Data Analysis in Fisheries Science
A look into the correlation of sea surface temperature and fish body size

Author Jacob Passfield
Supervisor Gustav Delius
Department of Mathematics
University of York
Heslington, York, UK

May 10, 2021

Acknowledgements

A special thank you to Gustav Delius for his supervision and for answering any questions I had promptly and thoroughly. Thank you to my father, Scott Paasfield, and Emily Mallion for their useful comments and proof reading and also to my mother Joanna Passfield, grandparents and friends who have supported me fully throughout my academic career. I am forever grateful.

Abstract

Using a data set of ten million observations, comprised of 335 species recorded around the Australian continent, this report uses R to identify whether there is a significant correlation between sea surface temperature and the body size of particular species. To do this we begin by discussing the simple linear regression model, which includes how to estimate the parameters of the model, namely the correlation coefficient, and the assumptions we make when using the model. Then we discuss the mixed-effects model, which introduces the idea of fixed and random effects, and we explain why this model is better suited to the purpose of our analysis. Both models are demonstrated using *whitefin surgeonfish*, a specific species found in the data set. Following this a suitable mixed-effects model is then applied to every species in the data. After the model is applied, the correlation coefficient is extracted from the data and tested at a significance level of 0.05. This analysis finds that while 34% of species do not have a significant correlation between temperature and body size, 21% of species have a significant positive correlation and 45% have a significant negative correlation. This is surprising as fishes are ectotherms and so are expected to be smaller at higher temperatures. Nevertheless calculating the correlation coefficients can give valuable insight into the potential effects of global warming.

Contents

1 Introduction

A fishery is defined as the industry of rearing and harvesting fish and other aquatic life. With an estimate of 179 million tonnes of fish produced worldwide in 2018, valued at 401 billion US dollars [1], it is no surprise that fisheries serve as a significant ground for economic growth. Beyond commercial advantages, however, the pressing issues of global warming make the analysis of fisheries science paramount for predicting the future of marine populations and, by doing so, protecting the worldwide economy.

An ecosystem is the combination of living organisms and their environment in which energy flows through a biological network (a food web). These networks are more complex in comparison to other ecosystems. Ocean organisms are highly sensitive to changes in their environment and because of this can be significantly affected by global warming [2]. For example, higher temperatures have modified the plankton ecosystem and has consequently decreased the survival rate of young cod in the North Sea [3].

Numerous articles cite that rising temperatures due to warming affect fish body size, with a prediction that body size will decrease between 14 and 24% by 2050 [4]. One study found that for every increase in water temperature by one degree Celsius, the size of fish decreased by between 20 and 30% [5].

Fishes are ectotherms, they cannot regulate their body temperature, so in warmer waters, their metabolism increases requiring more oxygen to support their bodily functions. Warmer waters are a result of anthropogenic emissions of carbon dioxide. Carbon dioxide is absorbed by our oceans, so increased emissions contribute to a higher concentration of it in our oceans. Because of this, less oxygen is available and so fish are not increasing in size simply because they cannot support themselves at bigger sizes. An implication of smaller species would be higher mortality rates as size would leave them at a higher risk of predation, seen with the young cod in the North Sea [5].

This does not apply to all fish species, however. A notable article by Audzijonyte et al (2020) [6] found that in warmer waters, smaller-sized species get smaller whilst the opposite occurs for larger species. Such alterations will have an impact on the food webs within marine ecosystems, although it is not extensively researched.

While we will not be looking at the evolution of fish sizes over time, in this report we will be analysing the correlation between sea surface temperature and fish body size, to therefore predict the potential effects of global warming. This is investigated using the same data used in Audzijonyte et al (2020), where the data consists of a variety of different fish species found around the continent of Australia.

This report uses mathematical modelling on the fish species to estimate model parameters that will give us an insight into the relationship between fish body size and sea surface temperature. Audzijonyte et al (2020) modelled the association of body size and temperature using hierarchical, mixed-effects modelling with Bayesian statistics. We will also use mixed-effects modelling in this report, but with frequentist statistics instead. Before this however, we discuss the simple linear regression model to demonstrate why using mixed modelling is better suited to this analysis. For example, mathematical models must verify a set of assumptions, otherwise we cannot have confidence in the parameters they estimate. All this is shown using a fish species found in the data set, commonly known as *whitefin surgeonfish* before a mixed-effects model is applied to every species found in the data set. Percentages, regarding the relationship between fish body size and temperature, are then calculated from the model results and discussed.

All analysis of the data is conducted using R, which is a powerful statistical tool. It contains a variety of packages that allow a large data set to be processed and modelled with ease. This is necessary when analysing ecological data, which can often be messy and complex. A basic understanding of R is not required to understand this report but the avid reader may want to read the following introductory guides *R for Data Science* and *Basic Fisheries Analysis with R*, [7] and [8], respectively. This will aid in clarifying the code presented at the end of the report, which is used to model and generate figures that highlight the relationship between sea surface temperature and fish body size.

2 Data

The data used in this report is the same data used in Audzijonyte et al (2020) obtained from the Reef Life Survey website [9]. It consists of approximately 9,500,000 observations from around the Australian continent, consisting of 335 different species of fish collected from approximately 21,000 surveys, which runs across a span of 26 years from 1992 to 2018. A survey refers to a specific data collection. Each survey has its own latitude and longitude coordinates, which is then given a unique site code and date.

The data within a survey is collected by two divers using a 50-metre transect line. Each diver records information of all fish seen 5 meters across and above one side of the entire transect line. Put simply: two divers record all observations seen within a $50 \times 5 \times 5 = 500m^3$ area. Observations outside this area are not recorded.

The taxonomic name (species name), species epithet, genus, family, order and size are all recorded for each observation of fish seen. The size of each fish is classed at the discretion of the diver in specific increments, from as small as 2.5 centimetres to as large as 187 centimetres. (For further details of data collection and survey methods visit the Reef Life Survey website) [9].

For their analysis, Audzijonyte et al (2020) assigned each survey to a 0.5 degree geographical cell, which they named geogroups. Then they calculated the annual mean sea surface temperature and added this average to the data set. So, two surveys in the same cell and the same year have the same mean sea surface temperature. There are 284 cells in the data set and their relative locations to Australian coastline can be seen in *Figure 1*.

Data exploration is a key step between data collection and model fitting. This step is to identify and correct any outliers or inconsistencies that could influence the analysis incorrectly. Fortunately, the data we use in this report has already undergone this step and the details of this are described in Audzijonyte et al (2020). However I will mention that Audzijonyte et al (2020) ensured that each of the 335 species had at least 1,000 observations over at least 10 geographic cells, over a period of 5 years or more. For example, the whitefin surgeonfish species has a total of 1,090 observations found in 30 different cells over the span of 10 years.

After data collection and exploration, the data is ready for analysis and mathematical models can be applied. To demonstrate the mathematical models used in this report, I created a data frame containing observations of *whitefin surgeonfish* only. We will use this data set to first illustrate the simple linear regression model, its assumptions and why we must reject the model and use mixed-effects modelling instead.

Then the data set is used to demonstrate mixed-effects modelling and why we can confidently use it to model the relationship between sea surface temperature and fish body size. We test various mixed-effects models and when the optimal model found, it is applied to every species in the original data set.

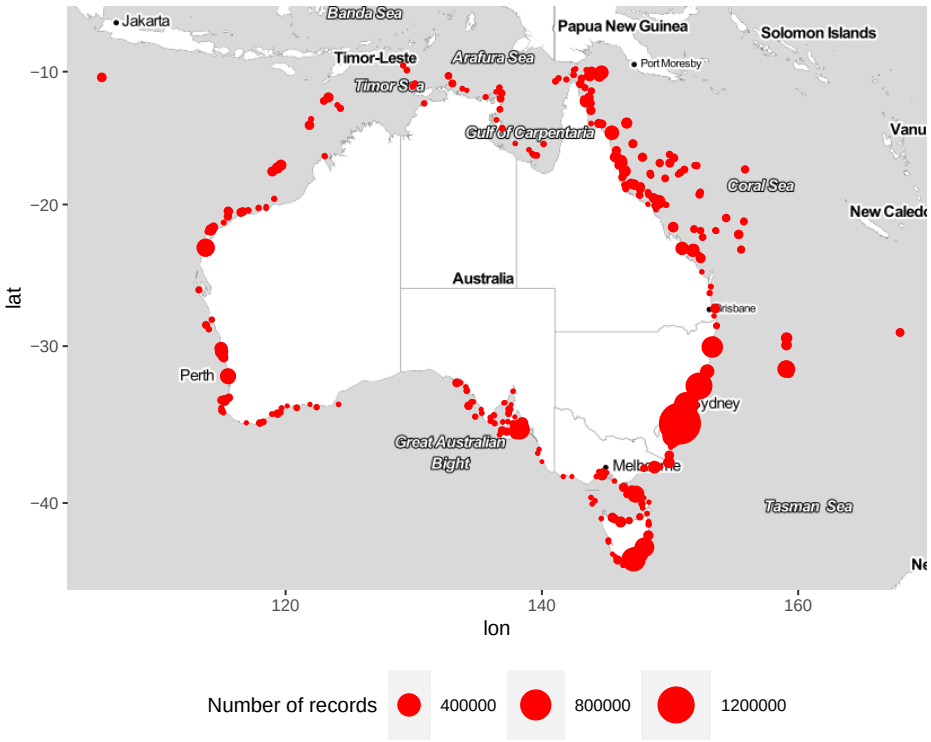

Figure 1: [10] The map of Australia showing the where each geographic cell can be found.
The size of each point corresponds to how many observations were recorded in each geographical cell.

3 Simple linear regression

The simple linear regression model can be perceived as the foundation of statistical modelling. It is often used on a sample to make observations about the entire population.

