KITTIWAKES OBSERVATION STATISTICAL ANALYSIS

1. INTRODUCTION

This analysis is to assist an ornithologist in examining the data collected on kittiwakes, a type of gull. There were 4 datasets provided namely observation data, historical data, measurement data, and location data. Observation data, as the name suggests, consists of the number of kittiwake sightings at an observation point taken at dawn, noon, mid-afternoon and dusk for 4 weeks (i.e., 28 days). Historical data contains breeding pairs at 4 sites over 5 years. Measurement data consisting of the weight (in grams), wing span (in centimetres) and culmen (beak length) (in millimetres) that are collected for 17 black-legged and 15 red-legged kittiwakes. Location data consists of the number of breeding pairs in 30 colonies, along with relevant covariate information.

We are requested to provide our insights to the ornithologist on several questions. Firstly, to provide a clear understanding of the observation data, we will conduct an exploratory data analysis of the same. 90% confidence interval has to be constructed for the mean number of kittiwakes observed at dawn. A confidence interval gives a range of plausible values to estimate some characteristic of the total population, such as the mean, when the true population parameter is unknown. It is calculated based on sample data provided and gives an estimate of the range within which the true population parameter might lie.

Secondly, to examine the historical data to check if the ornithologist's hypothesis of the decline in kittiwake numbers over time is independent of the site is true. The ornithologist would also like an estimate for the breeding pairs present at site C in 2006.

Thirdly, using the measurement data, to provide a visual summary of the data and examine if there is independence between wing span and culmen length for each sub-species. Furthermore, to assess if there is evidence of a difference in weights between the two sub-species and if there is a significant difference overall.

Fourthly, using the Location data, the ornithologist requires us to fit linear statistical models for the data and choose the most appropriate linear model for the data.

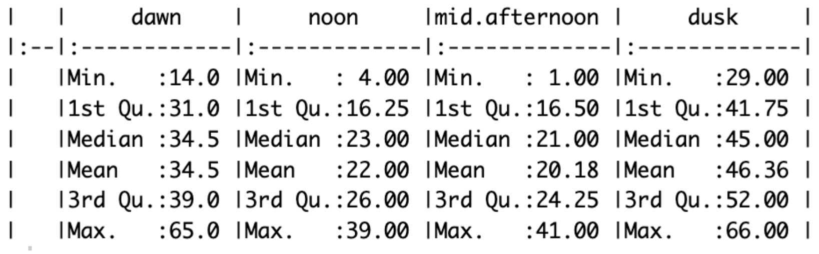
This analysis aims to assist the ornithologist in gaining valuable insights from the collected data and providing statistical solutions to all the questions put up using R Studio.

2. MAIN BODY

2.1. Observation Data

We use the observation data to conduct exploratory data analysis on the same for a better understanding of the data. In statistics, exploratory analysis refers to the process of examining and summarizing the data to get initial insights and understanding of the data. It is mainly done by using various graphical and numerical techniques to find the patterns, relationships, and distributions within the data without coming to any formal statistical conclusions or decisions.

