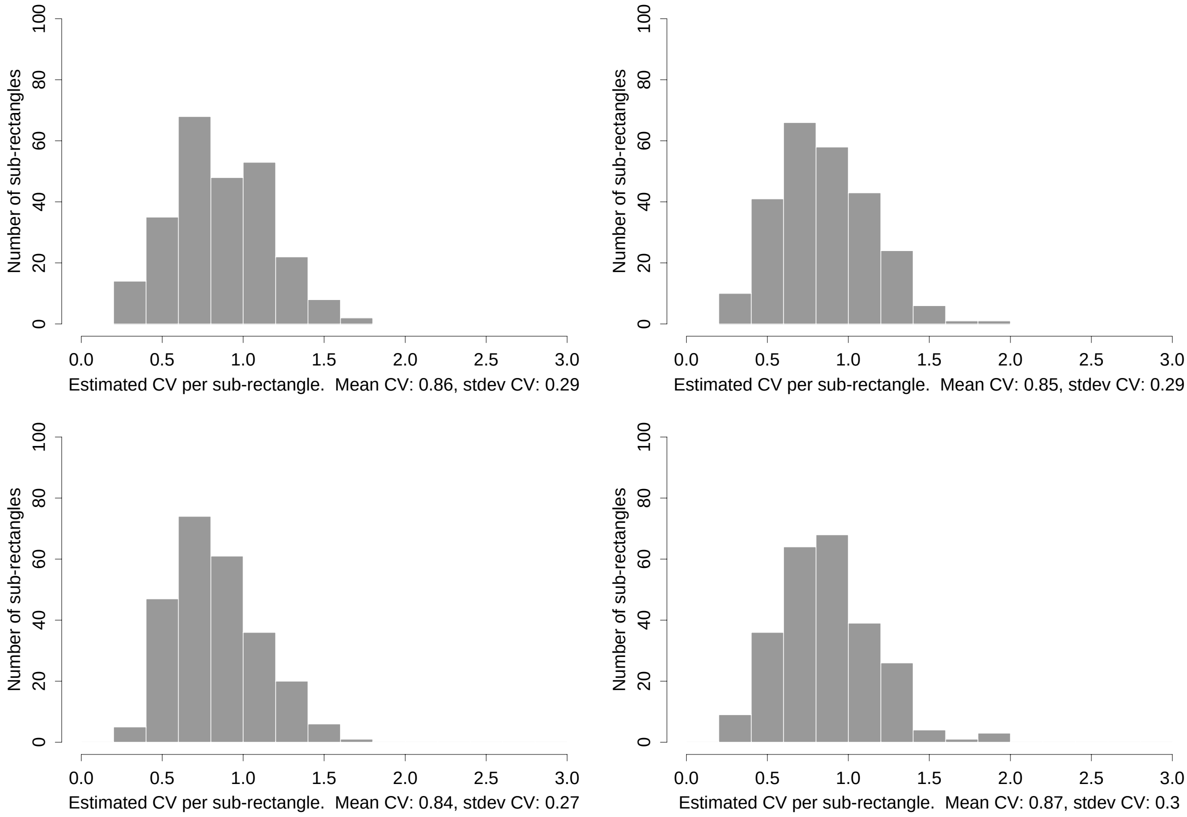

Figure 4.6: Histograms of estimated CVs, $s_j/\overline{y}_j$, for 250 simulated groups of 5 $G(1.05, 0.95)$ distributed random variables. Each graph shows a different simulation.

4.3 The variance of log transformed data

For a generalized linear model where the log transformed response variable $\log(Y_i)$ is normally distributed it is commonly assumed that the variance is constant for all i. A statistical test called Bartlett's test can be conducted in order to test this assumption.

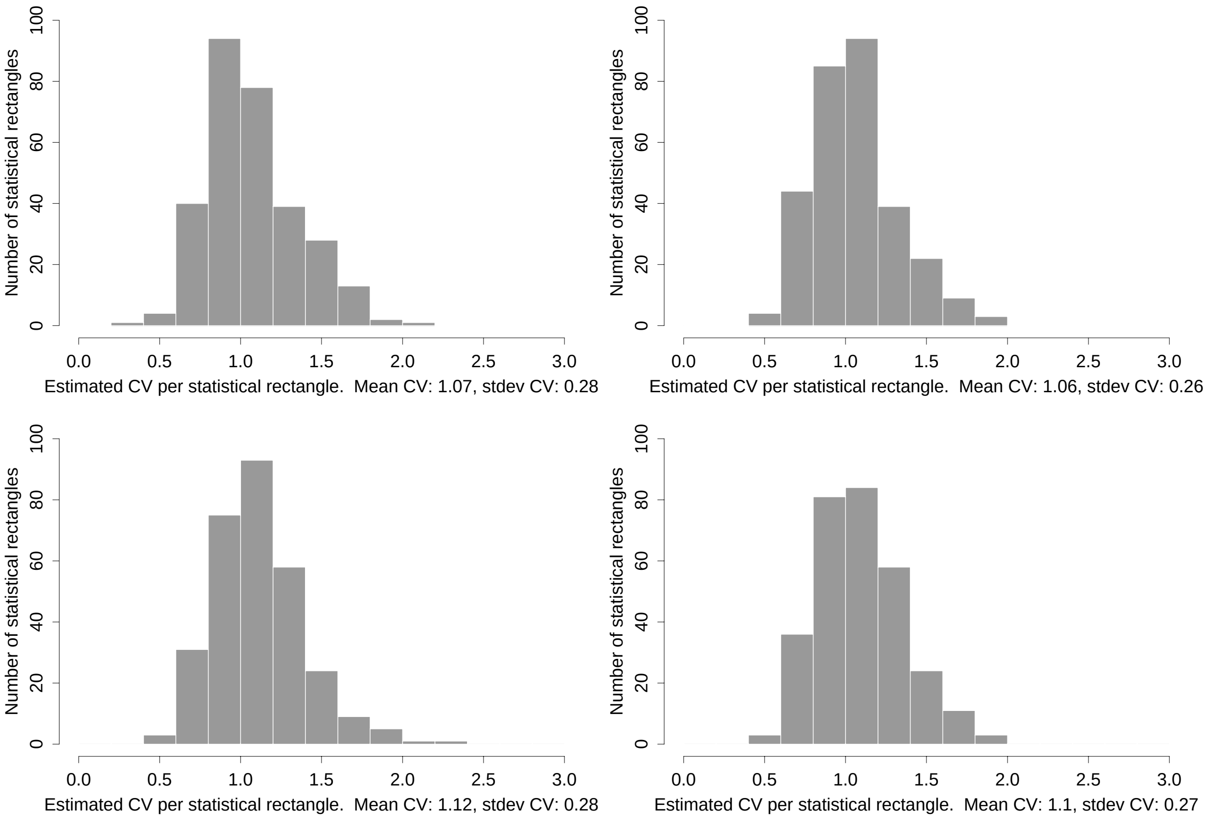

Figure 4.7: Histogram of estimated CVs, $s_j/\overline{y}_j$, for 300 simulated groups of 10 $G(0.70, 1.43)$ distributed random variables. Each graph shows a different simulation.

Under the assumption of normally distributed random variables the Bartlett's test can be used to test the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_J^2$. The test statistic is (Conradsen 1999, p. 482):

$$(N - J)\log(s_p^2) - \sum_{j=1}^{J}(n_j - 1)\log(s_j^2) \quad \sim \quad \chi^2(J-1)_{1-\alpha}$$

where

$$s_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(\log(y_{ji}) - \overline{\log(y_j)})^2 \quad \text{and} \quad s_p^2 = \frac{1}{N-J}\sum_{j=1}^{J}(n_j - 1)s_j^2\ .$$

The index $j = 1, 2, \ldots, J$ denote sub-rectangles or statistical rectangles. It should be noted here that this test is very sensitive to the normality assumption (Montgomery 2001, p.82).

The hypothesis of equal variances is rejected on a 5% significance level in both cases but the case of sub-rectangles is closer to acceptance. For sub-rectangles with 5 or more observations the test statistic is $398.1 > \chi^2(240)_{0.95} = 277.1$ and for statistical rectangles the test statistic is $631.7 > \chi^2(292)_{0.95} = 332.9$.

If the Bartlett test is conducted for each year separately the hypothesis is rejected on $\alpha = 5\%$ significance level in 8 years of 17. The test statistics in the case of sub-rectangles are shown in table 4.2.

As noted in chapter 2, the survey area is divided into main areas and sub areas (also called stratum) based on biological and hydrographical considerations. These considerations can contain

Year	Test stat.	*df*	$\chi^2(df-1)_{0.95}$		Year	Test stat.	*df*	$\chi^2(df-1)_{0.95}$	
1985	17.44	15	23.68	OK	1994	27.22	14	22.36	
1986	20.17	15	23.68	OK	1995	6.09	14	22.36	OK
1987	30.15	15	23.68		1996	34.31	14	22.36	
1988	30.19	14	22.36		1997	27.24	14	22.36	
1989	27.22	14	22.36		1998	16.27	11	18.31	OK
1990	23.21	15	23.68	OK	1999	14.40	13	21.03	OK
1991	25.33	15	23.68		2000	13.13	14	22.36	OK
1992	9.42	15	23.68	OK	2001	26.20	14	22.36	
1993	13.93	15	23.68	OK					

Table 4.2: Bartlett test statistics for equality of variance between sub-rectangles and 95% quantile of the corresponding χ^2 distribution for each year. The years where the hypothesis is accepted are marked with 'OK'.

Area	Test stat.	*df*	$\chi^2(df-1)_{0.95}$	
Stratum 1	251.52	151	179.58	
Stratum 2	40.55	26	37.65	
Stratum 3	16.59	17	26.30	OK
Stratum 6	73.36	34	47.40	
Northern area	386.10	228	263.15	
Southern area	9.67	13	21.03	OK

Table 4.3: Bartlett test statistics for equality of variance between sub-rectangles and 95% quantile of the corresponding χ^2 distribution for different areas. The areas where the hypothesis is accepted are marked with 'OK'.

reasons for the unequal variances. The density patterns and behavior of cod can vary between different survey areas. For example, young fish are known to be more abundant in the northern area and older fish are more abundant in the southern area (Stefánsson 1988). If the variances are found to differ between in different areas this could be incorporated in the model by introducing different weights for different areas, or simply by fitting separate models for each area. If the Bartlett test is conducted separately for the two main areas the hypothesis is accepted for the southern area which only contains one sub-rectangle. The test statistics are shown in table 4.3. In light of this it would be interesting to do the Bartlett test for the smaller strata in the northern area. The strata are shown on a map to the left of figure 4.8 and on the right the sub-rectangles that contain more than 5 observations are shown. Strata number 4 and 5 did not contain any sub-rectangles that contained more than 5 observations. Thus, the test was only conducted for strata number 1,2,3 and 6 and the test statistics are shown in table 4.3. The hypothesis is accepted in only one case, in stratum 3 where the data contain only one sub-rectangle.

Since the number of observations is almost the same in each sub-rectangle (174 sub-rectangles have 5 observations, 50 have 6 obs. and 17 have 7 obs.), F-tests for significant factors in the GLM model will only be slightly affected by the violation of the assumption of constant variances (Montgomery 2001, p. 80). Furthermore, the Bartlett's test is very sensitive to the normality assumption which can effect the test results. For further discussion of the effects of inequality of variance in regression analysis see Scheffé (1959, chapter 10).

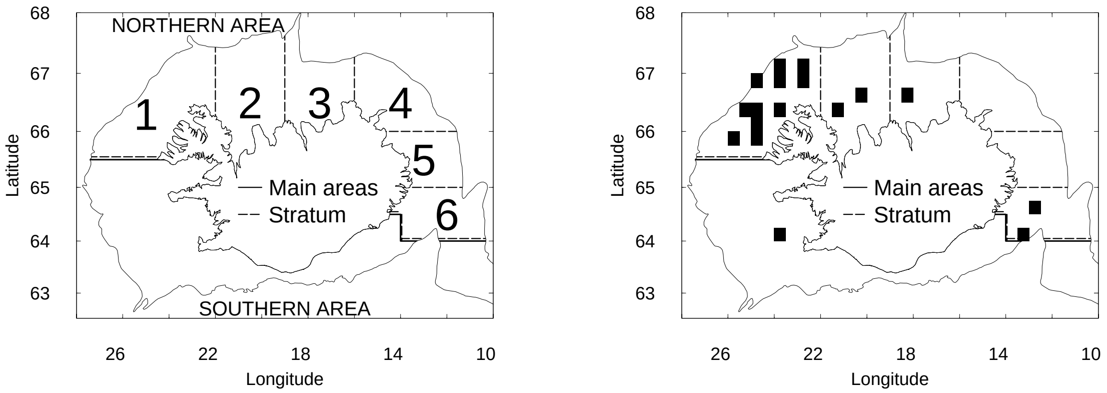

Figure 4.8: The survey area divided into main areas, northern and southern. The northern area is divided into 6 strata. On the map to the right the sub-rectangles that contain more than 5 observations are shown.

4.4 Goodness of fit

In this section a goodness-of-fit test with help of a generalized linear model is used to distinguish between the two proposed probability distributions, the gamma distribution and the log-normal distribution.

Let Y_i be a random variable that represents the number of cod caught in a tow, $i = 1, 2, \ldots, N$. It is assumed that either

$$Y_i \sim G(r, \mu_i/r) \quad \text{or} \quad \log(Y_i) \sim N(a_i, b^2)$$

The mean, μ_i or a_i, is assumed to be different between observations, depending on which sub-rectangle and which year the observation i comes from. The dispersion parameter on the other hand, $\frac{1}{r}$ or b^2, is assumed to be constant between sub-rectangles and years (see the investigation of these assumptions in sections 4.2 and 4.3). The effects of sub-rectangles and years are assumed to be multiplicative on the original scale of number of cod and hence additive on the log scale. This leads to the log link if Y_i is gamma distributed and the identity link if $\log(Y_i)$ is normally distributed.

A linear model where the mean, μ_i, is dependent of which sub-rectangle and which year the observation i comes from can be written in parametric form as:

$$\log(\mu_i) = \beta_0 + \sum_{y=1}^{A} \alpha_y \delta_y(i) + \sum_{j=1}^{J} \beta_j \delta_j(i) + \sum_{y=1}^{A} \sum_{j=1}^{J} \gamma_{yj} \delta_{yj}(i), \quad i = 1, \ldots, N \tag{4.4}$$

The parameter β_0 is the intercept (the grand mean), A is the total number of years ($A = 17$), J is the total number of sub-rectangles, N is the total number of observations and

$$\delta_y(i) = \begin{cases} 1 & \text{if observation } i \text{ comes from year } y, \\ 0 & \text{otherwise.} \end{cases}$$

$$\delta_j(i) = \begin{cases} 1 & \text{if observation } i \text{ comes from sub-rectangle } j, \\ 0 & \text{otherwise.} \end{cases}$$

$$\delta_{yj}(i) = \begin{cases} 1 & \text{if observation } i \text{ comes from year } y \text{ and sub-rectangle } j, \\ 0 & \text{otherwise.} \end{cases}$$

It is also useful to look at each year separately. Let Y_{yi}, $i = 1, 2, \ldots, n_y$, represent the number of cod caught in tow i in year y and let μ_{yi} denote the expected value of Y_{yi}. For a fixed year y the model becomes

$$\log(\mu_{yi}) = \beta_0 + \sum_{j=1}^{J} \beta_j \delta_j(i), \quad i = 1, \ldots, n_y \tag{4.5}$$

where n_y denotes the number of observations in year y.

The models in (4.4) and (4.5) were fitted separately for each of the proposed distributions. In order to get better estimates of the μ's, only data from sub-rectangles with 5 or more observations where included. In what follows, the fitted values and the estimated dispersion parameters are used to test the goodness-of-fit of the distributions.

If $Y_i \sim G(r, \mu_i/r)$ then

$$W_i = \frac{Y_i}{\mu_i} \sim G(r, 1/r), \quad i = 1, \ldots, N$$

This fact can be used to test the goodness-of-fit of the gamma distribution on the data at hand. Using the fitted values $\widehat{\mu_i}$ and the estimated dispersion parameter $1/\widehat{r}$ obtained from model (4.4), assuming gamma distributed errors, the hypothesis

$$H_0 : \widehat{W_i} = \frac{Y_i}{\widehat{\mu_i}} \sim G(\widehat{r}, 1/\widehat{r}), \quad i = 1, \ldots, N \ . \tag{4.6}$$

can be tested with the Kolmogorov-Smirnov test.

If $Z_i = \log(Y_i) \sim N(a_i, b^2)$ then

$$Z_i - a_i \sim \text{Normal}(0, b^2) \quad \text{and} \quad W_i = e^{Z_i - a_i} \sim LN(0, b^2) \quad i = 1, \ldots, N$$

This fact can be used to test the goodness-of-fit of the log-normal distribution on the data at hand. Using the fitted values $\widehat{a_i}$ and the estimated dispersion parameter $\widehat{b^2}$ obtained from the model (4.4), assuming normally distributed errors the hypothesis

$$H_0 \ \widehat{W_i} = e^{Z_i - \widehat{a_i}} \sim LN(0, \widehat{b^2}), \quad i = 1, \ldots, N \tag{4.7}$$

can be tested with the Kolmogorov-Smirnov test.

Both hypothesis (4.6) and (4.7) are rejected when data from all years are included in one model (4.4) but the log-normal distribution is closer to acceptance. The Kolmogorov statistic for the gamma distribution is $D_n = 0.075$ and $D_n = 0.056$ for the log-normal distribution but the 95% quantile of the distribution of D_n is $1.36/\sqrt{n} = 1.36/\sqrt{1289} = 0.038$. Figure 4.9 shows the cumulative distribution functions (CDFs) for the hypothesized distributions $G(1.07, 0.93)$ and $LN(0, 1.05)$ along with the empirical CDFs of corresponding W_i. On this graph the hypothesized distribution seem to fit well to the data but due to large number of observations (1289) the Kolmogorov-Smirnov test is rejected.

On the other hand both hypothesis (4.6) and (4.7) are accepted when data for each years are considered separately and model (4.5) is used. The hypothesized and empirical CDFs each year are shown in figures A.2 and A.3. In table 4.4 the p-values of the Kolmogorov test statistics are listed for each year. The log-normal distribution gives a higher p-value in all years except 1985 and 1987. Therefore, the log-normal distribution is preferred over the gamma distribution in the following analysis of this thesis.

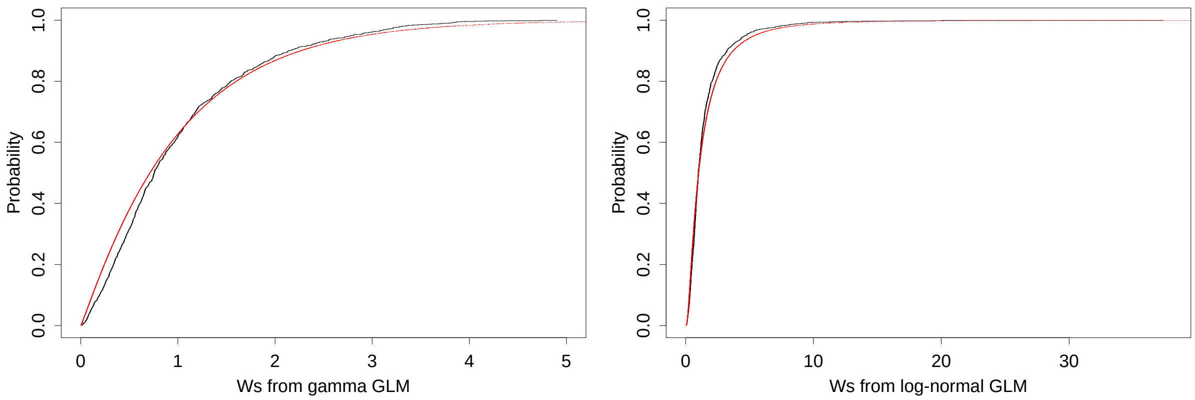

Figure 4.9: CDFs for the hypothesized distributions $G(1.07, 0.93)$ and $LN(0, 1.05)$ (in red color) along with the empirical CDFs of corresponding W_i (in black color). Although these graphs indicate a good fit the Kolmogorov test is rejected in both cases.

Year	gamma	log-normal	Year	gamma	log-normal
1985	0.3321	0.2869	1994	0.3267	0.3893
1986	0.1875	0.6959	1995	0.4359	0.7097
1987	0.3565	0.1982	1996	0.0938	0.2857
1988	0.5322	0.6667	1997	0.1725	0.4851
1989	0.3941	0.4948	1998	0.5321	0.8703
1990	0.4714	0.7693	1999	0.6767	0.9760
1991	0.3284	0.7975	2000	0.3034	0.6911
1992	0.3790	0.8302	2001	0.1214	0.4780
1993	0.3561	0.4042			

Table 4.4: p-values of the Kolmogorov test statistics per year. Only sub-rectangles with 5 or more observations are included. The log-normal distribution gives a better fit except in years 1985 and 1987.

4.5 Conclusions

Based on the goodness-of-fit tests performed in section 4.4 the log-normal distribution will be used in the remaining analysis of this thesis. There are, however, some unsolved issues that require further investigation.

