INTRODUCTION

The analysis presented in this report examines four distinct datasets related to kittiwakes, focusing on observation, measurement, historical and location data. The primary objective is to provide insights into the various aspects of kittiwake behavior, popular trends, physical attributes and habitat influences. We will address the following research questions:

1. Provide an exploratory analysis of the observation and construct a 99% confidence interval for the mean number of kittiwakes observed at dusk.
2. Does the historical data support the hypothesis that the decline in kittiwake numbers over time, is independent of site? Estimate the number of breeding pairs at site A in 2014.
3. For the measurement data,
    a. Provide a visual summary of the data.
    b. Check the independence of wingspan and culmen length for each sub-specie.
    c. Is there evidence that the weights of birds of the two sub-species are different?
    d. Is there evidence that there is a difference between the two sub-species?
4. For the location data,
    a. Fit a linear model to predict the number of breeding pairs.
    b. Fit a linear model to the logarithm of number of breeding pairs.
    c. Choose the most appropriate linear model.
    d. Comment on model fit and effect of the selected covariates.
    e. Provide an 80% confidence interval for the number of breeding pairs when coastal direction = West, sandeel = 2.93, temperature = 27.7 and cliff height = 3.55.

TASK 1 – OBSERVATION DATA

The observation data contains the number of kittiwakes sighted at different times of the day. The task was to do an exploratory analysis and to calculate a 99% confidence interval for the mean number of kittiwakes at dusk.

METHODOLOGY

1. The summary statistics of the data including mean and standard deviation were calculated to analyze the data and boxplots and histograms were used for visual analysis.
2. The following formula was used to calculate the 99% confidence interval for the mean number of kittiwakes at dusk.

    Confidence Interval = Sample Mean ± critical value * √ (sample variance / n)

    Where n is the number of samples and the critical t value is calculated for 99% confidence level and n-1 degrees of freedom.

    A t-test was also conducted in R in order to verify the calculated confidence interval.

RESULTS AND INTERPRETATION

1. Exploratory Analysis

After performing an initial data analysis, the following summary statistics were obtained.

Dawn	**Noon**	**Afternoon**	**Dusk**
Min. :34.00	Min. :14.00	Min. :30.00	Min. :41.00
1st Qu. :43.75	1st Qu.:35.00	1st Qu.:35.00	1st Qu.:48.75
Median :49.00	Median :40.50	Median :41.00	Median :53.00
Mean :50.86	Mean :40.18	Mean :40.82	Mean :53.68
3rd Qu. :59.25	3rd Qu.:45.25	3rd Qu.:43.50	3rd Qu.:55.25
Max. :67.00	Max. :64.00	Max. :67.00	Max. :77.00

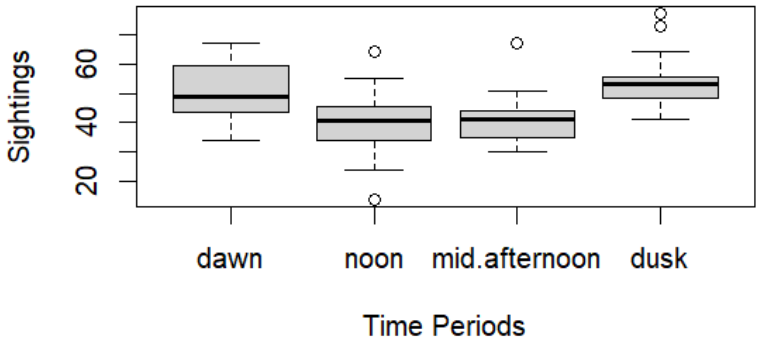
The standard deviation was calculated as 9.32, 10.55, 7.93 and 8.12 for dawn, noon, afternoon and dusk respectively.

We observe that both the mean and median increase in the following order.