In this step, we first read the data in the observation data CSV file. We print out the head of the data (i.e., the first few rows of the dataset) to look into the sample of the data that has been collected by the ornithologist. As explained earlier, the data consists of 28 rows (i.e., 28 days of data) of the number of kittiwake sightings at dawn, noon, mid-afternoon and dusk. The values are positive integer numbers representing the number of kittiwakes spotted at that particular time of the day. To know the basic statistical details about the data, we further use the summary function in R to obtain the basic summary of the dataset read. The summary of the dataset gives us the minimum and maximum number of Kittiwakes spotted at that particular hour, the 1st and 3rd quartile value of the entire column in the data, the average (statistical mean) number of Kittiwake sightings and the median of the dataset. The median gives us the central tendency of the data. It refers to the middle term of the data. For example, To calculate the median here, the number of kittiwake sightings at dawn in 28 days is sorted in ascending order and then the 50th percentile of the data is given out as output.



```
|   |     dawn     |     noon      |mid.afternoon |     dusk      |
|:--|:-------------|:--------------|:-------------|:--------------|
|   |Min.   :14.0 |Min.   : 4.00 |Min.   : 1.00 |Min.   :29.00 |
|   |1st Qu.:31.0 |1st Qu.:16.25 |1st Qu.:16.50 |1st Qu.:41.75 |
|   |Median :34.5 |Median :23.00 |Median :21.00 |Median :45.00 |
|   |Mean   :34.5 |Mean   :22.00 |Mean   :20.18 |Mean   :46.36 |
|   |3rd Qu.:39.0 |3rd Qu.:26.00 |3rd Qu.:24.25 |3rd Qu.:52.00 |
|   |Max.   :65.0 |Max.   :39.00 |Max.   :41.00 |Max.   :66.00 |
```

Fig2.2.1. Summary Table of Observation data

Consider the first column of the table, representing the summary values at dawn.

i. Min. – the minimum value of the data.

ii. 1st Qu. – the value below which the 25% of the data falls. To compute it, the data values are first ordered from smallest to largest. Then the middle value of the lower portion of the ordered data is identified. This middle value is the median of the lower half, which is used in the calculation.

iii. Median – Median refers to the middle value of the data. It is calculated using the formula ((n+1)/2)th term after sorting the data in ascending order, where n is the number of observations.

iv. Mean – Mean refers to the simple average value of all the data. It is calculated by summing the values of all the observations in the column and dividing it by the total number of observations.

v. 3rd Qu. – the value below which 75% of the data falls. It is calculated by finding the median of the upper half of the data after sorting it.

vi. Max. – the maximum value of the data.

Post looking into the summary table in detail, we plot histograms for each column of the data. This graph plotting technique will help us visualise the data and gain insights about the number of kittiwake sightings across 4 weeks.

The graphs plotted at each stage are shown below:

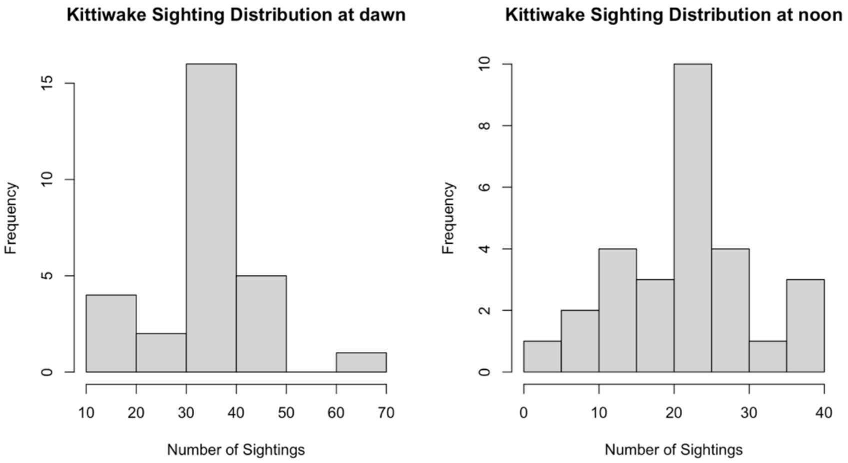


Fig2.2.2. Histogram for noon and dawn data

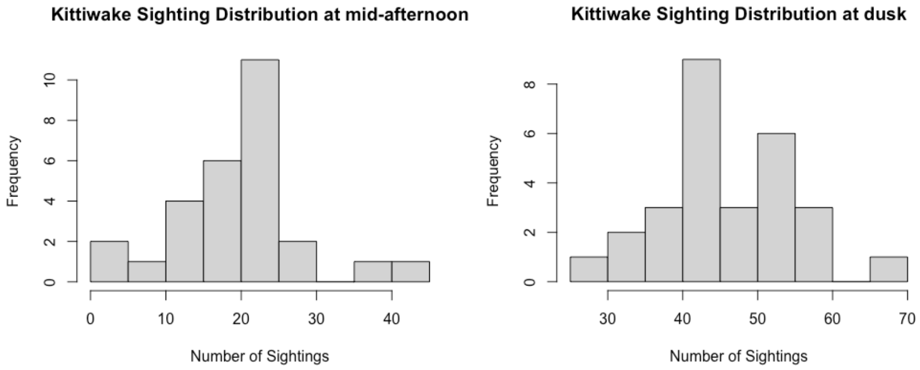


Fig2.2.2. Histogram for mid afternoon and dusk data

Further to construct a 90% confidence interval for the mean number of kittiwakes observed at dawn, we use t-test function in R. 90% confidence interval provides the range of values where we are 90% sure that the original value will lie. The t-test is used to compare the means of two groups or samples.

$$t = \frac{\bar{d} - \mu_d}{\sqrt{s_d^2/n}} :$$

where:

- t is the t-value or test statistic.
- x̄1 and x̄2 are the sample means of the two groups.
- s1 and s2 are the sample standard deviations of the two groups.
- n1 and n2 are the sample sizes of the two groups.

The t test in R gives us a 90% confidence interval of 31.17841 to 37.82159 with a standard mean of 34.5, to visualise the data a histogram with the mean and interval is plotted

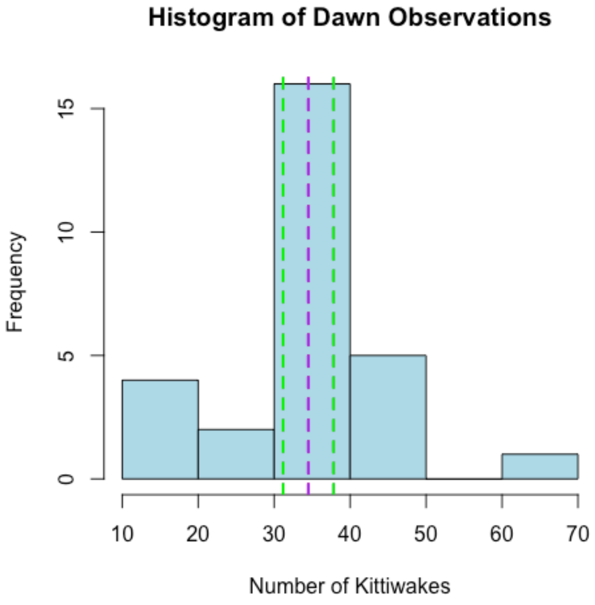


Fig2.1.6. 90% confidence histogram

2.2. Historical Data

Our next task is to check if the ornithologist's hypothesis of the decline in kittiwake numbers over time is independent of the site is true and to estimate the number of breeding pairs at site C in 2006.

For the first part of the task, we use the chi-square test. The chi-square test is a statistical test used to determine if there is a significant relation between two categorical variables. In our case, we must check if the ornithologist's hypothesis is true. The chi-square test gives us a p-value as output. The p-value indicates the significance of the association between the two variables.
In our case, if the p-value is greater than 0.05 we cannot reject the null hypothesis and conclude that the decline in kittiwake numbers over time is independent.

- Null Hypothesis (H0): There is no association between the year and the site of observation. In other words, the decline in kittiwake numbers is independent of the site.
- Alternative Hypothesis (H1): There is an association between the year and the site of observation. This means the decline in kittiwake numbers is not independent of the site.

From our code, we obtain a p-value of 0.6596, which is greater than the conventional alpha level of 0.05, so we fail to reject the null hypothesis (i.e., we cannot surely say that the ornithologist's hypothesis of the decline in kittiwake numbers over time is independent of the site is false).

From the results obtained, we conclude that there is no sufficient evidence to conclude that there is an association between the year and the site of observation. Therefore, based on this analysis, the decline in kittiwake numbers over time appears to be independent of the site and the ornithologist's hypothesis to be true.