First of all the variance of $\log(Y_i)$ can not be assumed to be constant, as shown in section 4.3. This will cause problems in drawing conclusions about the predicted values on the original scale:

$$\widehat{\mu_i} = e^{\widehat{a_i}+\frac{1}{2}\widehat{b^2}}, \quad \widehat{b^2} = \frac{D(\log(\mathbf{y}), \mathbf{x}_i\widehat{\boldsymbol{\beta}})}{N-J}.$$

The predicted values $\widehat{\mu_i}$ will be biased if $\widehat{b^2}$ is a biased estimate of $\text{var}(Y_i)$. And if the variance of Y_i is not constant, $\widehat{b^2}$ will be a biased estimate of $\text{var}(Y_i)$. However, as noted before, unequal variances will only slightly effect the F-tests for significant factors in a GLM. Since finding such factors is the main goal of this thesis this will not cause a serious problem here, but this bias has to be considered if the log-normal distribution is used in a catch forecasting model. It could, for example, be interesting to test the assumption of constant variances within the strata area with test methods that are more robust to departures from normality. Fitting a model separately for each area might prevent the problem of unequal variances from arising.

Secondly, the relationship between the observed variances and means of Y_i does not fully agree with the assumed variance function $\mathrm{var}(Y_i) = \phi E(Y_i)^2$. As shown in section 4.1, the power of $E(Y)$ is significantly higher than 2. Assuming a wrong variance function affects the calculation of the deviance $D(\mathbf{y}, \widehat{\boldsymbol{\mu}})$ and hence will affect the F tests for significant factors in a GLM. The question of how much this violation of the assumed variance function will affect the following analysis will not be investigated in this thesis.

A useful byproduct of this chapter is that sub-rectangles are found to be better suited as homogeneous areas than statistical rectangles since, according to the analysis in section 4.2 the CV can be assumed to be constant between sub-rectangles but not between statistical rectangles.

Chapter 5

Linear models for cod catch data

In this chapter, linear models for the cod catch data are investigated. Since the log-normal distribution was selected in chapter 4, the log transform of the number of cod in tow is considered here as the response (the dependent variable) and the errors are assumed to be normally distributed. The analysis of deviance therefore becomes just an analysis of variance, the deviance simply becomes the sum of squared errors.

In section 5.1, the model that was used in the goodness-of-fit process (4.4) is considered. That is a fully qualitative model using location (sub-rectangles) and time as main effects.

In section 5.2, several covariables available in the groundfish survey dataset are considered. A model containing a selected set of these covariates i investigated. That model is mostly quantitative with a single qualitative factor.

In section 5.3, locally weighted regression is used to estimate a temperature surface which is then used to estimate the size of the temperature gradient in each data point. This quantity is then considered as a new covariate in the model built in section 5.2.

5.1 A qualitative model

A generalized linear model (GLM) with sub-rectangles and years as qualitative factors is the most straightforward model available for the groundfish survey data. As argued in chapter 4, tows within one sub-rectangle in the same year can be assumed to have the same expected number of cod. The sub-rectangle factor represents a spatial effect, the effect of the habitation conditions of that area for the cod. These conditions are likely to be controlled by environmental effects such as depth, temperature, amount of food available etc. In section 5.2, an attempt will be made to substitute the sub-rectangle factor with several environmental variables that are recorded in the survey. The year factor represents the inter-annual fluctuations in the cod stock size and in catchability. An interaction between sub-rectangles and years represents the difference in variation between sub-rectangles for different years. In other words, the habitation conditions do not change in the same way between years in two different sub-rectangles.

5.1.1 The model

The log number of cod in tow i, $\log(\mu_i)$ is assumed to depend on the sub-rectangle and year of the ith tow:

$$H_0: \log(\mu_i) = \beta_0 + \sum_{y=1}^{A} \alpha_y \delta_y(i) + \sum_{j=1}^{J} \beta_j \delta_j(i) + \sum_{y=1}^{A} \sum_{j=1}^{J} \gamma_{yj} \delta_{yj}(i) + \epsilon_i, \quad i = 1, \ldots, N \tag{5.1}$$

The parameter β_0 is the intercept (the grand mean), A is the total number of years ($A = 17$), J is the total number of sub-rectangles, N is the total number of observations and

$$\begin{aligned}
\delta_y(i) &= \begin{cases} 1 & \text{if observation } i \text{ comes from year } y, \\ 0 & \text{otherwise.} \end{cases} \\
\delta_j(i) &= \begin{cases} 1 & \text{if observation } i \text{ comes from sub-rectangle } j, \\ 0 & \text{otherwise.} \end{cases} \\
\delta_{yj}(i) &= \begin{cases} 1 & \text{if observation } i \text{ comes from year } y \text{ and sub-rectangle } j, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

The errors ϵ_i are assumed to be normally distributed with mean 0 and constant variance. This is the same model as (4.4).

5.1.2 Results

The model (5.1) was fitted to cod catch data from the Icelandic groundfish surveys. In this analysis the dataset includes all sub-rectangles in a given year that have 3 or more observations. The total number of observations here is $N = 4802$ and there are $J = 94$ different sub-rectangles. The position of these sub-rectangles is shown in figure 5.1. Not all stations are visited every year of the survey so the number of different combinations of sub-rectangles and years are 1294 but not $94 \cdot 17 = 1598$.

The analysis of variance is shown in tables 5.1 and 5.2. In both tables, the terms are added sequentially (first to last) but the main effects are in reverse order. Successive testing of the models is done relative to H_0. This model accounts for 63.4% of the total variation using $1293/4801 = 26.9\%$ of the total degrees of freedom.

Source of variation	Df	SS	% explained	SS/Df	F-Test	p-value
Sub-rectangles	93	3771.9	37.2	40.56	38.31	0
+ Years	16	284.6	2.8	17.79	16.80	0
+ Interaction	1184	2380.5	23.5	2.01	1.90	0
Total model	1293	6437.0	63.4			
Residuals	3508	3713.8		1.059		
Total	4801	10150.7				

Table 5.1: Analysis of variance table for the qualitative GLM of log transformed cod catch data. The terms are added sequentially (first to last). 63.4% of the total variation is explained by this model. Both the main effects of sub-rectangles and years are significant and so is the interaction between them.

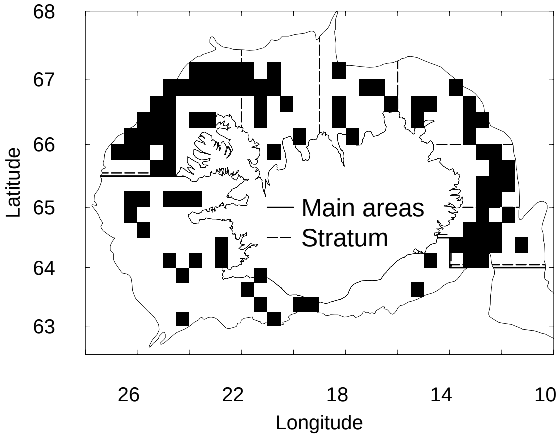

Figure 5.1: The survey area divided into main areas, northern and southern. The northern area is divided into 6 strata. The sub-rectangles that contain more than 3 observations are shown on the map.

Source of variation	Df	SS	% explained	SS/Df	F-Test	p-value
Years	16	313.3	3.1	19.58	18.50	0
+ Sub-rectangles	93	3743.1	36.9	40.25	38.02	0
+ Interaction	1184	2380.5	23.5	2.01	1.90	0
Total model	1293	6437.0	63.4			
Residuals	3508	3713.8		1.059		
Total	4801	10150.7				

Table 5.2: Analysis of variance table for the qualitative GLM of log transformed cod catch data. The terms are added sequentially (first to last). The main effects are in reverse order from table 5.1.

Both the main effects of sub-rectangles and years are significant and so is the interaction between them. Different order of main effects only slightly changes the sum of squares. This indicates that these effects are nearly orthogonal, i.e. no parts of the spatial effects of sub-rectangles can be explained with a time effect and vice versa. Most of the explained variation comes from the sub-rectangle effect, which explains about 37% of the variation, while the year effect explains only about 3%. The interaction between sub-rectangles and years explains a lot of variation, or 23.5%, but also uses the majority of the total degrees of freedom of the model.

Figures 5.2 - 5.4 plot the standardized residuals that can be used as an informal model control. Figure 5.2 shows scatter plots of standardized residuals and absolute standardized residuals versus the fitted values, which are satisfactory on the whole satisfactory. The residuals should not show any structure and they should have mean zero and unit variance. The red lines show the mean values and variances, calculated for groups of 50 residuals. The plot of standardized residuals (upper right hand corner) shows no curvature in the mean. The variance, although not entirely stable, is generally close to unity and furthermore there is no obvious structure in the variances. On this plot, parallel lines can be detected. The data (number of cod) is discontinuous so when response residuals, $y - \widehat{\mu}$, are plotted against fitted values $\widehat{\mu}$, the contours of fixed y will be parallel straight lines with a slope of -1 (McCullagh & Nelder 1989, p.399). These lines are therefore perfectly normal, they simply denote observation where 1, 2, 3, etc. cod were caught, the standardization has only skewed them a little. On the plot of absolute standardized residuals (lower right hand corner) there is no obvious curvature in the mean. This indicates that the variance function is adequate for these modeling purposes. Furthermore, these plots do not show a strong structure in range, except perhaps that for fitted values 0-1, where the range is smaller than for other fitted values.

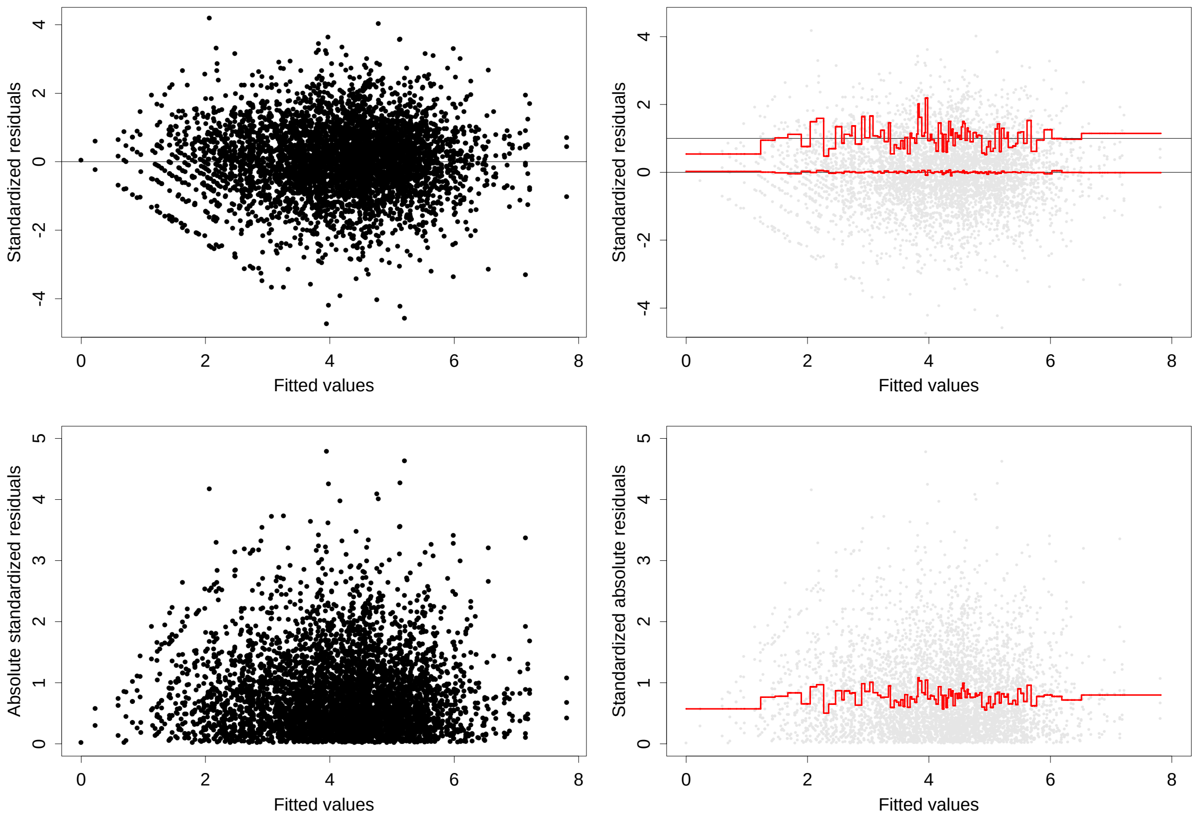

Figure 5.2: Scatter plots of standardized residuals (upper row) and absolute standardized residuals (lower row) versus the fitted values. The residuals should be without structure, with mean zero and unit variance. The red lines on the plot in the upper right hand corner show the mean values and variances, calculated for groups of 50 residuals, and the line on the plot in the lower right hand corner show the mean values, also calculated for groups of 50 residuals.

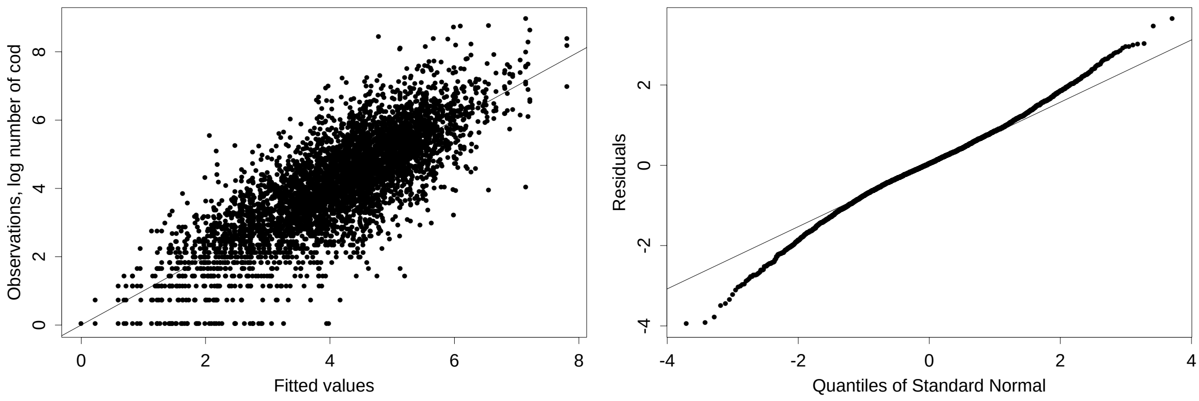

Figure 5.3: On the left: Scatter plot of observations versus fitted values along with the $y = x$ line. On the right: Normal probability plot of residuals.

The plot to the left in figure 5.3 shows observations versus the corresponding fitted values. These points should be scattered around the $y = x$ line, which is also shown on the plot. The plot is satisfactory although it reveals that the model fails to fit the lowest and the highest observed values. The plot to the right in this same figure shows a normal probability plot of the residuals. The residuals fail to be normally distributed. It seems like the tails, especially for low residual values, are thicker than could be expected for a normally distributed variable.

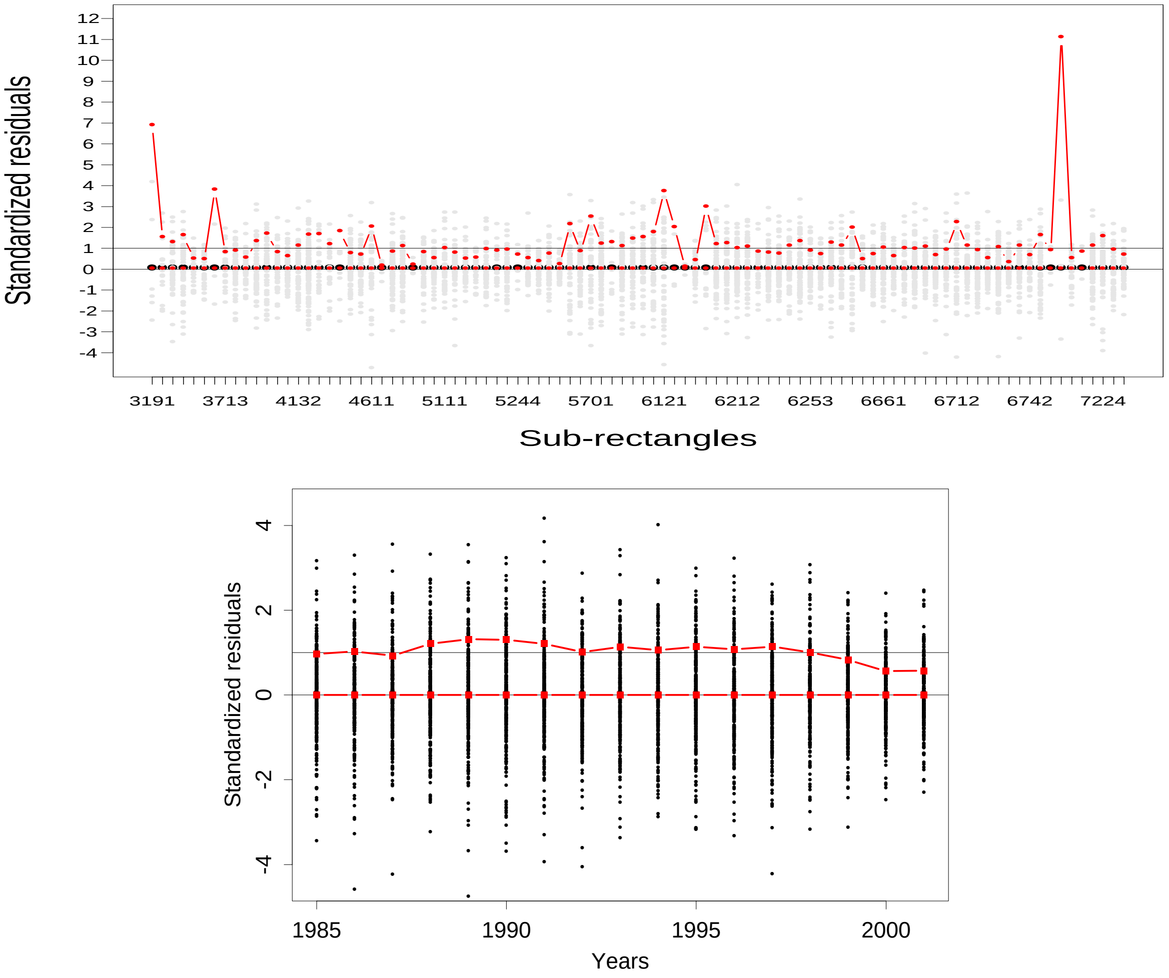

Figure 5.4: Standardized residuals (black points) of the qualitative GLM per sub-rectangle and per year. The mean and variances for each level are shown with a red square and are connected by a line for clarity.

Figure 5.4 shows the standardized residuals per sub-rectangle and years. As for figure 5.2 the residuals should be without structure, with mean zero and unit variance. The mean and variances are shown with a red square and are connected with a line. Actually, the residuals on these plots will automatically have mean zero because the fitted values are simply the average catch in the corresponding sub-rectangle and year[1].

The variances vary a lot between sub-rectangles and can not be considered to be constant, which is in analogy to the test result in section 4.3. However, there is no systematic structure in these

[1] And since $\overline{\epsilon_i} = \frac{\sum_{i=1}^{n} \epsilon_i}{n} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})}{n} = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\sum_{i=1}^{n} \overline{y}}{n} = \overline{y} - \frac{n\overline{y}}{n} = 0$.

differences or in the range of the residuals. This could simply be because the numeric order of sub-rectangles has no physical meaning. In figure B.1 the sub-rectangles are grouped according to strata. This does not reveal any obvious structure within strata but there seem to be different variances between strata. The different variances within strata can possibly be due to different conditions within sub-rectangles. There could, for example, be more rapid changes in depth or temperature in some sub-rectangles than in others, which can result in more variable catches in these sub-rectangles.

The variances are more stable between years, but they show an interesting trend. In years 1988 to 1991 the variance is higher than for the years before and after, there is no obvious change in range however. Another change in variances occurs in years 1999 to 2001, the variance becomes 50% lower than before and the range of residuals seems to be smaller for these years.

5.1.3 Conclusions

The analysis of variance shown above confirms that this simple model containing only location and time describes the cod catch data quite well. But the various residual plots also shown above reveal some shortcomings of the model. The variance of the residuals can not be assumed to be constant between sub-rectangles or strata, there is some unexplained trend in the variance and the range of residuals between years and the residuals can not be assumed to be normally distributed.

The departures from normality are of little concern because the analysis of variance is robust to the normality assumption, i.e. the F-test are only slightly affected (Montgomery 2001, p.77).

The difference in the variances of the residuals is another matter. According to Montgomery (2001, p.80) the F-test is only slightly affected in a balanced fixed effects model (equal number of observations for each level). The years have nearly equal sample sizes so the trend observed in the residual variances between years is not a serious problem for the analysis above. However, in unbalanced designs or in cases where one variance is very much larger than the others the problem is more serious. If factor levels having the larger variances also have smaller sample sizes, the significance levels are higher than anticipated. Conversely, if factor levels with larger variances also have the larger sample sizes, the significance levels are smaller than anticipated. Both of these cases occurs for levels of sub-rectangles. Whether the significance levels of the F-tests for the effect of sub-rectangles on the analysis above is seriously affected is hard to tell.