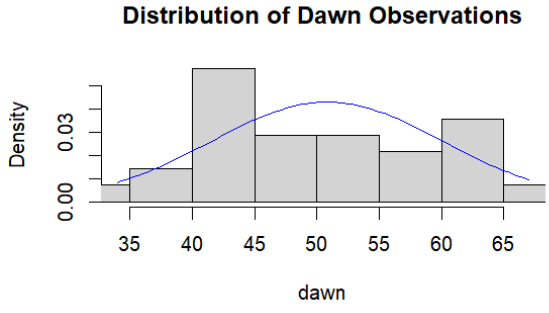
Noon < Afternoon < Dawn < Dusk

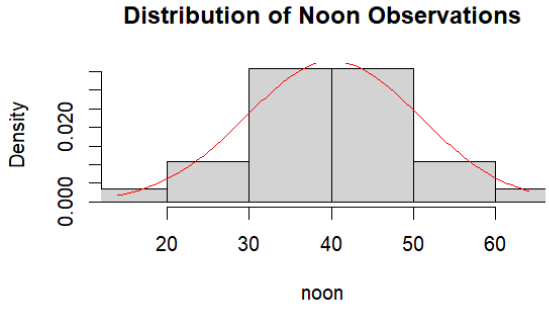
For a more clear analysis, we can compare the boxplots for each time of the day.

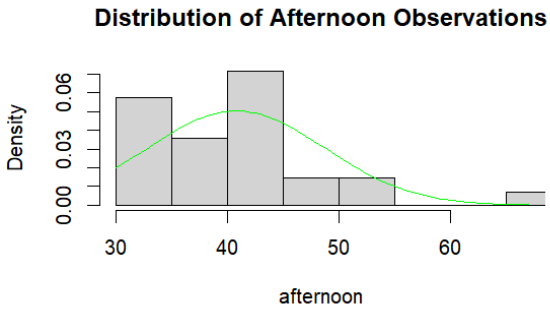
Although, dawn appears to have the highest variance, we obtained a higher standard deviation for noon. This is because of the outliers that can be seen on the boxplot. Afternoon and noon have similar mean and median values however, noon has a higher range and variance than afternoon.

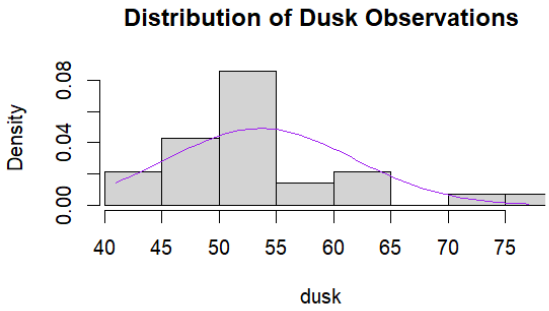


From the following histograms, we can see that the sightings at each time of the day follow a normal distribution.

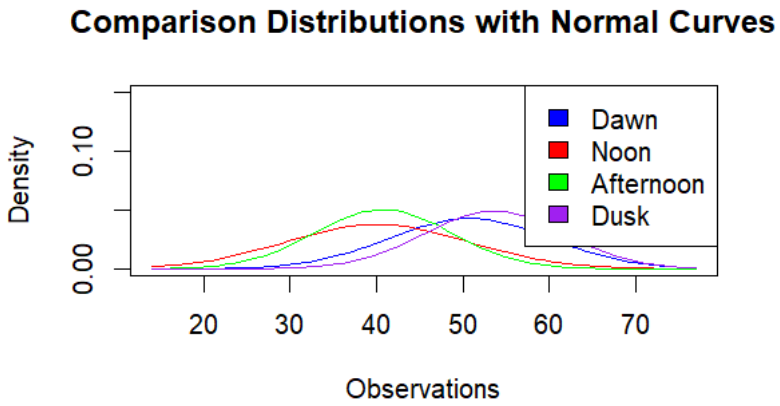








The distribution curves are compared in the following plot, where we can see that the curves for noon and afternoon are towards the left of the other two curves. This suggests that generally the sightings at noon and afternoon are less than the sightings at dawn and dusk.