To estimate the number of breeding pairs at site C in 2006, we use a linear regression model.
Linear regression is a statistical technique that is used to find the relationship between a dependent variable and one or more independent variables. We create the model in R using the Site C data from the historical dataset provided and then predict the value of the number of breeding pairs at site C in 2006. We obtained the results as the number of breeding pairs at site C in 2006 is 53, the model mapped the estimate with the nearest year with available value. The dot plot for the prediction is as below.

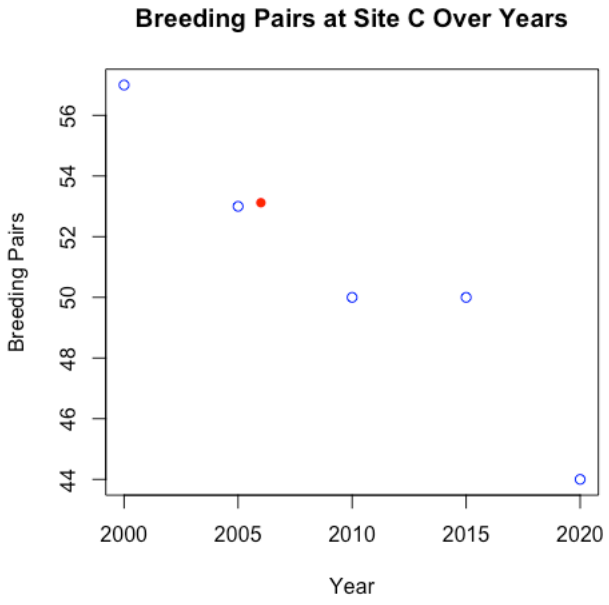


Fig2.2.1. box plot of the prediction

2.3.Measurement Data

To start with, we derive the visual summary of the measurement data using the scatter plot which allows to see the relationships between different pairs of variables, it shows how each pair correlates with each other with each dot showing the data point for example, the intersection of weight column and wingspan row shows the relationship between weight and wingspan

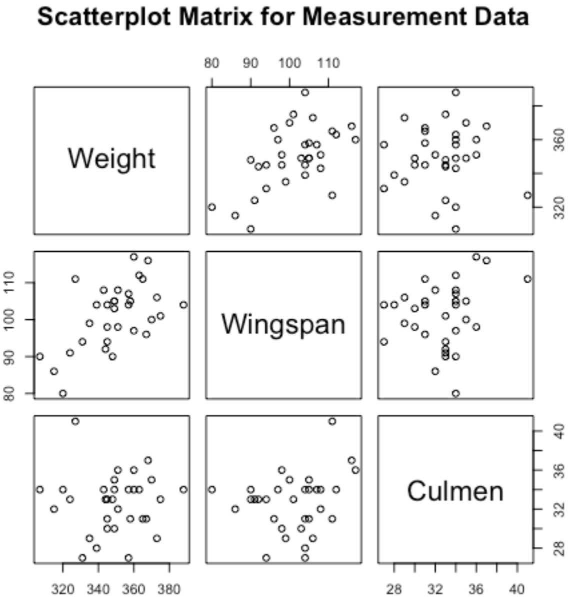


Fig2.3.1 scatter plot of the measurement data

To understand the data box plots for each individual feature for the sub specie is plotted. Starting with Weight the plot is given below.

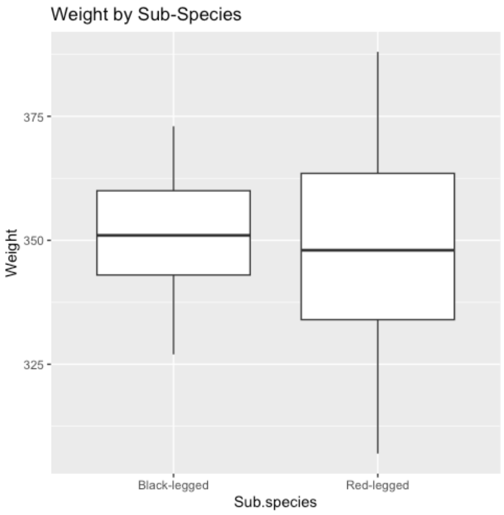


Fig2.3.1 box plot of weight

It shows that the weight for red-legged has a much greater range than the black legged but they have a similar mean, moving forward to the wingspan.

