Analysis of Kittiwake Ecology.

Introduction:

This report will illustrate the analysis of kittiwake, a type of gull species, on four different types of data, namely observation, historical, measurement and location data to aid the ornithologist. The observation dataset provides the number of kittiwake sightings at each period of the day. The period includes dawn, noon, mid-afternoon and dusk of the day over a period of 28 days. The second dataset, historical data contains the number of breeding pairs at five different sites named from A to E in six different years. The next dataset is measurement data, which has two different types of kittiwake species, red-legged and black-legged kittiwakes. And also their weights, wingspans, and culmen lengths for both species. Finally, the location dataset contains significant features like mean summer temperature, cliff height, sandeel concentration and costal direction that contribute to the number of breeding pairs in 26 colonies. These datasets depict the ecosystem of kittiwake and also explains kittiwake's attributes and their environment of their ageing.

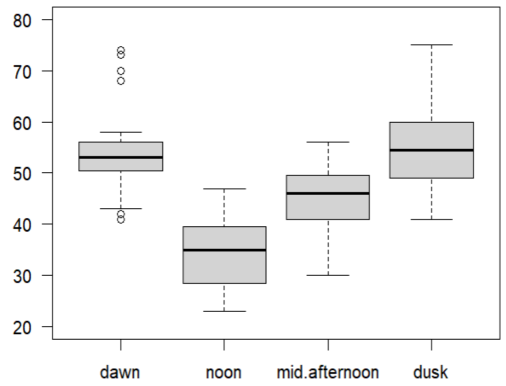
Main body:

The ornithologist has several challenges in every information he collected about kittiwake, they will be explained in detail and addressed in four sections, each for a separate dataset. Each section will explain the dataset and interpret the ornithologist's questions and methods used to solve his issue by presenting the results in the form of graph, wherever possible and interpreting the results thoroughly.

1. Observation data:

Ornithologist recorded the number of kittiwake he saw at each period of the day for 4 weeks or for 28 days. Dawn, noon, mid-afternoon and dusk are periods he chose to monitor the kittiwake. Now, ornithologist wants (a) exploratory analysis on the data he observed and also (b) calculate an interval with 80% confidence for the average number of kittiwakes observed at mid-afternoon period.

a) Let's explore the data ornithologist collected by constructing the box plot and bar plot for all periods of the day. Figure (1.1) shows the box plot for the observation dataset. The maximum number of kittiwakes can be seen at dawn and dusk with average over 50 numbers. At mid-afternoon, an average below 50 kittiwakes are observed. And the least number of kittiwakes are monitored at the noon period of the day. Furthermore, there are few kittiwake observed at both ends in the dawn subdivision of boxplot, which indicates that in some days there were abnormal number kittiwakes observed at dawn. And figure (1.2) represents the bar plot for the total number of kittiwake observations at each period of the day. There are around 1500 kittiwakes seen at dawn and at dusk in the 4 week period over 28 days. At noon, the kittiwakes are seen the least when compared to different periods of the day with less than 1000 kittiwakes.



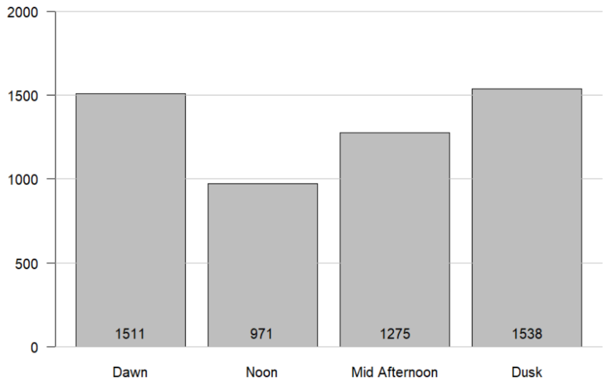


Figure 1.1 Box plot for observation data. Figure1.2 Bar plot for observation data.

b) To find out the average number of kittiwakes observed at mid-afternoon with 80 percent confidence interval, one sample t-test is used. One sample t-test is a statistical measure used to compare the average of the sample to another value. Based on one sample t-test result obtained using mid-afternoon observation data, the interval ranges from 44.02711 to 47.04432. That is with 80 percent confidence, ornithologist can observe an average of 44 to 47 number of kittiwakes at mid-afternoon.

