# Data Analysis in Fisheries Science
# *A look into how sea surface temperature affects fish body size*

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

April 19, 2021

**Acknowledgements**

Hello Gustav! I hope you have had a lovely Easter break despite everything that is going on.

There are bound to be grammatical inaccuracies, sentences that do not make sense and errors with referencing so apologies in advance. Also there are some sections I have not looked over in days (weeks..) but will do so after submission. I just wanted some version of a complete draft finished in order to get feedback on the whole report and not just a few sections. Still, I hope it makes sense sense and you enjoy reading what I have managed to come up with.

Best,
Jacob.

**Abstract**

Ten million observations comprised of 335 species were recorded around the continent of Australia and it was found that 59% of species shrink with warming, whilst 39% increase in size and only 2% of species unchanged. This is surprising as ectotherms are expected to shrink with warming. Nevertheless the consequences of such changes in fish body size may be detrimental to the future structure of marine ecosystems. These results were generated using linear mixed-effects modelling, but not before an in-depth discussion of the simple linear regression model and its limitations.

# 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 economic 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 global 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. Thus ocean organisms are highly sensitive to changes in their environment and so are affected by global warming [2]. For example, higher temperatures have modified the plankton ecosystem and the survival rate of young cod has decreased 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 20 to 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 getting bigger 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 smaller-sized fish get smaller whilst bigger fish get larger in warmer waters. Such alterations will have an impact on the food webs within marine ecosystems, although it is not extensively research.

Using the same data and methodology used in Audzijonyte et al (2020), this report investigates how sea surface temperature affects the body size of different species of fish around the continent of Australia.

Mathematical modelling on fish population dynamics must be taken to do this and model parameters are to be estimated from collected fisheries data. Mathematical models must satisfy a set of assumptions to feel confident in the accuracy of the estimates found.

Audzijonyte et al (2020) modelled the association of body size and temperature using hierarchical, mixed-effects modelling with Bayesian statistics. This report also uses mixed-effects modelling but with frequentist statistics instead. To make this report more user-friendly it begins with a discussion of simple linear regression, shown with a particular fish species, commonly known as *Pearly Wrasse*, found in the data, to demonstrate why mixed modelling is better suited to this analysis. A basic understanding of statistical concepts is required, however. Then percentages based on the model results about the effects of temperature on fish body sizes are presented and discussed.

All analysis of the data is conducted using R. R 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 is often messy and complex. A basic understanding of R is not required in understanding 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 help understand the code, given at the end of the report, used to model and generate figures that show the relationship between sea surface temperate 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, roughly, 21,000 surveys across the span of 26 years from 1992 to 2018.

It is important to mention that a survey refers to a specific data collection. Each survey consists of its latitude and longitude which is given a unique site code if a survey has not already been conducted there. The date of the survey is also noted.

The data within a survey is collected by two divers using a 50-metre transect line. Each diver records all observations seen 5 meters across and above one side of the entire transect line. Put simply, two divers record all observations seen within a $50 * 5 * 5 = 500m^3$ 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 specified 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].

To analyse how the size of fish change with temperature, Audzijonyte et al (2020) firstly grouped each survey into 0.5 degree geographical cells, called geogroups, and then the yearly mean sea surface temperature was calculated and included in the data set. This means that two surveys with a latitude and longitude within 0.5 degrees of each other are grouped together in the same geographical cell and observations found in the same cell in the same year will have the same mean sea surface temperature. There are 284 cells in the data set and their locations relative to Australia can be seen in *Figure 1* below.

After the data has been collected and before fitting models, the next step in the analysis is exploring the data. This step is to identify and correct any outliers or inconsistencies that could influence the analysis incorrectly. Fortunately, the data in this report has already undergone this stage and the details of how are described in Audzijonyte et al (2020). However, it is important to note that in order to analyse the relationship between temperature and fish body size, Audzijonyte et al (2020) ensured each of the 335 species had at least 1000 observations over at least 10 geographic cells and over at least 5 years. To understand this further, a data frame of how many occurrences each species was seen was created. It was then found that the species that had been recorded the least was *halichoeres margaritaceus*, commonly known as *Pearly Wrasse*, and had only 1008 observations in 37 different geographic cells over the span of 11 years.

With the data ready for analysis, mathematical models can be used to describe

or predict certain aims. A data frame containing only *Pearly Wrasse* observations was created to illustrate the relationship between fish body size and sea surface temperature using simple linear regression first, and then mixed-effects modelling. The mixed-effects model is then implemented to every species in the entire data set.

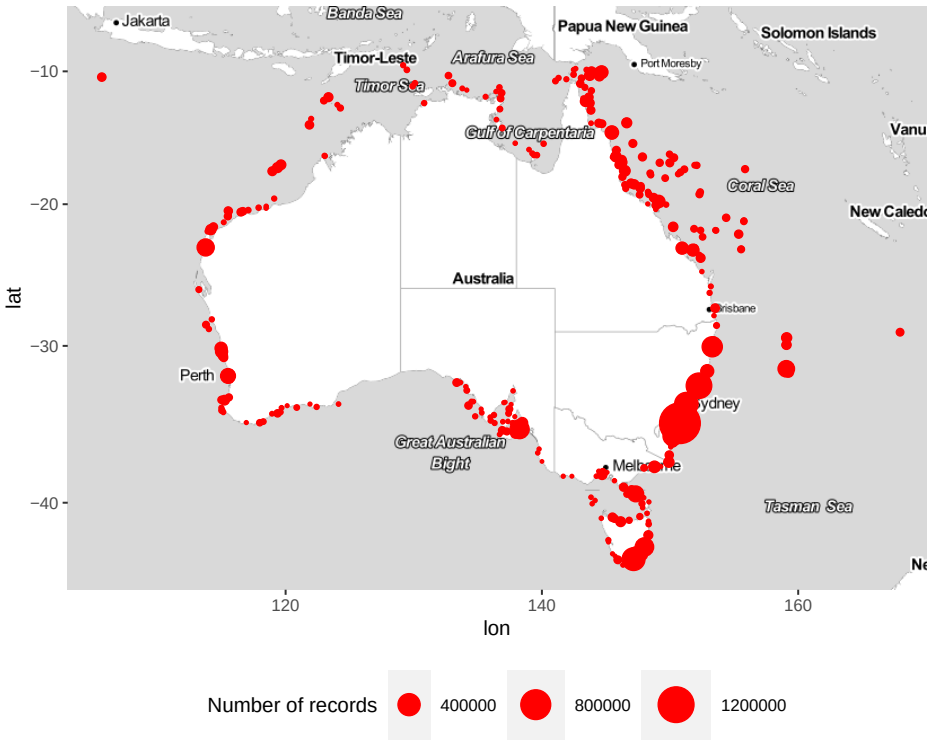


**Figure 1:** [10] The map of Australia showing the where each geographic cell can be found.
The size of each point corresponds to how many observations were recorded in each geographical cell.

# 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 observation 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*

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

*for* $i = 1, \ldots, n$*. The model parameters* $\alpha$ *and* $\beta$ *represent the intercept and slope of the* ***regression line*** $y = \alpha + \beta x$ *which is normal with mean* $\alpha + \beta x$ *and variance* $\sigma^2$*. The* $\epsilon_i$ *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 $\epsilon_i$. 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 $\sigma^2$. 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 $\alpha$, $\beta$, $\sigma^2$ 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 $\alpha$ is the value of $Y$ when $X = 0$ and $\beta$ 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 $\alpha$, $\beta$ and $\sigma^2$ in the simple linear regression model, the maximum likelihood principle is used. Firstly however, a formal definition of an estimate and the introduction to what it means for an estimate to be unbiased.