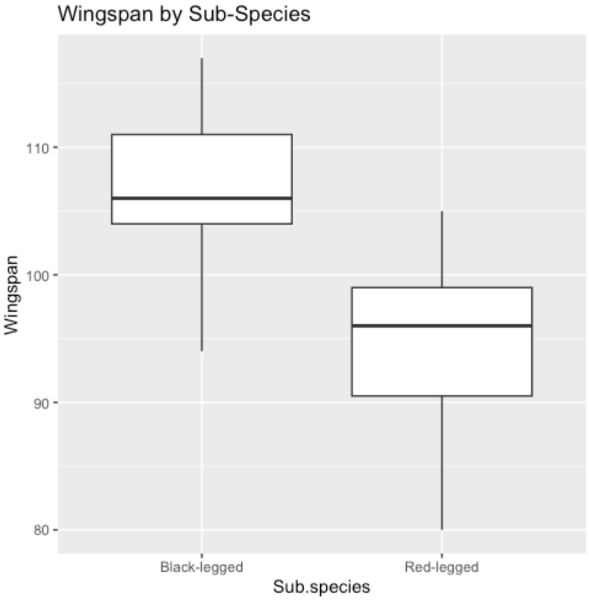


Fig2.3.2 box plot of wingspan

The wingspan for both species have a significant difference in their mean and minimum values, black legged have a larger wingspan. Finally, for the culmen length the box plot is.

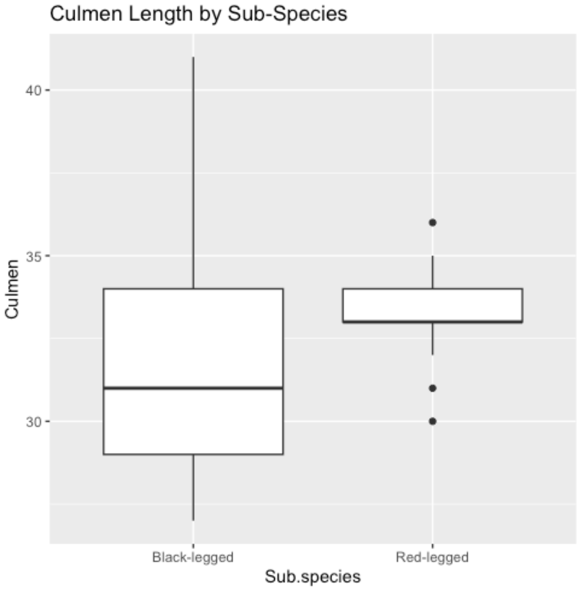


Fig2.3.2 box plot of culmen length

Culmen length for red legged have a very squished quartile range. Q1 and mean are almost the same and has a very small range compared to the black legged.

To check if the wing span and culmen length of each sub-species are independent, we perform a correlation test between the wingspan and the culmen length for each sub-species.
Correlation is a measure of the linear relationship between two variables.
The result of the correlation test includes the correlation coefficient (r), the p-value, and other relevant information.

The mathematical formula of the Correlation coefficient (r) is as follows:

$$r = \frac{n\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{n\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot n\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where:

- X and Y are the individual data points of the two variables being analysed.
- X̄ and Ȳ are the means (averages) of the X and Y variables, respectively.
- Σ represents the summation symbol, indicating that the sum of all the values for each variable should be taken.