2. Confidence Interval

The 99% confidence interval for dusk was calculated as (49.425, 57.932). This suggests that the mean of the number of kittiwakes sighted at dusk lies between this interval. The estimated mean of the population is 53.678, which was obtained by applying a t-test.

TASK 2 – HISTORICAL DATA

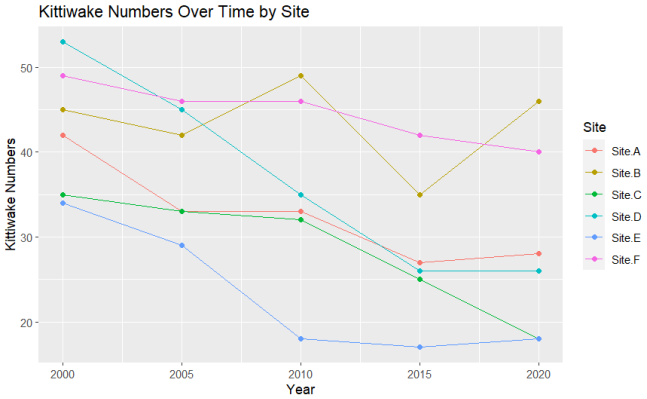
The historical data contains the number of breeding pairs at 6 sites in 5 different years. The task was to find out if the decline in kittiwake numbers over time is independent of site and to estimate the number of breeding pairs at site A in 2014.

METHODOLOGY

1. In order to test if the decline is independent over time, a chi-square test was conducted on the data to see if the number of kittiwakes at each state follow the same distribution.
2. A linear model was trained on the Site A data, to predict the breeding pairs in 2014.

RESULTS AND INTERPRETATION

Initially a graph was plotted for all the sites in order to visually analyze the trend in the the decline. As we can see sites A, C, E and F appear to have similar decline. However, the slope for site D is steeper and site B has several ups and downs.



A linear model was fit on the data for each site in order to obtain the slopes. The slopes were calculated as -0.68, -0.10, -0.84, -1.46, -0.88 and -0.44 for the sites A, B, C, D, E and F respectively. This difference in slopes can be by chance or due to the dependence on the sites.

In order to confirm if the breeding pairs at all the sites follow a similar distribution, a chi-square test was conducted.

a. Null Hypothesis: All sites follow the same distribution. (Decline is independent of site)
b. Alternate Hypothesis: All sites do not follow the same distribution. (Decline is not independent of the site)

An X statistic = 15.65 for 20 degrees of freedom and a p-value of 0.7381 was obtained. Since the p-value > 0.05 (5% significance level), we do not have evidence to reject the null hypothesis. Therefore, we conclude that the decline in kittiwake numbers is independent of the site.

To predict the breeding pairs at site A in 2014, a linear model was trained on the data for Site A, which had a p-value of 0.035 < 0.05 (assuming 5% significance level). Therefore, a linear model is an appropriate choice to make the prediction.

A prediction was made by the model for the year 2014 and it was predicted that 29.88(rounded to 30) kittiwakes were sighted at Site A in 2014.

TASK 3 – MEASUREMENT DATA

The measurement data contains the weight, wingspan and culmen length for black-legged and red-legged kittiwakes. The task was to plot a visual summary of the data, check if wingspan and culmen length of each sub-specie are independent, check if the weights of birds of the two sub-species are different and to find out if there is an overall difference between the two sub-species.