All things considered, it can be concluded that the majority of the variation in the cod catch data can by explained be a spatial effect and a time trend. This spatial effect will be investigated further in the next section.

5.2 A quantitative model

Although the model in 5.1 fits the data quite well and explains the majority of the variation in the data it comes with a high cost. The high number of parameters makes the model very unpractical. Furthermore, the main effects, sub-rectangles and years, do not provide other information than that the expected cod catch depends on location and time. A more informative model is one that relates some environmental variables to the expected catch, variables that can on biological grounds be expected to effect the behavior of cod for example. These environmental covariates can be thought of as substitutes for the sub-rectangles effect as they explain why the fish is more likely to be at one place rather than another.

5.2.1 Selecting covariates

The advantage of using the groundfish survey data collected by the MRI is that it contains many recorded measurements for every tow, which can be tested within the GLM framework. Even so, it can not be guaranteed that this dataset contains all the relevant factors. (It would for example be interesting to investigate if there is a correspondence between the tides and the expected catch.) But in this thesis, only the records in the groundfish survey database will be considered.

Figures 5.5-5.9 and B.2-B.3 display box plots of the log number of cod in tow versus several possible covariates. The shaded boxes show the middle 50% of the data (from the 25% to the 75% percentile) and the white small box shows the mean value. Dotted lines are drawn to the extreme points but are not made longer than 1.5 times the height of the shaded box and data outside that range are shown with a horizontal line. The width of the boxes is made proportional to the square root of the number of observations for the box. These plots can be examined in order to see how the mean value of log number of cod is affected by the covariate in question, i.e. they help examine the mean structure of the cod catch data. Such plots can of course only show one covariate at a time and are therefore not helpful when the combined effect of several factors are examined. However, the plots can give a good hint of which covariates are likely to explain the variability in the data and help suggesting the relationship between the response and the covariate.

The environmental measurements that are most likely to being able to substitute the spatial effect of sub-rectangles are depth and sea temperature as they directly effect the habitat conditions for the fish in the sea. Other measurements that might also have such spatial effects and are considered here are wind speed and direction, wave height, amount of clouds and a weather factor. The position parameters, latitude and longitude, can then be used to represent what remains of the spatial effect. The year effect is also considered here, but in order to reduce the number of parameters of the model it is considered to be a quantitative covariate. Variables that effect the efficiency of a single tow are trawl station data like the towing time, towing length and the vessel used. These variables are supposed to be constant in the groundfish survey but need to be considered all the same.

Depth and temperatures

The depth is measured in the groundfish survey in both the beginning and end of each tow. To get one representative depth quantity for each tow, the mean value of those two is used in this analysis. A box plot of log number of cod in tow versus depth is shown in the upper left hand corner of figure 5.5. From this plot it can be concluded that big cod catches are more likely at depth below 50 meters than in more shallow places, but when the depth is below 400 meters the expected catch becomes low again. On the whole, the log number of cod seems to be related to depth squared and a second degree polynomial in depth is therefore included in the model.

The measurements of 500 meters and higher do not fit into this polynomial form but since there are so few observations they can not be expected to reveal the true variation for these depth quantities.

Figure 5.5 also shows box plots of the log number of cod versus the bottom temperature, surface temperature and air temperature. The optimal bottom temperature for cod catching seems to be between 0 and 4°C. The expected catch is lower for bottom temperatures over 4°C and under 0°C even though there are not many observations for bottom temperature under 0°C to confirm this. On the whole, the log number of cod in tow seems to be related to the bottom temperature squared and a second degree polynomial is therefore included in the model.

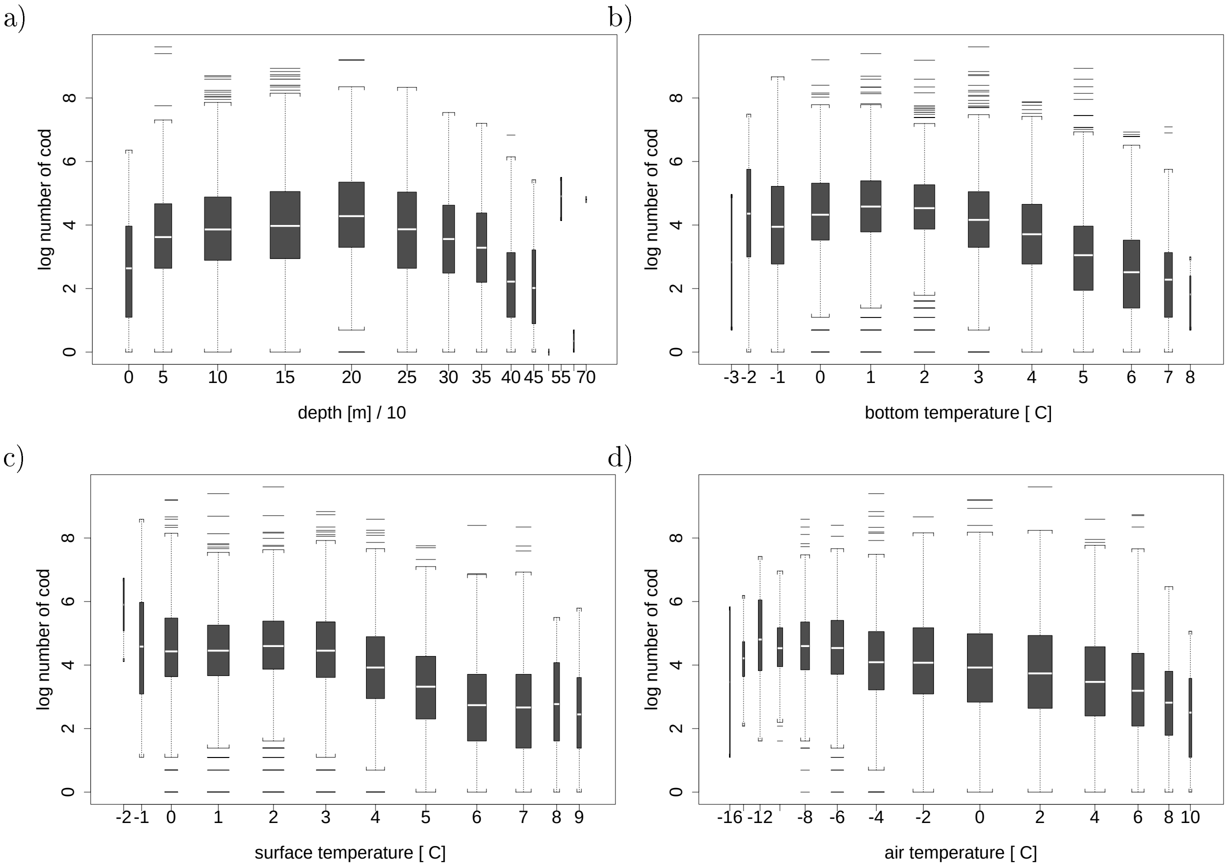

Figure 5.5: Box plot of log number of cod in tow versus a) depth, b) bottom temperature, c) surface temperature and d) air temperature. The temperatures (measured in °C) are rounded down to the nearest integer and the depth measurements are divided into intervals of length 50 meters by the transform floor(depth/50). Number of missing records are a) 150, b) 797, c) 590 and d) 1141. a) The response seems to be dependent on the depth squared. b) The response seems to be dependent of the bottom temperature squared. c) The response decreases with increasing surface temperature. d) The response decreases with increasing air temperature.

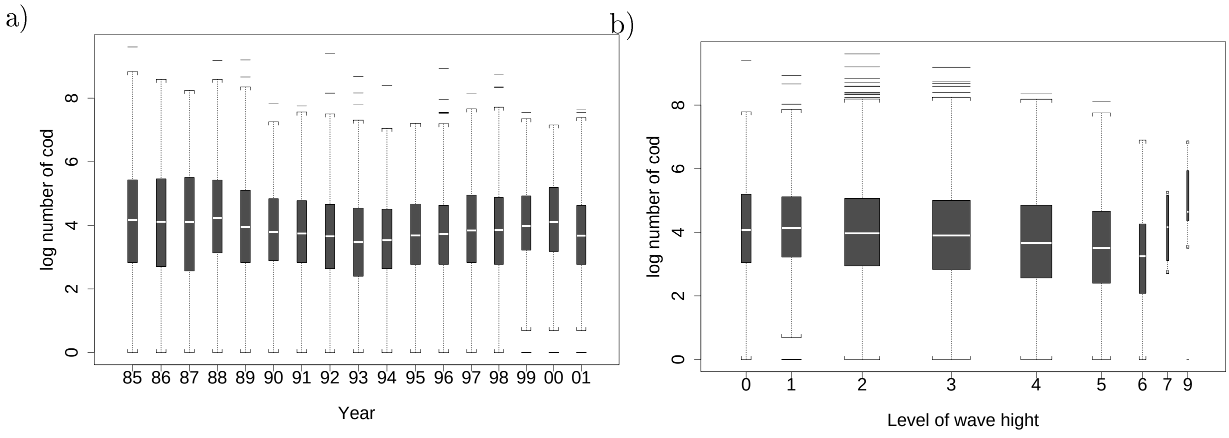

Figure 5.6: Box plot of log number of cod versus the a) years and b) increasing levels of wave height. a) The relationship between the response and the years seems to fit a 4-degree polynomial. b) Except for values of 7 and 9 the response seems to be dependent on the squared wave height level.

The log number of cod in tow seems to fall with increasing surface temperature. Actually the response is rather stable for the temperatures −1°C to 3°C, then falls between 3°C and 6°C, and becomes stable again for surface temperatures over 6°C. It is not obvious how to capture this trend so the surface temperature is included just as a single quantitative covariate in the model. The log number of cod decreases with increasing air temperature except perhaps for the few observations of −16°C to −13°C. The air temperature is therefore included as a single quantitative covariate in the model.

It would be rational to assume that there is a positive correlation between these three temperature variables, i.e. if the surface temperature is low then the air temperature would be expected to be low also and vice versa. Therefore it is likely that whilst one temperature has been included in the model the others will not have significant effects on the cod catch.

Wave height and other spatial variables

To the right of figure 5.6 is a box plot of the log number of cod versus wave height. The scale is not linear and the levels 0 to 9 represent increasing wave height as follows:

Level	0	1	2	3	4	5	6	7	8	9
Wave height [m]	0	0-0.1	0.1-0.5	0.5-1.25	1.25-2.5	2.5-4	4-6	6-9	9-14	Not rec.

If the values of 7 and 9 are excluded, the response seems to decrease with increasing wave height. Because the mean value for level 1 is a little higher than for level 0, the response could be dependent on the wave height squared and a second degree polynomial of the wave height is therefore included in the model. Naturally, records with the value 9 are excluded from the analysis since they represent missing records. Records the value 7 and higher are also excluded for the following reason: The survey is not supposed to be continued in such bad weather.[2] Given

[2] The vessels are supposed to stop trawling when the wind force exceeds 8 on the Beaufort scale (see section

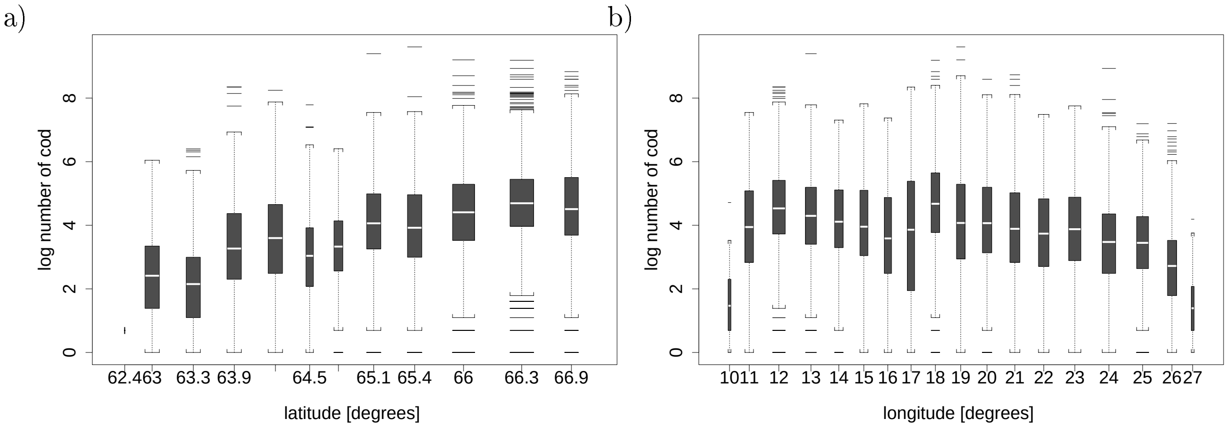

Figure 5.7: Box plot of log number of cod in tow versus a) latitude (from south to north) and b) longitude (from east to west). a) The latitudes are divided into intervals of length 0.3 degrees by the transform floor(latitude/3000). The relationship between the response and the latitude seems to fit a 4 degree polynomial. b) The longitudes are divided into intervals of length 1 degree by the transform floor(longitude/10000). The relationship between the response and the longitude seems to fit a 2 or 4 degree polynomial.

that the weather is that bad, the expedition directors are more likely to continue fishing if they are getting big catches. That makes a bias in the dataset as there are only big catches recorded in such bad weather.

The spatial variables wind speed, wind direction, weather factor, cloud factor and barometric pressure are shown on figures B.2, B.3 and B.4. Figure B.4 also shows the time at the beginning of tows. These covariates are not included in the model but some explanations and considerations can be found in Appendix B.2.

Latitude, longitude and years

For each tow in the survey, the latitude and longitude is recorded for the position of both the beginning and the end of each tow. The record is a 6 digit number. The first two digits represents the degree, the two in the middle represent the minutes (0-59) and the last two a hundreds part of a minute. The minutes where transformed to the interval of $[0, 99]$ to get a continuous 6 digit number for position. To get one representative two-dimensional position variable, the mean values of both the latitude and longitude are used.

Box plots of the log number of cod in tow versus latitude and longitude are shown on figure 5.7. The general trend on the latitude plot is that cod is more likely to be caught in the northern survey areas (high values of latitude) than in the southern survey area. But this trend is not smooth enough to be described with a straight line, a 4 degree polynomial would be more appropriate. The trend is not as obvious on the longitude plot, but cod seem to be equally likely to be caught in the eastern survey areas (low latitude values) as in western areas. An interesting peek occurs around 18 degrees (ca. the middle of the survey area). I should be noted here that at longitude 18 degrees there are considerably more stations located in the northern area than in

2.1) and according to the survey handbook (Einarsson et al. 2002) the expected wave height in such weather is over 5.5 meters.

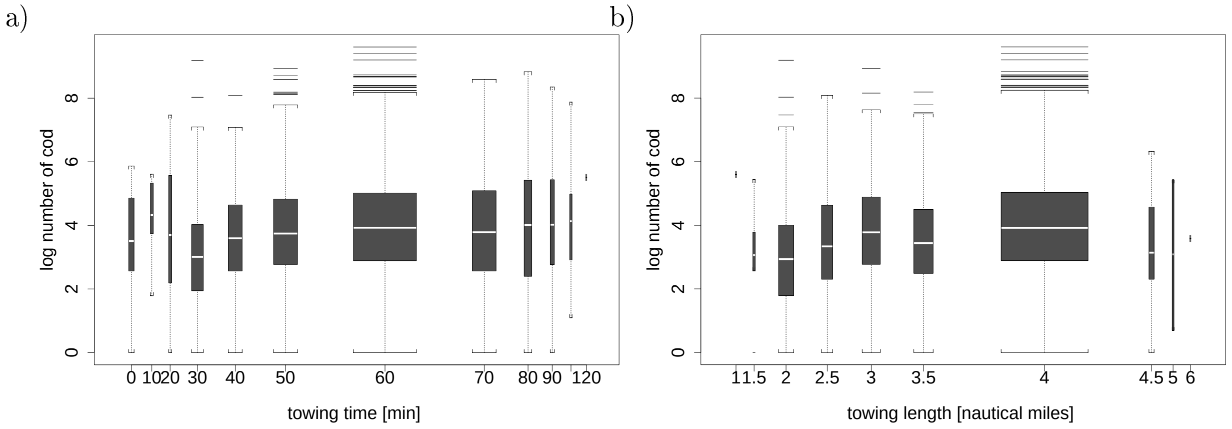

Figure 5.8: Box plot of log number of cod in tow versus a) towing time rounded down to the next whole multiple of 10 minutes and b) towing length in nautical miles rounded down to the nearest half integer. a) The response seems to be related to the towing time squared. b) The response seems to be related to the towing length squared.

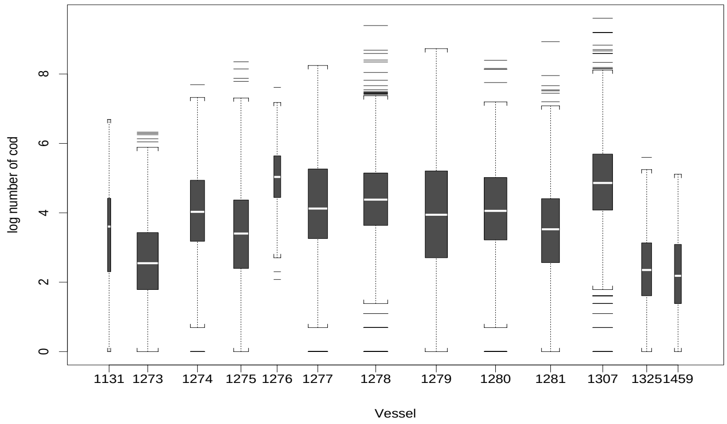

Figure 5.9: Box plot of log number of cod in tow versus the vessels. There seems to be a significant difference between the vessels.

the southern area. This peek is therefore just consistent with the latitude trend. On the whole, the relationship between the log number of cod and longitude can be described with a 2 or 4 degree polynomial.

A box plot of the log number of cod in tow versus the years is shown on the left side of figure 5.6. A 4 degree polynomial seems to fit the inter annual fluctuations for these years. It should be noted here that such a polynomial is not suitable for prediction of future years because it will simply predict a decreasing (or increasing) stock for all eternity. Since there was a significant interaction between sub-rectangles and years in the model in section 5.1 such an interaction can not be ignored in this model. Therefore, a 4 degree mixed polynomial in latitude, longitude and years is included in the model instead of one in the years and another one in latitude and longitude.

Trawl station data

Box plots of log number of cod versus towing time and towing length are presented in figure 5.8. The log number of cod seems to be related to the squared towing time. An exception of this is low towing times, but since the towing time should in fact be constant, the low towing times are probably mistakes in recording. Therefore towing times under 20 min are not included in this analysis. The log number of cod also appears to be related to the squared towing length. Second degree polynomials of both towing time and towing length (separate) are included in the model. In advance it would be expected that these two covariates are positively correlated and thus that once one has been included in the model, the other would not be found significant, it affects the the cod catch in the same way as the first covariate.

A box plot of the log number of cod in tow versus the vessels is shown on figure 5.9. The effects of a vessel on the log number of cod in tow seem to be significantly different. It is not obvious, though, whether this is due to different efforts of the vessels or simply different locations that each vessel was sent to in the survey. Nevertheless, a qualitative factor representing the different vessels is included in the model.

5.2.2 The model

In accordance to the considerations about possible covariates in the previous section, the following model will be investigated:

$$\begin{aligned} H_0 : \log(Y_i) \;=\;& \beta_0 + \text{poly}(\text{Depth}_i, 2) + \text{poly}(\text{Bottom temp}_i, 2) + \text{Surface temp}_i + \\ & \text{Air temp}_i + \text{poly}(\text{Wave height}_i, 2) + \text{poly}(lat_i, lon_i, \text{Year}_i, 4) + \\ & \text{poly}(\text{Towing time}_i, 2) + \text{poly}(\text{Towing length}_i, 2) + \text{factor}(\text{Vessel}) + \epsilon_i \end{aligned} \tag{5.2}$$

where $\log(Y_i)$ denotes the log number of cod in tow i, β_0 is an intercept and ϵ_i is assumed to be normally distributed with mean 0 and constant variance. The notation 'poly$(x_1, \ldots, x_p, n)$' denotes an orthonormal polynomial of degree n (the degree of combined terms does not exceed n) in the variables $x_1, \ldots, x_p$. There are mainly computational reasons for using orthonormal polynomials instead of ordinary polynomials (like $x + x^2 \cdots + x^n$) as it ensures that the model matrix will have full rank. The notation 'factor(Vessel)' denotes that the vessel effect is qualitative. This term could be written

$$\sum_{v=1}^{V} \beta_v \delta_v(i), \quad \text{where} \quad \delta_v(i) = \begin{cases} 1 & \text{if tow } i \text{ is performed by vessel } v, \\ 0 & \text{otherwise.} \end{cases}$$

in analogy with the notation of the model (5.1).