2. Historical data:

In historical dataset, ornithologist collected the number of breeding pairs of kittiwakes at different sites from site A to site E during the time span of 2000 to 2016 with interval of four years. There are six observation for each site. So in total there are thirty six observations in the dataset. a) Ornithologist wants to check a hypothesis that decline in number of kittiwake is independent of site. And also b) predict the number of breeding pairs at site D in the year 2014.

a) To check the hypothesis statement that decline in number of kittiwakes is independent of site, Pearson's Chi-squared test is used. It is a statistical test used to examine the probability between difference sets of data. Assuming the null hypothesis (H0) as the decline in kittiwake number is independent of site and alternate hypothesis (H1) as decline is dependent of site. And also a significance level (α) of 95 percentage is predefined, so α is set to 0.05.To apply Pearson’s chi-square test, expected observation value of number of breeding pair at every site in each year is required. In order to calculate expected value, first calculate the sum of number of breeding pair at each site and also calculate the sum of number of breeding pair at each year. Then multiply both these values and divide it by the sum of observed number of breeding pair. Once the expected value is calculated. Calculate the Pearson’s chi-square test for expected value and observed value of number of breeding pairs. Based on the test result, the test statistic value was 11.171 and p-value was 0.9417.The p-value is a measure that helps to determine how strong is our evidence is against the null hypothesis. Since the p-value is greater significance level, the null hypothesis is rejected. Therefore, the decline in kittiwake number is independent of site.

b) To estimate the number of breeding pairs at site D in the year 2014, a linear regression analysis is used. A Linear regression algorithm is used to predict the value of a feature based on the value of another feature. It assumes a linear relationship between dependent feature and independent feature and constructs a linear decision boundary based on the independent feature. This decision boundary is used to predict the value of the unseen data, which is to predict the future data.

$$Y_i = \beta_0 + \beta_1 X_i$$

Figure 2.1 Linear decision boundary.

Figure (2.1) shows the equation of linear decision boundary, where X and Y are the independent and dependent feature and i indicates the data points at each feature. The β0 and β1 are the intercept and slope of the linear decision boundary. Now a linear regression model is built on the historical dataset using Site.D data as dependent feature and year data as independent feature. A linear regression model is constructed with intercept of 1682.3810 and slope of -0.8143. This model is used to predict the number of breeding pairs. At site D in the year 2014, the model predicted 42 number of breeding pairs and with interval prediction, the model estimated that the interval ranges from 28 to 56 number of breeding pairs.

3. Measurement data:

The measurement data has information about two sub-species: red-legged and black-legged kittiwakes. And their attributes such as weight is measured in gram, wing span is measured in centimetre and culmen beak length is measured in millimetre. There are 16 black-legged and 15 red-legged observations in the dataset. Ornithologist wants to a) visually summarize the dataset and find out, b) is wing span and culmen beak length independent for each sub species. Furthermore, ornithologist want to know c) is there enough evidence that the weight of each sub-species are different and also d) is there evidence that two sub-species is different from each other. All these tasks will be addressed in the further segment.

a) In the measurement data, there are two different sub-species. They are red-legged kittiwakes and black-legged kittiwakes. Firstly, let’s summarize the data collected by ornithologist. Figure (3.1) shows represents the box plot for each measurement of both sub-species. The average weight of black-legged kittiwake is lower than the average weight of red-legged kittiwake. The black-legged kittiwake weight ranges around 355 to 395 grams whereas for red-legged kittiwake ranges around 350 to 405 grams. For wingspan attribute, the wingspan value of black-legged kittiwake ranges approximately from 90 to 105 centimetre with average wingspan little above 95 centimetre and on the other hand, the wingspan value of red-legged kittiwake ranges approximately from 85 to 110 centimetre with average wingspan little below 100 centimetre. And for the attribute culmen length, the average beak length for black-legged kittiwake was around 40 millimetre and its length ranges roughly from 37 to 43 millimetre. Similarly, the average beak length for red-legged kittiwake was around 34 millimetre and its length ranges roughly from 30 to 38 millimetre.