**Definition 3.2** *Let the data set $(x_1, \ldots, x_n)$ be modelled by random variables $(X_1, \ldots, X_n)$. An **estimate** of a model parameter $\theta$ is a vale $t$ that is only dependent on the given data set. For example, $t = h(x_1, \ldots, x_n)$ where $t$ is defined by some function of the data set. The random variable $T$ is called an **estimator** when $T = h(X_1, \ldots, X_n)$.* [11].

Now 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].

**Definition 3.3** *Let the data set $(x_1, \ldots, x_n)$ be modelled by random variables $(X_1, \ldots, X_n)$ with a joint distribution and a parameter $\theta$. The **likelihood** $L(\theta)$ is the probability of observing the data. Since $X$ is continuous,*

$$L(\theta) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n).$$

*The **maximum likelihood estimate** for $\theta$ is a vale $\hat{\theta}$ that maximises $L(\theta)$. For example, $\hat{\theta} = h(x_1, \ldots, x_n)$. Writing $\hat{\theta}$ in this way makes $\hat{\Theta} = h(X_1, \ldots, X_n)$ the **maximum likelihood estimator** for $\theta$.*

The maximum likelihood principle can only be applied if the distribution of the residuals $\epsilon_i$ is known. Since definition 3.1 assumes that

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

Then the distribution of $Y_i$ can be defined as:

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2).$$

Now maximum likelihood can be used to estimate the parameters $\alpha$ and $\beta$ by

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

where

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

This 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

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

The log likelihood $l(\alpha,\beta)$ 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 $S(\alpha,\beta)$ is found from the condition

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

For $\alpha$,

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

Thus

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

where $\sum_{i=1}^{n}$ is written as $\sum$ to save space. $\beta$ 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 for $\sigma^2$, the maximum likelihood estimator 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 calculates these estimates easily so there is no need for any handwritten work.

## 3.2 Model validation

Once the estimates are obtained, the model must be validated in order to avoid erroneous conclusions. This is done by checking the model satisfying certain assumptions, so to estimate population parameters the sample data must follow the four following assumptions which are: normality, homogeneity, independence and fixed X [12]. Otherwise, the model must be rejected. Note that these assumptions can be verified using statistical tests but this report uses graphs to visualise the assumptions to make it more user-friendly.

### 3.2.1 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 (especially in this report's analysis) that the size of the measurement error $x$ is small when compared with random fluctuations in $Y$ [12].

### 3.2.2 Normality

Normality implies that for each value of $X$ the observations will be normally distributed if sampling is repeated numerous times in exactly the same environmental conditions [12]. This can be checked with a histogram by plotting the values of the response variable at each $X$ value, however there are often not enough samples at each $X$ value to do this. Luckily normality can be checked by making a histogram of the residuals against the explanatory variable and is why $\epsilon_i \sim N(0, \sigma^2)$ is defined in 3.1. Inspecting the residuals to assess normality is also better to do since the effect of the response variable can be explained by the explanatory variable whilst the residuals are not [13]. The residuals are normally distributed if the histogram follows a bell-shaped curve with a peak at one observed value in the horizontal axis.

### 3.2.3 Homogeneity

If the spread of the data is not the same at each value of the explanatory variable $X$ then there is a violation of homogeneity, called *heterogeneity* [13]. Notice that the $\sigma^2$ in definition 3.1 does not have an index $i$ attached to it, thus the variance is constant across all the residuals, $\epsilon_i$. Homogeneity prevents having to estimate the variance for each residual as it is constant over all residual, and so constant over all $Y_i$, meaning only one variance has to be estimated [12].

Violations of homogeneity can be inspected by plotting the fitted values of $Y_i$ against the residuals. If the data is not heterogeneous the residuals will be spread equally along the horizontal axis and not show any non-linear patterns and fortunately R can produce a line, determined by the spread of the residuals, to demonstrate this. A flat and horizontal line would indicate no violations of homogeneity meaning the use of a linear regression model is appropriate for analysis [14].

### 3.2.4 Independence

Data is independent if, for example, two different observations are not influenced by each other. Violation of independence could be the result of an improper model or 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 very close to each other, it is expected to find the same observations in both surveys (spatial). Also it is highly likely observations recorded in a survey on one day are expected to be recorded in a survey conducted the following day (temporal) [12]. The independence of the data can be checked by plotting the residuals against the values of the explanatory variables and by understanding how the data was collected. In the plot however, the spread of the residuals would be constant across the horizontal axis if the data is independent.

### 3.2.5 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, or if the homogeneity assumption is violated then perhaps the variance of the residuals is not constant and so $\sigma^2$ could be changed accordingly. But often times this is not the case and so other models must be used, especially when the independence assumption is violated and new models are introduced. For instance, adding more explanatory variables to account for the dependency of the data turns simple linear regression into multiple linear regression, or mixed-effects models can be used, which can also be used to try and explain why the variance of the residuals is not constant.

## 3.3 Pearly Wrasse

The association between the body size of *Pearly Wrasse* and sea surface temperature was modelled using the simple linear regression model. For an observation $i$ the size class is given:

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

where $y_i$ is the body length of a specific *Pearly Wrasse*, $x_i$ is the scaled mean annual sea surface temperature and $\alpha$ and $\beta$ are regression coefficients (the intercept and slope) to be estimated. The residuals, $\epsilon_i$ are drawn from a normal distribution with a mean of zero and a variance $\sigma^2$.

It must be noted that before the analysis of the *Pearly Wrasse* the independent variables (the values of the annual mean sea surface temperature) was centered and scaled to have a mean of zero and a standard deviation of one (unit-variance). Scaling the explanatory variable ensures estimates are of the same scale leading to easier analysis.

Now applying the model to R estimates the linear regression equation as:

$$y_i = 6.05 - 0.89x_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim N(0, 2.538^2)$$

which is plotted in *Figure 2* below.

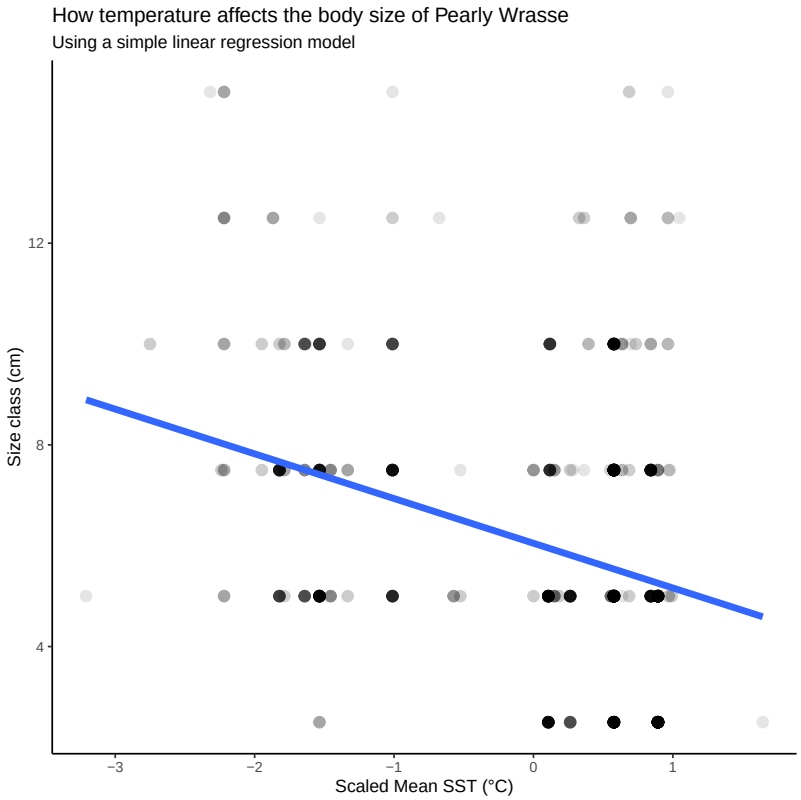


**Figure 2:** A scatter plot illustrating the scaled mean sea temperature against the size class of observed Pearl Wrasse. A regression line is added on top to model the relationship. The points are shaded according to how many observations have the same size at the same temperature. The darker the point, the more observations are recorded at that point.