5.2.3 Results

The model (5.2) was fitted to cod catch data from the Icelandic groundfish surveys. In this analysis, all data where any of the covariates have missing records are naturally excluded. The total number of observations here is 7066 and their locations in the survey area are shown in figure 5.10.

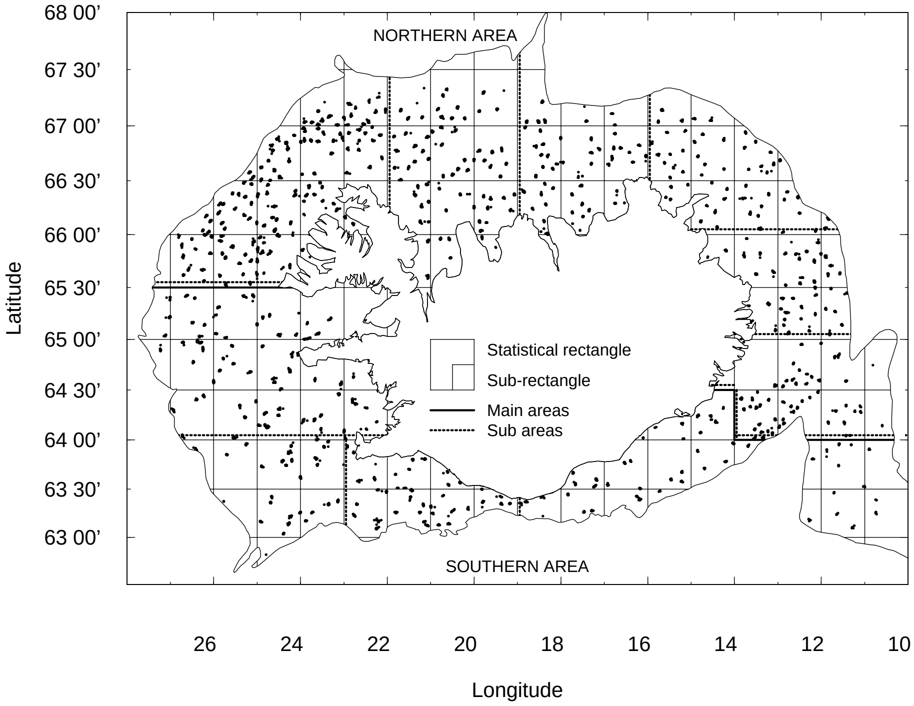

Figure 5.10: A map of the survey area showing the 500 meters contour line, main areas, sub areas and statistical rectangles. The points denote the (middle) locations of stations for all 7066 observations used in the quantitative model analysis.

The results of analysis of variance (ANOVA) is shown in table 5.3. The terms are added sequentially (first to last) and successive testing is done relative to H_0. Tables B.1 and B.2 show the same kind of ANOVAs but the terms are added in a different order. All covariates are significant although their contribution to variance explanation is quite different. The complete model H_0 explains 45.2% of the total variation.

Table 5.4 contains information on models that include only one of the terms in the original model. The fourth column (% explained) shows the percentage of the total variation that each term explains when it is the only (or first) term in the model, and the fifth column shows the residual sum of squares (RSS) of each model. The sixth column shows the Mallows' C_p statistic. This is defined as

$$C_p = \text{RSS} + 2\sigma^2 p$$

where p is the total degrees of freedom used by the model and σ^2 is the dispersion parameter, estimated from the original H_0 model. This C_p statistic is helpful for comparing different models. One criterion for a good model is a low residual sum of squares, i.e. a high proportion of the total variation explained, and another is a low number of parameters. Models with low C_p are therefore desirable.

Table 5.5 shows the change in the residual sum of squares (RSS) if one term is removed from the model, along with Mallows' C_p statistic. Tables B.3 to B.5 show t-tests for the hypothesis that one parameter of the model is zero.

A model containing only the 4-degree polynomial in latitude, longitude and year explains 41.1%

of the total variation (table 5.4). Such a model is comparable to the qualitative model (5.1) in the sense that both models include only spatial and time effects and interaction thereof. The polynomial model does not reduce the variation as much as the qualitative model but considering the number of degrees of freedom (34 versus 1293) it performs very well. Most of the t-tests for the parameters of the latitude, longitude and year polynomial (see table B.5) are rejected on 5% significance level. Since only a few of the tests for 4 degree terms are accepted on a 5% significance level, there are no grounds for reducing the degree of this polynomial.

The spatial terms, depth, the three temperature variables and wave height, do indeed explain a part of the spatial effects in the data but not all. These terms explain 27.3% of the variation (see table 5.3) and when the polynomial in latitude, longitude and year is added it reduces the variation by 16.8%. Altogether, these terms explain 44.1% of the variation, slightly more than the latitude, longitude and year polynomial alone so these terms explain something that the big polynomial can not (the extra 3%). But there is obviously room for improvement here, that is, to find more covariates that can explain what is left of the spatial effect of the 4 degree polynomial. One attempt to do this is made in section 5.3.

The temperature terms contribute by far the most of the spatial terms to variance reduction but the effects of these three terms are highly correlated. The bottom, surface and air temperatures explain 20.6%, 18.0% and 4.9% of the total variation when fitted separately (see table 5.4). When put together in one model the total variance reduction is only 23.14% (see table B.1) which means that the temperature effects are, not surprisingly, positively correlated. It does not make much difference whether the air or surface temperature comes first (table B.1 versus B.2), the remaining effect of surface temperature after the bottom temperature has been accounted for is a good 2% and under 0.5% for air temperature. The hypothesis that the parameter for air temperature are zero (see table B.4) can not be rejected and in table 5.5 this term is the only term that reduces the C_p (though it is almost not measurable). Due to this and the fact that air temperature contributes very little to the reduction of variance it can be excluded from the model. The surface temperature on the other hand with the good 2% of the total variation can not be disregarded. Even though the majority of the effect of surface temperature can be explained by the bottom temperature there is still some part left. This might be interpreted this way: In places where the bottom temperatures are equal, the variation in cod catch can partly be explained by surface temperature, perhaps because pelagic species like capelin which cod feeds on pursue certain surface temperatures.

The depth and wave height terms also make significant contribution to the variance reduction. The depth term explains 5.1% of the total variance when it comes as the first term (table 5.3). This decreases to 3.4% when the depth effect is tested after the temperature effects (see table B.2). That means that some of the depth effect can be explained by difference in temperature. Surprisingly, the sum of squares for the depth term actually increases when the depth effects is tested after the wave height effect (See table B.2). This means that there is some kind of negative correlation between the effects of these two terms, i.e. the latter term corrects some of the errors not captured by the former term. The t-tests in table B.4 show that the second order term in the wave height polynomial in not needed.

The vessel effect turns out to be mostly explained by the locational terms, i.e. these effects are mostly due to the fact that the vessels are sent to different survey areas. The vessel effect explains a lot of the total variation when fitted alone (21.0, see table 5.4). But when this effect becomes last in the model (table 5.3) this decreases to only 0.7% and it does not matter whether this effect comes before or after towing time and length (table B.1). This means that almost all the

Source of Variation	Df	SS	% explained	SS/Df	F-test	p-value
poly(Depth, 2)	2	893.9	5.1	447.0	324.0	0.00000
+ poly(Bottom temp., 2)	2	3471.1	19.7	1735.6	1258.0	0.00000
+ Surface temperature	1	290.9	1.6	290.9	210.9	0.00000
+ Air temperature	1	18.0	0.1	18.0	13.1	0.00030
+ poly(Wave height, 2)	2	137.5	0.8	68.7	49.8	0.00000
+ poly(Latitude, Longitude, Year, 4)	34	2961.8	16.8	87.1	63.1	0.00000
+ poly(Towing time, 2)	2	52.4	0.3	26.2	19.0	0.00000
+ poly(Towing length, 2)	2	15.5	0.1	7.7	5.6	0.00369
+ factor(Vessel)	12	131.3	0.7	10.9	7.9	0.00000
Total model	58	7972.4	45.2			
Residuals	7007	9666.7	54.8	1.380		
Total	7065	17639.2				

Table 5.3: Analysis of variance table for the quantitative GLM of log transformed cod catch data. The terms are added sequentially (first to last) and they are all significant. 45.2% of the total variation is explained by this model.

Term	Df	SS	% explained	RSS	Cp
<none>				17639	17642
poly(Depth, 2)	2	893.9	5.1	16745	16752
poly(Bottom temperature, 2)	2	3640.7	20.6	13999	14006
Surface temperature	1	3178.1	18.0	14461	14466
Air temperature	1	862.3	4.9	16777	16782
poly(Wave height, 2)	2	327.2	1.9	17312	17319
poly(Latitude, Longitude, Year, 4)	34	7252.4	41.1	10387	10469
poly(Towing time, 2)	2	218.4	1.2	17421	17428
poly(Towing length, 2)	2	218.3	1.2	17421	17428
factor(Vessel)	12	3698.9	21.0	13940	13971

Table 5.4: Results for fitting a GLM of log transformed cod catch data with only one term at a time.

difference in vessel effects seen on figure 5.9 is due to the fact that different vessels are sent to different survey areas, an effect that has already been explained by former terms in the model. This term is on the other hand found to be significant, t-tests for the vessel parameters being zero are all rejected on 5% significance level (except one, see table B.3) and dropping the term increases the C_p (table 5.5) so this term will not be omitted from the model.

The towing time and length contribute very little to the variance reduction (see tables 5.3 and 5.4), which is consistent with the fact that the towing speed is standardized in the survey. There is, not surprisingly, a strong positive correlation between the effects of these two terms. It does not matter whether towing time or length comes first (table B.1 versus B.2) the second term almost disappears and is close to being non-significant. In other words, towing time and towing length have the same effects on the expected number of cod in tow. Dropping one of them has almost no effect on the C_p statistic so one of these terms can be excluded from the model. Furthermore, t-tests for the parameters of these terms being zero are accepted on 5% significance level for both the towing time parameters but not the towing length parameters. It is therefore

Term	Df	SS	RSS	Cp
<none>			9667	9805
poly(Depth, 2)	2	247.7	9914	10048
poly(Bottom temperature, 2)	2	172.6	9839	9973
Surface temperature	1	12.0	9679	9815
Air temperature	1	2.5	9669	9805
poly(Wave height, 2)	2	14.2	9681	9815
poly(Latitude, Longitude, Year, 4)	34	2183.3	11850	11909
poly(Towing time, 2)	2	12.7	9679	9813
poly(Towing length, 2)	2	12.7	9679	9813
factor(Vessel)	12	131.3	9798	9908

Table 5.5: Effect on the residual sum of squares (RSS) and Mallows' statistics (C_p) of dropping one term out of the quantitative GLM. Air temperature is the only term that does not increase C_p.

suggested here that the towing time term is omitted.

5.2.4 Model adequacy checking

Figures 5.11 to 5.13 plot the standardized residuals that can be used as an informal model control. Figure 5.11 plots the standardized residuals and the absolute standardized residuals versus the fitted values. The residuals should be without structure, with mean zero and unit variance. The red lines show the mean values and variances, calculated for groups of 100 residuals. There is not a systematic structure in the range of the residuals on these two plots except that the range seems to be smaller for high fitted values (5.5-6.5) than others.

On the plot of standardized residuals (upper right corner of figure 5.11) there is not a serious curvature in the mean except that it is quite high for the lowest fitted values (0 to 1.5). The parallel lines that can be detected on this plot are perfectly natural, as described in section 5.1.2. The variances of the residuals is not entirely stable and furthermore, they are generally greater than unity.

On the plot of absolute standardized residuals (lower right hand corner of figure 5.11) the mean values seem to decrease with increasing fitted values, although the trend is not very strong. The variances of standardized residuals (upper right corner) also display a negative trend. This is surprising, since it might indicate that the variance function increases too rapidly with the mean, which contradicts the results of section 4.1.

Figure 5.13 shows the standardized residuals per otolith strata and years. As for figure 5.11 the residuals should be without structure, with mean zero and unit variance. The mean and variances are shown with a red square and are connected with a line. There is almost no curvature in the means in neither of the plots. The variances, however, are generally higher than unity and different between strata. This trend could possibly be eliminated by weighting observations in the model according to the variation in cod catch within each strata. The residual variances also differ between years.

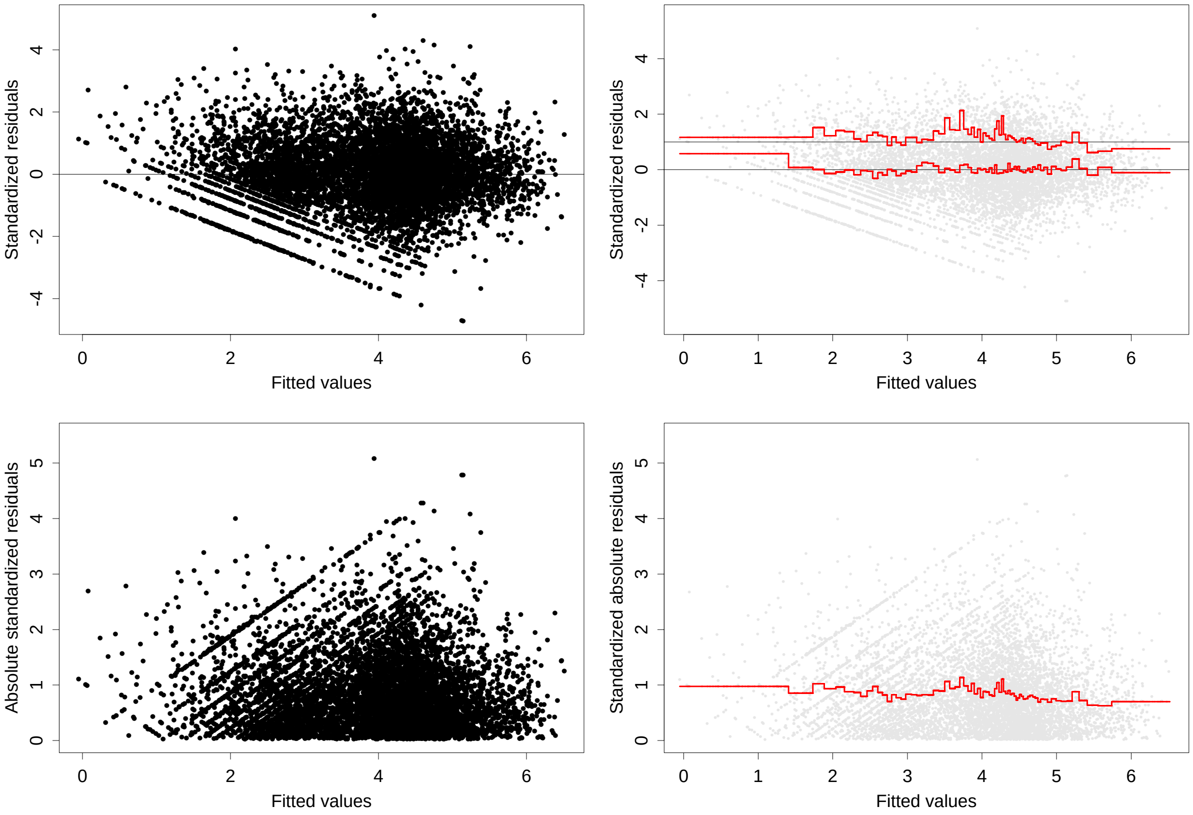

Figure 5.11: Scatter plots of standardized residuals (upper row) and absolute standardized residuals (lower row) versus the fitted values. The residuals should be without structure, with mean zero and unit variance. The red lines on the plot in the upper right hand corner show the mean values and variances, calculated for groups of 100 residuals, and the line on the plot in the lower right hand corner show the mean values, also calculated for groups of 100 residuals.

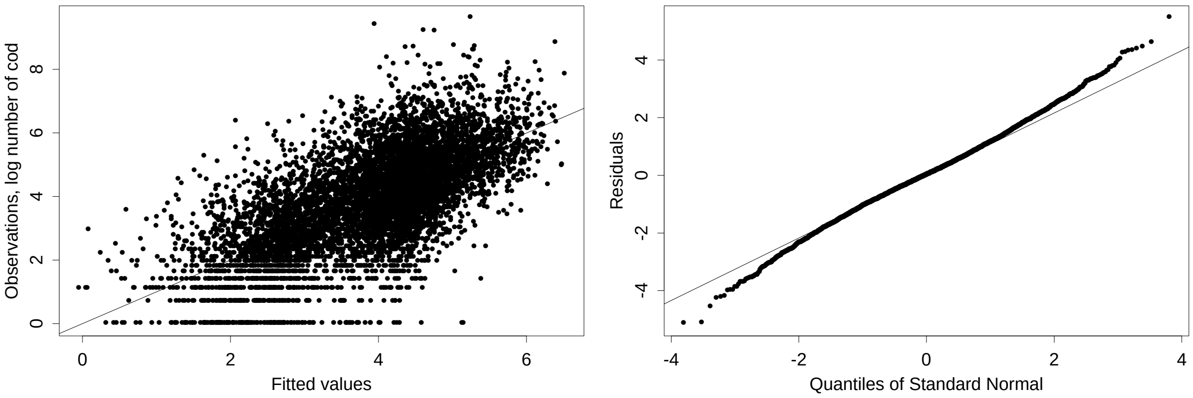

Figure 5.12: On the left: Scatter plot of observations versus fitted values along with the $y = x$ line. On the right: Normal probability plot of residuals.

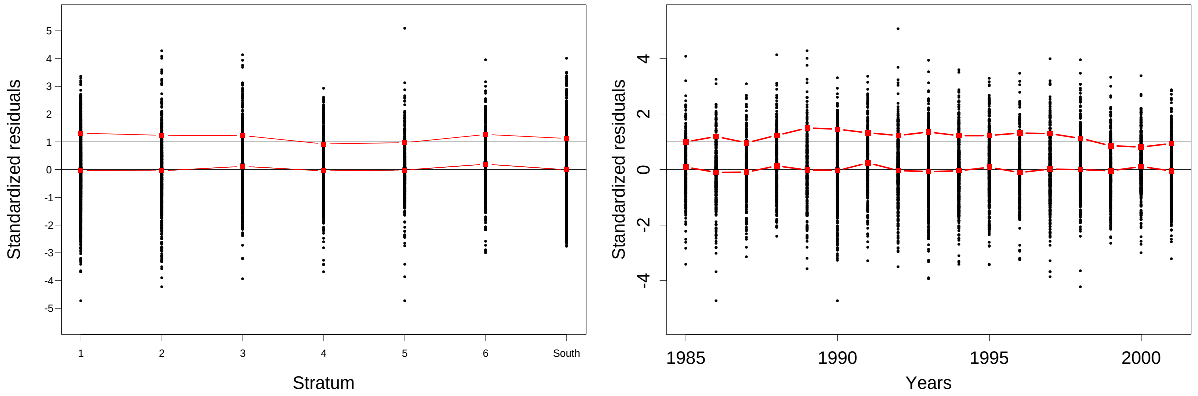

Figure 5.13: Standardized residuals (black points) of the quantitative GLM per otolith strata and per year. The mean and variances for each level are shown with a red square and are connected by a line for clarity.

The plot to the right in figure 5.12 shows a normal probability plot of the residuals. This plot indicates that the residuals fail to be normally distributed.

The plot to the left of figure 5.12 presents the observations versus the corresponding fitted values. These points should be scattered around the $y = x$ line. It is clear from this plot that the model fails to fit the low observations, e.g. fitted values corresponding to tows containing only one cod (the lowest parallel line) range from about 0 to 5 on log scale. Furthermore, fitted values of about 4 occur for the whole range of observed values. This indicates that there is a considerable amount of variation in the data that is yet unexplained. The ill behaving variances in the other plots can of course simply be revealing this fact.

A simple version of cross validation can be used in order to investigate the performance of model (5.2) in predicting new data. The dataset is divided into two subsets and the model (5.2) fitted to both of them separately. The model parameters obtained by fitting dataset 1 are then used to predict the expected log number of cod by using dataset 2 as the new data and vice versa.

In order to get two subsets that both represent the variability of data, tows in a sub-rectangle in one year were divided randomly into two equally large groups. The results of analysis of variance for these two datasets are presented in tables 5.6 and 5.7. It turns out that the terms explain the same percentage of variation in these datasets as they did for the whole dataset (see table 5.3) and the total model explains the same percentage of variation (44.5% and 46.5% compared to 45.2%).

Let $\widehat{\mu}^{kl}$ denote predicted values of the model fitted by data set k, predicting dataset l. It is interesting to examine how much of the variation in dataset 2 is explained by the fitted values $\widehat{\mu}^{12}$ and compare it to the R^2 value of the model fitted by dataset 2. The explanation percentage can be defined as:

$$1 - \frac{\sum_{i \in \text{dataset } 2} (y_i - \widehat{\mu}_i^{12})^2}{\sum_{i \in \text{dataset } 2} (y_i - \overline{y})^2}$$

where $\overline{y}$ is the grand mean of dataset 2. Table 5.8 shows this ratio along with the R^2 values for both datasets. The results are that the explanation percentages are very close to the R^2 values in both cases. It can therefore be concluded that prediction within the scope of the model is stable.