Definition 3.1 *For a bivariate data set* $(x_1, y_1), \ldots, (x_n, y_n)$ *drawn from an independent and identical distributed sample* $(X_1, Y_1, \epsilon_1), \ldots, (X_n, Y_n, \epsilon_n)$*, the* ***simple linear regression model*** *is defined as*

$$Y_i|\{X_i = x_i\} = \alpha + \beta X_i + \epsilon_i \quad \textit{where} \quad \epsilon_i \sim N(0, \sigma^2)$$

for $i = 1, \ldots, n$*. The model parameters* α *and* β *represent the intercept and slope of the* ***regression line*** $y = \alpha + \beta x$*.* $Y_i|\{X_i = x_i\}$ *is the conditional probability distribution of* Y_i *given* $\{X_i = x_i\}$ *which is is normal with mean* $\alpha + \beta x$ *and variance* σ^2*. The* ϵ_i *are called the* ***residuals****.* [11]*.*

The X-variable is called the explanatory variable and is independent of the other variable, whereas the Y-variable is called the response variable and is dependent on the values of the explanatory variable. Because of this, X and Y are also known as the independent and dependent variable, respectively.

In a perfect scenario, all the points x_i and y_i in the data set would lie along the regression line. Since this is not the case, the random fluctuations (the information that cannot be explained) between points is accounted for in the residual term ϵ_i. The residuals can be thought graphically as the vertical displacement between the regression line and a specific data point [11]:

$$\epsilon_i = y_i - \alpha - \beta x_i, \quad \text{for} \quad i = 1, \ldots, n.$$

The residuals are assumed to be drawn from a normal distribution with a mean of zero and and variance σ^2. This makes the residuals pivotal in assessing whether the simple linear regression model can be used on sample data to estimate the population data [12].

The population parameters and population variance α, β, σ^2 are unknown and need to be estimated from a given sample. The sample variance is denoted by s^2 and estimates for the population parameters are denoted by a hat. For example, the slope of the regression line would be written as $\hat{\beta}$ [12]. To calculate such estimates, R uses maximum likelihood estimation.

Note the intercept α is the value of Y when $X = 0$ and β is the amount by which Y increase when X increases by 1-unit, [12], thus the slope describes the relationship between the X and Y variables [13].

3.1 Maximum likelihood estimation

A sample of the population is collected to obtain estimates of the true population parameters and variance. To estimate α, β and σ^2 in the simple linear regression model, the maximum likelihood principle is used.

The maximum likelihood principle is stated as "given a data set, choose the value of the model parameters in such a way that the data is most likely" [11] and can only be applied if the distribution of the residuals ϵ_i is known. Definition 3.1 makes the assumption that

$$\epsilon_i \sim N(0, \sigma^2)$$

which implies the distribution of Y_i given that $\{X_i = x_i\}$ can be defined as:

$$Y_i|\{X_i = x_i\} \sim N(\alpha + \beta x_i, \sigma^2).$$

Now maximum likelihood can be used to estimate the parameters α, β and σ^2 by

$$L(\alpha, \beta, \sigma) = f_{Y_1}(y_1) \cdots f_{Y_n}(y_n)$$

where

$$f_{Y_i}(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right).$$

The factorisation of the likelihood can be done because Y_i is assumed to be independent. Then working with log likelihood

$$\log(f_i(y_i)) = \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2}$$

gives

$$\log L(\alpha, \beta, \sigma) = l(\alpha, \beta, \sigma) = \sum_{i=1}^{n} \log(f_i(y_i)) = n \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$

The log likelihood $l(\alpha, \beta, \sigma)$ is maximised for any fixed $\sigma > 0$ when $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ is minimal. Let this be a new function, written as

$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

and note that

$$S(\alpha, \beta) = \sum_{i=1}^{n} \epsilon_i^2.$$

As a result, this way estimation is also called *the method of least squares*.

The minimum of the quadratic function $S(\alpha, \beta)$ can be found using the condition that

$$\frac{\partial S}{\partial \alpha}(\hat{\alpha}, \hat{\beta}) = 0 = \frac{\partial S}{\partial \beta}(\hat{\alpha}, \hat{\beta}).$$

For α,

$$\frac{\partial S}{\partial \alpha}(\hat{\alpha}, \hat{\beta}) = 2n\hat{\alpha} - 2\sum y_i + 2\hat{\beta}\sum x_i = 2n(\hat{\alpha} - \bar{y}_n + \hat{\beta}\bar{x}_n) = 0$$

Thus

$$\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n$$

where $\sum_{i=1}^{n}$ is written as $\sum$ to save space. β is calculated similarly and is given by

$$\hat{\beta} = \frac{\sum(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum(x_i - \bar{x}_n)^2}.$$

Finally the maximum likelihood estimate for σ is found by differentiating $l(\alpha, \beta, \sigma)$ with respect to σ such as

$$\frac{\partial}{\partial \sigma} l(\alpha, \beta, \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

and by setting the equation to zero, the maximum likelihood estimate s^2 for σ^2 is given by

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

The estimates $\hat{\alpha}$, $\hat{\beta}$ and s^2 are known as *the least square estimators* for the model parameters and population variance [11]. Fortunately R easily calculates these estimates for us, so there is no need for any handwritten work.

3.2 P-values

We use p-values in this report to assess whether there is a significant correlation between mean sea surface temperature and the size of a fish species.

We make two claims, one called the *null hypothesis* and the other called the *alternative hypothesis*, and use p-values to determine which claim to have confidence in. In our analysis the null hypothesis, denoted H_0, states that there is no correlation between sea surface temperature and fish body size, such as $\beta = 0$. The alternative hypothesis, denoted H_1, states that there is a correlation between the two, such as $\beta \neq 0$ [14]. Our hypotheses can be formulated as follows

$$H_0 : \beta = 0, \quad \text{and} \quad H_1 : \beta \neq 0.$$

We test these hypotheses at a significance level of 0.05, which means we want to be at least 95% confident that there is a relationship between temperature and body size. A p-value that is less than 0.05, such as $p < 0.05$, indicates that there is a significant correlation between the two variables, so we reject the null hypothesis, $\beta \neq 0$, and we can have confidence in the $\hat{\beta}$ estimated by R. Alternatively, a p-value that is greater

than 0.05, such as $p \geq 0.05$, indicates that there is not a significant correlation between the variables and so we accept the null hypothesis that $\beta = 0$. Although calculating p-values are challenging, we can rely on R to produce these values on our behalf.

3.3 Model validation

Once the estimates are obtained, we must verify whether the model can be used in our analysis, otherwise we risk making erroneous conclusions. We do this by verifying whether the data follows the following five assumptions made when we use the simple linear regression model. These are fixed X, linearity, normality, homoscedasticity and independence [12]. If the data does not satisfy these assumptions, we cannot use the model.

3.3.1 Fixed X

A fixed X implies the exact values of the explanatory variables are measured without error. Any error in the measurements of X leads to biased regression parameters making any conclusions inaccurate [13]. This is often the case in ecology so it is usually assumed that the size of the measurement error, x, is small when compared to the random fluctuations in Y [12]. This is assumed in our data set, too; given that the size of fish are only recorded in given size classes, in increments from 2.5 to 187 centimetres, we have a good idea of how much error there is in the measurements, which is clearly visible in the graphs presented later.

3.3.2 Linearity

This assumption ensures that there is a linear relationship between the explanatory variable X and the response variable Y. We can check that our data satisfies this assumption by creating a plot of the residuals against the fitted values. Fitted values are simply the predicted mean response values made by the model. To satisfy the assumption, the points in the graph would fall along a straight line. In this report, I add a blue line to this plot to check this easily. A flat, horizontal blue line would indicate that the linearity assumption has been met [15].

3.3.3 Normality

The normality assumption is the assumption that the residuals are normally distributed with a mean of zero and a variance σ^2, which is stated in definition 3.1. This assumption can be identified by making a histogram of the residuals [13]. The residuals are normally distributed if the histogram follows a bell-shaped curve that peaks at zero along the horizontal axis.

3.3.4 Homoscedasticity

If the spread of the data is not the same at each value of the explanatory variable X then there is a violation of homoscedasticity, called *heteroscedasticity* [13]. Notice that the σ^2 in definition 3.1 does not have an index i attached to it, thus the variance is constant across all the residuals, ϵ_i. Homogeneity prevents having to estimate the

variance for each residual as it is constant over all the residuals, and so constant over all Y_i, meaning only one variance has to be estimated [12].

We can check this assumption by using the residuals versus the fitted values plot. If this assumption is met then the spread of the points in the plot would be equally spread along the horizontal axis, the range of fitted values.

3.3.5 Independence

This assumption means the model assumes that observations within a data set are independent. Data is independent if, for example, two different observations are not influenced by each other. Violations of this assumption could be the result of fitting an improper model to the data, or simply because of the nature of how the data was collected, which is common for spatial and temporal data [13]. For example, when conducting two surveys at two different sites in close proximity to one another, it is likely that the same observation is recorded in both surveys (spatial). It is also likely that an observation recorded in a survey conducted on one day is recorded in a survey conducted on the following day (temporal) [12]. Thus understanding the nature of our data set is crucial is identifying whether this assumption is violated or not, which is discussed later, using the *whitefin surgeonfish* data set as an example.

3.3.6 Next steps

Usually only one or two assumptions are violated and sometimes the linear regression model can be adjusted depending on the data set. For example, if the normality assumption is violated then the response variable could be transformed mathematically. Often, however, other models must be employed, especially when the independence assumption has been violated. For instance, adding more explanatory variables to account for the dependency of the data, turns simple linear regression into multiple linear regression; alternatively, mixed-effects models can be applied which is what we will go on to use in our analysis.

3.4 Whitefin surgeonfish

We model the association between the sea surface temperature and body size of *whitefin surgeonfish* using the simple linear regression model. For an observation i, the size class is given by

$$y_i = \alpha + \beta x_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim N(0, \sigma^2)$$

where y_i is the body length of a specific *whitefin surgeonfish*, x_i is the scaled mean annual sea surface temperature, which I have scaled and centred to have a mean of zero and a standard deviation of one for easier computation. α and β are regression coefficients (the intercept and slope) that are to be estimated. The residuals, ϵ_i are drawn from a normal distribution with a mean of zero and a variance σ^2.

Using R to apply the model to the data, the linear regression equation is estimated as:

$$y_i = 20.79 + 0.96x_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim N(0, 5.837^2).$$

The given p-value for the correlation coefficient β is 8.13×10^{-8}. Since $8.13 \times 10^{-8} < 0.05$, we can confidently conclude that there is a significant positive correlation of 0.96 between the explanatory and response variable.