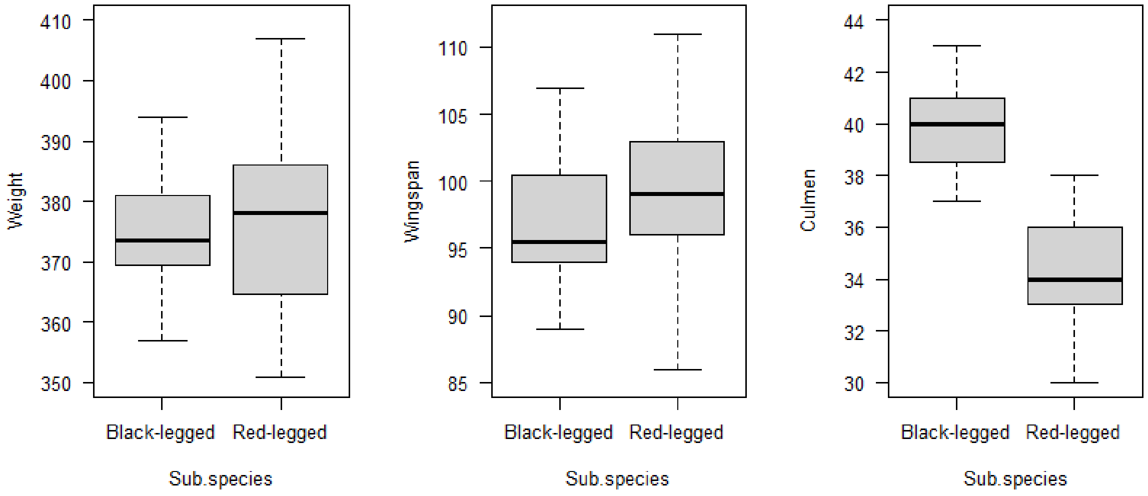


Figure 3.1 Box plot for each measurements of both sub-species.

Furthermore, Figure (3.2) and (3.3) shows the scatter plot for red-legged and black-legged species respectively. Both plots provide an overview of how the attributes like weight, wing span and culmen is spread in two dimensional space for both sub species. The weight and culmen length of red-legged species and also weight and wing span of red-legged species tends to have a weak linear relationship between them. On the other hand, there is no relation observed in the scatter plot of black-legged species. That is the data points were scattered randomly.

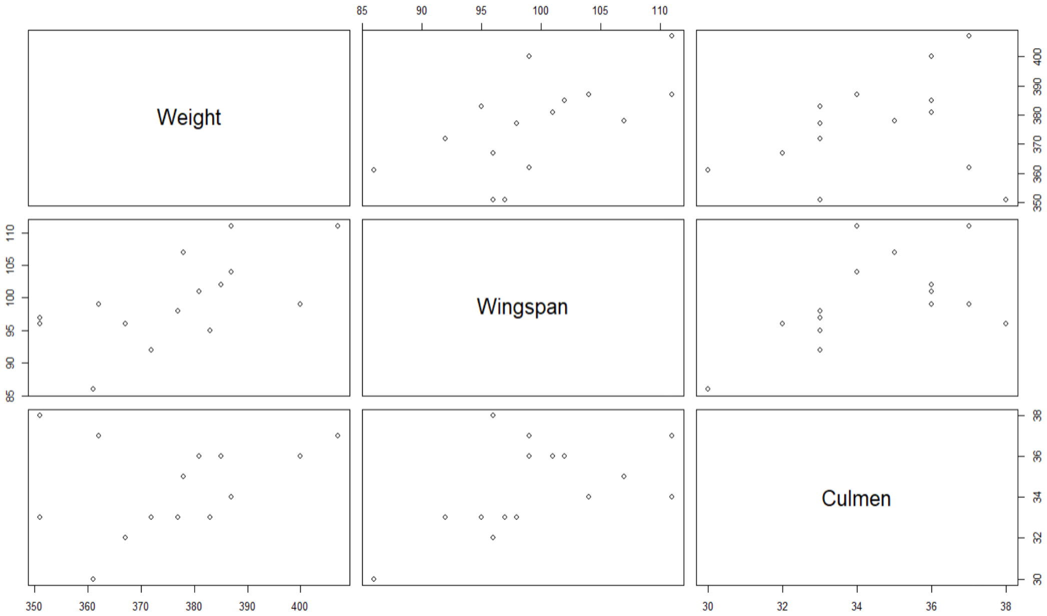


Figure 3.2 Scatter plot for red-legged species.