Source of Variation	Df	SS	% explained	SS/Df	F-test	p-value
poly(Depth, 2)	2	433.0	4.9	216.5	154.6	0.00000
+ poly(Bottom temp., 2)	2	1736.7	19.8	868.3	619.9	0.00000
+ Surface temperature	1	119.0	1.4	119.0	85.0	0.00000
+ Air temperature	1	3.4	0.0	3.4	2.4	0.11815
+ poly(Wave height, 2)	2	89.1	1.0	44.5	31.8	0.00000
+ poly(Latitude, Longitude, Year, 4)	34	1420.4	16.2	41.8	29.8	0.00000
+ poly(Towing time, 2)	2	27.8	0.3	13.9	9.9	0.00005
+ poly(Towing length, 2)	2	1.2	0.0	0.6	0.4	0.64800
+ factor(Vessel)	12	62.5	0.7	5.2	3.7	0.00001
Total model	58	3893.0	44.5			
Residuals	3473	4865.1	55.5	1.401		
Total	3531	8758.1				

Table 5.6: Analysis of variance table for dataset 1. 44.5% of the total variation is explained by this model.

Source of Variation	Df	SS	% explained	SS/Df	F-test	p-value
poly(Depth, 2)	2	460.9	5.2	230.5	168.6	0.00000
+ poly(Bottom temp., 2)	2	1734.0	19.5	867.0	634.1	0.00000
+ Surface temperature	1	175.0	2.0	175.0	128.0	0.00000
+ Air temperature	1	17.3	0.2	17.3	12.7	0.00038
+ poly(Wave height, 2)	2	50.7	0.6	25.3	18.5	0.00000
+ poly(Latitude, Longitude, Year, 4)	34	1560.9	17.6	45.9	33.6	0.00000
+ poly(Towing time, 2)	2	27.1	0.3	13.6	9.9	0.00005
+ poly(Towing length, 2)	2	21.8	0.2	10.9	8.0	0.00035
+ factor(Vessel)	12	80.0	0.9	6.7	4.9	0.00000
Total model	58	4127.7	46.5			
Residuals	3475	4750.9	53.5	1.367		
Total	3533	8878.6				

Table 5.7: Analysis of variance table for dataset 2. 46.5% of the total variation is explained by this model.

	Explanation % for dataset 1	Explanation % for dataset 2
Model fitted on dataset 1	44.5	45.3
Model fitted on dataset 2	43.3	46.5

Table 5.8: R^2 (the diagonal) compared to the percentage of variation explained by the model fitted on the other dataset.

5.2.5 Conclusions

The quantitative model (5.2) has revealed how a number of environmental variables effect the expected cod catch. The bottom temperature polynomial explains 20% of the total variation in the cod catch data and is the most important environmental factor. Depth and surface temperature are also important.

The majority of the total variation is, however, still unexplained. The groundfish survey data does not offer more covariates but other data, like weather and ocean current data could possibly be linked to the cod catch data. In the next section, a new covariate is introduced, a temperature gradient estimated from temperature records in the groundfish survey data.

The assumption of constant variance is not fulfilled by this model and the variances of the standardized residuals turned out to be higher than unity in general. The reason for these departures from model assumptions can of course be many and "assumptions like this can be violated in many more ways than they can be satisfied" (Scheffé 1959, p.331). These differences in residual variances can for example simply be revealing the fact that the majority of the data is still unexplained.

As noted before (section 5.1.3), the F-tests are are robust to the assumption of constant variances. The results concerning the environmental effects can therefore be considered valid.

A simple cross validation check indicates that the model gives a stable prediction within the scope of the model.

5.3 Estimated temperature gradient

It has been suggested that the food for cod, such as capelin, may aggregate in frontal regions where cold sea meets with warmer sea, i.e. where there is sudden change in temperature (Vilhjálmsson 1994). Therefore, it could be expected that cod would pursue such temperature fronts. The results in section 5.2 seem to support this suggestion, at least to some extent, since the bottom temperature explains a large portion of the total variance.

In light of this, it would be interesting to estimate the size of the gradient of bottom temperature at each data point and examine if that quantity becomes a significant term in the model of section 5.2. This is done in the following pages with a locally weighted regression.

5.3.1 An additive model for temperature

Here, the procedure used to obtain the estimates of the gradient vector will be described. The main idea is to use the measured bottom temperature to estimate a surface over the whole survey area so that temperature estimates on a fine grid can be used to estimate the gradient vector.

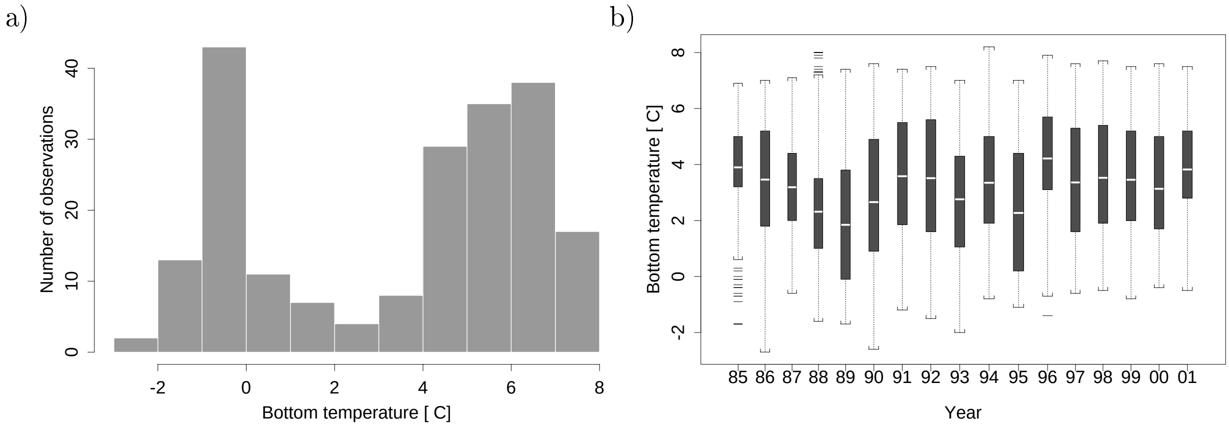

Figure 5.14: a) Histogram of bottom temperature only for tows where no cod was caught. This distribution is consistent with the trend observed in figure 5.5. b) Box plot of the bottom temperature versus the years. Tows where no cod was caught are also included.

The data

The data used is all the groundfish survey data from 1985 to 2001 where the bottom temperature was measured. Here tows with zero number of cod are also included, simply to obtain more temperature measurements. The total number of observations for this analysis are 8705, 512 per year on average (see table 5.9). The histogram in figure 5.14 shows how the tows with no cod are distributed over temperature. This plot confirms the trend observed in figure 5.5 that cod is more likely to be caught where the temperature is $0 - 4°\text{C}$ (the tows with no cod are unlikely) than where the temperature is lower or higher (the tows with no cod become more likely).

A box plot of the bottom temperature versus the years is shown in figure 5.14. It is obvious from this plot that the temperature varies greatly between years and this effect can not be ignored. But instead of including a year effect into the analysis, a separate temperature surface will be estimate for each year of the survey.

A grid containing all the survey area was constructed by distributing 101 points equally over the latitude range and 101 points over the longitude range. That makes a total of 10201 points on the grid. This grid used is shown on figure 5.15 along with the data points from the year 1995.

The additive model

In order to obtain a bottom temperature surface over the survey area, an additive model with one, two-dimensional, term was used:

$$t_i = g(lat_i, lon_i) + \epsilon_i \tag{5.3}$$

The t_i represents the bottom temperature and ϵ_i is assumed to be Normally distributed with mean 0 and constant variance. The function g of latitude and longitude is a second degree loess smoother. That is, $\widehat{g}(lat_i, lon_i)$ is the fitted value at t_i of the weighted regression

$$t_i = \beta_1 lat_i + \beta_2 lon_i + \beta_3 lat_i^2 + \beta_4 lon_i^2 + \beta_5 lat_i \cdot lon_i + \epsilon_i$$

with weights as defined in (3.9). The scope parameter was set $f = 0.2$.

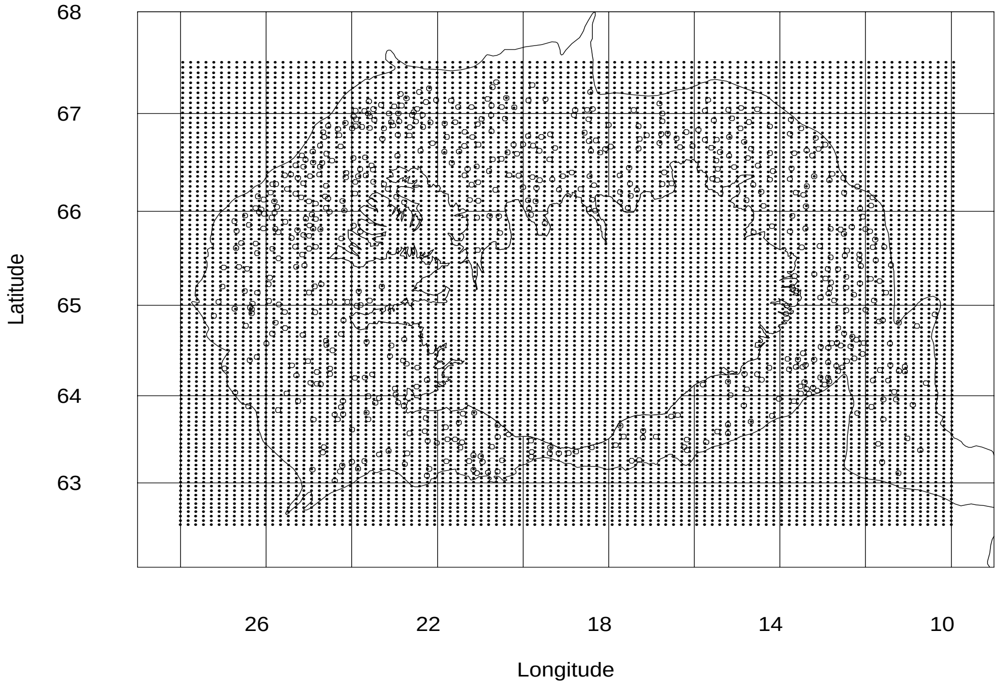

Figure 5.15: The grid (dots) and stations where bottom temperature was recorded in 1995 (circles).

Results for the additive model

Model (5.3) was fitted on 17 datasets, each containing the recorded bottom temperatures, latitudes and longitudes of that year along with the grid points, which were given the temperature value of 0°C. A point in these datasets (t_i, lat_i, lon_i) was given a weight of 1 if was a record from the survey data but a weight of 10^{-10} if it was a grid point. The effect of these weights is that they are simply multiplied to the built-in weights of the loess smoother. The grid points therefore have negligible effect on the fitted values, except at points that are far from the survey data, where the fitted temperatures are not needed anyway. On the other hand, the survey data are dominant for the fitted values at the grid points and hence the fitted values at the grid points provide a smooth surface of the temperature over the survey area for every year.

In table 5.9, the approximate degrees of freedom used by the loess model are given and the ratio between the residual deviance of the model (5.3) and the total deviance for each year. This percentage shows how much of the total temperature variance the model explains, and gives an idea of how well the smoothed surface represents the data. It appears that the model does not explain much of the variation, only 16% on average. These numbers suggest that a considerably better model could be made to fit the temperature data. The scope parameter could for example be altered or another type of smoother could be used. This model will, however, not be investigated further in this thesis.

Figures B.5 to B.7 show contour plots of the smoothed temperature surface of model (5.3) for every year of the survey. The plots all show the same general temperature trend. The bottom temperature is higher (mostly > 3°C) in the south and east of Iceland than in the north and

Year	1985	1986	1987	1988	1989	1990	1991	1992	1993
Total obs.	434	485	426	426	500	541	532	544	568
Approx. df	23.50	24.10	24.04	22.98	22.93	23.95	24.02	23.93	23.80
% explained	38.88	24.48	20.18	17.25	8.35	15.76	19.21	11.95	11.60

Year	1994	1995	1996	1997	1998	1999	2000	2001	Total
Total obs.	591	591	537	470	499	515	515	531	8705
Approx. df	24.33	23.65	23.29	23.48	22.83	23.54	23.29	23.23	
% explained	14.69	12.03	15.59	8.90	9.17	11.66	12.68	18.19	

Table 5.9: For each year, this table shows the number of station where the bottom temperature was measured, the approximate degrees of freedom used by the loess model and the proportion of total variation explained by the loess model.

west of the the country (mostly < 3°C).

The gradient

Once the estimated temperature $\widehat{t_i} = \widehat{g}(lat_i, lon_i)$ has been obtained on the grid points for each year of the survey, the gradient vector can estimated. The definition of the gradient vector of a two dimensional function $f(x, y)$ is:

$$\nabla f(x,y) = \left(\lim_{h \to 0} \frac{f(x+h,y) - f(x,y)}{h}, \lim_{k \to 0} \frac{f(x,y+k) - f(x,y)}{k} \right)$$

In analogy to this definition the squared length of the gradient vector at (lat_i, lon_i) can be estimated by the difference between adjacent points in the grid:

$$||\widehat{\nabla g}(lat_i, lon_i)|| = \sqrt{(\widehat{g}(lat_{i+1}, lon_i) - \widehat{g}(lat_i, lon_i))^2 + (\widehat{g}(lat_i, lon_{i+1)} - \widehat{g}(lat_i, lon_i))^2}$$

Ideally the $\widehat{g}(lat_{i+1}, lon_i) - \widehat{g}(lat_i, lon_i)$ should be divided with the distance between lat_{i+1} and lat_i but since this distance is the same for all latitudes this will only have a scalar effect on the size of the gradient and can therefore be omitted.

Finally, a tow in the survey data is assigned the gradient length value of the grid point that is nearest to the position of the tow. This gradient value is then considered as a new covariate in the quantitative GLM model (5.3).

A gradient vector has the property that the maximum rate of increase and decrease in any direction from a point (a, b) is $||\nabla f(a, b)||$. If the suggestion is correct, the expected number of cod would be high when $||\widehat{\nabla g}(lat, lon)||$ is high and low when $||\widehat{\nabla g}(lat, lon)||$ is near 0. Figure 5.16 shows a box plot of the log number of cod versus the estimated length of the gradient vector. There is a weak positive trend on this plot for gradient lengths 0 to 0.3 but the log number of cod decreases for longer gradient vectors. To be on the safe side, a third degree polynomial in the gradient length is included in the quantitative model.

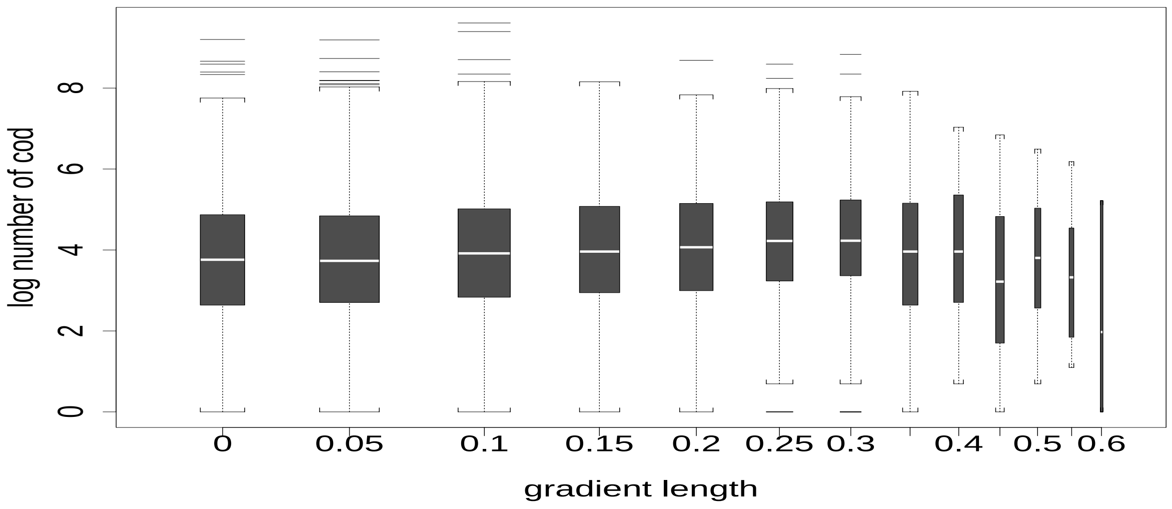

Figure 5.16: Box plot of the log number of cod versus the estimated length of the temperature gradient vector.

5.3.2 The improved quantitative model

This model is the same as model (5.2) except that the surface temperature and towing time terms are omitted and a third degree polynomial in gradient length is included after the other two temperature terms, i.e.

$$\begin{aligned} H_0 : \log(Y_i) \; = \; & \beta_0 + \text{poly}(\text{Bottom temp}_i, 2) + \text{Surface temp}_i + \\ & \text{poly}(\text{Gradient}_i, 3) + \text{poly}(\text{Depth}_i, 2) + \text{poly}(\text{Wave hight}_i, 2) + \\ & \text{poly}(lat_i, lon_i, \text{Year}_i, 4) + \text{poly}(\text{Towing length}_i, 2) + \text{factor}(\text{Vessel}) + \epsilon_i \end{aligned} \quad (5.4)$$

$\log(Y_i)$ denotes the log number of cod in tow i, β_0 is an intercept and ϵ_i is assumed to be normally distributed with mean 0 and constant variance.

5.3.3 Results

The model (5.4) was fitted to the cod catch data from the Icelandic groundfish surveys along with the estimated gradient values. The resulting analysis of variance (ANOVA) is shown in table 5.10. Table 5.11 shows information on model that only includes the gradient polynomial, table 5.12 shows the change in the residual sum of squares if one term is removed from the model along with Mallows' statistic and table 5.13 shows t-tests for removing the parameters belonging to the gradient polynomial from the model.

The new covariate is found to be significant although its contribution to variance reduction is small. As the first term in a model, the gradient term explains 1.0% of the total variation (table 5.11). When added to a model containing only a bottom temperature term (table 5.10) this percentage is nearly the same, 0.9%. Removing the gradient term does not reduce the C_p statistic (see table 5.12) and the t-test shown in table 5.13 reject the hypothesis that the parameters are zero on a 5% significance level.

Source of Variation	Df	SS	% explained	SS/Df	F-test	p-value
poly(Bottom temp., 2)	2	3640.7	20.6	1820.3	1325.2	0.00000
+ Surface temperature	1	410.9	2.3	410.9	299.1	0.00000
+ poly(Gradient, 3)	3	153.5	0.9	51.2	37.3	0.00000
+ poly(Depth, 2)	2	608.9	3.5	304.4	221.6	0.00000
+ poly(Wave height, 2)	2	126.2	0.7	63.1	45.9	0.00000
+ poly(Latitude, Longitude, Year, 4)	34	2886.6	16.4	84.9	61.8	0.00000
+ poly(Towing length, 2)	2	56.0	0.3	28.0	20.4	0.00000
+ factor(Vessel)	12	131.5	0.7	11.0	8.0	0.00000
Total model	58	8014.1	45.4			
Residuals	7007	9625.0	54.6	1.374		
Total	7065	17639.2				

Table 5.10: Analysis of variance table for the quantitative GLM of log transformed cod catch data where a polynomial in gradient length has been included. The terms are added sequentially (first to last) and they are all significant. 45.3% of the total variation is explained by this model.

Term	Df	SS	% explained	RSS	Cp
<none>				17639	17642
poly(Gradient, 3)	3	177.0	1.0	17462	17472

Table 5.11: Results for fitting a GLM containing only a gradient length polynomial to the log transformed cod catch data.

Term	Df	SS	RSS	Cp
<none>			9625	9763
poly(Bottom temp., 2)	2	172.1	9797	9931
+ Surface temperature	1	19.2	9644	9780
+ poly(Gradient, 3)	3	57.1	9682	9813
poly(Depth, 2)	2	251.5	9877	10010
+ poly(Wave height, 2)	2	11.2	9636	9770
+ poly(Latitude, Longitude, Year, 4)	34	2157.5	11783	11841
+ poly(Towing length, 2)	2	52.0	9677	9811
+ factor(Vessel)	12	131.5	9756	9867

Table 5.12: Effect on the residual sum of squares (RSS) and Mallows' statistics (C_p) of dropping one term out of the model

	Value	Std. Error	t value	Pr(>|t|)
poly(Gradient, 3)1	8.4535	1.7930	4.7147	0.0000
poly(Gradient, 3)2	-3.3220	1.4061	-2.3625	0.0182
poly(Gradient, 3)3	-4.4651	1.2434	-3.5910	0.0003

Table 5.13: T-test for the hypothesis that the model parameters belonging to the gradient polynomial are zero.

Figures B.9 and B.8 show plots of residuals analogous to those in figures 5.11 and 5.12. These plots show exactly the same patterns so the discussion made in section 5.2.3 is also relevant here.

5.3.4 Conclusions

The effects of the polynomial in the estimated gradient length does not live up to expectations as it explains about 1% of the total variation in the cod catch data. There are of course several possible reasons for this.

The loess model used to estimate the bottom temperature surface does explains a very small portion of the total variation in the temperature data. The surface is therefore too smooth and does not capture local changes in temperature but only a global trend over the whole survey area.