The estimated linear regression equation is plotted in *Figure 2*.

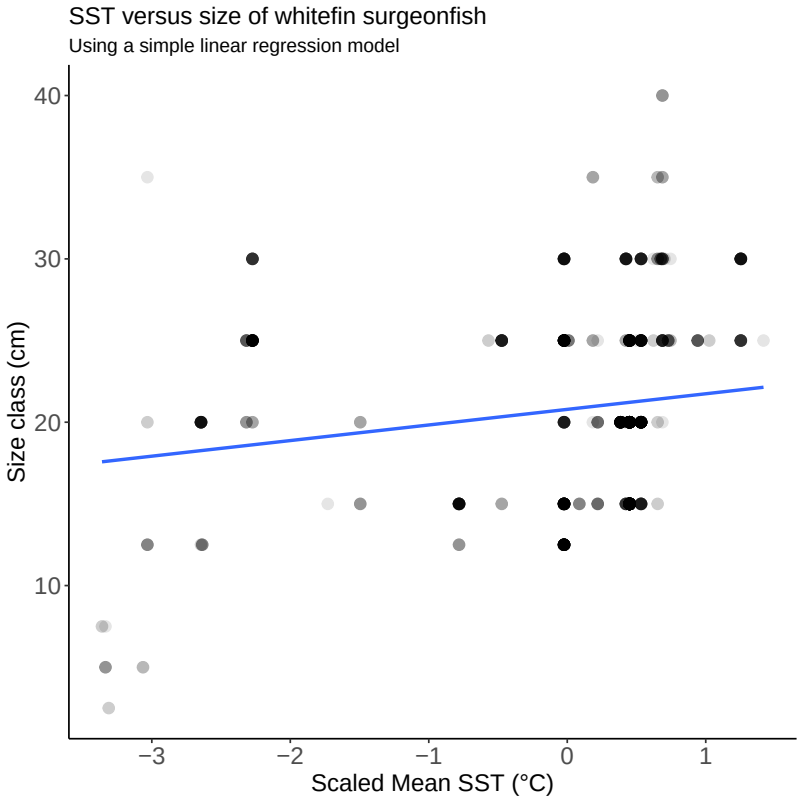

Figure 2: A scatter plot showing the size class of observed *whitefin surgeonfish* against scaled sea surface temperature with a regression line added to visualise the relationship. The points are shaded according to how many observations were found with the same size class at the same temperature; the darker the point, the more observations are recorded at that point.

Figure 2 visualises the positive correlation between the size of *whitefin surgeonfish* and scaled mean sea surface temperature. This indicates that larger observations were recorded in warmer waters and smaller observations were recorded in colder waters. The data set must verify the assumptions made when using the simple linear regression model before we can stand by this conclusion, however. Furthermore, this figure clearly shows that observations are recorded in given size classes, so with a good idea of how much error there is in the measurements, we can assume the assumption of a fixed X is met. *Figure 3* includes two plots to aid the verification of the linearity, homoscedasticity and normality assumptions.

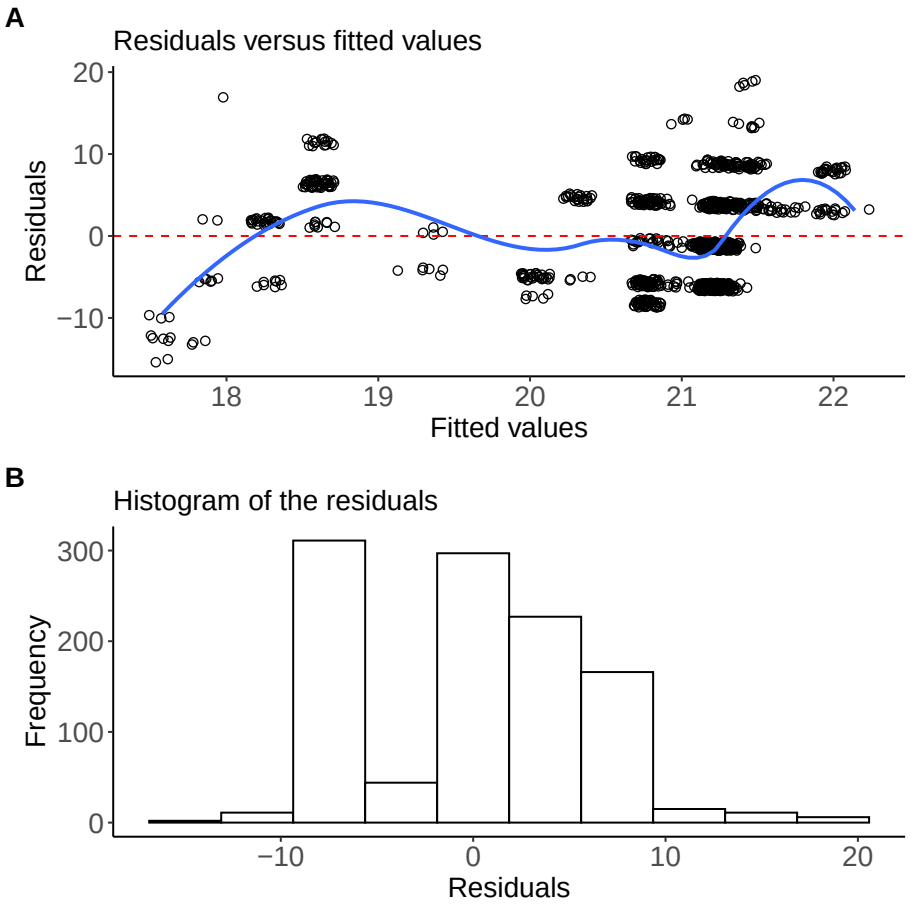

Figure 3: Model validation graphs.
A: Residuals versus the fitted values (linearity, homoscedasticity). The points in this graph have "jittered" to show how many points have the same values; without the jittered effect these points would lie on top of each other
B: Histogram of the residuals (normality).

With the blue line not lying flat and horizontal in *Figure 3A* there is a violation of the linearity assumption. Also, the spread of the points are not equally spread along the range of the fitted values, so the assumption that the residuals have a constant variance is not true and homoscedasticity has been violated. *Figure 3B* show that the residuals do not follow a normal distribution because there is not a peak at zero and if we were to draw a curve following the distribution, it would not be bell-shaped. Thus, the normality assumption is not met. In addition, the nature of the data violates the assumption that the data is independent. In the *whitefin surgeonfish* data set, observations are recorded in 30 different geographical cell over 10 years and in 84 different surveys. Thus it is perfectly plausible that observations found within each cell,

survey and year are more similar (correlated) to each other than to observations found in different cells, years and surveys. With all the assumptions violated, we cannot use the simple linear regression model on this data set and so it is rejected.

Choosing a better model requires an understanding as to why the previous model failed. With observations found in the same cell, year and survey more likely to be correlated to each other than to observations found in a different cell, year and survey, *Figure 4* assesses whether this is true or not. We demonstrate the dependencies between observations using cell only and draw similar conclusions about year and survey.

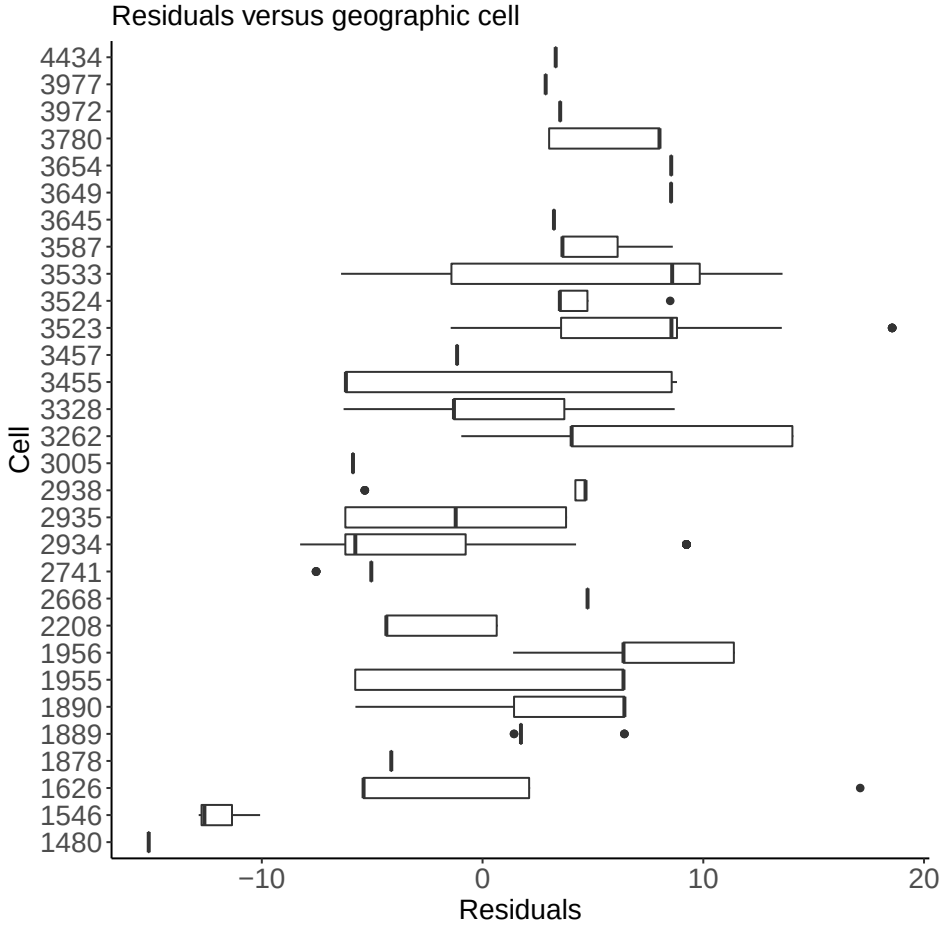

Figure 4: A box plot showing the spread of the residuals *whitefin surgeonfish* in each geographic cell.