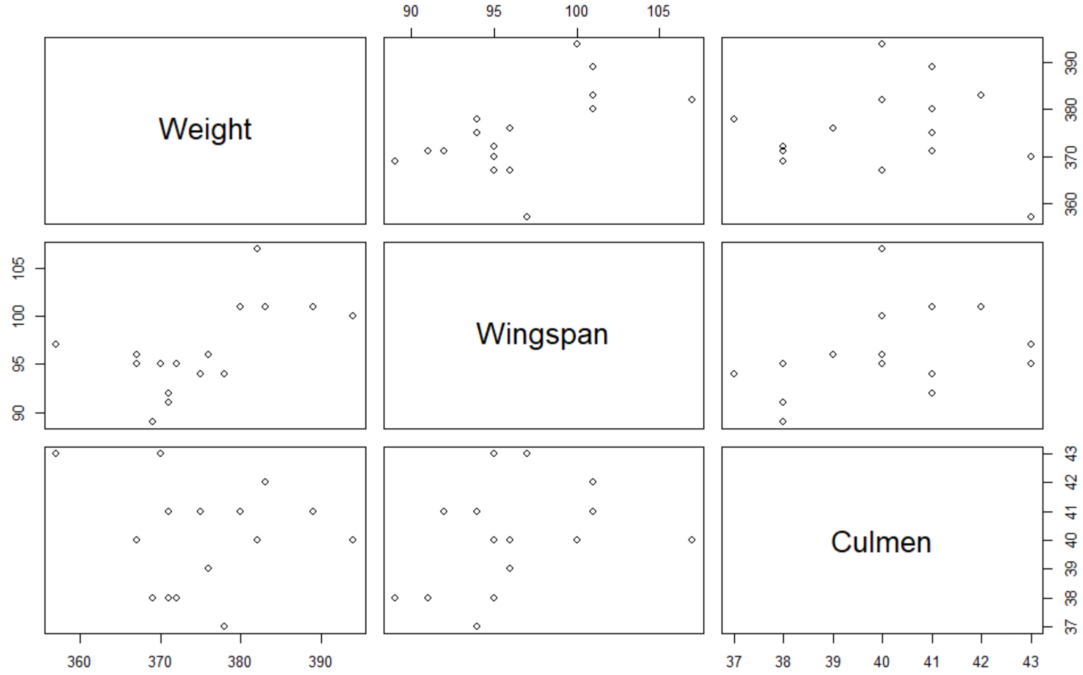


Figure 3.3 Scatter plot for black-legged species.

b) To find out, if wing span and culmen length is independent for each sub-species, Pearson's product-moment correlation is utilized. It is a measure of relationship between two features and it is also used to capture the correlation between two features. Calculate the Pearson's product-moment correlation for wing span and culmen length for both the species. Based on the result obtained from the test, the correlation value for black legged kittiwake's wingspan and culmen length is 0.3698665. Similarly, the correlation value for red legged kittiwake's wingspan and culmen length is 0.5267068. Therefore for both sub species the wing span and culmen length is independent but it has very weak dependency in black legged and moderate relation in red legged.

c) To check the weights of kittiwakes for two sub-species are different, two sample t-test is used. Two sample t-test is a method used to test whether the average population of two groups are same or different. In order to use two sample t-test for hypothesis testing, null and alternative hypothesis has to be assumed. Null hypothesis (H0) is attribute of two sub-species is same. Alternative hypothesis (H1) is attribute of two sub-species is different. And also a significance level (α) of 95 percentage is predefined, so α is set to 0.05. Here, attribute is assumed as weight of the kittiwake. Now two sample t-test is calculated for the weight of red-legged kittiwakes and black-legged kittiwakes. From the test result, the test statistic value is 0.3721 and p-value is 0.7459. Since p-value is greater than α value, there is not enough evidence to reject the null hypothesis at the 5% significance level. Therefore the weight of red-legged kittiwakes and black legged kittiwakes are not different, in other words it is same.