*Figure 2* illustrates that *Pearly Wrasse* is smaller in warmer waters and is confirmed by the negative estimated slope. However, the assumptions must be validated before this conclusion can be made. *Figure 3* includes four plots to aid the verification of the model.

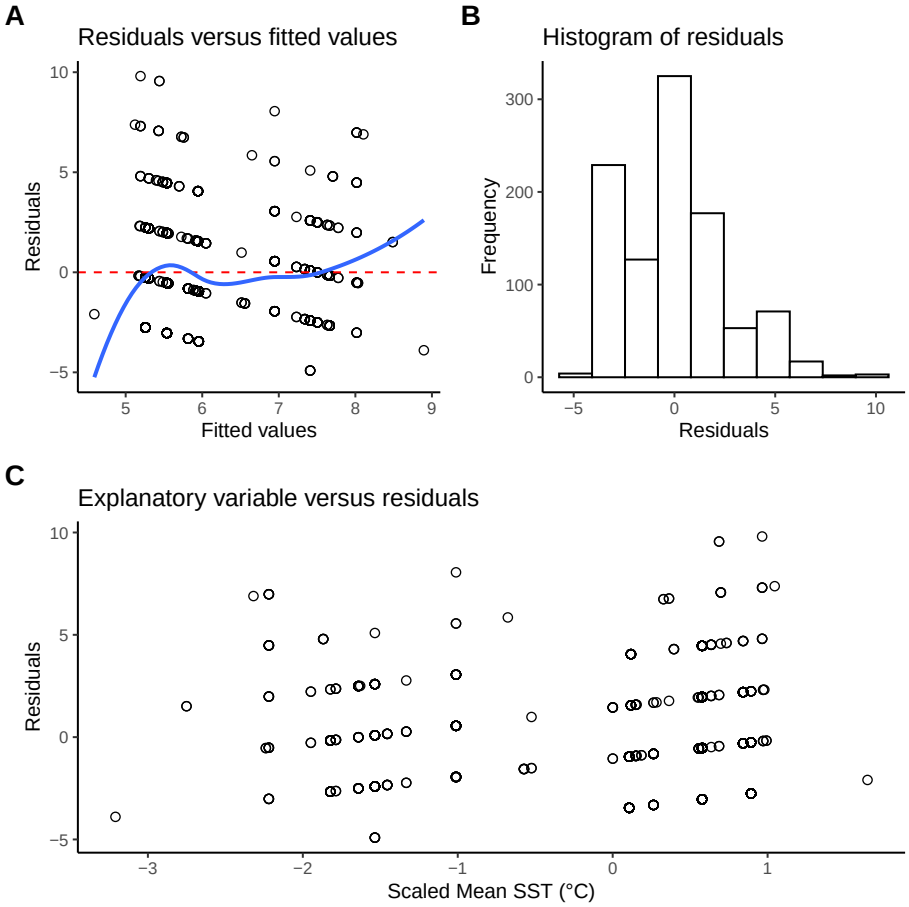


**Figure 3:** Model validation graphs.
**A:** Residuals versus fitted value (homogeneity). **B:** Histogram of residuals (normality).
**C:** Residuals versus scaled mean sea surface temperature (independence).

*Figure 3A* shows that there is a violation of homogeneity because the blue line does not lie flat. Therefore the assumption that the residuals have a constant variance is not true. *Figure 3B* shows that the residuals roughly follow a normal distribution, but there is room for improvement. *Figure 3C* shows that the spread of the residuals is not completely the same at each scaled temperature value indicating a violation of independence. With each assumption violated in some way, the simple linear regression model should be rejected.

Choosing a better model requires an understanding as to why the previous model failed. In the Pearly Wrasse data set data was collected over 11 years, were found in 37 different geographical cells and in 79 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. This is visualised, focusing on the differences in geographic cells only, in *Figure 4* and *5*.

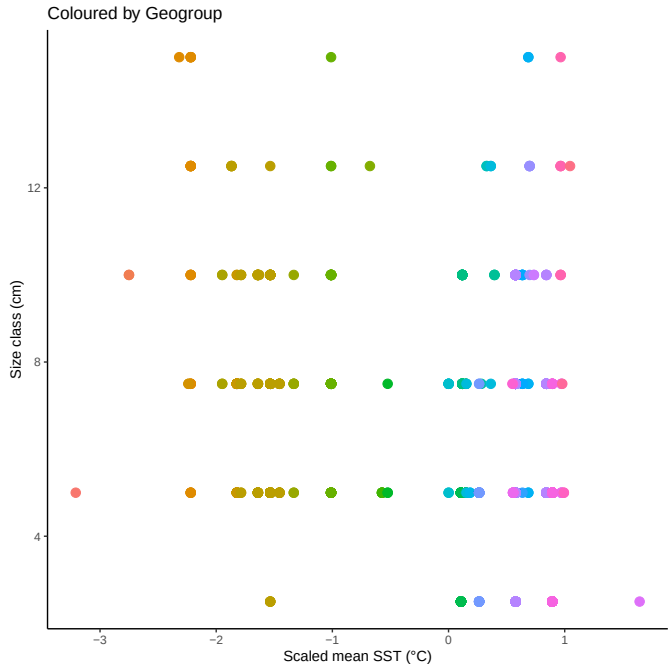


**Figure 4:** A scatter plot illustrating the scaled mean sea temperature against the size class of observed *Pearly Wrasse*. Points are coloured according to which geographic cell was found in.

In *Figure 4* points coloured in pink are plotted to the right of the horizontal axis, whilst points coloured in orange are plotted to the left. There is little to no overlap between different coloured points, indicating that geographical cells vary in both the size of *Pearly Wrasse* recorded and the sea surface temperature at which they were found. This idea is reiterated using box plots in *Figure 5* below.

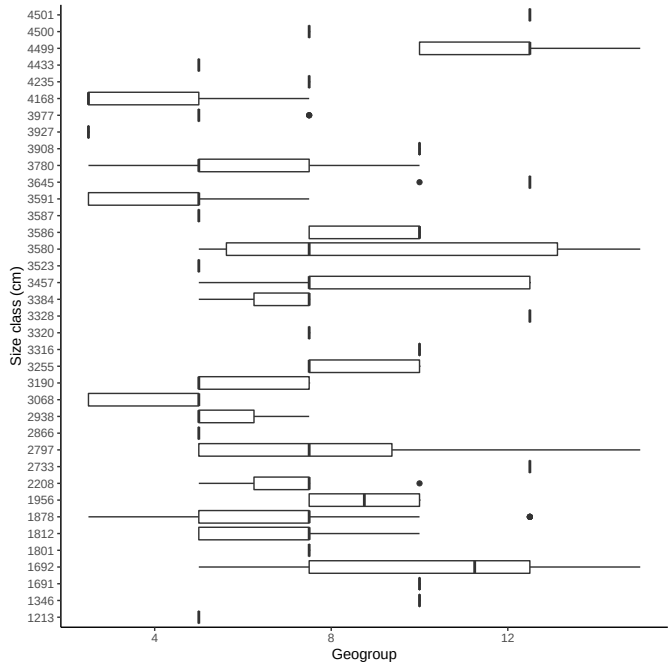


**Figure 5:** A box plot showing the spread of the size *Pearly Wrasse* in each geographic cell.

The spread of each box plot would be similar if the data was independent, which is visibly not the case for the *Pearly Wrasse* data in *Figure 5*. Both figures confirm that

observations from within each of the cells are not independent and ignoring this could lead to erroneous conclusions.

Now the *Pearly Wrasse* data could be separated by geographical cell and then many simple linear regression models could be fit for each cell, but *Figure 6* highlights why this is a bad idea.