Figure 4 highlights that the spread of the residuals is not the same in all cells. Various cells have no spread, cells 3005 and 4434 for example, which could indicate that only one observation of *whitefin surgeonfish* was recorded in the cell. For other cells there is less of spread in the residuals when compared to others. For example, there is less spread in cell 1546 than in cell 3353. Thus *Figure 4* confirms that observations in the data set are not independent and not taking the dependency of the data into account will lead to erroneous conclusions. The same conclusions can be made about year and survey when plotting year and survey against the residuals, too.

To account for the dependency of geographic cell, the *whitefin surgeonfish* data set could be separated by cell and many simple linear regression models can be fitted to each cell. However, *Figure 5* highlights why this is a bad idea.

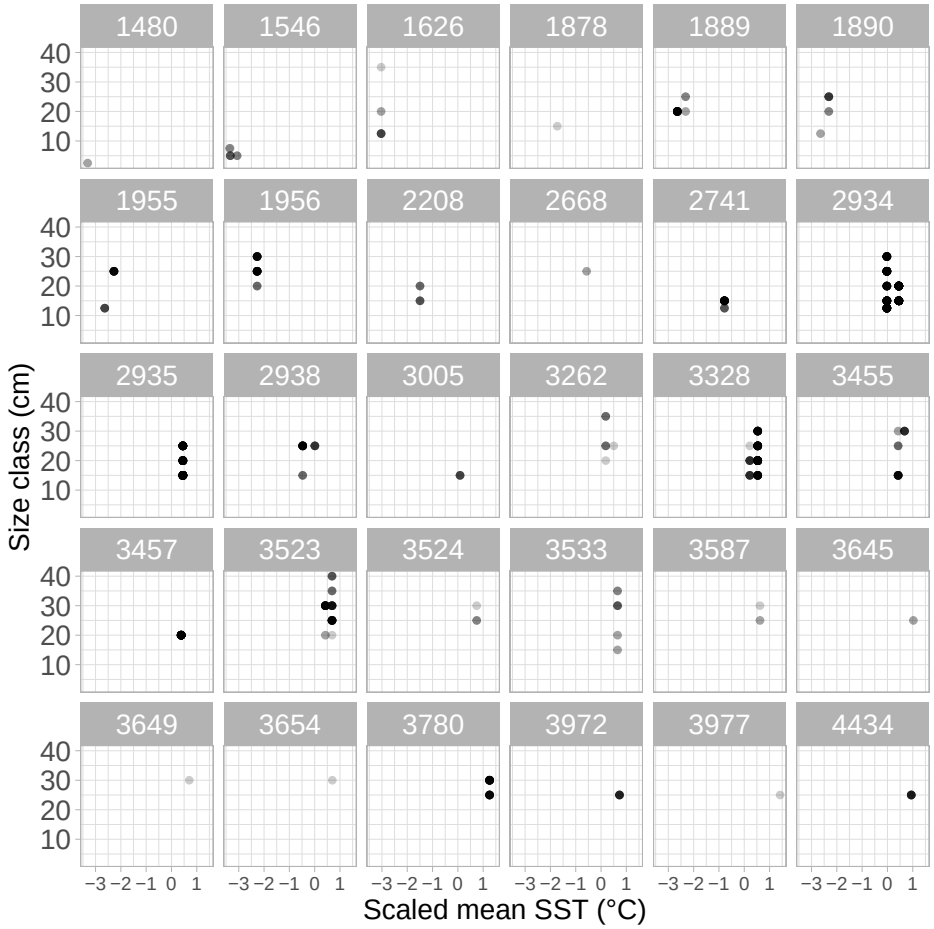

Figure 5: A scatter plot for each geographical cell illustrating the scaled mean sea temperature against the size class of observed *whitefin surgeonfish*. Points are shaded according to how many observations have the same size at the same temperature; the darker the point, the more observations are recorded at that point. Note the x-axis size is small to prevent overlapping numbers.

Running separate analyses for each cell would mean that the simple linear regression model would have to be applied 30 times. With each model requiring an estimate for α, β and σ^2, a total of $30 \times 3 = 90$ parameters would need to be estimated, which increases the risk of making errors. This may be appropriate to do if the sample size in each cell was large enough, but *Figure 5* clearly shows that this is not the case. Furthermore, focusing only on each cell ignores the variation coming from each year and survey, so violations of independence may still present. However, including these variables would continue to inflate the number of estimated parameters, whilst, simultaneously, decreasing the sample size to a negligible number. Finally, the aims of our analysis would require this procedure to be applied to the remaining 334 species found in the original data set. So, the number of models made and the number of parameters that need to be estimated increases exponentially, which carries huge risks of errors and is computationally time-consuming. Most importantly though, this would not help answer the original aim of this report. So overall, running multiple models for each cell, year and survey is not plausible and so we use mixed-effects modelling for our data set.

4 The mixed-effects model

A linear mixed-effects model is one extension of the simple linear regression model. The model incorporates both fixed and random effects and the presence of both of these effects lead to the name *mixed effects* [16].

Fixed effects contain explanatory variables and population parameters, such as α and β, that need to be estimated. For example, the terms α and βx in the simple linear regression model can be thought of as fixed effects in the mixed-effects model. Random effects are not included in a simple linear regression model, however. These effects are determined by the nature of a data set and by the aims of a particular analysis [16]. Furthermore, a random effect is a categorical grouping variable, for example, a cell groups different observations of fish together. It also attempts to explain the unexplained information captured within the residuals. [17].

The advantages of using a mixed-effects model over a simple linear regression model is that we can account for the dependency in the data using only one model. This means that fewer parameters have to be estimated and we can use the entire sample size given [17], immediately overcoming the issues faced in the previous section.

Zuur et al (2009) [13] presents the linear mixed-effects in matrix form as follows:

$$\mathbf{Y_i} = \mathbf{X_i}\,\boldsymbol{\beta} + \mathbf{Z_i}\,\mathbf{b_i} + \boldsymbol{\epsilon_i},$$

which is repeated with the dimensions added for clarity [18]:

$$\mathbf{Y_i} \quad = \quad \underbrace{\mathbf{X_i}}_{n \times p} \quad \underbrace{\boldsymbol{\beta}}_{p \times 1} \quad + \quad \underbrace{\mathbf{Z_i}}_{n \times q} \quad \underbrace{\mathbf{b_i}}_{q \times 1} + \quad \boldsymbol{\epsilon_i}$$

where all terms are $n \times 1$ column vectors. n is the total number of observations, p is the number of explanatory variables and q is the number of random effects. $\mathbf{Y_i}$ is the response vector. The term $\mathbf{X_i}\,\boldsymbol{\beta}$ is the fixed effects term. In the case where $p = 2$ the first element of β is the estimated intercept and the second element is the estimated correlation coefficient (the slope). The unfamiliar term $\mathbf{Z_i}\,\mathbf{b_i}$ is the random effects term where $\mathbf{Z_i}$ is a design matrix and $\mathbf{b_i}$ is a vector of random effects. The $\boldsymbol{\epsilon_i}$ are the residuals capturing the unexplained information [13].

A very simple linear mixed effects model with only a single explanatory variable and with a single random effect affecting only the intercept, can be written as [16]:

$$y_{ij} = \underbrace{\alpha + \beta x_i}_{\text{fixed}} + \underbrace{b_j + \epsilon_{ij}}_{\text{random}}$$

where y_{ij} is the response variable for an observation i in group j for $i = 1, \ldots, r$ and $j = 1, \ldots, s$; x_i is the explanatory variable, b_j is the random effects term for the group j and the ϵ_{ij} is the residual term.

In every version of the linear mixed-effects model, the residuals are still assumed to be normally distributed, but now the random effects are assumed to be normally distributed, as well. For example [17]:

$$\epsilon_{ij} \sim N(0, \sigma_\epsilon^2) \quad \text{and} \quad b_i \sim N(0, \sigma_b^2).$$

Random effects depend on what the aim of the analysis is: one aim may treat a variable as a random effect but another aim may have the variable as a fixed effect. A random effect should have at least five levels to it, or similarly, at least five groups. For example, in the *whitefin surgeonfish* data set, cell can be considered as a random effect because it is a 30 level factor. There are 30 different cells recorded in the data set that splits observations into 30 different groups. If a variable does not have at least five levels to it then the use of a mixed-effects model may not be appropriate. For instance, when analysing a group of people, the sex of a particular person cannot be treated as a random effect because sex is a two level factor (male or female). Sex must be treated as a fixed effect instead and therefore a simpler model can be used in place of a mixed-effects model [17].

There are different types mixed-effects models: the random intercept model, the random intercept and slope model, and the random effects model [13]. We use the random intercept model in this report which allows the intercept to change between the different levels of a random effect but the slope of the regression line remains the same between them. We also use this model because R can create a scatter plot similar to the one created using the simple linear regression model [16].

4.1 Types of random effects

The two types of random effects are: nested and crossed. Distinguishing between a random effect that is nested or crossed is determined by the aims of analysis.

4.1.1 Nested random effects

The concept of nested random effects can be described as similar to Russian nesting dolls. Where a smaller doll is nested inside a bigger doll, a random effect can be nested within another random effect. However, a random effect can only be nested within another random effect if it only appears within a particular group (or level) of the random effect it is supposedly nested within. For example, a smaller doll is a "nested random effect" if it is nested within only one bigger doll [19].

A basic linear mixed-effects model with one nested random effect can be written as:

$$\begin{gathered} y_{ijk} = \alpha + \beta x_i + b_j + b_{jk} + \epsilon_{ijk}, \\ i = 1, \ldots, p, \quad j = 1, \ldots, q, \quad k = 1, \ldots, r \\ b_j \sim N(0, \sigma_j^2), \quad b_{jk} \sim N(0, \sigma_{jk}^2), \quad \epsilon_{ijk} \sim N(0, \sigma_\epsilon^2). \end{gathered}$$

where the model has a nested random effect at two levels: the effect b_j for a group j and the effect b_{jk} for a group k nested within the group j [16].

4.1.2 Crossed random effects

A random effect is crossed if it appears in more than one other random effect. For example, a smaller doll is a "crossed random effect" if it can be nested within more than one bigger dolls [19].