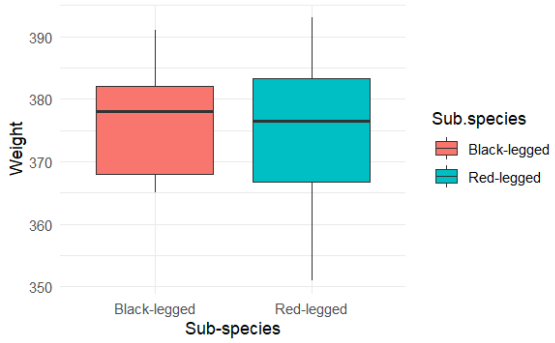
METHODOLOGY

1. Boxplots were used to present the visual summary of the measurement data. Separate boxplots were plotted for weight, wingspan and culmen for a visual comparison between the sub-species for each attribute.
2. As we are dealing with numerical data, we use correlation test to check the independence between wingspan and culmen for each sub-specie. Two separate correlation tests were conducted to check the independence in case of either sub-specie. A scatter plot between wingspan and culmen was also plotted to analyze the relation between them.
3. A t-test comparing the mean weight of each sub-specie was used to find out if the two sub-species differ in weight.
4. In order to check the overall difference between the two sub-species, three t-tests comparing the weight, wingspan and culmen length of each sub-specie were conducted.

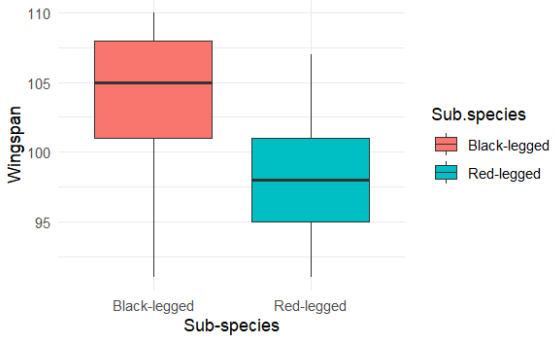
RESULTS AND INTERPRETATION

1. Visual Summary

From the first boxplot, we can see that the weights of both the sub-species are quite similar. The median for the black-legged species is slightly higher than the red-species. However, the weight of the red-species has a higher inter-quartile range and variance.



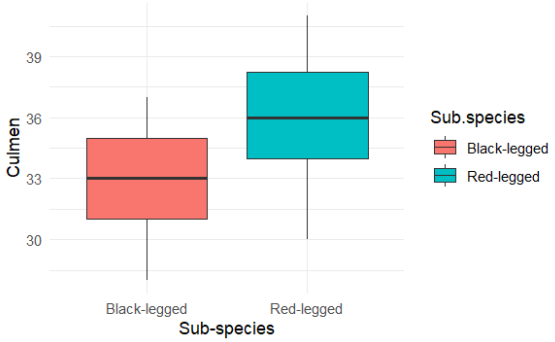
The second boxplot suggests that the black-legged species generally have a higher wingspan than the red-legged species. The inter-quartile range and variance is slightly higher incase of black-legged species.



According to the third boxplot, the culmen length of the red-legged species is generally higher than the black-legged species. The inter-quartile range and variance for the red-legged species is also higher than the black-legged species.

2. Independence of Wingspan and Culmen Length

In order to check the independence of wingspan and culmen length for each specie, two separate correlation tests were conducted.



The first correlation test was conducted between the wingspan and culmen length of the black-legged species.

a. Null Hypothesis (H0): Correlation is equal to 0. (The wingspan and culmen length of the black-legged sub-species are independent)
b. Alternate Hypothesis (H1): Correlation is not equal to 0. (The wingspan and culmen length of the black-legged sub-species are not independent)

The correlation test resulted in a p-value = 0.0009971, which is less than 0.05 (assuming 5% significance level). Therefore, we reject the null hypothesis and conclude that the wingspan and culmen of black-legged species are not independent.

The correlation is estimated as 0.80, which is closer to 1 and lies within the 95% confidence interval (0.45, 0.94).