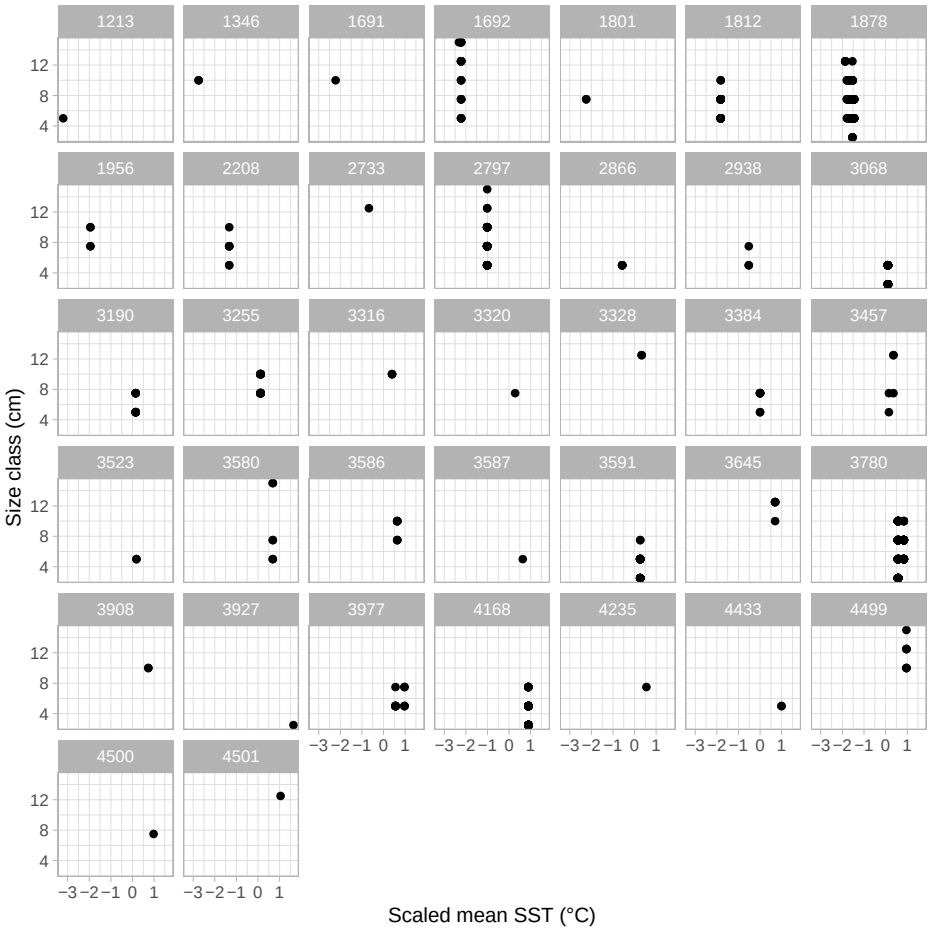


**Figure 6:** A scatter plot for each geographical cell illustrating the scaled mean sea temperature against the size class of observed *Pearly Wrasse*.

Running separate analyses for each cell means the simple linear regression model would have to be applied 37 times. With each model needing an estimate for $\alpha$, $\beta$ and $\sigma^2$ a total of $37 * 3 = 111$ parameters would need to be estimated, which carries the risk of making errors. This may be appropriate to do if the sample size in each cell was large enough, but *Figure 6*, and even *Figure 5*, clearly shows that this is not the case. Furthermore, focusing only on each cell ignores the variation from year and survey so there may still be violations of independence. Including these variables, however, would continue to inflate the number of estimated parameters, whilst, simultaneously, decreasing the sample size to an negligible number. Finally, this procedure would have to be applied to the other 334 species found in the data set. This means thousands of models would need to be made and thousands of parameters would need to be estimated. But most importantly doing this would not help answer the original aim of this report. Thus running multiple models for each cell, year and survey is not plausible and where mixed-effects modelling is better suited for our analysis.

# 4 The mixed-effects model

Linear mixed-effects modelling 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* [15].

Fixed effects contain population parameters that need to be estimated, the $\alpha + \beta x$, the regression coefficients and the explanatory variables, in the simple linear regression model can be thought of as fixed effects and these two terms are still included in mixed-effects modelling. Random effects are not used in simple linear regression modelling, however. These effects are based on the sample of individual observations that are drawn from a population at random [15] and are categorical variables that need to be controlled [16].

The advantages of mixed-effects modelling is that all the data is used, any correlation between observations is accounted for and fewer parameters have to be estimated [16], immediately overcoming 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 to the vectors and matrices for clarity [17]:

$$\overbrace{\mathbf{Y_i}}^{n_i \text{ x } 1} = \overbrace{\underbrace{\mathbf{X_i}}_{n_i \text{ x } p}\ \underbrace{\boldsymbol{\beta}}_{p \text{ x } 1}}^{n_i \text{ x } 1} + \overbrace{\underbrace{\mathbf{Z_i}}_{n_i \text{ x } q}\ \underbrace{\mathbf{b_i}}_{q \text{ x } 1}}^{n_i \text{ x } 1} + \overbrace{\boldsymbol{\epsilon_i},}^{n_i \text{ x } 1}$$

where $\mathbf{Y_i}$ is the response vector, the matrix term $\mathbf{X_i}\,\boldsymbol{\beta}$ is treated as a fixed effects term where $\mathbf{X_i} = \begin{pmatrix} 1 & X_i \end{pmatrix}$ and $\boldsymbol{\beta}$ has two elements: the first being the intercept and the second being the usual $\beta$ (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]. The $n_i$ corresponds to the total number of observations.

Alternatively a linear mixed-effects model can be simply written as [15]:

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

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

In every version of the linear mixed-effects model, not only are the residuals still assumed to be normally distributed, but now the random effects are assumed to be normally distributed, too. For example [16]:

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

It is important to stress that random effects depend on what the aim of the analysis is: one aim may treat a variable as a random effects but another aim may have the variable as a fixed effect. Furthermore the term *random* refers to how certain variables

are grouped and is not related to the idea of randomness in mathematics. Finally a given random effect should have at least five levels to it. For example in an analysis of a group of people, the sex of a particular person could not be treated as a random effect because sex is a two level factor (male or female) and so it must be treated as a fixed effect instead [16].

There are different types mixed-effects models, for example, the random intercept model, the random intercept and slope model and the random effects model [13]. This report focuses on the random intercept model which allows the intercept to change between each the different levels of a random effect but the slope of the regression line remains the same between them. This model is used because it allows R to create a scatter plot that looks similar to the scatter plot created using the simple linear regression model [15].

## 4.1 Types of random effects

Not only are there different types of mixed-effects models, but there are different types of random effects, also; the two types of random effects are: crossed random effects and nested random effects.

### 4.1.1 Crossed random effects

A crossed random effect is a variable whose observations can occur between different levels of the particular variable. For example, the geographical cell can be thought as a crossed effect as fish species occur in more than one different cells.

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

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

### 4.1.2 Nested random effects

Hierarchical, or multilevel, linear mixed-effects modelling is a mixed-effects model that contains nested random effects. Nested random effects can be thought similar to Russian nesting dolls: some variables in the data set are nested within other variables.

Now a variable is considered a nested effect if it only appears within a specific level of another variable [16]. For example, the variable year can be thought as a random effect nested within the variable cell, and this is the same for survey. Survey is a variable that is nested within year and nested within cell.

Although the variable year may be thought as a crossed random factor initially, for this analysis it is nested. Even though the same year can be recorded in two different cells, the mean sea surface temperature is a yearly average, specific to each cell. This makes the year specific to each cell, thus year is nested within cell.

A basic linear mixed-effects model can be written as:

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

where the model has a nested random effect at two levels: the effect $b_j$ for a variable $j$ and the effect $b_{jk}$ for a variable $k$ nested within the variable $j$ [15].