A basic linear mixed-effects model with two crossed random effects can be written as:

$$y_{ijk} = \alpha + \beta x_i + b_j + b_k + \epsilon_{ijk},$$
$$i = 1, \ldots, p, \quad j = 1, \ldots, q, \quad k = 1, \ldots, r$$
$$b_j \sim N(0, \sigma_j^2), \quad b_k \sim N(0, \sigma_k^2), \quad \epsilon_{ijk} \sim N(0, \sigma_\epsilon^2).$$

where the model has two crossed random effects: the effect b_j for a group j and the effect b_k for a group k [16].

4.1.3 Whitefin surgeonfish

In the *whitefin surgeonfish* data set, the data is grouped into 30 geographical cells, 10 years and 84 surveys. So, cell, year and survey can be treated as random effects with 30, 10 and 84 levels, respectively.

Survey can be considered as a nested random effect within cell because a particular survey only appears in one cell. For example, the survey labelled *R2002761* only appears in cell 1480. For a similar reason, survey can be considered nested within year as well.

Year can be considered as a nested random effect within cell, although it may appear that year is a crossed random effect with cell because the same year can appear in two different cells. This is because for a particular observation of a *whitefin surgeonfish*, the corresponding mean sea surface temperature is calculated as the yearly average sea surface temperature recorded in the cell in which the observation was found. For example, observations recorded in cell 1878 in 2016 have a mean sea surface temperature value of 23.3° C, whereas cell 1692 in 2016 has a value of 21.9° C. Therefore, since the temperature is not the same for every year in the data set, year can be considered nested within cell. In light of this, survey can also be considered nested within year nested within cell.

4.2 Maximum likelihood and restricted maximum likelihood estimation

In order to obtain the estimates of the population parameters, α and β, and the variances of the residuals and the random effects, σ_ϵ^2 and σ_b^2, R uses restricted maximum likelihood estimation (REML). REML is preferred over the use of maximum likelihood (ML) estimation as the calculated estimates are normally less biased, whereas estimates calculated using ML are known to be biased [17]. An estimate is biased if its expected value differs from the population parameter it is estimating and so a biased estimate is ”incorrect” by a certain amount.

The estimated variance using ML estimation is biased because ML does not take into consideration the fact that the intercept and slope, α and β, are also estimated. Thus, REML finds less biased estimators. Furthermore, the estimates of α and β found using REML are not the same as the estimates using ML, but they are still unbiased and so in all instances, the use of REML is preferred. For further details regarding REML estimation please refer to *Mixed effects models and extensions in ecology with R* [13].

4.3 P-values

For mixed-effects models we continue to check whether there is a significant correlation between the size of a species and the mean sea surface temperature by using p-values at a significance level of 0.05. However, p-values cannot be calculated exactly for mixed models but any approximation will be good enough for our analysis. Again, R produces these values on our behalf.

4.4 Choosing the optimal model

In this report many linear mixed-effects models are created and then compared with each other in order to find the optimal model. Fortunately R generates an AIC value and a BIC value that can both be used to compare different models.

AIC stands for *Aikaike Information Criterion* and its value is generated by the following formula:

$$AIC = -2\log L + 2p,$$

where p is the number of parameters in the model and L is the likelihood.

BIC stands for *Bayesian Information Criterion* and its value is generated by the following formula:

$$BIC = -2\log L + p\log n,$$

where p is the number of parameters in the model, n is the total number of observations and L is the likelihood.

Without going into too much detail here, both criteria assess the of goodness-of-fit of each model and penalises a model for the number of parameters included. The model with the lowest AIC and/or BIC value is better than the other models and should be chosen for analysis. However, this does not mean the model is the best it can be; these values simply act as a way to compare different models together. Now, although R produces an AIC and a BIC value, this report only uses the BIC value as a comparative measure because this criterion carries a larger penalty for the number of parameters included in each model [19].

4.5 Model validation

The assumptions made when using a mixed-effects model can be verified in the same way as the simple linear regression model. Linearity and homoscedasticity can be verified by plotting residuals against the fitted values and normality can be checked by making a histogram of the residuals. The fixed X assumption for our data set is still met when using a linear mixed-effects model for the same reasons stated previously and the dependency of the data is accounted for when using a mixed-effects model. The only difference between validating the assumptions of a simple linear regression model and a mixed-effects model is that we must verify that the random effects follow a normal distribution. We achieve this by creating a quantile-quantile plot (a QQ-plot) of the random effects. The random effects are normally distributed if the points in a QQ-plot lie along on a straight line, which will be shown later.

4.6 Whitefin surgeonfish

The investigation on the relationship between the sea surface temperature and the body size of *whitefin surgeonfish* is continued, but this time using mixed-effects modelling. We do this in attempts to control the dependency of geographic cell, year and survey in the data set and explain some of the unexplained information previously captured in the residuals. To find approximations for the regression coefficients and the variance of the residuals and random effects, we use restricted maximum likelihood estimation. Recall that there are 1,090 records of *whitefin surgeonfish* found in 30 cells and in 84 surveys over 10 years.

In all the models that follow, y is the body length of a specific *whitefin surgeonfish* observation, x_i is the scaled mean annual sea surface temperature for $i = 1, \ldots, 1090$ and α and β are regression coefficients, treated as fixed effects, which are to be estimated. Each model is assigned a numeric value to compare the BIC values more easily.

The first three equations tested consider cell, year and survey as a singular random effect. For an observation i in either group j for cell, k for year and l for survey, the size of *whitefin surgeonfish* is modelled by:

$$y_{ij} = \alpha + \beta x_i + b_j + \epsilon_{ij}, \quad j = 1, \ldots, 30 \tag{1}$$

where the geographic cell is treated as a random effect.

$$y_{ik} = \alpha + \beta x_i + b_k + \epsilon_{ik}, \quad k = 1, \ldots, 10 \tag{2}$$

where the year is treated as a random effect

$$y_{il} = \alpha + \beta x_i + b_l + \epsilon_{il}, \quad l = 1, \ldots, 84 \tag{3}$$

where the survey is treated as a random effect.

The next model tested considers the entire nested structure of the data set.

$$\begin{gathered} y_{ijkl} = \alpha + \beta x_i + b_j + b_{jk} + b_{jkl} + \epsilon_{ijkl}, \\ j = 1, \ldots, 30, \quad k = 1, \ldots, 10, \quad l = 1, \ldots, 84 \end{gathered} \tag{4}$$

where the model has a random effect at three levels: the effect b_j for the cell, the effect b_{jk} for the year nested within cell and the effect b_{jkl} for the survey nested within year nested within cell.

The last three equations tested consider different combinations of the nested random effects as follows:

$$y_{ijk} = \alpha + \beta x_i + b_j + b_{jk} + \epsilon_{ijk}, \quad j = 1, \ldots, 30 \quad k = 1, \ldots, 10 \tag{5}$$

where the model has a random effect at two levels: the effect b_j for the cell, the effect b_{jk} for the year nested within cell.

$$y_{ijl} = \alpha + \beta x_i + b_j + b_{jl} + \epsilon_{ijl}, \quad j = 1, \ldots, 37 \quad l = 1, \ldots, 84 \tag{6}$$

where the model has a random effect at two levels: the effect b_j for the cell, the effect b_{jl} for the survey nested within cell.

$$y_{ikl} = \alpha + \beta x_i + b_k + b_{kl} + \epsilon_{ikl}, \quad k = 1, \ldots, 10 \quad l = 1, \ldots, 84 \tag{7}$$

where the model has a random effect at two levels: the effect b_k for the year, the effect b_{kl} for the survey nested within year.

For each model, the residuals and random effects are assumed to be drawn from a normal distribution with a mean zero and their own variance σ^2.

Using R to apply the model to the data, each resulting BIC value is extracted and presented in the *Table 1*.

BIC value						
1	**2**	**3**	**4**	**5**	**6**	**7**
6458.854	6646.322	5343.703	N/A	6370.123	5341.405	5341.479

Table 1: The BIC values for each of the 7 linear mixed-effects models applied to the *whitefin surgeonfish* data set.

Reading from *Table 1*, the sixth model has the lowest BIC value, suggesting that this model should be used for our analysis. This is not surprising because a survey informs us of the specific year and cell that it was conducted in.

Note the fourth model does not have a BIC value because R stated that the boundary is singular. This warning occurs when one or more estimated variances are near to zero and is the result of overfitting, where the structure of the random effects is too complex for the data. To prevent this error, the R documentation suggests to avoid fitting complex models and so I disregarded this model.

Before instilling confidence in this model, the data must validate similar assumptions made when using the simple linear regression model.

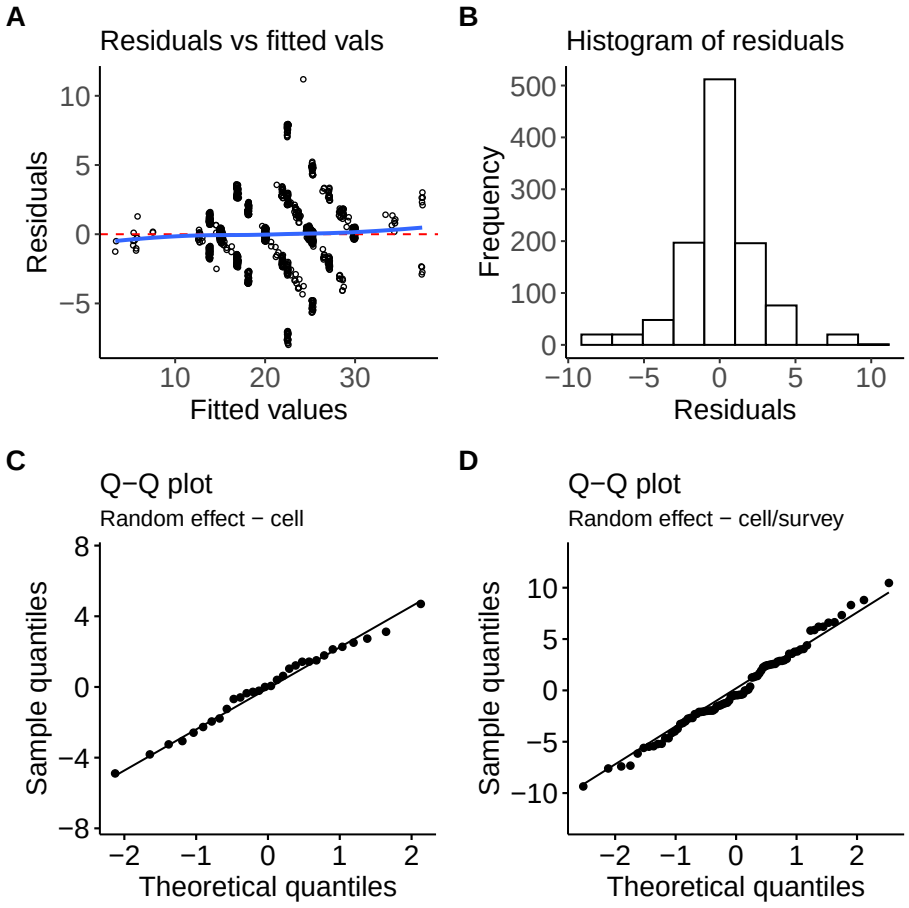

Figure 6: Model validation graphs.
A: Residuals versus fitted value (linearity and homogeneity). Points are jittered to show points with the same value. **B:** Histogram of residuals (normality of residuals).
C: Q-Q plot of random effect of cell (normality of random effects).
D: Q-Q plot of random effect survey nested within cell (normality of random effects).