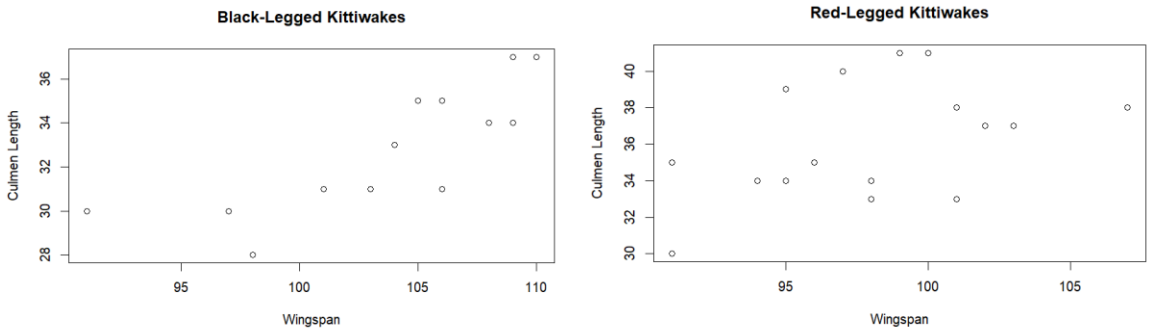
The second correlation test was conducted between the wingspan and culmen length of the red-legged sub-species.

a. Null Hypothesis (H0): Correlation is equal to 0. (The wingspan and culmen length of the red-legged sub-species are independent)
b. Alternate Hypothesis (H1): Correlation is not equal to 0. (The wingspan and culmen length of the red-legged species are not independent)

The correlation test resulted in a p-value = 0.089, which is greater than 0.05 and less than 0.1. Therefore, in case of 5% significance level, we do not reject the null hypothesis and conclude that the wingspan and culmen length of the red-legged sub-species are independent of each other. However, in case of 10% confidence interval, we would reject the null hypothesis.

The correlation was estimated to be 0.4389, with 95% confidence interval (-.007, 0.77)

Scatter plots were plotted between the wingspan and culmen of each specie to observe the correlation. We can clearly see that the quantities are correlated in case of black-legged sub-species however in case of red-legged sub-species, a strong correlation can not be observed.



3. Difference in Weights

A t-test was conducted to compare the weights of the two sub-species.

a. Null Hypothesis (H0): True difference in means is equal to 0. (The weights of black-legged and red-legged species are not different)
b. Alternate Hypothesis (H1): True difference in means is not equal to 0. (The weights of black-legged and red-legged species are different)

We obtain a t-value of 0.294 with degrees of freedom = 26.3 and a p-value of 0.77.

Since the p-value > 0.05 (assuming 5% significance level), we do not reject the null hypothesis. We conclude that the weights of the sub-species are not different. This can be verified from the boxplots in the visual summary section, which show similar weights for both sub-species.

4. Difference in the 2 sub-species

In order to check if the 2 sub-species are different, we conducted 3 t-tests comparing the weight, wingspan and culmen length of each sub-specie.

The first t-test was conducted between the weights of the 2 species. The results are explained in the previous part. We have evidence that the weights of both the species are similar.

The second t-test was conducted between the wingspan of the 2 sub-species.

a. Null Hypothesis (H0): True difference in means is equal to 0. (The wingspan of the two sub-species are not different)

b. Alternate Hypothesis (H1): True difference in means is not equal to 0. (The wingspan of the two sub-species are different)

We obtain a t-value of 2.97 with degrees of freedom 22.4 and a p-value of 0.0069.

Since the p-value < 0.05 (assuming 5% significance level), we reject the null hypothesis. Therefore, we conclude that the wingspans are different for each sub-specie.

The third t-test was conducted between the culmen lengths of the sub-species.

a. Null Hypothesis (H0): True difference in means is equal to 0. (The culmen length of the two sub-species are not different)
b. Alternate Hypothesis (H1): True difference in means is not equal to 0. (The culmen length of the two sub-species are different)

We obtain a t-value of -3.05 with degrees of freedom 26.7 and a p-value of 0.005.

The p-value is less than 0.05 (assuming 5% significance level). Therefore, we reject the null hypothesis and conclude that the culmen lengths for the 2 sub-species are different.

After performing these t-tests, we know that the species have similar weights but differ in wingspan and culmen lengths. These differences can also be observed through the boxplots in the visual summary. Therefore, we conclude that the 2 sub-species are different.