## 4.2 Maximum likelihood and restricted maximum likelihood estimation

In order to obtain the estimates of the population parameters, $\alpha$ and $\beta$, and the variances of the residuals and the random effects, $\sigma_\epsilon^2$ and $\sigma_b^2$, 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 [16].

**Definition 4.1** *Let the data set $(x_1, \ldots, x_n)$ be modelled by random variables $(X_1, \ldots, X_n)$. The **expected value** $E[T]$ is the mean of the random variable $T$. An estimator $\hat{\theta}$ is **unbiased** if $E[\hat{\theta}] = \theta$, regardless of what the value of $\theta$ is, otherwise it is biased. The **bias** of $\hat{\theta}$ is the difference $E[\hat{\theta}] - \theta$ where the model parameter does not equal the expected value of the estimation parameter.* [11].

Recall that the ML estimator for $\sigma^2$ in the simple linear regression model is given by

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

This estimate is biased by a factor of $(n-p)/n$ where $p$ is the number of explanatory variables in the simple regression model. The estimated variance using ML estimation is biased because the fact that $\alpha$ and $\beta$ (the intercept and slope) are also estimated is ignored. Thus REML estimation is used to find better and less biased estimators [13].

Wanting to avoid having to estimate the intercept and slope, REML avoids this by applying a matrix multiplication to the matrix $\boldsymbol{\beta}$ in the vector and matrix notation of the linear mixed-effects model, given above. The consequential mathematical rigour is omitted from this report but in the end REML makes the $\mathbf{X_i}\boldsymbol{\beta}$ term disappear, avoiding having to make estimates based on estimates. After the term is removed, ML estimation is then used to calculate the estimates for the resulting parameters. The estimates for the intercept and slope are not the same as the ML estimators but they are unbiased and so their use is preferred [13].

## 4.3 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:

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

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. There is no absolute value that will 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 uses only the BIC value as a comparative measure. This is because this criterion carries a larger penalty for the number of parameters included in each model [18].

## 4.4 Model validation

Linear mixed-effects models can be validated in the same way a simple linear regression model can be validated. To inspect homogeneity of the residuals, plot the residuals against the fitted values by assessing non-linear patterns in the residuals and that the residuals are equally spread along the horizontal axis (R produces a line to help this: the flatter the line, the better). To assess normality of the residuals, plot a histogram of the residuals and inspect whether the graph follows a bell-shaped curve. Finally for the Independence assumption, plot the residuals against explanatory variables and make sure the residuals are equally spread along the horizontal axis.

## 4.5 Pearly Wrasse

The investigation on how sea surface temperature effects the body size of *Pearly Wrasse* is continued, but this time using linear mixed-effects modelling with restricted maximum likelihood estimation to find estimates for the regression coefficients and the variances of the residuals and random effects. Thus the investigation changes slightly to how sea surface temperature affects fish body size after controlling the variation of geographic cell, year and survey.

Note that there are 1008 records of *Pearly Wrasse* found in 37 cells over 11 years in 79 different surveys and that each model is assigned a numeric value to compare the BIC values neatly.

In all the models that follow, $y$ is the body length of a specific *Pearly Wrasse* observation, $x_i$ is the scaled mean annual sea surface temperature for $i = 1, \ldots, 1008$ and $\alpha$ and $\beta$ are regression coefficients, treated as fixed effects, which are to be estimated.

Now the first three equations that are tested include cell, year and survey as a separate and singular crossed random effect. So for an observation $i$, the size of *Pearly Wrasse* is modelled by:

$$y_{ij} = \alpha + \beta x_i + b_j + \epsilon_{ij}, \quad j = 1, \ldots, 37 \tag{1}$$

where the geographic cell is treated as a random effect.

$$y_{ij} = \alpha + \beta x_i + b_j + \epsilon_{ij}, \quad j = 1, \ldots, 11 \tag{2}$$

where the year is treated as a random effect

$$y_{ij} = \alpha + \beta x_i + b_j + \epsilon_{ij}, \quad j = 1, \ldots, 79 \tag{3}$$

where the survey is treated as a random effect.

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

$$\begin{gathered} y_{ijkl} = \alpha + \beta x_i + b_j + b_{jk} + b_{jkl} + \epsilon_{ijkl}, \\ j = 1, \ldots, 37, \quad k = 1, \ldots, 11, \quad l = 1, \ldots, 79 \end{gathered} \tag{4}$$

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 include different combinations of nested random factors as follows:

$$y_{ijk} = \alpha + \beta x_i + b_j + b_{jk} + \epsilon_{ijk}, \quad j = 1, \ldots, 37 \quad k = 1, \ldots, 11 \tag{5}$$

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.

$$y_{ijk} = \alpha + \beta x_i + b_j + b_{jk} + \epsilon_{ijk}, \quad j = 1, \ldots, 37 \quad k = 1, \ldots, 79 \tag{6}$$

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

$$y_{ijk} = \alpha + \beta x_i + b_j + b_{jk} + \epsilon_{ijk}, \quad j = 1, \ldots, 11 \quad k = 1, \ldots, 79 \tag{7}$$

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

In all model equations, along with the residuals, each random effect is assumed to be drawn from a normal distribution with a mean zero and variance $\sigma^2$.

Applying each model into R and then extracting the BIC values leads to the following table below:

| **BIC value** | | | | | | |
|---|---|---|---|---|---|---|
| **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| 4416.070 | 4747.702 | 4375.641 | N/A | 4394.367 | 4369.391 | 4380.878 |

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

Reading from *Table 1*, the sixth model has the lowest BIC value, suggesting that this model should be used for this analysis. This does not come at a surprise since survey is specific to a certain day in a certain year in a certain cell. In addition (omitted from this report) the box plot of year against the response variable does not show as much difference in the spread of the data when compared to the drastically different spread of data for cell and survey.

Note that the fourth model does not have a BIC value because the model failed to converge in R. This means that restricted maximum likelihood could not find a minimum to form an estimate. This convergence error could be overcome by using maximum likelihood to find estimates instead, but this would lead to biased estimates and so it was decided to omit this model from the analysis.

Before instilling confidence in this model, the model must be validated under the same assumptions used for validating the simple linear regression model.

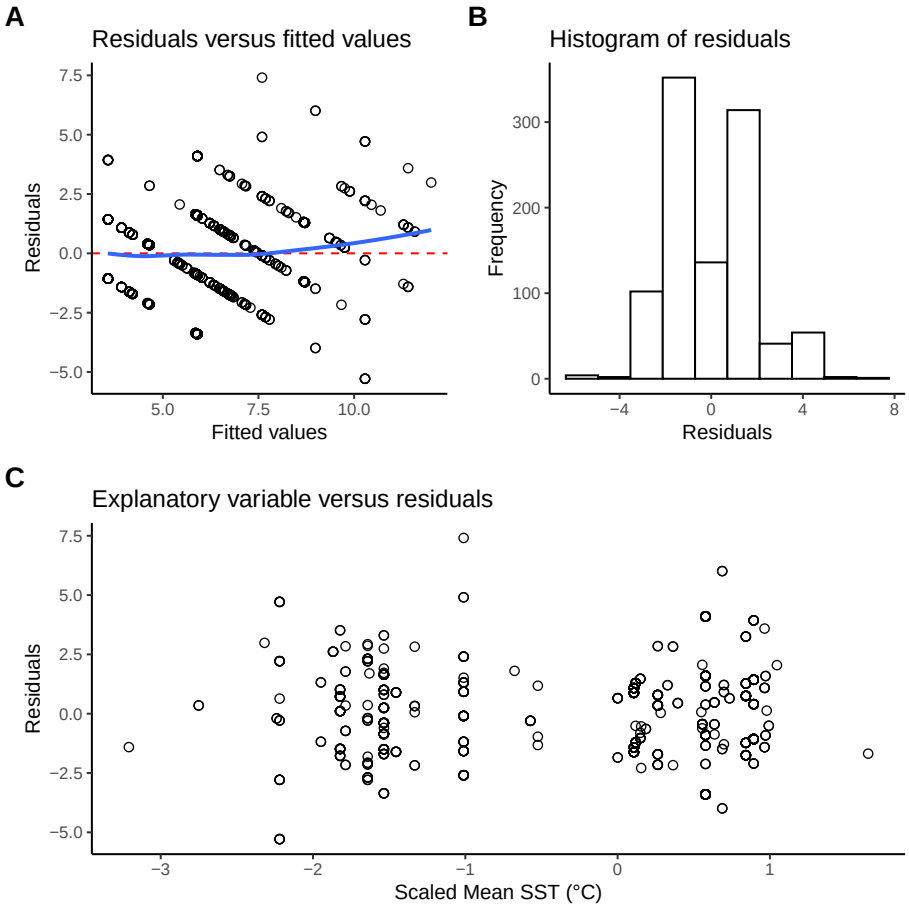


**Figure 7:** Model validation graphs.
**A:** Residuals versus fitted value (homogeneity). **B:** Histogram of residuals (normality).
**C:** Residuals versus scaled mean sea surface temperature (independence).