This formula calculates the covariance between the X and Y variables, which measures how the variables change together. The covariance is then divided by the product of the standard deviations of X and Y to obtain the correlation coefficient.

The correlation function cor.test() in R gives us the correlation coefficient (r), the p-value and other relevant information. In our test, we will be focusing only on the correlation coefficient (r) value and p-value.
The correlation coefficient (r) is a value that ranges from -1 to 1.

- r = 1 - perfect positive correlation,
- r = −1 - perfect negative correlation,
- r = 0 - no linear relationship.

The p-value is a measure that supports us to determine the strength of evidence against the null hypothesis. In the case of correlation, the p-value represents the probability of observing a correlation as extreme as, or more extreme than, the one observed, assuming the null hypothesis is true. In correlation, the null hypothesis is that there is no correlation between the two variables (i.e., wingspan and culmen length).

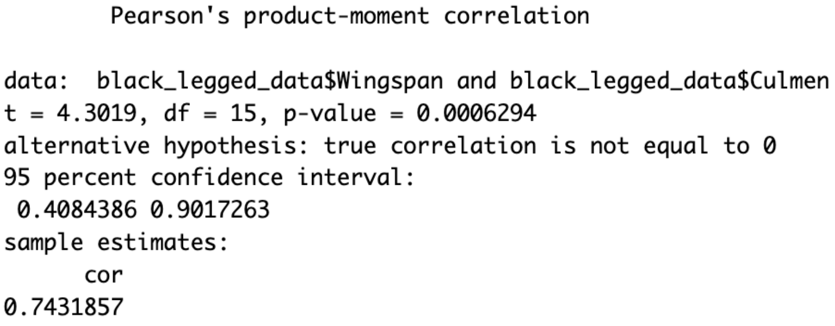
In the context of our analysis, we have two sub-species of Kittiwakes. Black-legged and Red-legged kittiwakes. We should check if the wing span and the culmen length of the respective species are independent or not. After performing the correlation test using cor.test() in R, we obtained the results as follows:

For sub-species black-legged kittiwake, the p-value obtained after is 0.0006294 which is lesser than 0.05.

- Null Hypothesis (H0): The wing span and the culmen length of the black-legged kittiwake are independent of each other.

- Alternate Hypothesis (H1): The wing span and the culmen length of the black-legged kittiwake are dependent on each other.

As the p-value obtained is less than 0.05 we can reject the null hypothesis and hence conclude the wing span and culmen length of the black-legged kittiwakes are dependent on each other.
The correlation coefficient value obtained is 0.7431857 which represents a moderately strong relation between the two variables. (i.e., if the value of wing span increases then the culmen length tends to increase and vice-versa).



```
        Pearson's product-moment correlation

data:  black_legged_data$Wingspan and black_legged_data$Culmen
t = 4.3019, df = 15, p-value = 0.0006294
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4084386 0.9017263
sample estimates:
      cor
0.7431857
```

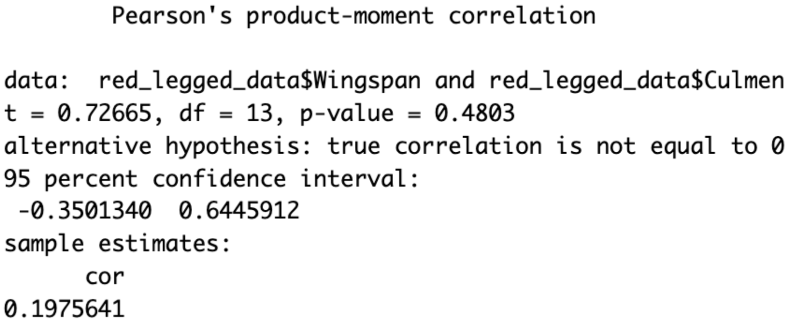
Fig.2.3.1. Correlation test for black-legged kittiwake

For sub-species red-legged kittiwake, the p-value obtained after is 0.4803 which is greater than 0.05.