TASK 4 – LOCATION DATA

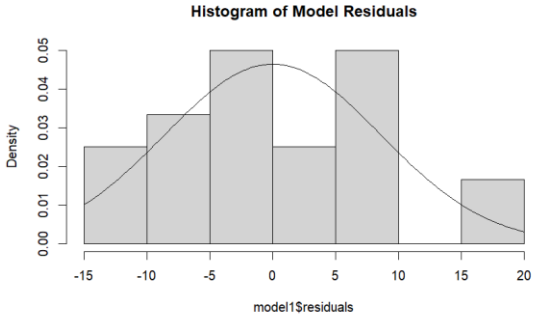
The location data contains the number of breeding pairs in 24 colonies along with the habitat conditions. The task was to fit two linear models on the number of breeding pairs and the logarithm of the breeding pairs respectively, choose the most appropriate model and comment on the model fit. An 80% confidence interval was to be predicted for the number of breeding pairs if coastal direction = West, sandeel = 2.93, temperature = 27.7 and cliff height = 3.55.

METHODOLOGY

1. A linear model was trained that used all the columns to predict the breeding pairs. The step function was then used to keep only the important covariates. A histogram was plotted for the model residuals to see if they follow the normal distribution. Then a scatter plot was plotted between the fitted values and the residual to make sure that the residuals don't follow a trend/pattern and the model assumptions are met.
2. A linear model was trained to predict the logarithm of the breeding pairs using all the columns. Then the step function was used to keep only the relevant covariates. A histogram was plotted to check if the model residuals follow the normal distribution. Then a scatter plot between the fitted values and the residuals was plotted to make sure that the residuals don't follow a trend/pattern and the model assumptions are met.
3. The AIC for both models was calculated the model with the lower AIC was selected.
4. A pair plot was plotted to examine the effect of the selected covariates on the number of breeding pairs.
5. The selected model was used to predict the 80% confidence interval using the predict function in R and setting the interval = "confidence" and level = 0.8

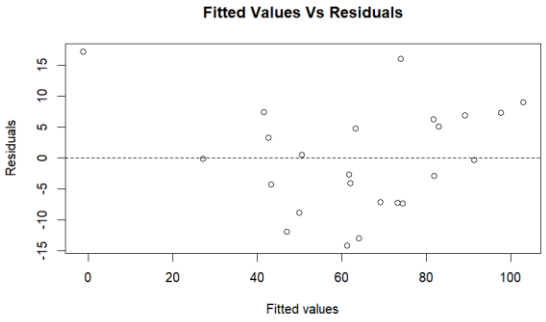
RESULTS AND INTERPRETATION

1. Linear Model for the number of breeding pairs



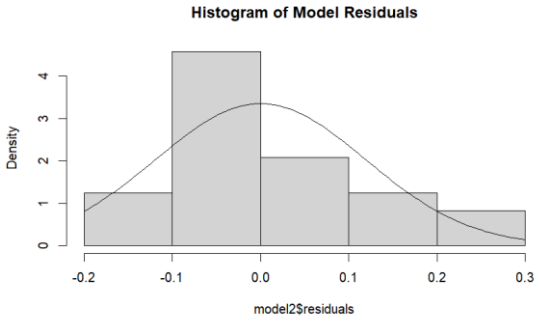
A linear model was trained using all the covariates (Coast direction, sandeel concentration, summer temperature and cliff height) to predict the breeding pairs. The model had a residual standard error 8.847 with 17 degrees of freedom and a p-value 5.622e-08.

The step function was then used to select the relevant covariates to reduce the AIC of the model. The covariates coast direction and temperature are removed because they increase the AIC and are not important to predict the number of breeding pairs. The final model depends on just two covariates sandeel concentration and cliff height and has residual standard error 8.994 on 21 degrees of freedom and a p-value 1.414e-10.