Another problem with this analysis is that the survey data are collected over a 2-3 weeks period by 5 different vessels. This means that records spatially close need not be close in time, although that is most often the case.

Chapter 6

Discussion and conclusions

In this final chapter an overview is given of the conclusions made in this thesis and possible directions of future work are discussed.

6.1 The probability distributions

Two probability distributions were proposed for describing the data, the gamma distribution and the log-normal distribution. In chapter 4 several issues concerning these distributions were investigated and a goodness of fit test was conducted in order to distinguish between them.

The variance function

The gamma distribution and the log-normal distribution have the same variance function, i.e. the variance is proportional to the mean squared. In section 4.1 this relationship was investigated by examining a plot of $\log(s^2)$ versus $\log(\overline{y})$. The sample variance, s^2, and the sample mean, $\overline{y}$, were calculated for small groups of data. The small groups were on the one hand those sub-rectangles where 5 or more tows are performed and on the other hand those statistical rectangles are 10 or more tows were performed. These plots showed a sample of points that were scattered around a straight line. A weighted linear regression of the log sample variance on the log sample mean gave a slope of 2.23, suggesting a power relationship for the variance function with index close to 2.

The s^2 gives biased estimates of the variance for both a gamma distributed variable and a log-normal distributed variable. It would therefore be interesting to examine the accuracy of the investigation described above by simulation. Small groups of gamma and log-normally distributed variables could be simulated and a plot of $\log(s^2)$ versus $\log(\overline{y})$ calculated from these simulations could be examined. Furthermore a regression could be made for these simulated samples. The group sizes would be a factor in such analysis, small sample sizes will give a more variable results.

Another approach to future research would be to assume the variance function $V(\mu) = \mu^{2.23}$ for the data and estimate the model parameters by constructing a quasi-likelihood function. Other variance functions are of course also possible. For example, $V(\mu) = \mu + \mu^2/k$, which is the variance function for the negative binomial distribution.

Constant coefficient of variation

A GLM with gamma distributed errors usually assumes a constant coefficient of variation (CV). This assumption was investigated in section 4.2. The distribution of estimated CVs, $s^2/\overline{y}$, calculated from the cod catch was compared to the distribution of $s^2/\overline{y}$ for simulated gamma distributed variables. As before, the sample variance, s^2, and the sample mean, $\overline{y}$, were calculated for small groups of data, the same groups as above. A two sample Kolmogorov-Smirnov test for the hypothesis that the $s^2/\overline{y}$ calculated from data and $s^2/\overline{y}$ calculated from simulated data follow the same distribution was accepted for sub-rectangles but not for statistical rectangles. This indicates that sub-rectangles are more homogeneous areas than statistical rectangles.

Constant variation of log transformed data

A GLM with normally distributed errors usually assumes a constant variance. This assumption for log transformed data was investigated in section 4.3 by means of Bartlett's test. These tests were made firstly for all the data, secondly within each survey year and thirdly within survey areas but most of the tests were rejected. It should of course be kept in mind that the Bartlett's test is very sensitive to the normality assumption.

Goodness of fit

In section 4.4 the two proposed distribution were compared via a goodness of fit test. In order to estimate the mean values and the shape parameter a GLM with sub-rectangles and years as factors was fitted to the cod catch data. One model where the errors were assumed to be gamma distributed and a logarithmic link was used. Another model for log transformed data where the errors were assumed to be normally distributed. In addition, separate models for each year including only one factor, the sub-rectangles, were also fitted.

The fitted values from these models where used to scale the observations. Then the hypothesis that the scaled data follow the $G(\widehat{r}, 1/\widehat{r})$ or the $LN(0, \widehat{\sigma^2})$ were tested. The shape parameters r and σ^2 were estimated by the GLM. Both hypothesis were rejected when data from all years were included in one model, but the log-normal distribution was closer to acceptance. On the other hand both hypothesis were accepted when data for each year were considered separately. The log-normal distribution gave higher p-values in most years. Therefore the log-normal distribution was preferred over the gamma distribution in the following analysis of the thesis.

6.2 The models

In section 5.1 a fully qualitative model with only spatial and time effects was investigated. Since the expected cod catch is believed to be constant at a given place and time this is the most straightforward model available. This model explained 63% of the total variation but used far too many parameters.

A quantitative model was investigated in section 5.2. The environmental data collected in the groundfish surveys were examined by means of box plots. Other data, like position (latitude and longitude) and trawl station data were examined in the same way. Based on these plots polynomials were proposed to describe the relationship between each variable and the cod catch. These polynomials were included in a quantitative model and tested for significance.

A second degree polynomial in bottom temperature explained 20% of the total variation in the cod catch data and was the most important environmental factor. Depth and surface temperature were also found important. On the whole, the quantitative model explained 42% of the total variation. The majority of the variation in cod catch is, therefore, still unexplained. The groundfish survey data does not offer more covariates but other data like weather and ocean current data could possibly be linked to the cod catch data. A search for other variables to explain what remains of the variation in cod catch could be an avenue of future research. One step down that avenue was taken in section 5.3.

In section 5.3 the size of the gradient vector of bottom temperature at each data point was estimated. Locally weighted regression (loess) was used to obtain estimated temperatures on a fine grid and the gradient norm estimated from these grid points. A third degree polynomial was suggested for the relationship with the cod catch and that polynomial was included in the quantitative model. The effect of this polynomial did, however, not live up to expectations since it explained less than 1% of the total variation in the cod catch data. There are of course several possible reasons for this.

The loess model used to estimate the bottom temperature surface explained a very low portion of the total variation in the temperature data, only a 16% on average. The surface was therefore probably too smooth to capture local changes in temperature. Furthermore, the data collected in the survey could be too sparse to give an exact surface estimate. Since the capelin is a pelagic species and it is known that cod preys on capelin, it might have been more appropriate to use surface temperature in this analysis.[1] Furthermore, satellite pictures of ocean surface temperature could be used in this purpose. That would at least give a more accurate estimate of the temperature fronts.

Model adequacy

Various plots of residuals from both of the models revealed some shortcomings of these models, which were more serious for the quantitative model. The most obvious one was that the residual variances could not be assumed to be constant. The models fail to predict the smallest and the largest observations and furthermore, some of the fitted values occurred for the whole range of observed values. The ill behaving variances can of course simply be revealing this fact.

A simple cross validation check indicated that the model gives a stable prediction within the scope of the model.

[1] A quick examination revealed however that surface temperature gradients do not explain more of the variation than the bottom temperature gradients.

6.3 Other considerations

An alternative to GLM

In this thesis, possible covariates were examined by means of box plots. Based on these plots polynomials of different degrees were suggested for the relationship with the response. The disadvantage of this method is however that it is difficult to guess the most appropriate degree of such polynomials. This is especially the case for combined effects, i.e. when the polynomial contains more than one variable. Another disadvantage is that the estimated surface of such a polynomial regression becomes very limited since the fit of one data point depends on all the independent data.

Generalized additive models (GAMs) provide a method where less restrictive function of covariates can be used for modeling data. In this method no rigid parametric assumption are made about the dependence of the response to the covariates. The loess smoother used in this thesis is one example of these models. GAMs are therefore an important area of future research of the groundfish survey data.

Catch forecasts

Some adjustments has to be made before the analysis in this thesis can be utilized in catch forecasting model for the commercial fleet. The data used in such a model would be the trawler reports described in chapter 1. They include the estimated catch in weight per species for every action of the Icelandic fleet for several years.

First of all, the data investigated here is the number of cod by tow. Analysis of the trawler reports would be in weights. The total weight per species is not recorded in the surveys but a sample of the catch is weighted and age determined for every station. The age-distribution obtained for the otolith strata can therefore be used to estimate the total catch in weight per station.

The actions recorded in the trawler reports are of course not standardized like those made in the survey. Several trawl station data would therefore be necessary in a forecasting model using these data. For example the gear used, towing time and length the actual vessel used and so forth. Furthermore, a more detailed time factor is necessary, a month or perhaps a week factor would have to be included.

The environmental variables that in this thesis were found to explain an important part of the variation in cod catch, could be included in a forecasting model based on trawler reports. The depth and position is recorded in the trawler reports but not ocean temperatures. If historical temperature data exists that can be used. Surface temperature can be used if the bottom temperature is not known since this factor can explain 18% of the total variation. Surface temperature can for example be obtained by satellite pictures.

Appendix A

Additional figures for chapter 4

Figure A.1 displays an unweighted regression for the analysis of the variance function in section 4.1.

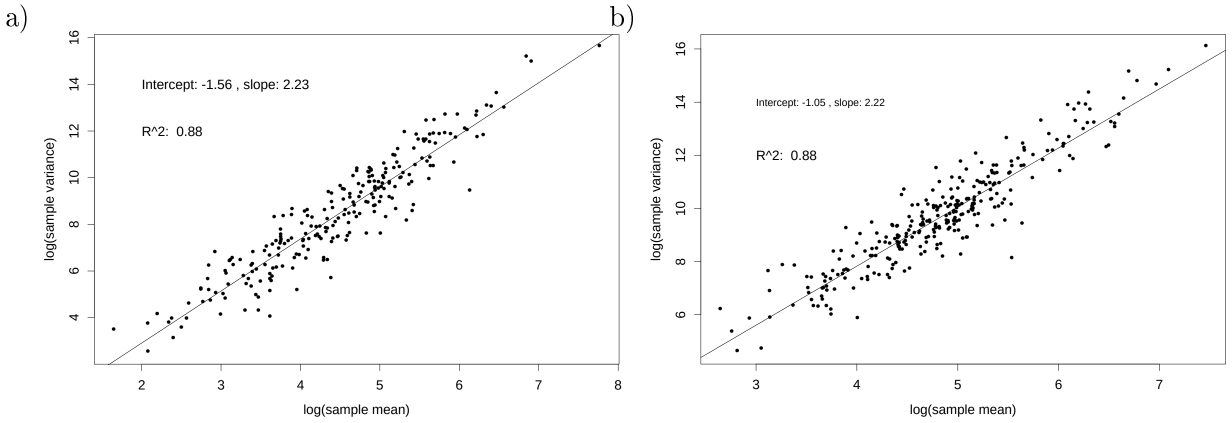

Figure A.1: Scatter plot of $\log(s_j^2)$ versus $\log(\overline{y}_j)$ for a) every sub-rectangle j that has 5 or more observations and b) every statistical rectangle j that has 10 or more observations. a) A classic regression (without weights) gave a slope of 2.23 ± 0.05 and an intercept of -1.6 ± 0.2 and the regression line is shown on the graph. b) A classic regression (without weights) gave a slope of 2.22 ± 0.05 and an intercept of -1.0 ± 0.3 and the regression line is shown on the graph.

Figures A.2 and A.3 on the following two pages show hypothesized cumulative distribution functions (CDFs) in red color and empirical CDFs for the goodness-of-fit tests in section 4.4 in black color.

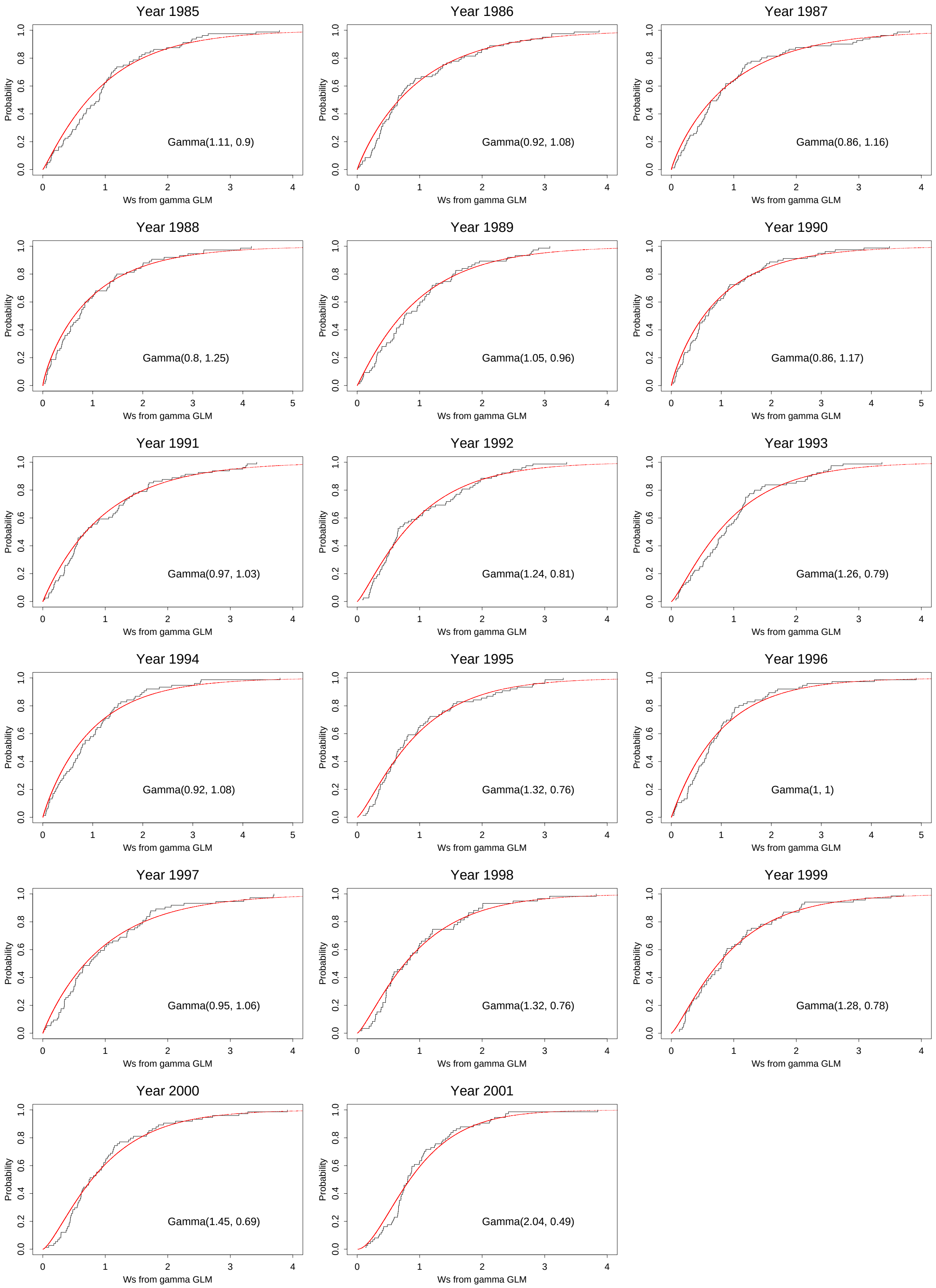

Figure A.2: CDFs for the hypothesized gamma distributions (in red color) along with the empirical CDFs of W_i estimated with parameters from a GLM assuming gamma distributed errors (in black color). All the hypothesized distributions are accepted but the log-normal distribution gives a better fit (see figure A.3).

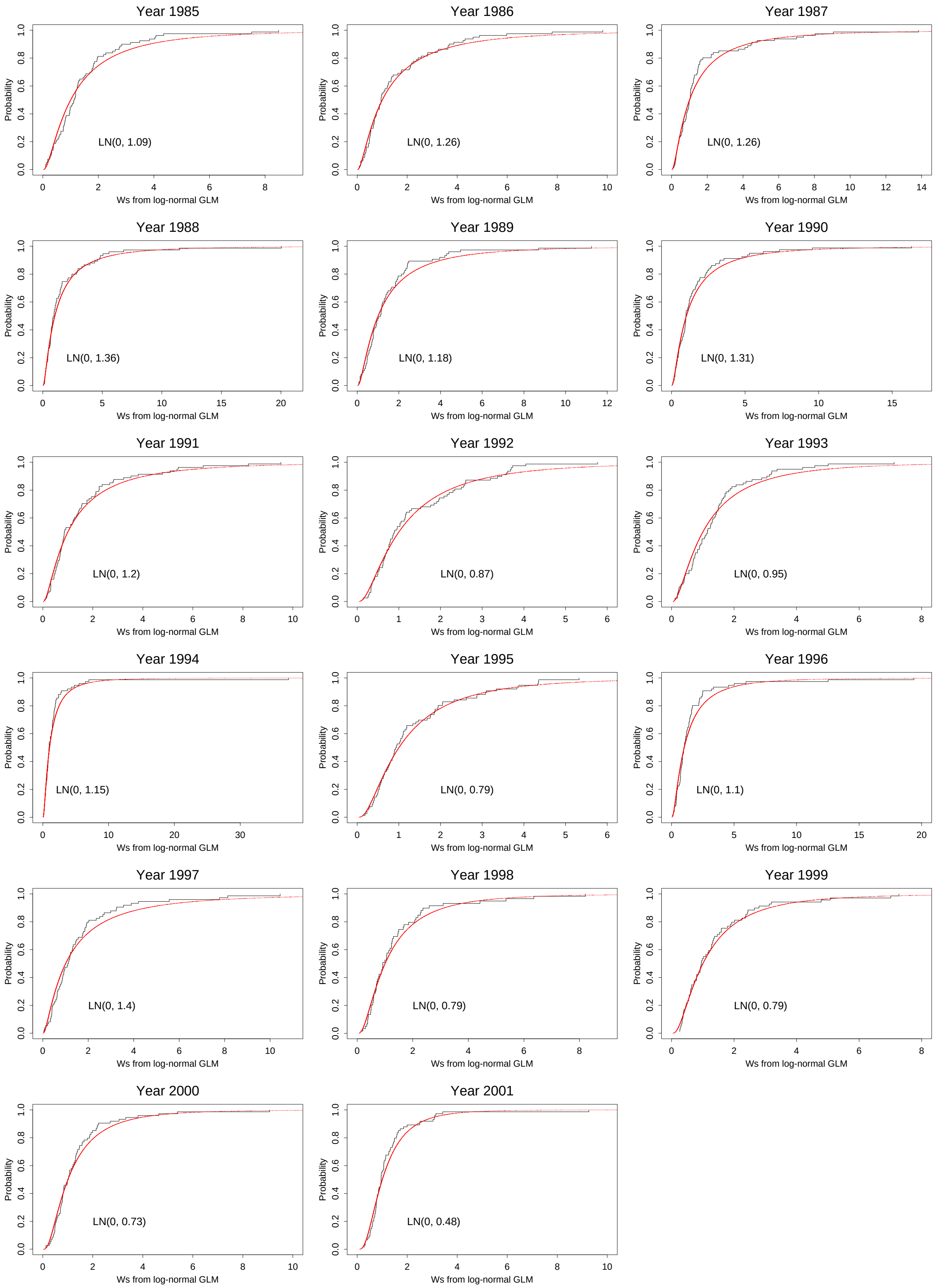

Figure A.3: CDFs for the hypothesized log-normal distributions (in red color) along with the empirical CDFs of W_i estimated with parameters from a GLM with log transformed data assuming normally distributed errors (in black color). All the hypothesized distributions are accepted and give a better fit than the gamma distribution (see figure A.2).

Appendix B

Additional figures and tables for chapter 5

B.1 Residuals per survey area (section 5.1.2)

Figure B.1 shows the standardized residuals of the qualitative model (5.1) per sub-rectangle divided by survey areas (strata). The plot on the right hand side in the lowest row shows the standardized residuals per stratum. As for figure 5.2 the residuals should be without structure, with mean zero and unit variance. The mean and variances for each sub-rectangle are shown with a red square and are connected by a line for clarity.

There is no systematic change in range or variances but the variances are not stable. Some sub-rectangles have very large variances but few observations, e.g. sub-rectangle 7184 in stratum 3 and sub-rectangle 3191 in the southern area. There are also sub-rectangles that have large variances and many observations, e.g. 5701 and 6712 in stratum 2 and 6121 in stratum 4. One possible explanation for this is that there can be different conditions within a sub-rectangle, for example changes in depth and temperature. The variances are more stable between strata but are still different.