- Null Hypothesis (H0): The wing span and the culmen length of the red-legged kittiwake are independent of each other.

- Alternate Hypothesis (H1): The wing span and the culmen length of the red-legged kittiwake are dependent on each other.

As the p-value obtained is greater than 0.05 we fail to reject the null hypothesis and hence conclude there is not enough evidence to predict the wing span and culmen length of the red-legged kittiwakes are independent. The correlation coefficient value obtained is 0.1975641 which represents a weak relation between the two variables.



```
              Pearson's product-moment correlation

data:  red_legged_data$Wingspan and red_legged_data$Culmen
t = 0.72665, df = 13, p-value = 0.4803
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3501340  0.6445912
sample estimates:
      cor
0.1975641
```

Fig.2.3.2 Correlation test for red-legged kittiwake

In the next part of our analysis, we check for evidence that the weights of birds of the two sub-species are different. To analyse the same we use a t-test. The t-test is explained in 2.1 of this article. T-value of 0.58682 and p-value of 0.5634 is resulted from the t-test. As the p-value is greater than 0.05 we fail to reject the null hypothesis that the weights of both species are same, we conclude that we do not have enough evidence to prove the difference.

Similarly, to check for evidence to prove difference in both the species, t-tests for both culmen length and wingspan are also performed. Test performed for wingspan yield a t-value of 5.323 and p-value of 1.204e-05, as the p-value is less than 5% we fail the null hypothesis and prove that there is a difference in wingspans, a positive t-value means that black legged have a larger wingspan than the red-legged. Test is repeated for culmen length, -1.2742 and 0.216 are the results for t-value and p-value respectively, p-value is greater than the critical alpha value hence fail to void the null hypothesis and claim that not evidence is present to prove the difference.

In conclusion, wingspan is different for both the species, black legged having larger wingspan than the red legged, whereas, the difference in weights and culmen length could not be proved.

2.4 Location Data

A model to predict the breeding pairs using different covariates is to be developed in this analysis, linear model is first made using the **lm()** function in R and the summary of results is printed. The lm function is a statistical tool used to easily understand and predict relationships between different variables. The created model has Residual standard error of 14.54, multiple R squared value 0.895 and F-statistic 32.6 with Coast.directionNorth, Coast.directionSouth, Coast.directionWest and cliff.height being variables that have an effect on breeding pairs. A few diagnostic plots are as below. The model has an AIC value of 253.7884 which describes how good the model is, the lower the better.

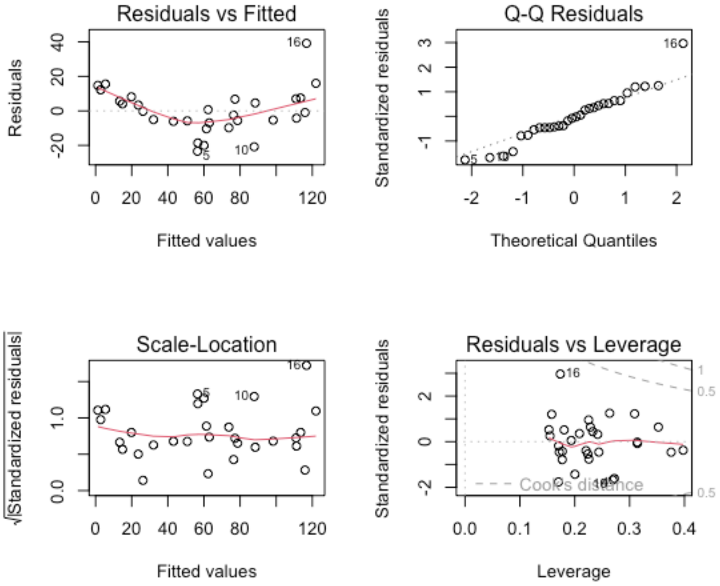


Fig 2.4.1: Plots for linear model

The residual vs fitted graph is plotted to check the constant variance of residuals, QQ plot shows if the residuals are normally distributed, scale-location graph is another way of checking for variance of residuals, finally, the residual vs leverage graph helps us understand influential data points and outliers that have effect on the predictions made by the model.