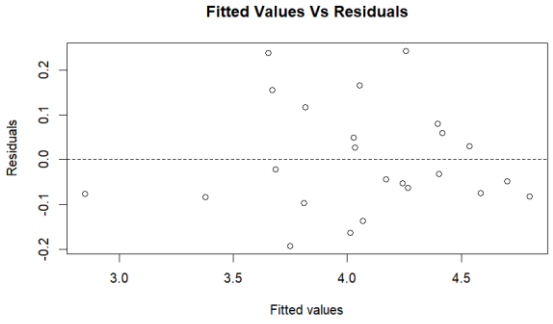
A histogram of model residuals was plotted and a normal distribution was observed. A scatter plot between the fitted values and the residuals was plotted and no pattern was observed in the residuals. The residuals were randomly scattered around 0. Thus, the model satisfies all the linear model assumptions.

2. Linear Model for the logarithm of breeding pairs



The logarithm of breeding pairs was calculated and the linear model was trained to predict the log(breeding pairs) using all the covariates. The model had a residual standard error 0.1315 on 17 degrees of freedom and a p-value 1.88e-09.

The step function was used to select the covariates that help reduce the AIC value. The covariates sandeel and cliff height were selected because the other covariates direction and temperature did not help in prediction. The fitted model has residual standard error 0.1245 on 21 degrees of freedom and a p-value 4.53e-13.



A histogram of the model residual follows a normal distribution. In a scatter plot between fitted values and residuals, the residuals are randomly scattered around 0. Thus, the model satisfies all the assumptions of the linear model.

3. The most appropriate linear model

In order to select the most appropriate linear model, the AIC of both the models was calculated. The AIC for model 1 was 178.34, while the AIC for the second model was -27.08.

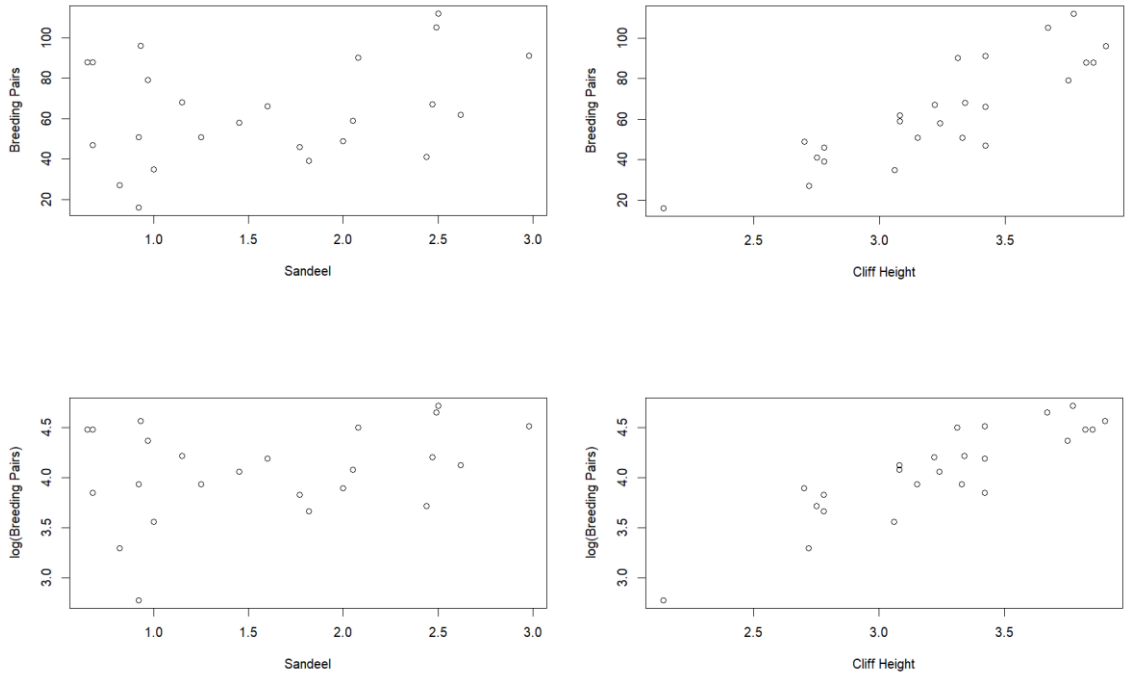
Since the AIC for the second model is smaller than the first model, therefore we selected the second model i.e. the model that predicts the logarithm of the number of breeding pairs.