The blue line lies much flatter in *Figure 7A* than it did in *Figure 3A*. This means homogeneity can be safely assumed and that the variance is constant across the residuals. *Figure 7B* shows a better normal distribution of residuals than it did in *Figure 3B* so normality is verified and the spread of the residuals is more equally spread across the entire horizontal line in *Figure 7C* than in *Figure 3B* so the independent assumption is not violated. This validates the model chosen and so the scatter plot of *Figure 8* can be made confidently.

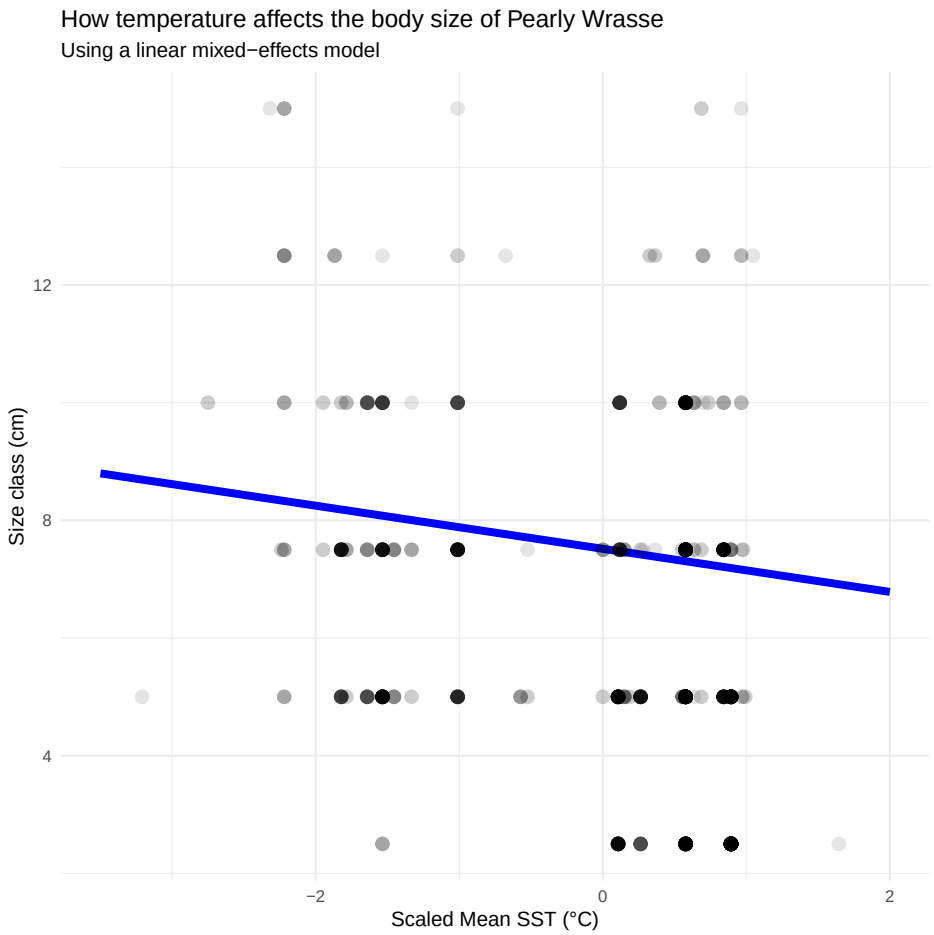


**Figure 8:** A scatter plot illustrating the scaled mean sea temperature against the size class of observed *Pearly Wrasse*. A regression line is added on top to model the relationship. The points are shaded according to how many observations have the same size at the same temperature. The darker the point, the more observations are recorded at that point.

*Figure 8* continues to show that *Pearly Wrasse* are smaller in warmer waters but now there can be confidence in this graph. The plotted line appears to be less steep than in *Figure 2* however. This is likely the result of the added random effects accounting for some of the unexplained information that was previously in the residuals. These arguments are supported by the summary of the model produced in R, which estimates the linear mixed-effects equation as:

$$y_{ijk} = 7.52 - 0.37x_i + b_j + b_{jk} + \epsilon_{ijk},$$
$$i = 1, \ldots, 1008, \quad j = 1, \ldots, 37, \quad k = 1, \ldots, 79$$
$$b_j \sim N(0, 1.264^2), \quad b_{jk} \sim N(0, 2.122^2), \quad \epsilon_{ijk} \sim N(0, 1.939^2).$$

Comparing this with the simple linear regression model above, $y_i = 6.05 - 0.89x_i + \epsilon_i$ for $i = 1, \ldots, 1008$ where $\epsilon_i \sim N(0, 2.538^2)$, the intercept is estimated to be higher in value, the slope is less negative and so less steeper and there is less variance in the residuals, mathematically confirming what was concluded above.

## 4.6 Every species

With a model selected it is time apply said model to each species in the data set. To do this, the mean annual sea surface temperature was firstly scaled and then the data was separated by species through a method called *nesting* in R. Nesting the data set created a tibble (a data frame) with 335 rows and 2 columns. For each species, the corresponding row had a tibble nested within it consisting of all the records of that particular species. The mixed-effects model found using the *Pearly Wrasse* data was then applied to each species. For an observation $i$ of a particular species, the size of the species is modelled by:

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

where $y_i$ 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, $b_{jk}$ is the random effect for the survey nested within the cell and $\alpha$ and $\beta$ are species specific regression coefficients to be estimated. $p$ is the total number of observations of the species, $q$ is number of geographic cells the species was found in, $r$ is the number of surveys the species was recorded in.

When the model was applied to each species however, two errors arose. The first being convergence. This was seen in one of the models tried on the *Pearly Wrasse* data set. In the model was disregarded to keep the analysis simple and to avoid biased estimated coefficients. The other error was that the model is singular. This means that a few of the variances estimated are near to zero. The R documentation suggests to avoid complex models as a way at preventing this error. Because of this, it was decided to choose a simpler model.

Looking back at *Table 1* the next lowest BIC value was the third model at 4375.641, which is very close to the BIC value of the current model at 4369.391. The third model treats survey as a singular random effect. The BIC value is not too surprising since survey was the lowest level of the nested data since survey was specific to year and cell. Thus it seems appropriate to use the third model for analysis going forward. Now for an observation $i$ of a particular species, the size of the species is modelled as:

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

where $y_i$ 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 survey and $\alpha$ and $\beta$ are species specific regression coefficients to be estimated. $p$ is the total number of observations of the species and $r$ is the number of surveys the species was recorded in.

Before applying the new model to every species, it was tested on the *Pearly Wrasse* data set to ensure the model's validity. The following figure is a scatter plot and the graphs to validate the model produced using R.