The blue line lies horizontal and flat in *Figure 6A*, thus the linearity assumption is met. Also, the spread of the residuals appear symmetric and apart from the points at the fitted value of approximately 22 having a larger range than the points at other fitted values, the remaining points are roughly spread equally along the horizontal axis. Thus, we can assume homoscedasticity. In *Figure 6B* the residuals follow a normal distribution. If we were to draw a curve of the distribution then it would bell-shaped with a peak at zero. The random effects of cell and survey nested within cell in *Figures 6C* and *6D* follow a normal distribution since the points in both graphs fall along the straight line. Therefore, the data satisfies the assumption made when using a mixed-effects model and so the scatter plot in *Figure 7* can be made.

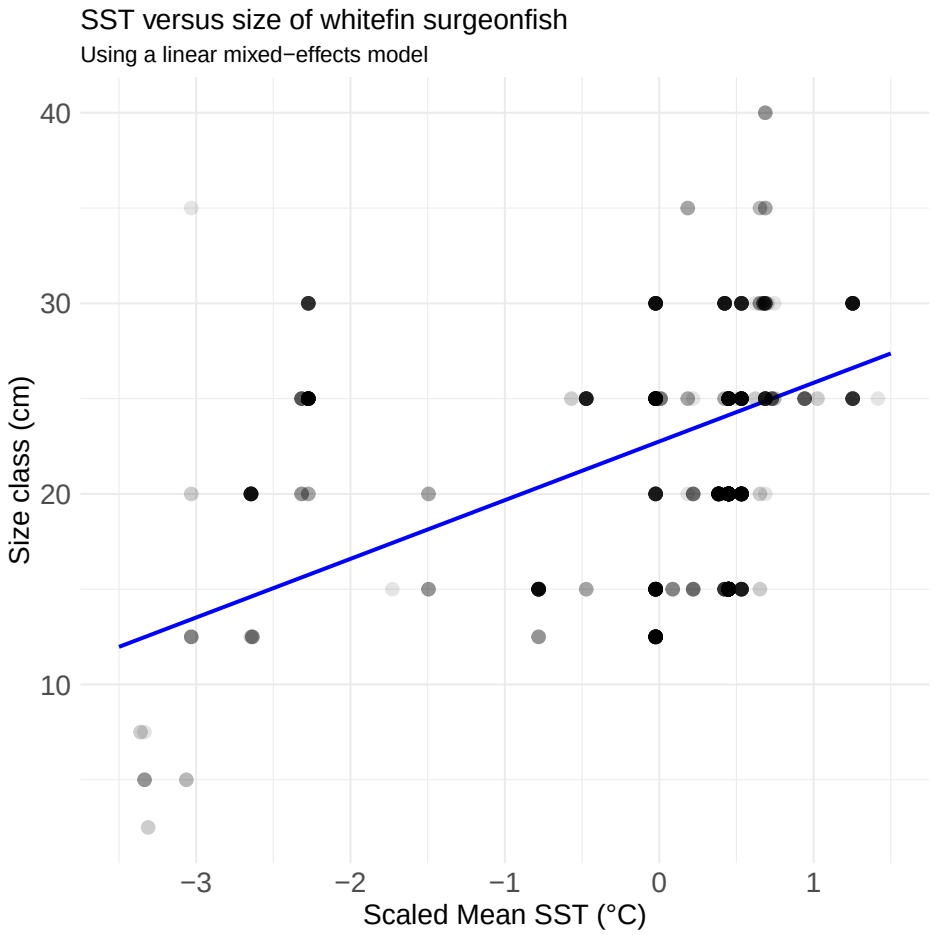

Figure 7: A scatter plot showing the size class of observed *whitefin surgeonfish* against scaled sea surface temperature with a regression line added to visualise the relationship. The points are shaded according to how many observations were found with the same size class at the same temperature; the darker the point, the more observations are recorded at that point.

Figure 7 continues to show a positive correlation between sea surface temperature and the body size of *whitefin surgeonfish*. However, the added regression line appears to be more steep than in *Figure 2*; likely the results of the random effects accounting for the dependency structure in the data and better explaining the previously unexplained information captured in the residuals. These arguments are supported by the summary of the model produced in R, which estimates the linear mixed-effects model as:

$$y_{ijl} = 22.75 + 3.08x_i + b_j + b_{jl} + \epsilon_{ijl},$$
$$i = 1, \ldots, 1090, \quad j = 1, \ldots, 30, \quad l = 1, \ldots, 84$$
$$b_j \sim N(0, 3.256^2), \quad b_{jl} \sim N(0, 4.615^2), \quad \epsilon_{ijl} \sim N(0, 2.423^2).$$

The given p-value for the correlation coefficient β is 8.86×10^{-6}. With $8.86 \times 10^{-6} < 0.05$, we can confidently conclude that there is a significant positive correlation of 3.08 between size and temperature.

Comparing this with the simple linear regression model, $y_i = 20.79 + 0.96x_i + \epsilon_i$ for $i = 1, \ldots, 1090$ where $\epsilon_i \sim N(0, 5.837^2)$, the intercept has a higher estimated value, the slope is steeper and there is less variance in the residuals, all likely because of the random effects, confirming what was concluded above.

4.7 Every species

We can now apply the model to every species in the data set. To do this, I firstly scaled the mean annual sea surface temperature and separated the data by species through a method called *nesting* in R. Nesting the data creates a tibble (a data frame) with 335 rows and 2 columns. Each row, corresponding to one species, has a tibble nested within it, which consists of all the records of that particular species. The mixed-effects model found with the *whitefin surgeonfish* data set was then applied to each species. Thus, for an observation i of a particular species in cell j and survey l, the size of the species is modelled by:

$$\begin{gathered} y_{ijl} = \alpha + \beta x_i + b_j + b_{jl} + \epsilon_{ijl}, \\ i = 1, \ldots, p, \quad j = 1, \ldots, q, \quad l = 1, \ldots, r \\ b_j \sim N(0, \sigma_j^2), \quad b_{jl} \sim N(0, \sigma_{jl}^2), \quad \epsilon_{ijl} \sim N(0, \sigma_\epsilon^2). \end{gathered}$$

where y_{ijl} is the body length of a specific species, x_i is the scaled mean annual sea surface temperature, b_j is the random effect for the geographic cell, b_{jl} is the random effect for the survey nested within the cell and α and β are species specific regression coefficients to be estimated. p is the total number of observations of a species recorded, q is number of geographic cells the species were found in, r is the number of surveys the species were recorded in.

However, when the model was applied to each species two errors arose. The first being that some models had a singular fit, which was seen in the fourth model tried on the *whitefin surgeonfish* data set. The other was that some models failed to converge, which indicates that restricted maximum likelihood could not find a minimum to form a trustworthy estimate, the result of overfitting the model with too many parameters. This convergence error could be overcome by using maximum likelihood to find the estimates instead, but this would lead to biased estimates and not prevent the singular fit error. Thus, I decided to use a simpler model.

Looking back at *Table 1* the next lowest BIC value was the seventh model at 5341.479 which considers year and survey nested within year as a random effect. The same errors occurred when I applied the model to the data. The next lowest BIC value was the third model at 5343.703, which is not too drastically different to the BIC values of the sixth and seventh model. The third model treats survey as a singular random effect. This model producing a the next lowest BIC value is not too surprising because a survey tells us information about the cell and year of a particular observation. Thus it seems appropriate to use the third model for our analysis going forward. Now for an observation i of a particular species in survey l, the size of the species is modelled as:

$$\begin{gathered} y_{il} = \alpha + \beta x_i + b_l + \epsilon_{il}, \\ i = 1, \ldots, p, \quad l = 1, \ldots, q, \\ b_j \sim N(0, \sigma_l^2), \quad \epsilon_{il} \sim N(0, \sigma_\epsilon^2). \end{gathered}$$

where y_{il} is the body length of a specific species, x_i is the scaled mean annual sea surface temperature, b_l is the random effect for survey and α and β are species specific regression coefficients to be estimated. p is the total number of observations of a species and r is the number of surveys the species were recorded in.

Before applying the new model to every species, I tested the model on the *whitefin surgeonfish* data set to ensure the model's validity. The following figure is a scatter plot

and the graphs to validate the model produced by R.