4. Model Fit and the effect of the selected covariates

The final model has just 2 covariates sandeel and cliff height, which have substantial influence on the prediction. The rest of the covariates cause an increase in the AIC, while both these covariates help reduce the AIC of the model.

For the model with the logarithm of breeding pairs, removing the coastal direction covariate reduces the AIC by about 4, while removing temp decreases the AIC by about 2. Further removing any covariate would increase the AIC, therefore we do not remove sandeel and height.

We can see in the following plots that the sandeel and cliff height are linearly related to the breeding pairs and cause it to increase. We can also see that they are more tightly related to the logarithm of the breeding pairs than the actual breeding pair number.



5. 80% Confidence Interval

The selected model is used to predict the interval for the logarithm of the number of breeding pairs, when the direction = West, sandeel = 2.93, temp = 27.7 and cliff height = 3.55. The interval parameter is set to “interval” and level = 0.8.

The confidence interval for the log of breeding pairs is (4.62, 4.77), while the predicted value is 4.697. We get the actual value of the breeding pairs by taking antilog of the values. Thus, the confidence interval for the breeding pairs is (101.456, 118.424), while the predicted value is 109.612.

CONCLUSION

1. An exploratory analysis of the observation data was done through statistical summary, boxplots and histograms. It was concluded that a fewer number of kittiwakes are sighed at noon and afternoon compared to dawn and dusk. The number of kittiwakes at dawn has the highest inter-quartile range. However, due to some outliers, noon appears to have the highest variance. It was observed that the data follows a normal distribution.
   The 99% confidence interval for mean number of kittiwakes at dusk is (49.425, 57.932)

2. A chi-square test on the historical data returned a p-value of 0.7381 which is lower than the conventional significance level of 0.05. Therefore, we conclude that the breeding pairs at all the sites follow a similar distribution and the decline in the breeding pairs is independent of the site.
   The linear model predicted that 29.88(rounded to 30) birds were present at site A in 2014.

3. Boxplots were used for the visual summary of the measurement data. It was observed that both the sub-species have similar weights. However, black-legged kittiwakes have greater wingspan and red-legged species have a higher culmen length.
   A correlation test was applied between wingspan and culmen length of black-legged and red-legged species. A p-value of 0.0009971 was obtained for the black-legged species. This p-value is smaller than the conventional significance level of 0.05. Therefore, the null hypothesis of independence is rejected.
   The p-value for red-legged species was 0.089 > 0.05. Thus, the null hypothesis is not rejected for 5% significance level but is rejected for 10% significance level. Thus, the wingspan and culmen length of red-legged species are independent in case of 5% significance level but not in case of 10% significance level.
   The t-test on the weights of the two sub-species returned a p-value of 0.77, which is greater than 0.05. Therefore, we do not reject the null hypothesis and conclude that the weights are not different for the sub-species.
   The t-tests applied on weight, wingspan and culmen of the two sub-species returned p-values of 0.77, 0.0069 and 0.005 respectively. We reject the null hypothesis in case of wingspan and culmen but not in case of weight. Therefore, it is concluded that the weights of the two species are similar but the wingspan and culmen are different. Thus, the species are quite different from each other.

4. The covariates sandeel and cliff height were selected for the linear model that predicts the breeding pairs. The model had a p-value of 1.414e-10.
   The linear model for the logarithm of the number of breeding pairs also had the covariates sandeel and cliff height and a p-value of 4.53e-13.
   The model for log of breeding pairs was chosen because of its lower AIC.
   The 80% confidence interval for the number of breeding pairs in the given conditions was predicted to be (101.456, 118.424).