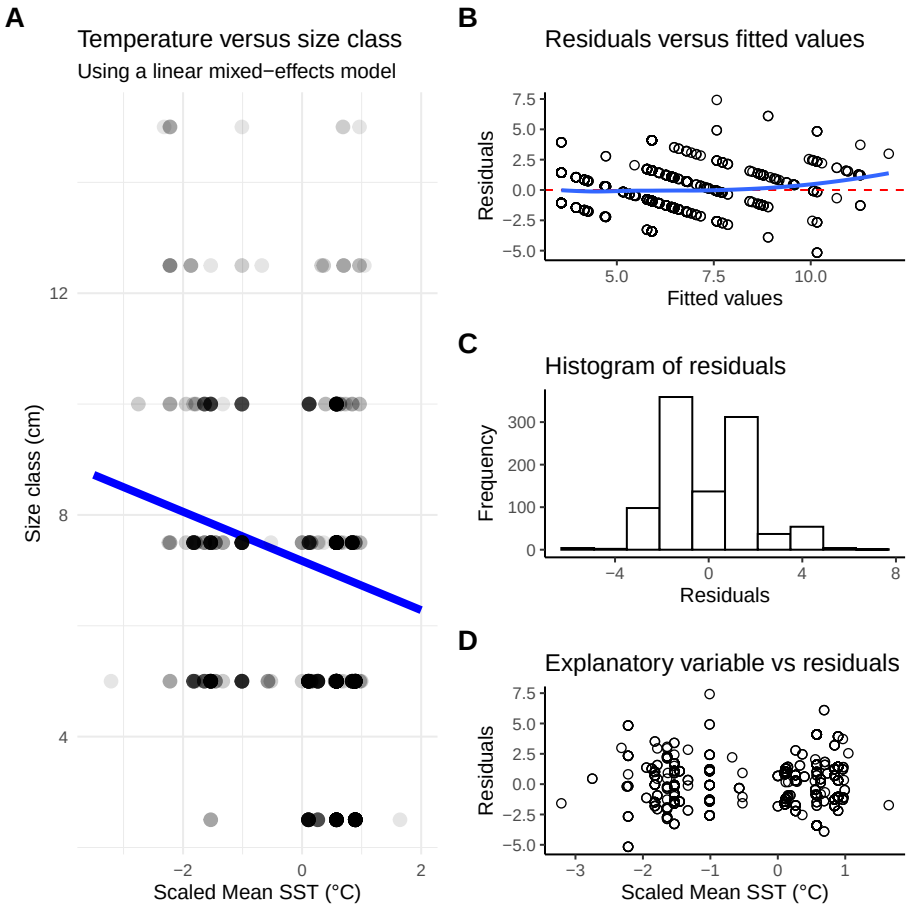


**Figure 9:** Scatter plot and model validation graphs.
**A:** Scatter plot of scaled mean sea surface temperature against size class of *Pearly Wrasse*
**B:** Residuals versus fitted value (homogeneity). **C:** Histogram of residuals (normality).
**D:** Residuals versus scaled mean sea surface temperature (independence).

*Figure 9A* continues to show that *Pearly Wrasse* are smaller in warmer waters and *Figure 9A*, *9B* and *9C* are show the assumptions are not invalidated. Thus this new model can be used in the following analysis. R estimates the new linear mixed-effects equation as:

$$y_{ijk} = 7.17 - 0.44x_i + b_j + \epsilon_{ij},$$
$$i = 1, \ldots, 1008, \quad j = 1, \ldots, 79$$
$$b_j \sim N(0, 4.782^2), \quad \epsilon_{ij} \sim N(0, 0.3749^2).$$

Comparing this with previous mixed model, $y_{ijk} = 7.52 - 0.37x_i + b_j + b_{jk} + \epsilon_{ijk}$ where $b_j \sim N(0, 1.264^2)$, $b_{jk} \sim N(0, 2.122^2)$, $\epsilon_{ijk} \sim N(0, 1.939^2)$, the intercept is estimated to be slightly lower in value, the slope is now more negative and so more steeper and there is more variance in the residuals and the random effects.

This new model was then applied to all the data without any occurrence of errors. Now wanting to know how each species body size responds to temperature, the estimated coefficient $\beta$ is of prime interest since its value indicates whether the size of the fish decreases or increase with an increase of temperature. Thus the coefficient for the scaled mean sea surface temperature, $\beta$ was extracted from each species and put into a new data frame. This data frame was then used to make a plot of the estimates.

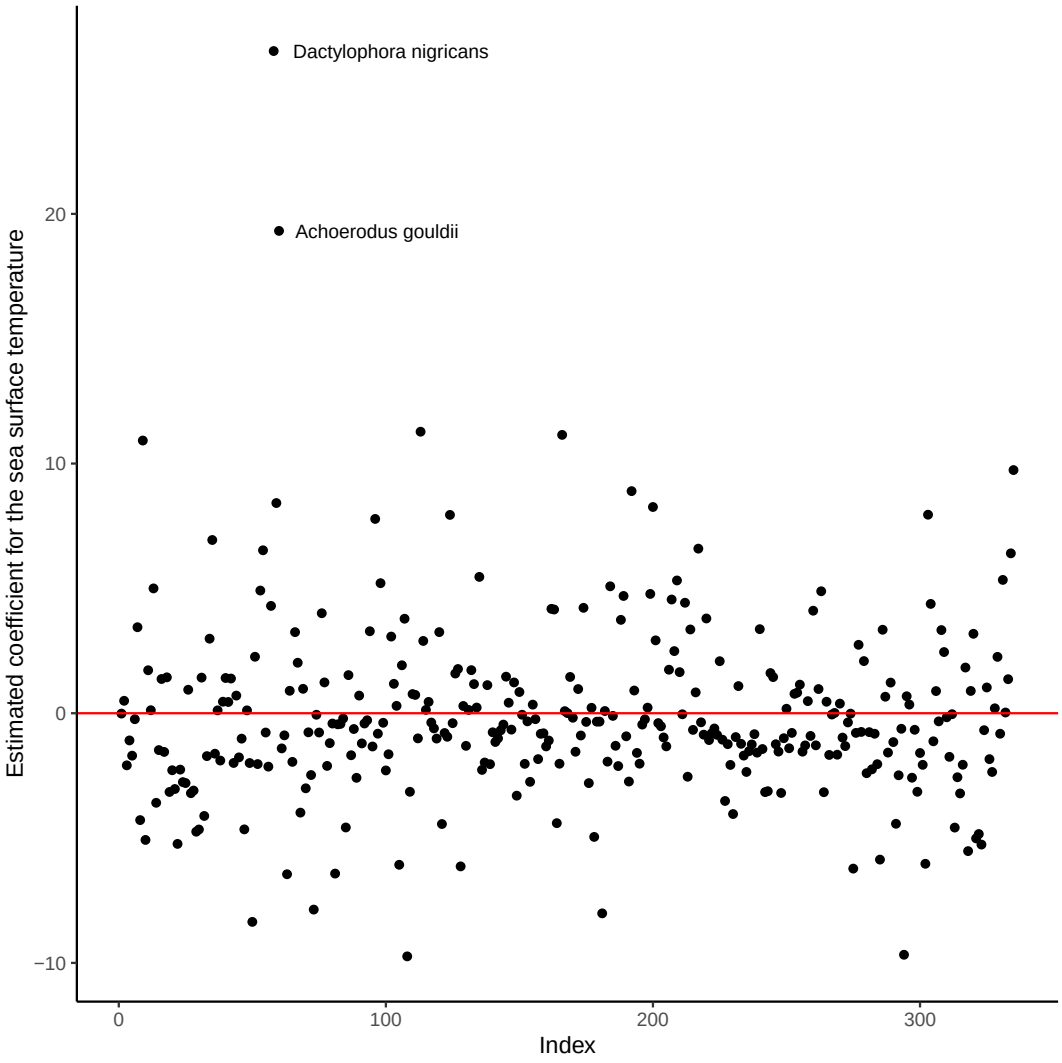


**Figure 10:** The estimated coefficients of the scaled mean sea surface temperature, $\beta$, for the 335 species in the data set. The closer a point to the red line, the closer $\beta$ is estimated with a value of zero. Points with an estimate of $\beta > 15$ are labelled.