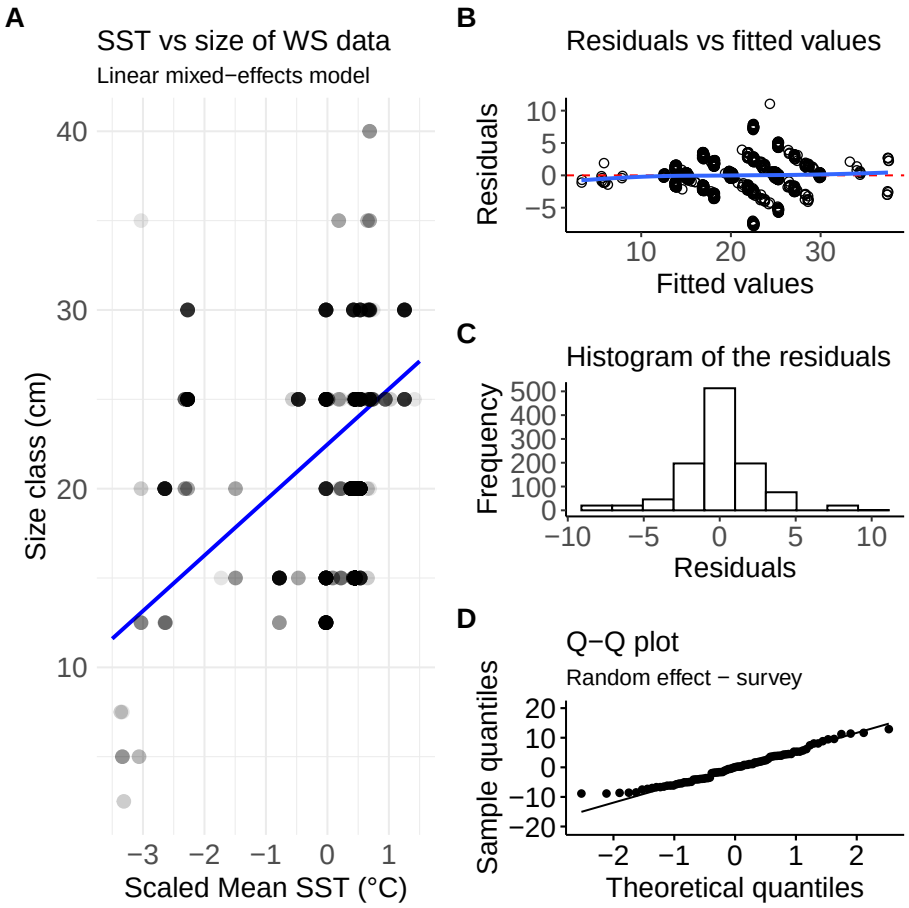

Figure 8: Scatter plot and model validation graphs.
A: Scatter plot of scaled mean sea surface temperature against size class of *whitefin surgeonfish*.
Points are shaded according to the number of observations were recorded with the same values.
B: Residuals versus fitted value (linearity and homogeneity).
Points are jittered to illustrate how many points have the same value.
C: Histogram of residuals (normality of the residuals).
D: QQ-plot of the random effect survey (normality of the random effect).

Figure 9A continues to show a positive correlation between temperature and the size of *whitefin surgeonfish* and *Figures 9B*, *9C* and *9D* show that the data satisfies the model assumptions. Thus the model can be used in our analysis. R estimates the new linear mixed-effects equation as:

$$y_{il} = 22.47 + 3.11x_i + b_l + \epsilon_{il},$$
$$i = 1, \ldots, 1090, \quad l = 1, \ldots, 84$$
$$b_l \sim N(0, 5.589^2), \quad \epsilon_{ij} \sim N(0, 2.424^2).$$

The given p-value for the correlation coefficient β is 1.03×10^{-9}. With $1.03 \times 10^{-9} < 0.05$, we can confidently conclude that there is a significant positive correlation of 22.47 between size and temperature.

Comparing this with the previous mixed model, $y_{ijl} = 22.75 + 3.08x_i + b_j + b_{jl} + \epsilon_{ijl}$ where $b_j \sim N(0, 3.256^2)$, $b_{jl} \sim N(0, 4.615^2)$, $\epsilon_{ijl} \sim N(0, 2.423^2)$, the intercept, slope and residual variance are all very similar in value, but there is now more variance in the random effect.

Fortunately there were no errors when this model was applied to every species in R and so we can start formulating our results. We want to identify the correlation between body size and temperature, thus the estimated correlation coefficient, β, for the scaled mean sea surface temperature is of prime interest. I extracted these values and their corresponding p-values and found that 114 species did not have a significant p-value, indicating that there is no significant correlation between size and temperature. With the 221 species that did have a significant p-value, I created *Figure 9* to show the range of estimated correlation coefficients.

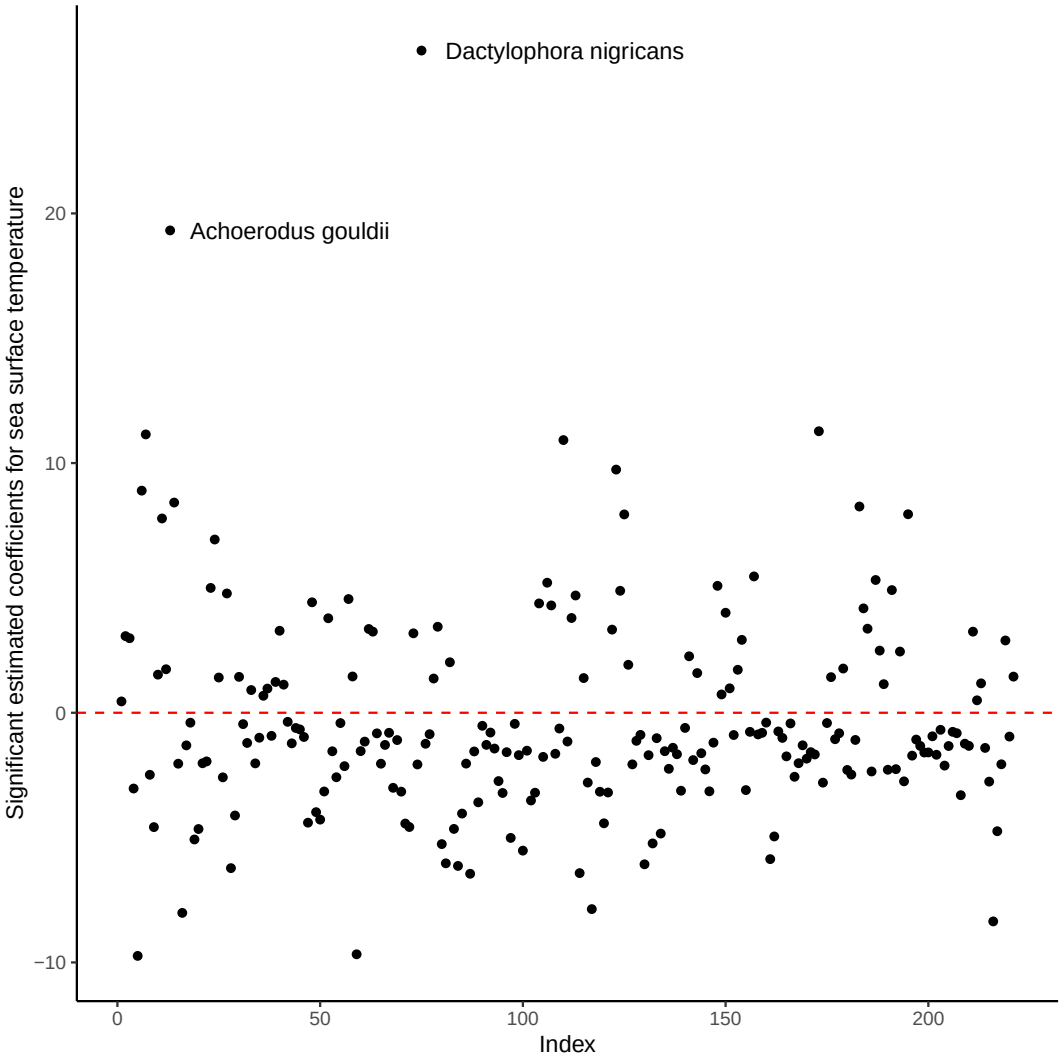

Figure 9: The estimated coefficients of the scaled mean sea surface temperature, β, for the 221 species in the data set that display a significant relationship between seas surface temperature and body size; the p-value is less than 0.05. The dashed red line is added to show the strength of the positive or negative correlation of the sea surface temperature. Points that seem extreme, for example with a $\beta > 15$, are labelled.

With many points below the red line in *Figure 9*, there are clearly many species that have a negative correlation between sea surface temperature and size, and there are many species with a positive correlation. The points with extreme values are the species with an estimated β greater than the value of 15 and such points are labelled accordingly. As a result, I created two separate data sets, one containing all the records of *Achoreodus gouldii* and the other containing all the records of *Dactlophora nigricans*. Then I created model validation graphs for each species to assess why these values have an absolute value that is larger than the other points.