d) To examine whether there is difference between two sub-species of kittiwakes, again two sample t-test is used for all attributes of kittiwakes. Since, it is already found that weight attribute of these two sub-species is same in previous segment. Let’s calculate two sample t-test for other attributes with same assumption of null and alternative hypothesis. Firstly, let’s assume attribute as wingspan and then two sample t-test for the wingspan of red-legged kittiwakes and black-legged kittiwakes with same five percent significant level is calculated. From the outcome of this result, the test statistic value is 1.5058 and p-value is 0.1429. Since p-value is greater than α value, there is not enough evidence to reject the null hypothesis at the 5% significance level. Hence, the wing span length of red-legged kittiwakes and black kittiwakes are same. Secondly, let’s assume last attribute as culmen length and then two sample t-test is calculated for culmen length of both species with same hypothesis assumption. Based on the result, the test statistic value is -7.8896 and p-value is 1.06e-08. Here, since the p-value is less than alpha. The null hypothesis is rejected. Thus there is significant difference in culmen length attribute of each sub-species. In conclusion, by comparing the result of each attribute of kittiwakes. The two sub species differ from each other in terms of culmen length attribute but on the contrary, they do not differ in terms of kittiwake’s weight and wing span attribute.

4. Location Data :

Location dataset contains the details about the number of breeding pairs of kittiwakes in 26 colonies noted along with the significant dependent features like mean summer temperature, cliff height measured in logarithm, sandeel concentration and coastal direction. Coastal direction is a categorical feature which has four types direction namely north, south, east and west. Ornithologist wants to a) create a linear model that predicts the number of breeding pair and also b) another linear model that predicts the logarithm of breeding pair. c) Then choose the best linear model for the location data and d) justify the model fit and effect of the selected covariates on the number of breeding pairs. And finally, e) predict the number of breeding pair with south as coastal direction, 2.56 as sandeel concentration, 20.8 as mean summer temperature and cliff height as 1.89 with 98 percent confidence interval.

Let’s solve each task one by one. a) Multiple Linear regression analysis is used to create a linear model for many independent features, referring the same concept used in 2(a). In 2a) the model is built on one independent features. In this task the model will take multiple independent features and use one target feature. That is, the model is built by assigning the total number of breeding as dependent features and rest other feature like cliff height, coastal direction, sandeel concentration and summer temperature as independent features. The linear model builds a decision boundary with β0 intercept and β1 co-efficient for each independent feature. In this model, the intercept is -393.0565 and co-efficient of north coastal direction is 25.6066, co-efficient of south coastal direction is -99.7269, co-efficient of west coastal direction is -53.1488, co-efficient of sandeel is 40.7747, co-efficient of summer temperature is -0.1404 and co-efficient of cliff height is 195.8277 is observed. The residual standard error is the average of predicted value that deviates from the actual value. The residual standard error for this model is 84.2 and AIC is 312.1539, where AIC is a mathematical method for evaluating how well a model fits the data it was generated from. Based on the summary of the model, cliff height independent feature has the highest contribution to predicting the number of breeding pairs.

b) Similar to previous task, a linear model is built to predict logarithm of breeding pair. The model is built by replacing the dependent feature with log of number of breeding pair. Here the intercept value is 0.990730 and co-efficient of coastal direction of north independent covariate if 0.990730, co-efficient of coastal direction south independent feature is 0.032404, co-efficient of coastal direction west independent feature is 0.106429, co-efficient of sandeel covariate is -0.272626, co-efficient of summer temperature is -0.009208 and co-efficient of cliff height is 1.236291 is observed. The residual standard error is 0.1415 and AIC is -20.03761 is obtained. From the summary of the model, the p-value is 1.953e-15 and residual standard error is 0.1415. The sandeel and cliff length covariate contribute the most to the prediction of number of breeding pairs.