With many points below the red line, there are clearly many species that decrease in size when the sea surface temperature increases, but *Figure 10* also shows that there are numerous species that increase in warmer waters. The points with extreme values are the species with an estimated $\beta$ greater than the value of 15. The points are labelled accordingly. As a result, model validation graphs were created for each of these two species to assess why these values have a larger absolute value than the other points.

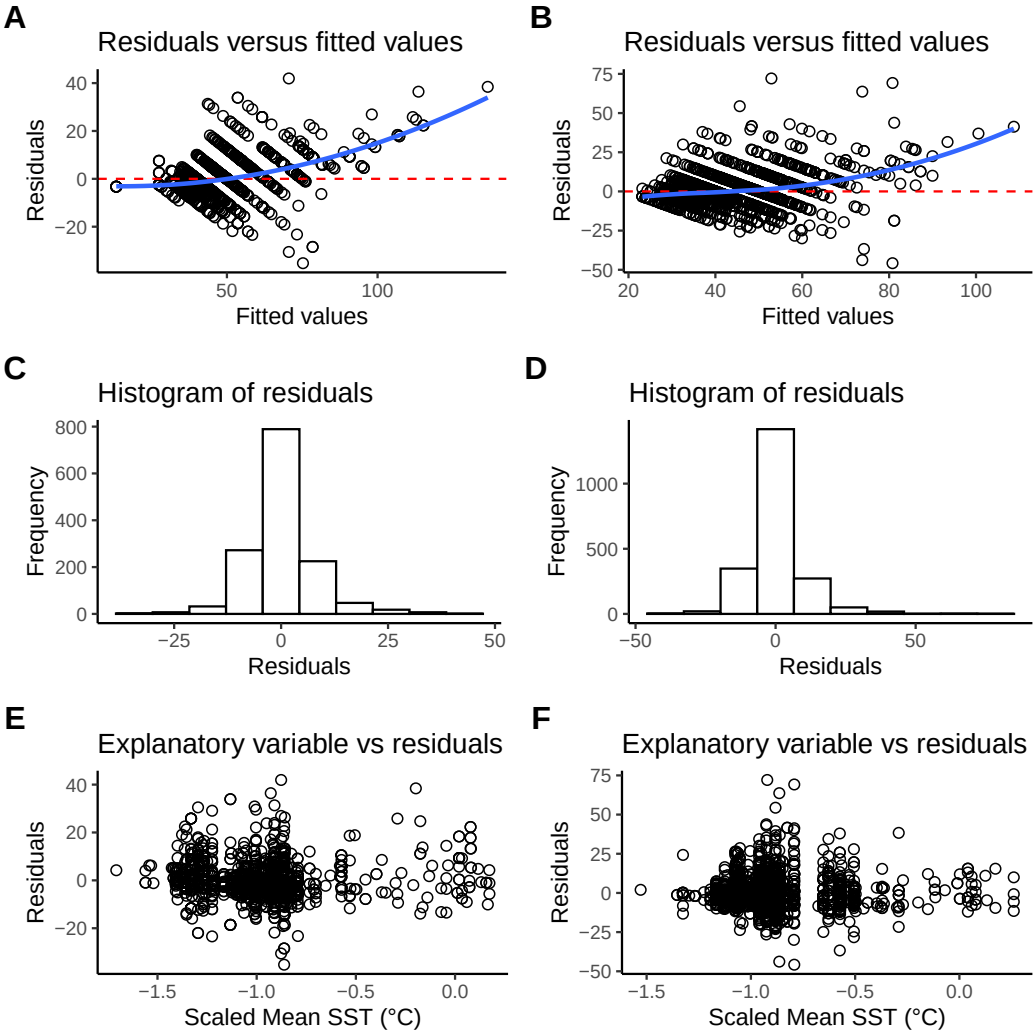


**Figure 11:** Model validation graphs for *Dactlophora nigricans* in **A, C** and **E** and *Achoreodus gouldii* in **B, D** and **E**.
**A, B:** Residuals versus fitted value (homogeneity). **C, D:** Histogram of residuals (normality).
**E, F:** Residuals versus scaled mean sea surface temperature (independence).

For both species, the normality assumption is met (*Figure 11C* and *11D*) and although in *Figure 11E* and *11F* there seems to be large clusters of the residuals, the residuals are relatively spread equally across the horizontal axis, thus the independence assumption is met. *Figures 11A* and *11B* indicate a slight violation of homogeneity since the blue line starts flat but rises as it moves across the horizontal axis. Now this could be fixed by changing the variance structure of the residual assumption, being that the residuals are drawn from a normal distribution with a mean of zero and a variance of $\sigma^2$, or with only a few points around the steep section of the blue line, the observations in the data set could be assessed and perhaps removed if the value seem extreme.

Removing observations must be taken with caution as extreme values are interesting, especially in ecology. Extreme values may not be the fault of measurement error but could show vital signs at how species are changing with warming, for example, and ignoring such observation could be detrimental at finding ways to manage the species. With only two species showing extremity, out of the 335 observed, it was decided to keep the species as they are and carry on with the analysis.

# 5 Results with discussion

The effects of sea surface temperature on fish body size is summarised in the table below. Note that each correlation coefficient, $\beta$, was rounded to the nearest tenth.

| Temperatures effect on fish body size | Percentage of species |
|---|---|
| Increase | 39 |
| Neither increase nor decrease | 2 |
| Decrease | 59 |

**Table 2:** The effect that increases of temperature has on the 335 different species body size, summarised into percentages, rounded to the nearest whole number.

The result of the analysis used in this report provides empirical solutions specifying the percentage of species, out of the 335 recorded in the data set, that increase, decrease or stay the same in body size when sea surface temperature is increased.

The majority of species decrease in body size at 59% but it is interesting to see that a large proportion of species increase in body size at 39%. Nearly all species are affected by increases in seas surface temperature with only 2% of species body sizes remaining unchanged.

Audzijonyte et al (2020) found that studies conclude that even small changes in the body size of species lead to an increased mortality rate, yet these studies seem to focus solely on the effects of species shrinking, therefore, these results show that a focus on what the effects of species increasing in size is paramount. With all being said however, it is undeniable that any changes in body size will alter marine ecosystems, such as: known food chains, predator to prey relationships, species redistribution and mortality leading to possible extinction and the details of this may be of interest to investigate.

Overall these results highlight that global warming will affect nearly all species of fish in some way and so it is paramount to find ways to manage the body size of fish species in a rapidly warming ocean.

# 6 Summary

This report began by introducing the simple linear regression model and created figures using R to visualise how to determine whether a model is appropriate for analysis. It was discovered that the data used in this analysis violating a set of assumptions and so the simple linear regression model had to be rejected.

Linear mixed-effects modelling was then used and had to be demonstrated on a complex data set and not a data set designed to demonstrate such modelling easily, as many articles and tutorials do.

Although this analysis was a much simplified version of Audzijonyte et al (2020), this report thoroughly explained each step of its analysis, making it easy to follow and consequently understand.

The resulting analysis discovered that even though the majority of species shrink with global warming, a large proportion of species increase in size instead. Since the effects of larger fish species is not well researched it is critical to investigate what this means for the future of marine ecosystems in order to protect marine wildlife from an anthropogenic-caused and ever-warming environment.

# Code and data availability

- Reef Life Survey
- Audzijonyte et al (2020) GitHub
- My github

(Will turn the list into a paragraph when all is done).

# 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] Bommae Kim, *Understanding Diagnostic Plots for Linear Regression Analysis* (2015), available at `https://data.library.virginia.edu/diagnostic-plots/`.

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

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

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

[18] 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.