Another model is created with taking a logarithm of the breeding pairs that handles the skewness of the data and stabilizing the variance. The generated model has Residual standard error of 0.1172, multiple R squared value 0.9774 and F-statistic 165.7. The model has the same variables effecting the number of breeding pairs but with different coefficients. The AIC for the logarithmic model is -35.49022. The similar graphs as the linear model are plotted

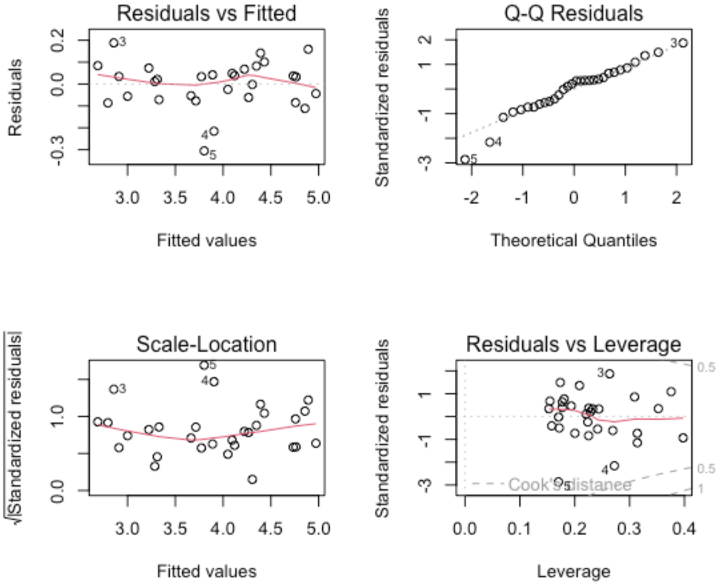


Fig 2.4.2: Plots for logarithmic linear model

Logarithmic model is a better prediction model as it has a lower AIC value (-35.49022), a very high R-square value (0.9842) indicating an excellent model fit and a small F-statistic p-value (2.2e-16) indicating a statistically significant model, all these factors play a major role in deciding the best model. The effects that the selected covariates have on the number of breeding pair are as follows:

- Coastal direction of North have a negative effect on the number of breeding pairs
- Coastal direction of South have a negative effect on the number of breeding pairs
- Coastal direction of West have a negative effect on the number of breeding pairs
- Sandeel concentration has a positive effect on the number of breeding pairs
- Cliff height has a positive effect on the number of breeding pairs
- Summer temperature has a negligible effect.

A positive effect means that increase in the quantity will increase the number of breeding pairs and vice versa, similarly, a negative effect means that increase in the quantity will decrease the number of breeding pairs.

The model is used to predict the number of breeding pairs with 98% confidence interval under the following conditions:

- coastal direction = East
- sandeel concentration = 2.21
- mean summer temperature = 24.4
- height = log (3.1)

The data is stored and passed to the model in R to predict the number of breeding pairs with 98% of confidence interval and we take the exponential of the final answer as we are using logarithmic model. The confidence interval resulted is 13.36 to 16.16

3. CONCLUSION

To conclude the analysis, a lot of information is gathered from the statistical analysis of the 4 types of data. A relationship could not be found between the decline of kittiwakes and sites using the historic data, hence to preserve all the decisions will be similarly applied to all the sites.

A relationship was found between the culmen length and wingspan for the black legged kittiwakes but no such evidence was found for the red legged. For each sub specie a difference for wingspan could be found but nothing could be proved for weights and culmen length. Best model to predict was selected to predict the number of breeding pairs of kittiwakes in presence of different covariates, for each covariate the effect on the breeding pair was studied too.

This analysis will help make better decisions to preserve the decline of kittiwakes and help scientists in their future research.