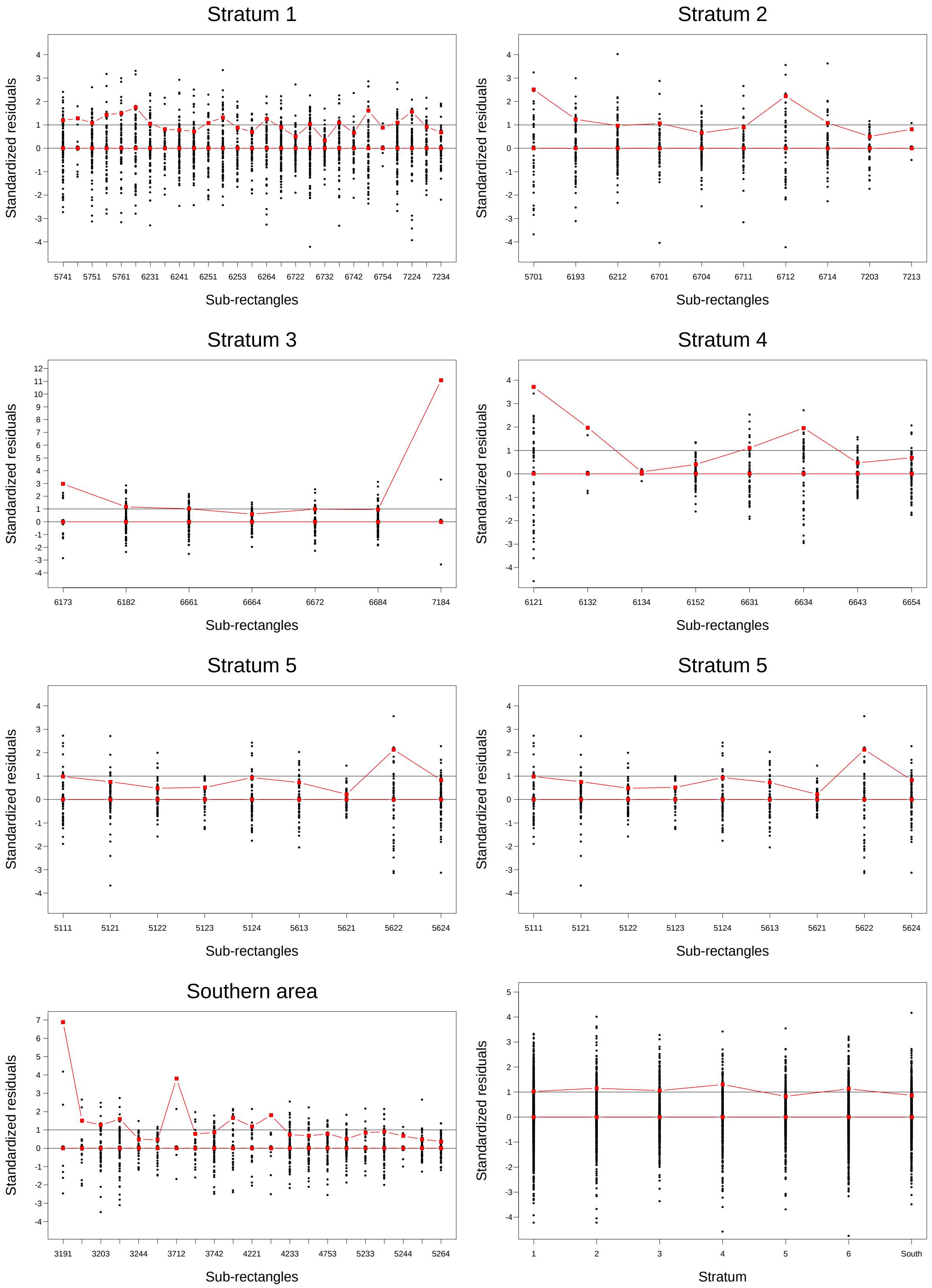

Figure B.1: Standardized residuals (black points) of the qualitative GLM per sub-rectangle divided by otolith strata. The mean and variances for each sub-rectangle are shown by a red square and are connected with a line for clarity. The plot in the last row to the right shows the residuals for all sub-rectangles within each strata.

B.2 Variables not included in the quantitative model (section 5.2.1)

Figure B.2 displays box plots of the log number of cod in tow versus the wind speed and the wind direction. A problem with these records is that according to the survey handbook (Einarsson et al. 2002), the possible values for wind speed are 0 to 12, measured in Beaufort scale, but a lot of higher values are actually recorded. This may be due to a switch from measuring wind speed in the Beaufort scale to measuring it in meters per second. If that is the case, there is no way of knowing whether a wind speed of, for example, 6 means 6 on Beaufort scale or $6m/s$. A similar situation occurs for the wind direction records. Most of the boxes which are thinner than the others represent numbers that are not defined by the handbook (Einarsson et al. 2002). Possibly they represent directions that are in between the two on either side. Since there seems to be some chaos in the recording of both the wind speed and direction and that there is no obvious relationship with the cod catch on the plots, these two variables are not included in the quantitative model.

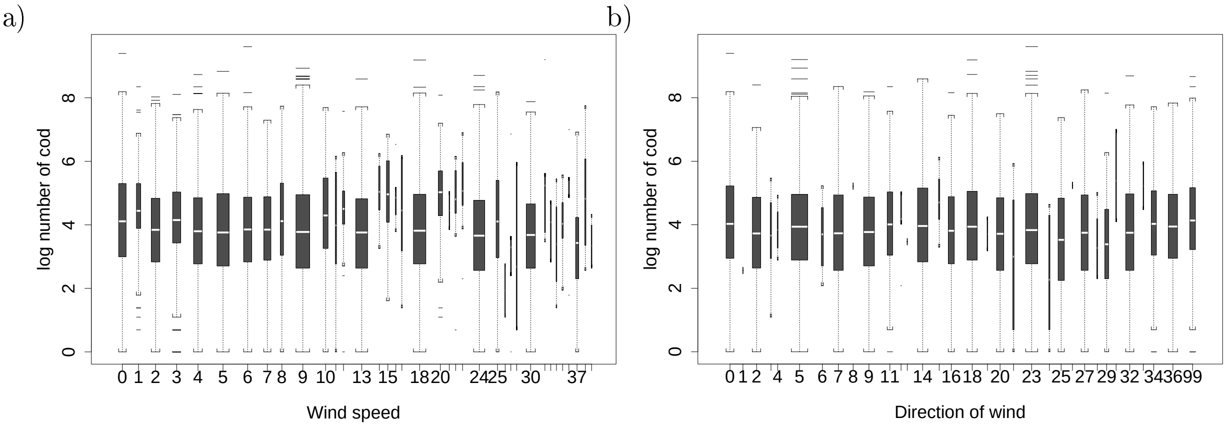

Figure B.2: Box plot of log number of cod in tow versus a) wind speed and b) direction of wind. a) There is uncertainty about on which scale the wind speed is measured (Beaufort scale or m/s). b) The levels 0 and 99 represent calm weather and changeable directions, 2 to 36 represents wind directions from NNE to N . Some recorded numbers (mostly the thin boxes) are not defined in the handbook.

Figure B.3 shows box plots of the log number of cod in tow versus a weather factor and cloud levels. The weather factor tells if it's raining, snowing etc., and clouds levels represents the proportion of sky covered in clouds. There is no striking relationship apparent on these figures and since there is no reason to believe in advance that these factors could have significant effect of the number of cod in a tow, they are not included in the model.

Figure B.4 shows box plots of log number of cod in tow versus barometric pressure and the time, in hours, when the tow began. Even though a quadratic relationship can be spotted for the barometric pressure, this variable is not included in the model because there were too many records missing. There is no obvious diurnal trend and this variables is not included in the model.

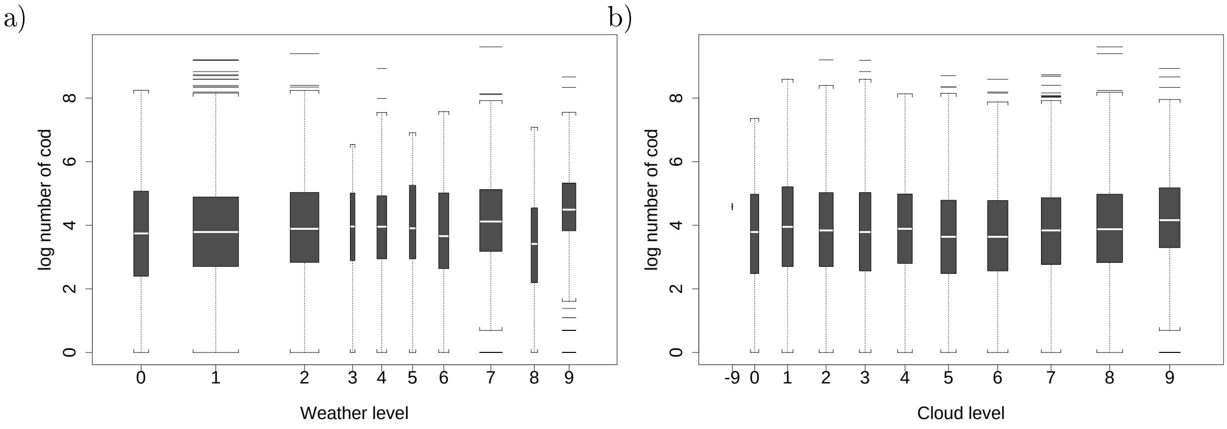

Figure B.3: Box plot of log number of cod in tow versus a) a weather factor and b) cloud levels. a) The levels 0-8 represent different weather conditions; bright weather, cloudy, overcast, sandstorm, fog, mist, rain, snow and rain showers. The 9th level stands for not recorded. b) The level n ($n = 0, \ldots, 8$) represent that $n/8$ portions of the sky are covered with clouds. The 9th level stands for not recorded.

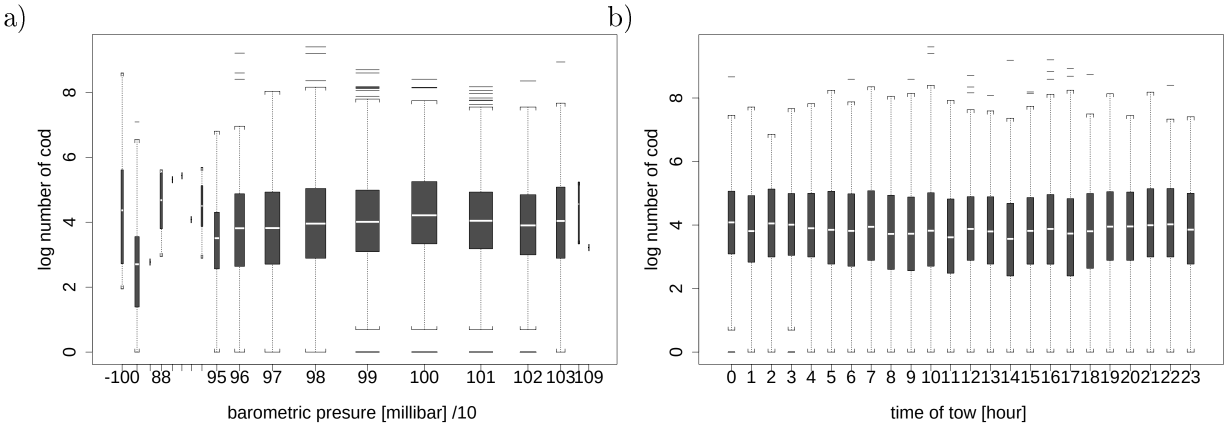

Figure B.4: Box plot of log number of cod in tow versus a) barometric pressure and b) the hour when the tow began. There are 3169 missing record of barometric pressure.

B.3 Additional tables for results of the quantitative model (Section 5.2.3)

The following tables B.1 and B.2 show ANOVAs for the quantitative GLM where the terms are in different order than in table 5.3. Tables B.3 and B.4 show t-tests for removing one parameter from the model.

Source of Variation	Df	SS	% explained	SS/Df	F-test	p-value
poly(Bottom temp., 2)	2	3640.7	20.6	1820.3	1319.5	0.00000
+ Surface temperature	1	410.9	2.3	410.9	297.8	0.00000
+ Air temperature	1	29.7	0.2	29.7	21.5	0.00000
+ poly(Depth, 2)	2	592.8	3.4	296.4	214.8	0.00000
+ poly(Wave height, 2)	2	137.5	0.8	68.7	49.8	0.00000
+ poly(Latitude, Longitude, Year, 4)	34	2961.8	16.8	87.1	63.1	0.00000
+ factor(Vessel)	12	136.1	0.8	11.3	8.2	0.00000
+ poly(Towing time, 2)	2	50.3	0.3	25.2	18.2	0.00000
+ poly(Towing length, 2)	2	12.7	0.1	6.3	4.6	0.01000
Total model	58	7972.4	45.2			
Residuals	7007	9666.7	54.8	1.380		
Total	7065	17639.2				

Table B.1: Analysis of variance table for the quantitative GLM of log transformed cod catch data. The terms are added sequentially (first to last) but in a different order from table 5.3. Depth comes after the temperatures and the vessel effects comes before towing time and length.

Source of Variation	Df	SS	% explained	SS/Df	F-test	p-value
poly(Bottom temp., 2)	2	3640.7	20.6	1820.3	1319.5	0.00000
+ Air temperature	1	79.2	0.4	79.2	57.4	0.00000
+ Surface temperature	1	361.3	2.0	361.3	261.9	0.00000
+ poly(Wave height, 2)	2	103.2	0.6	51.6	37.4	0.00000
+ poly(Depth, 2)	2	627.0	3.6	313.5	227.3	0.00000
+ poly(Latitude, Longitude, Year, 4)	34	2961.8	16.8	87.1	63.1	0.00000
+ factor(Vessel)	12	136.1	0.8	11.3	8.2	0.00000
+ poly(Towing length, 2)	2	50.3	0.3	25.2	18.2	0.00000
+ poly(Towing time, 2)	2	12.7	0.1	6.4	4.6	0.01005
Total model	58	7972.4	45.2			
Residuals	7007	9666.7	54.8	1.380		
Total	7065	17639.2				

Table B.2: Analysis of variance table for the quantitative GLM of log transformed cod catch data. The terms are added sequentially (first to last). The order differs from table B.1 in that the air and surface temperatures have switched places, wave height is now before the depth and the towing time and length have been switched.

Coefficient	Value	Std. Error	t value	Pr(>|t|)
factor(Vessel)1273	1.1536	0.2790	4.1345	0.0000
factor(Vessel)1274	1.1838	0.2681	4.4155	0.0000
factor(Vessel)1275	1.2459	0.2905	4.2883	0.0000
factor(Vessel)1276	-0.4652	0.3545	-1.3122	0.1895
factor(Vessel)1277	1.5880	0.2888	5.4984	0.0000
factor(Vessel)1278	1.2455	0.2765	4.5051	0.0000
factor(Vessel)1279	1.1586	0.2743	4.2235	0.0000
factor(Vessel)1280	1.0027	0.2712	3.6977	0.0002
factor(Vessel)1281	0.8733	0.2652	3.2932	0.0010
factor(Vessel)1307	1.1140	0.2828	3.9390	0.0001
factor(Vessel)1325	1.1235	0.2951	3.8074	0.0001
factor(Vessel)1459	1.4287	0.3442	4.1513	0.0000

Table B.3: T-tests for the hypothesis that vessel parameters in the quantitative GLM are zero. The effect of a vessel is relative to the effect of vessel 1131 which is included in the intercept. These parameters are all found to be significant on the 5% significance level, except for vessel 1276. It makes no sense, however, to exclude the effect on one vessel only.

Coefficient	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	2.8671	0.2672	10.7302	0.0000
poly(Depth, 2)1	0.5679	1.7691	0.3210	0.7482
poly(Depth, 2)2	-19.9251	1.4894	-13.3778	0.0000
poly(Bottom temperature, 2)1	7.7329	2.7915	2.7702	0.0056
poly(Bottom temperature, 2)2	-16.1365	1.6004	-10.0828	0.0000
Surface temperature	-0.0458	0.0155	-2.9551	0.0031
Air temperature	-0.0059	0.0044	-1.3440	0.1790
poly(Wave height, 2)1	-4.1341	1.3438	-3.0765	0.0021
poly(Wave height, 2)2	0.7525	1.4939	0.5037	0.6145
poly(Towing time, 2)1	0.6450	3.7699	0.1711	0.8642
poly(Towing time, 2)2	4.8670	2.8844	1.6873	0.0916
poly(Towing length, 2)1	6.3089	2.7645	2.2821	0.0225
poly(Towing length, 2)2	-3.5838	1.3477	-2.6592	0.0079

Table B.4: T-tests for the hypothesis that parameters in the quantitative GLM are zero. Based on these tests the air temperature term, the towing time term and the second order wave height term can be omitted from the model. It should be noted, however, that these tests are not valid simultaneously for all parameters.

Coefficient	Value	Std. Error	t value	Pr(>|t|)
poly(Lat, Lon, Year, 4)1.0.0	60.6716	3.0839	19.6735	0.0000
poly(Lat, Lon, Year, 4)2.0.0	-12.0800	2.0503	-5.8920	0.0000
poly(Lat, Lon, Year, 4)3.0.0	-10.7460	1.8129	-5.9275	0.0000
poly(Lat, Lon, Year, 4)4.0.0	2.8393	1.5972	1.7777	0.0755
poly(Lat, Lon, Year, 4)0.1.0	-24.0321	2.9641	-8.1078	0.0000
poly(Lat, Lon, Year, 4)0.2.0	-12.2279	2.2544	-5.4240	0.0000
poly(Lat, Lon, Year, 4)0.3.0	-19.9232	1.8937	-10.5206	0.0000
poly(Lat, Lon, Year, 4)0.4.0	-9.4316	1.9616	-4.8081	0.0000
poly(Lat, Lon, Year, 4)0.0.1	-11.8645	1.9203	-6.1786	0.0000
poly(Lat, Lon, Year, 4)0.0.2	8.5198	1.8822	4.5265	0.0000
poly(Lat, Lon, Year, 4)0.0.3	-8.4709	1.6912	-5.0089	0.0000
poly(Lat, Lon, Year, 4)0.0.4	-9.4647	1.4174	-6.6776	0.0000
poly(Lat, Lon, Year, 4)0.1.1	528.2320	169.9462	3.1082	0.0019
poly(Lat, Lon, Year, 4)0.2.1	358.2318	157.1354	2.2798	0.0227
poly(Lat, Lon, Year, 4)0.3.1	336.1328	150.5563	2.2326	0.0256
poly(Lat, Lon, Year, 4)0.1.2	-1429.4737	160.6377	-8.8987	0.0000
poly(Lat, Lon, Year, 4)0.2.2	484.5911	144.3935	3.3560	0.0008
poly(Lat, Lon, Year, 4)0.1.3	201.1883	148.6102	1.3538	0.1758
poly(Lat, Lon, Year, 4)1.1.0	-784.0654	224.5959	-3.4910	0.0005
poly(Lat, Lon, Year, 4)2.1.0	257.6674	172.0477	1.4977	0.1343
poly(Lat, Lon, Year, 4)3.1.0	-1540.8319	149.9445	-10.2760	0.0000
poly(Lat, Lon, Year, 4)1.2.0	-547.4902	202.8398	-2.6991	0.0070
poly(Lat, Lon, Year, 4)2.2.0	-56.5769	208.4136	-0.2715	0.7860
poly(Lat, Lon, Year, 4)1.3.0	285.8522	155.9633	1.8328	0.0669
poly(Lat, Lon, Year, 4)1.0.1	-650.3923	182.1696	-3.5703	0.0004
poly(Lat, Lon, Year, 4)2.0.1	411.2486	142.5959	2.8840	0.0039
poly(Lat, Lon, Year, 4)3.0.1	249.0225	132.6895	1.8767	0.0606
poly(Lat, Lon, Year, 4)1.0.2	1244.5853	147.6033	8.4320	0.0000
poly(Lat, Lon, Year, 4)2.0.2	-41.6809	129.5669	-0.3217	0.7477
poly(Lat, Lon, Year, 4)1.0.3	-670.4026	134.9444	-4.9680	0.0000
poly(Lat, Lon, Year, 4)1.1.1	23049.7946	16283.8343	1.4155	0.1570
poly(Lat, Lon, Year, 4)2.1.1	-27700.9674	15575.9505	-1.7784	0.0754
poly(Lat, Lon, Year, 4)1.2.1	-10521.3914	15815.4562	-0.6653	0.5059
poly(Lat, Lon, Year, 4)1.1.2	-37833.5335	14067.3085	-2.6895	0.0072

Table B.5: T-tests for the hypothesis that parameters for the 4 degree polynomial in latitude, longitude and year in the quantitative GLM are zero. Most of the parameters are significantly different from zero on the 5% significance level. Some of the tests for 4 degree terms are accepted on the 5% significance level but there are no grounds for reducing the degree of this polynomial.

B.4 Contour plots of bottom temperatures (section 5.3.1)

Figures B.5 to B.7 show contour plots of the fitted bottom temperature surfaces obtained by a loess smoother. The plots all show the same general trend, the bottom temperature is higher (mostly > 3°C) in the south and east of Iceland than in the north and west of the the country (mostly < 3°C).