c) To choose the best linear model with best covariates, Step function is utilised. It is function available in R language that helps to choose the best model by removing the features which is not contributing to the construction of decision boundary in multiple regression linear algorithm. Applying the step function in the linear regression model created in 4(a) that predicts the number of breeding pair. It resulted the model with AIC of 234.37 and residual standard error is 0.7599. Based on the summary of step function model, the cliff height covariate contribute the most to the prediction of number of breeding pairs.

d) Comparing the model obtained from step function and the model generated in 4(a) the AIC value is decreased from 312.1539 to 234.37 and residual standard error is also decreased from 84.2 to 0.7599 and also the p value is 1.248e-05 which is way lower than the significance level making it more robust and best model among all others to predict the number of breeding pairs. Figure 4.1 shows the scatter plot for predicted values and residuals of the model obtained from step function. The data points are well randomly distributed and does not have any patterns, which indicates the model is fitting the independent features well.

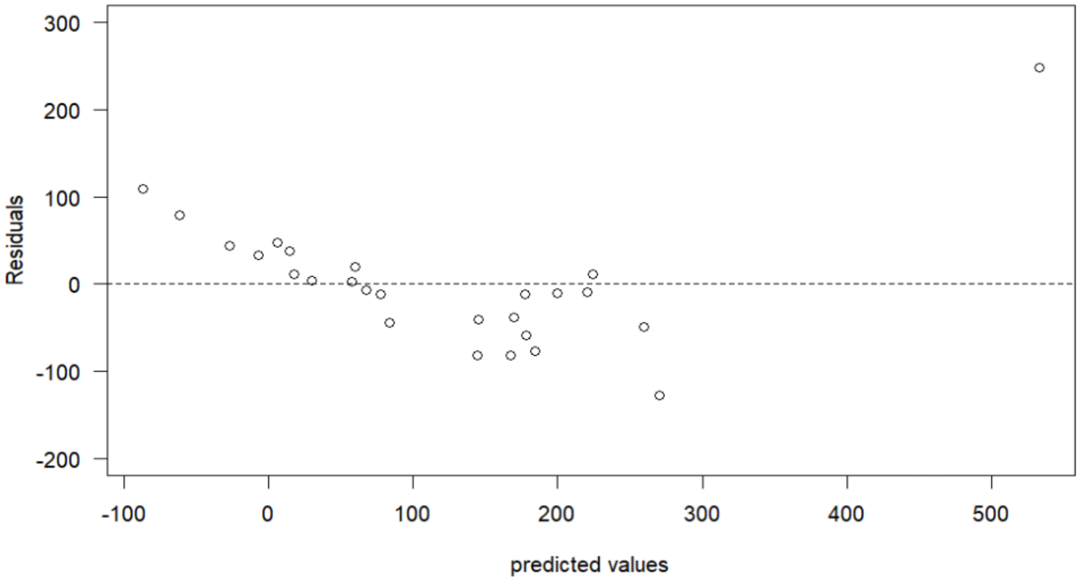


Figure 4.1 Scatter plot for predicted values and residuals.

e) The final task is to predict the number of breeding pair with south as coastal direction, 2.56 as sandeel concentration, 20.8 as mean summer temperature and cliff height as 1.89 with 98 percent confidence interval. Using the model obtained from step function predict the number of breeding pair. The value resulted from the model is -21 number of breeding pairs and with 98 percent confidence interval, the value ranges from -140.913 to 98.76552 number of breeding pairs.

Conclusion:

In conclusion, the report has in-depth analysis of kittiwakes based on the dataset provided by the ornithologist. The analysis of observation data highlighted the most number of kittiwakes can be monitored at dawn and dusk. The historical data revealed that reduction in number of kittiwakes is independent of site over time. The measurement data gives the information about the attribute of two sub species in kittiwake. And, the difference is only between the two subspecies is culmen length; the weight and wing span are almost the same. Furthermore, linear models were constructed in the location based data and in which cliff length covariate contribute the most to the prediction of number of breeding pairs. In order to further understand the kittiwake behaviours, collect more information about their types of food they eat and intake of food at each seasons.