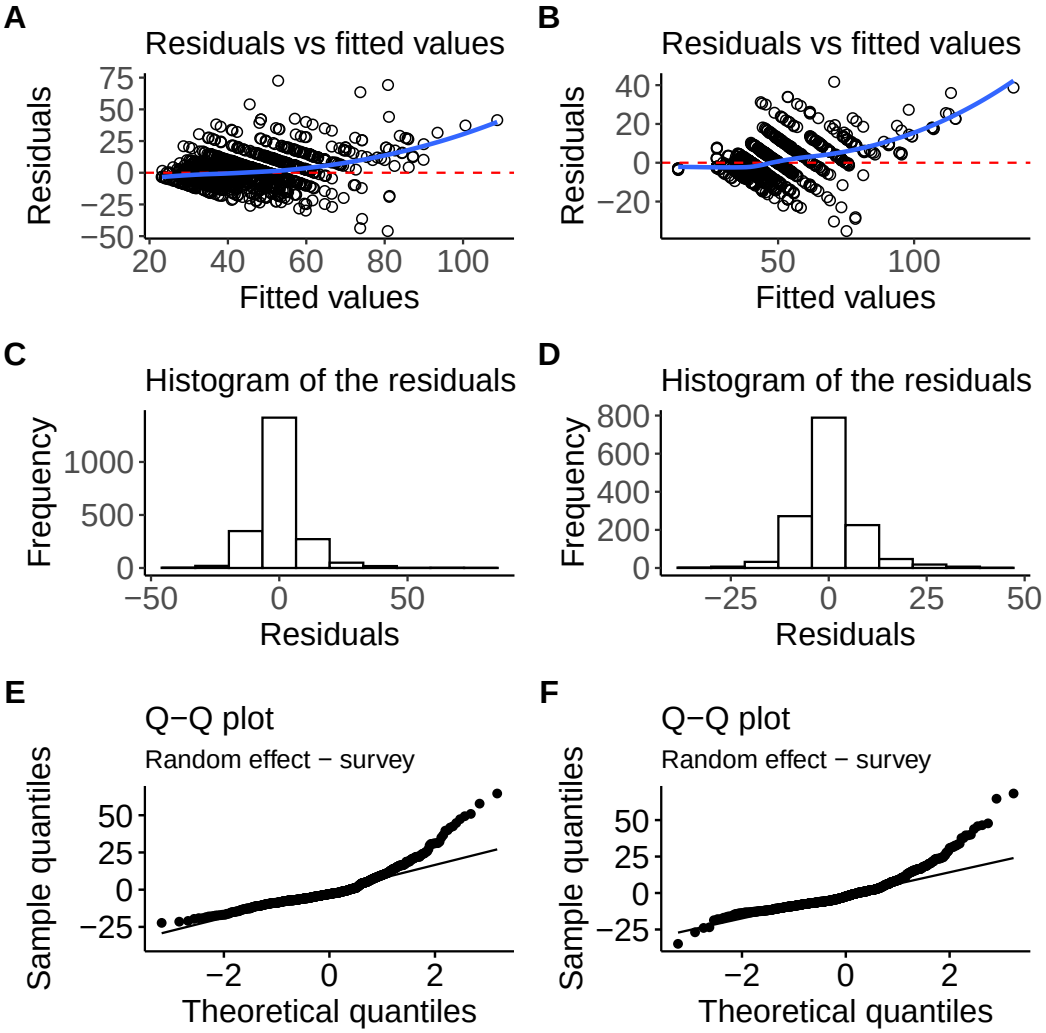

Figure 10: Model validation graphs for *Achoreodus gouldii* in **A, C** and **E** and *Dactlophora nigricans* in **B, D** and **E**.
A, B: Residuals versus fitted value (linearity and homogeneity). Points are jittered to show points with the same value.
C, D: Histogram of residuals (normality of residuals).
E, F: Residuals versus scaled mean sea surface temperature (independence).

The normality assumption is met for both *Achoreodus gouldii* and *Dactlophora nigricans*. If we were to draw a curve for both distributions shown in *Figure 10C* and *10D*, both would be bell-shaped with a peak at zero. There is slight violations of linearity with the blue line in *10A* and *10B* starting off flat and then increasing along the horizontal axis. In addition, the spread of these points are not entirely spread equally, so there is a slight violation of homoscedasticity. Finally, most points in *Figure 10E* and *10F* lie along the straight line but start to trail off at the right-hand side of the horizontal axis, indicating a slight violation of the normality of the random effects assumption. With both data sets violating the assumptions made when using a mixed-effects model, the seemingly extreme estimates of β are understandable. We could exclude these species from our results but removing observations must be taken with caution; extreme values are interesting, especially in ecology. They may not be the fault of measurement error but could show a significant relationship between temperature and body size; ignoring such species could be detrimental at finding ways to support the future of those species. With only two species showing extremity, out of the 221 observed, and with both species only slightly violating the assumptions made when using a mixed-effects

model, I decided to keep the species as they are and include them in my results.

5 Results with discussion

The relationship between sea surface temperature and the body size of species is summarised in *Table 2*. These results were generated by testing the correlation coefficient, β, at a significance level of 0.05.

Correlation coefficient	Percentage of species
Positive	21
Not significant	34
Negative	45

Table 2: The correlation coefficient between the sea surface temperature and fish body size for the 335 different species in the data. Percentages are rounded to the nearest whole number.

The results of this analysis provide empirical solutions specifying the percentage of species, out of the 335 species recorded in the data set, that either have an insignificant correlation coefficient, or a significant positive or negative correlation.

The majority of species have a negative correlation coefficient indicating that smaller fish are found in warmer waters. There is also a sizable proportion of species that are neither bigger or smaller in warmer waters. Moreover, it is surprising how 21% of species are bigger in warmer waters, especially when many studies focus only on the effects of smaller-sized fish.

These results are very different from the results discovered in Audzijonyte et al (2020). They found that 45% of species are smaller in warmer waters, whilst 55% are bigger. This may be the result of them using Bayesian statistics and this report using frequentist statistics, instead. Although, there are potential weaknesses in our analysis. Temperature may not be the only explanation as to why more smaller fish are found in warmer waters. For example, it could be the result that a growing fish naturally migrates from warmer waters to colder waters. Also, with the survival rate of young cod decreased, perhaps other species are not getting the chance to grow to its maximum size so the population of bigger fish is getting smaller.

Other flaws in our analysis are that there are different sample sizes between species. The *whitefin surgeonfish* had only 1,090 observations, whilst *yellowtail horse mackerel* had 914,584 observations, thus if more data was available for each species then perhaps the results may be different and more species would have had a significant correlation coefficient. In addition, using the mixed-effects model was not verified for every species. With *achoredus gouldii* and *dactlophora negricans* showing minor violations of a few assumptions made when using the mixed-effects model, it is perfectly plausible that other species also violate some of the assumptions, too.

However, these results still have value. For example, with nearly half the species in the data set found at smaller sizes in warmer waters, research into how much a fish will decrease is justified. Furthermore, our results support the claim made in Audzijonyte et al (2020). They claimed that not all species shrink with global warming. Although we did not investigate how temperature affects the size of a species, the 21% of species showing a positive correlation could indicate that some fish will get bigger with warming.

Thus, we know that global warming will affect nearly all species of fish in some way. Even so, it is paramount to find ways to manage the body sizes of fish to sustain the livelihood of species in a rapidly warming ocean.

6 Summary

This report began by introducing the simple linear regression model and used R to create figures to visualise how to determine whether a model is appropriate for analysis. Using the *whitefin surgeonfish* species, we discovered that the data violates a set of assumptions that we made using model. Subsequently, we discovered that a mixed-effects model was better suited to our analysis and we applied an optimal mixed model to all the 335 species found within the original data set.

Although the analysis was a much simplified version of Audzijonyte et al (2020), this report explained each step of the analysis thoroughly, so it was easier to follow and understand. In addition, mixed-effects models were introduced using a real and complex data set, which is unlike many introductory tutorials and articles that use a data set specifically designed to introduce mixed-effects models easily.

The resulting analysis identified that while the majority of species' body sizes are negatively correlated with sea surface temperature, a surprising amount of species showed a positive correlation, or showed no correlation at all.

With these patterns identified, we can infer that global warming will have an affect on the majority of fish body size. While many studies have been conducted about the effect of smaller-size species, the effect of larger-sized species has not been well researched. It is therefore critical to investigate what this could mean for the future of marine wildlife and marine ecosystems to protect species from an anthropogenic-caused, ever-warming environment.

Code and data availability

All the data and code used in this report is available at:
`https://github.com/jacobpassfield/final-year-project`.
In addition the code and data used in Audzijonyte et al (2020) is avaliable at:
`https://github.com/astaaudzi/RLSfishSize`.

References

[1] FAO, *The State of World Fisheries and Aquaculture 2020. Sustainability in action.*, FAO, 2020.

[2] GG Matishov, *Effects of Global Warming on Marine Ecosystems*, Climate Change, Human Systems and Policy. EOLSS Publishers, Paris, France (2009), 188–204.

[3] Chen-Yi Tu and K. Chen and Chih-hao Hsieh, *Fishing and temperature effects on the size structure of exploited fish stocks*, Scientific Reports **8** (2018).

[4] William WL and Sarmiento Cheung Jorge L and Dunne, *Shrinking of fishes exacerbates impacts of global ocean changes on marine ecosystems*, Nature Climate Change **3** (2013), no. 3, 254–258.

[5] Daniel and Cheung Pauly William WL, *Sound physiological knowledge and principles in modeling shrinking of fishes under climate change*, Global change biology **24** (2018), no. 1, e15–e26.

[6] Asta Audzijonyte, Shane A. Richards, Rick D. Stuart-Smith, et al., *Fish body sizes change with temperature but not all species shrink with warming*, Nat Ecol Evol **4** (2020), 809–814.

[7] Garrett Grolemund and Hadley Wickham, *R for Data Science*, 2016.

[8] Derek Ogle, *An introduction to basic fisheries analysis with R*, 2015.

[9] *Reef Life Survey*, available at `https://reeflifesurvey.com/`.

[10] David Kahle and Hadley Wickham, *ggmap: Spatial Visualization with ggplot2*, The R Journal **5** (2013), no. 1, 144–161, available at `https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf`.

[11] F.M Dekking, C. Kraaikamp, Lopuhaa H.P., and Meester L.E., *A Modern Introduction to Probability and Statistics*, Springer-Verlag London, 2005.

[12] Alain F. Zuur, Elena N. Ieno, and Graham M. Smith, *Analyzing ecological data*, Springer, New York, 2007.

[13] Alain F. Zuur, Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev, and Graham M. Smith, *Mixed effects models and extensions in ecology with R*, Statistics for Biology and Health, Springer, New York, 2009.

[14] Boundless Statistics, *Hypothesis Testing: Correlations*, available at `https://courses.lumenlearning.com/boundless-statistics/chapter/hypothesis-testing-correlations/`.

[15] Zach, *The Four Assumptions of Linear Regression* (2020), available at `https://www.statology.org/linear-regression-assumptions/`.

[16] José Pinheiro and Douglas Bates, *Mixed-Effects Models in S and S-PLUS*, Statistics and Computing, Springer, New York, NY, 2000.

[17] Gabriela Hajduk, *Introduction to linear mixed models* (2019), available at `https://ourcodingclub.github.io/tutorials/mixed-models/`.

[18] UCLA: Statistical Consulting Group, *Introduction to linear mixed models*, available at `https://stats.idre.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/`.

[19] Yury Zablotski, *Mixed Effects Models 2: Crossed vs. Nested Random Effects* (2019), available at `https://yury-zablotski.netlify.app/post/mixed-effects-models-2/#crossed-and-nested-at-the-same-time`.

[20] Phillip M Alday, Matthias Schlesewsky, and Ina Bornkessel-Schlesewsky, *Electrophysiology reveals the neural dynamics of naturalistic auditory language processing: event-related potentials reflect continuous model updates*, Eneuro **4** (2017), no. 6, available at `https://www.biorxiv.org/content/10.1101/062299v1.supplementary-material`. Note that only the supplementary materials was used in this report, which can be found using the link provided.