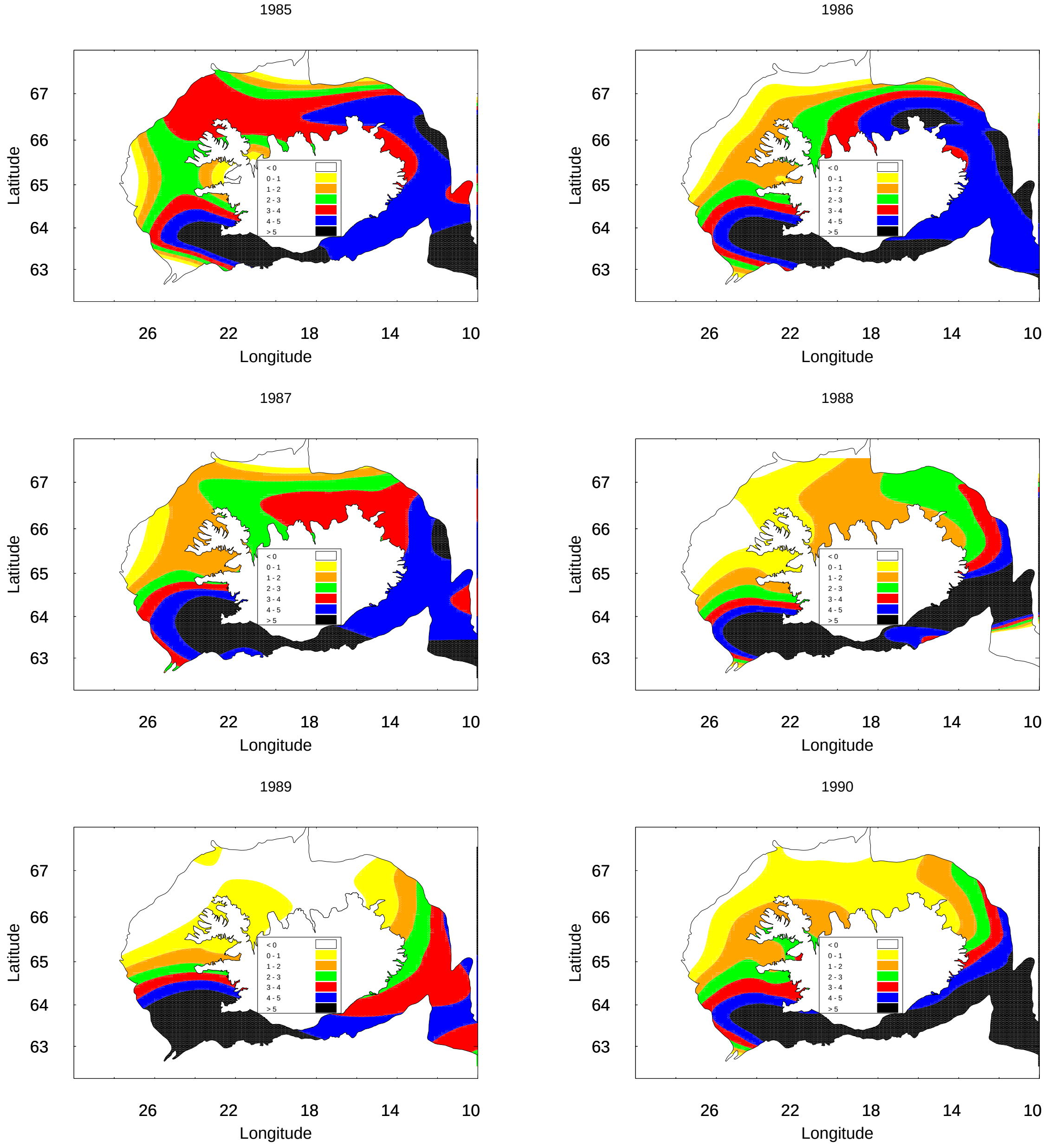

Figure B.5: Contour plots for years 1985 - 1990 of the smoothed temperature surface inside the 500m depth line. Darker colors represent higher temperatures than brighter colors.

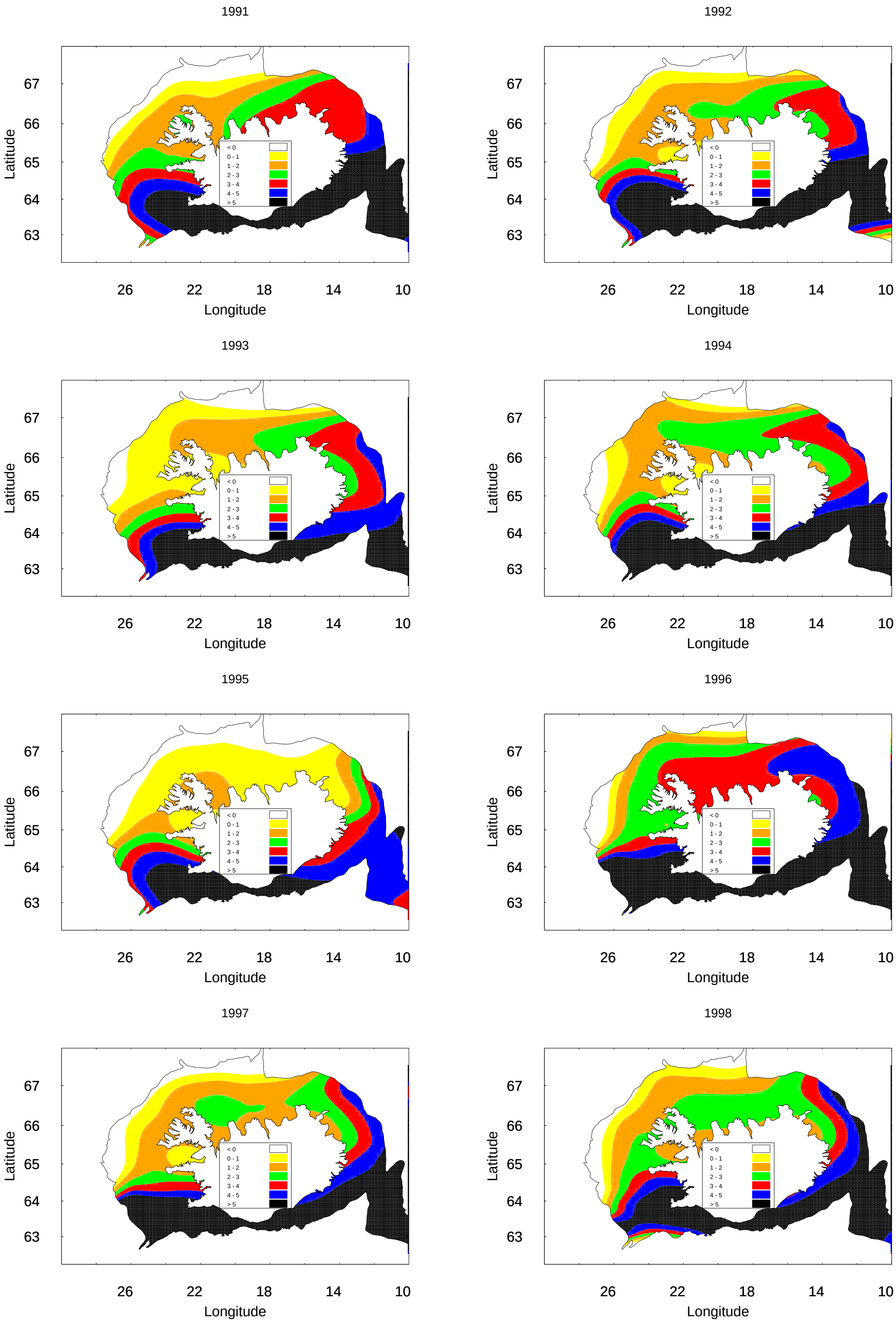

Figure B.6: Contour plots for years 1991 - 1998 of the smoothed temperature surface inside the 500m depth line. Darker colors represent higher temperatures than brighter colors.

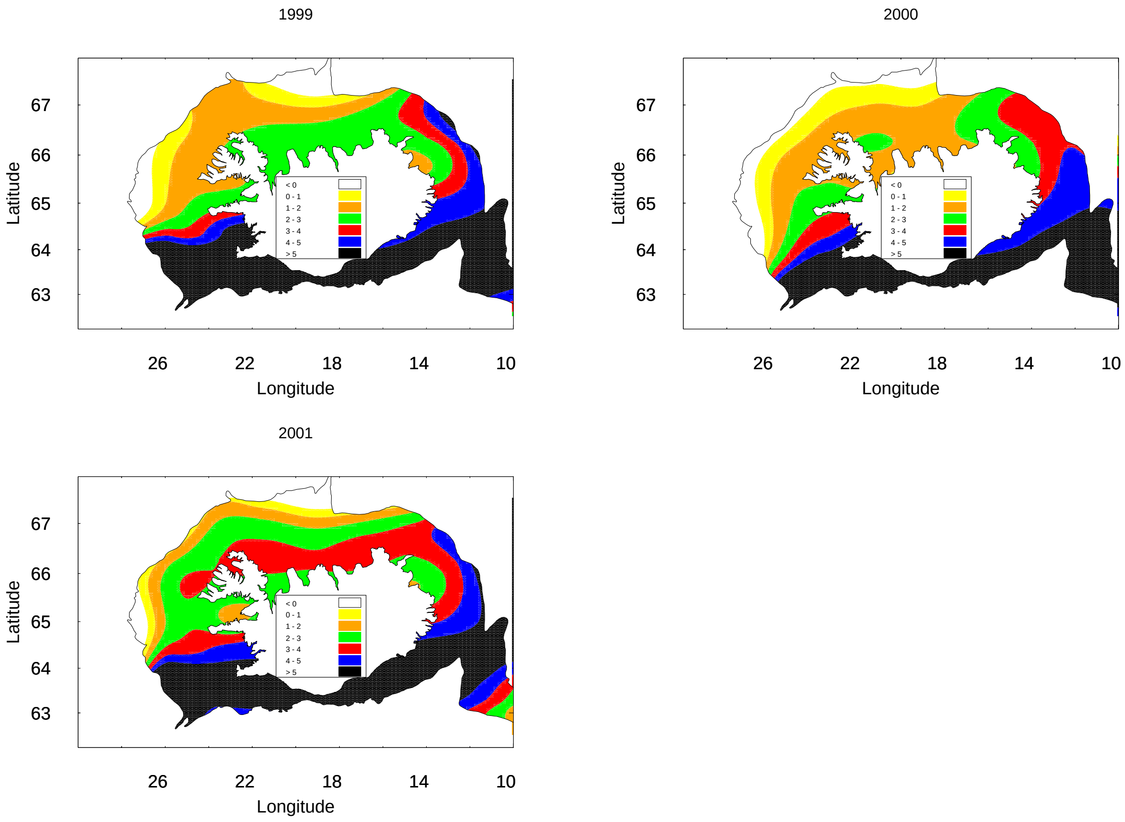

Figure B.7: Contour plots for years 1999 - 2001 of the smoothed temperature surface inside the 500m depth line. Darker colors represent higher temperatures than brighter colors.

B.5 Residual plots for the model containing a gradient term (section 5.3.3)

Figures B.8 and B.9 show plots of the standardized residuals that can be used for an informal model control. These plots show exactly the same patterns as figures 5.11 and 5.12.

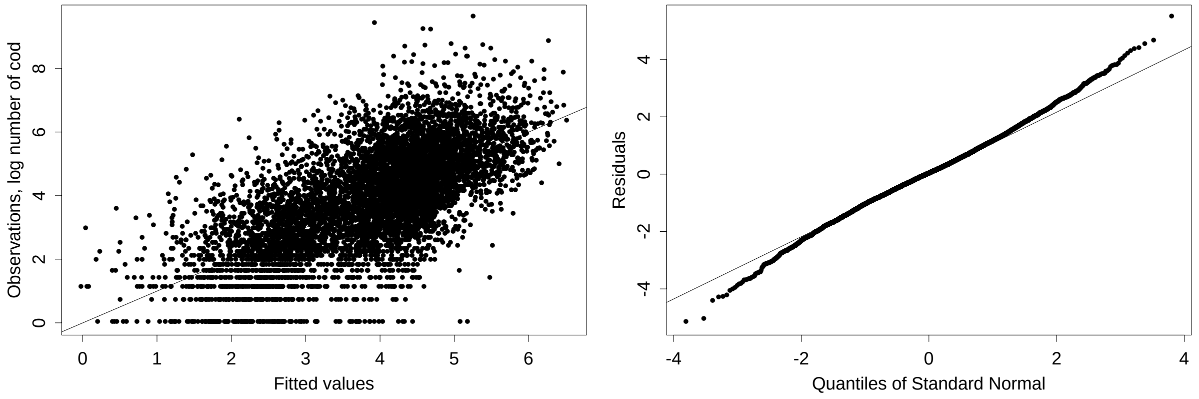

Figure B.8: To the left: Scatter plot of observations versus fitted values along with the $y = x$ line. To the right: Normal probability plot of residuals.

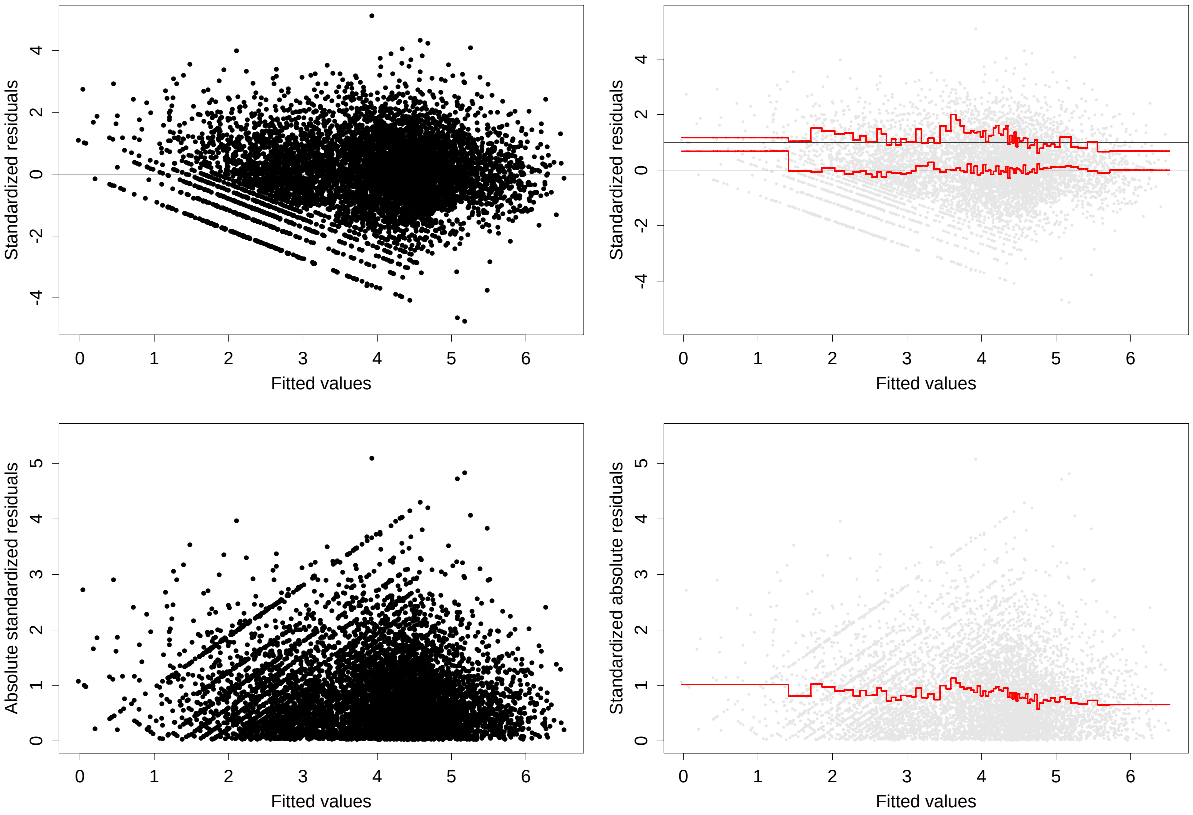

Figure B.9: Scatter plots of standardized residuals (upper row) and absolute standardized residuals (lower row) versus the fitted values. The residuals should be structureless, with mean zero and unit variance. The red lines on the plot in the upper right hand corner show the mean values and variances, calculated for groups of 100 residuals, and the line on the plot in the lower right hand corner show the mean values, also calculated for groups of 100 residuals.

Bibliography

Anon. (1992), Report of the workshop on the analysis of trawl survey data, C.M. 1992/D:6, International Council for the Exploration of the Sea.

Anon. (2001), State of Marine Stocks in Icelandic Waters 2000/2001. Prospects for the Quota Year 2001/2002, MRI Technical Report 87, Marine Research Institute, Reykjavík.

Anon. (2002), dst2. Development of structurally detailed statistically testable models of marine populations. QLK5-CT199-01609. Progress Report for 1 January 2001 to 31 December 2001., MRI Technical Report 87, Marine Research Institute, Reykjavík.

Atkinson, A. C. (1982), 'Regression Diagnostics, Transformations and Constructed Variables', *Journal of the Royal Statistical Society* **44**(1), 1–36.

Bjarnason, K. G. (1997), Upplýsingakerfi skipstjóra, aflakort og aflaspár, Master's thesis, Háskóli Íslands, Verkfræðideild, Reykjavík. (in Icelandic).

Cleveland, W. S. & Devlin, S. J. (1988), 'Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting', *Journal of the American Statistical Association* **83**(403), 596–610.

Conradsen, K. (1999), *En introduktion til statistik*, Vol. 1A and 1B, Institute of Mathematical Modelling, Technical University of Denmark, Lyngby. 545 p. (in Danish).

Einarsson, S. T., Jónsson, E., Björnsson, H., Pálsson, J., Schopka, S. A. & Bogason, V. (2002), *Handbók um stofnmælingu botnfiska á Íslandsmiðum 2002*, Hafrannsóknastofnunin (The Marine Research Institute in Iceland). (in Icelandic).

Firth, D. (1988), 'Multiplicative Errors: Log-normal or Gamma?', *Journal of the Royal Statistical Society* **50**(2), 266–268.

Goñi, R., Alvarez, F. & Adlerstein, S. (1999), 'Application of generalized linear modelling to catch rate analysis of Western Mediterranean fisheries: the Castellón trawl fleet as a case study', *Fisheries Research* **42**, 291–302.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall/CRC, London. 335 p.

Lo, N. C., Jacobson, L. D. & Squire, J. L. (1992), 'Indices of Relative Abundance from Fish Spotter Data based on Delta-Lognormal Models', *Canadian Journal of Fisheries and Aquatic Sciences* **49**(12), 2515–2526.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, 2 edn, Chapman & Hall/CRC, London. 511 p.

Montgomery, D. C. (2001), *Design and Analysis of Experiments*, 5 edn, John Wiley & sons, New York. 684 p.

Myers, R. A. & Pepin, P. (1986), The Estimation of Population Size From Research Surveys Using Regression Models, C.M. 1986/D:9, International Council for the Exploration of the Sea (ICES).

Nelder, J. A. & Wedderburn, R. W. M. (1972), 'Generalized Linear Models', *Journal of the Royal Statistical Society* **135**(3), 370–384.

Pennington, M. (1983), 'Efficient Estimators of Abundance, for Fish and Plankton Surveys', *Biometrics* **39**, 281–286.

Pennington, M. (1996), 'Estimating the mean and the variance from highly skewed marine data', *Fishery Bulletin* **94**(3), 498–505.

Pálsson, O. K., Jónsson, E., Schopka, S. A., Stefánsson, G. & Steinarsson, B. . (1989), 'Icelandic Groundfish Survey Data Used to Improve Precision in Stock Assessment', *Journal of Northwest Atlantic Fishery Science* **9**(1), 53–72.

Sakuma, K. M. & Ralston, S. (1995), 'Distributional patterns of late larval groundfish off central California in relation to hydrographic features during 1992 and 1993', *California Cooperative Oceanic Fisheries Investigations Reports* **36**, 179–192.

Scheffé, H. (1959), *The Analysis of Variance*, John Wiley & sons, New York. 477 p.

Smith, S. J. (1990), 'Use of Statistical Models for the Estimation of Abundance from Groundfish Trawl Survey Data', *Canadian Journal of Fisheries and Aquatic Sciences* **47**(5), 894–903.

Stefánsson, G. (1988), A statistical analysis of Icelandic trawler reports, 1973-1987, C.M. 1988/D:13, International Council for the Exploration of the Sea.

Stefánsson, G. (1996), 'Analysis of groundfish survey abundance data: combining the GLM and delta approaches', *ICES Journal of Marine Science* **53**(3), 577–588.

Stefánsson, G. & Pálsson, O. K. (1997), 'Statistical evaluation and modelling of the stomach content of Icelandic cod (Gadus morhua)', *Canadian Journal of Fisheries and Aquatic Sciences* **53**(2), 89–93.

Steinarsson, B. & Stefánsson, G. (1986), Comparison of random and fixed trawl stations in Icelandic groundfish surveys and some computational considerations, C.M. 1986/D:13, International Council for the Exploration of the Sea.

Thyregod, P. (1998*a*), *En introduktion til statistik*, Vol. 3B, Institute of Mathematical Modelling, Technical University of Denmark, Lyngby. p. 103-352 (in Danish).

Thyregod, P. (1998*b*), *Fordelinger med anvendelser i statistik*, Institute of Mathematical Modelling, Technical University of Denmark, Lyngby. (in Danish).

Venables, W. N. & Ripely, B. D. (1997), *Modern Applied Statistics with S-PLUS*, 2 edn, Springer, New York. 548 p.

Vilhjálmsson, H. (1994), 'The Icelandic Capelin Stock', *Journal of the Marine Research Institute, Reykjavik* **XIII**(1), 1–281.

Wiens, B. L. (1999), 'When Log-Normal and Gamma Models Give Different Results: A Case Study', *The American Statistician* **53**(2), 89–93.

Ye, Y., Al-Husaini, M. & Al-Baz, A. (2001), 'Use of generalize linear models to analyze catch rates having zero values: the Kuwait driftnet fishery', *Fisheries Research* **53**, 151–168–93.