A Statistical Analysis of Bird Extinction Factors
First Year Project

Cristina Avram, Shanon Marietta Badger, Gabriela Zhelyazkova,
Daniil Strelan, Alexis Serruys

March 18, 2023

Abstract

The aim of this report is to present a statistical investigation of a data-set on breeding pairs of land-bird species collected from 16 islands around Britain over several decades. The theme of the project is "factors affecting extinction," where the aim is to determine the relationship between various factors and the extinction times of bird populations. This report will take you on a statistical journey through the data-set, guided by a list of items that mimics the investigation process.

The project is motivated by the need to understand the factors that influence the extinction of bird populations by analyzing a data-set that measures breeding pairs of birds. This report seeks to identify factors that affect extinction in bird populations through a statistical investigation. It provides a step-by-step analysis of the data-set, and the findings can help us understand the relationship between various factors and the extinction times of bird populations.

1 - Introduction

Since the early 1980's, environmentalism has been a concern amongst scientists worldwide. It was in this era that the damage humans can do to their environment and the species they must coexist with first became a major concern and methods to combat extinction became prominent. (Source: 1980s Environmentalism and How the Reagan-Era Shaped the Natural World, Backstory 0322, 4.24.2020). One of the questions asked was how certain factors affected the extinction of species.

A journal concerning this was published in American Naturalist 132 in 1988 authored by biologists Stuart L. Pimm and H. Lee Jones and ornithologist Jared Diamond. Their research focused on birds species and attempted to outline a connection between the time until a species became extinct, the size and migratory status of the birds and the number of breeding pairs in the area. Their data-set was limited to a series of 16 islands around Britain and attempted to determine how long a species of birds could survive in that environment.

This project will test the researchers' assumptions by examining their data-set.

2 - Methodology

This was a guided project, using a series of tasks as a guideline (Project Factors Affecting Extinction, Carlsen). It began with fitting a model to data provided. The data set was examined for outliers and to determine any possible transformations to use on the model. Various transformations were attempted on the data to find the best fitting model, followed by determining if other variables, specifically the one called Pairs should be transformed. Finally, the data was subdivided into four combinations of subsets to determine if their effect on the model was equal and a reduced model was created based on the findings from these experiments.

3 - The Data

To start with, a linear relationship with the data provided by the researchers was assumed and tested using the program R-Studio and the programming language R.

The variables in this data-set are the average time of extinction (Time), average number of nesting pairs (Pairs), species size (Size), denoted by "L" for large or "S" for small and migratory status (Status) of the species (migrant ("M") or resident ("R")). It was desirable to know if there exists a relationship between the extinction time and the number of nesting pairs as well as whether migratory status and size affected the data.

Checking that the data was clean and that no data was missing was necessary even though there should have been no anomalies due to the data being provided from a previous study. However it is good practice to always check the data and clean it as necessary, as one can not assume clean data will be provided. Box-plots, scatter charts and histograms were created from the original data-set allowing confirmation of the data-set's linearity. It was determined that, based on histograms alone, no single two variables were linearly related. Nor was it possible to determine outliers solely from scatter charts. Further investigation would be required to determine the outliers and how to transform the data into a useful form.

After the initial investigation, a standard multiple regression model was created using R to fit the data. This was done with all the variables in combination and this model was inspected for the assumptions of linearity. This generated a model with a low p value (0.0001097), but an unsatisfactory R-squared value of 0.2645. The plots of the model were also enough to reveal that there was no linear relationship. In particular, Figure 1 shows a distinct cluster of data-points around a line that skews markedly from 0, when to prove linearity there needs to be a nearly horizontal line at 0 with no discernible clusters of data.

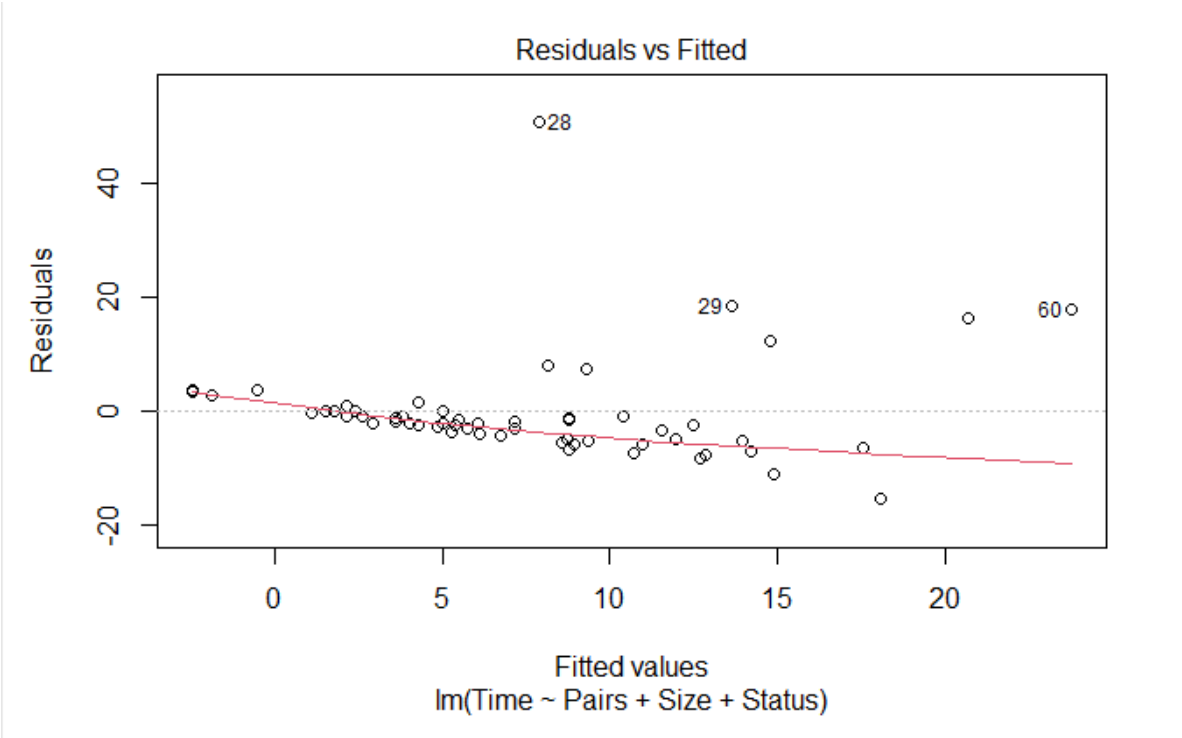


Figure 1: Residuals vs Fitted plot of our first model

4 - Examining and Transforming the Original Model

After examining the full model in detail alongside its values (Adjusted R-squared, p-value) and plots (Residuals vs. Fitted, Cook's Distance, Q-Q Plot) could be evaluated. The p-value is low (0.001446), which means that such an extreme observed outcome would be very unlikely under the null hypothesis. The low adjusted R-squared value (0.2542) indicates that the model explains only 25% of the data.

Another function arguing against efficiency of the model is the Residuals vs. Fitted plot, Figure 2, which is the most preferred plot used for residuals analysis. It is used to detect non-linearity, unequal error variances, and outliers. (Source: Pennsylvania State University). The line should be approximately horizontal to 0, but it is not. Based on the plot, three influential points(outliers) can be seen (points' indexes: 28, 29, 60). To avoid corrupting the data-set by removing too many outliers, the significant points would be removed after some essential transformations.

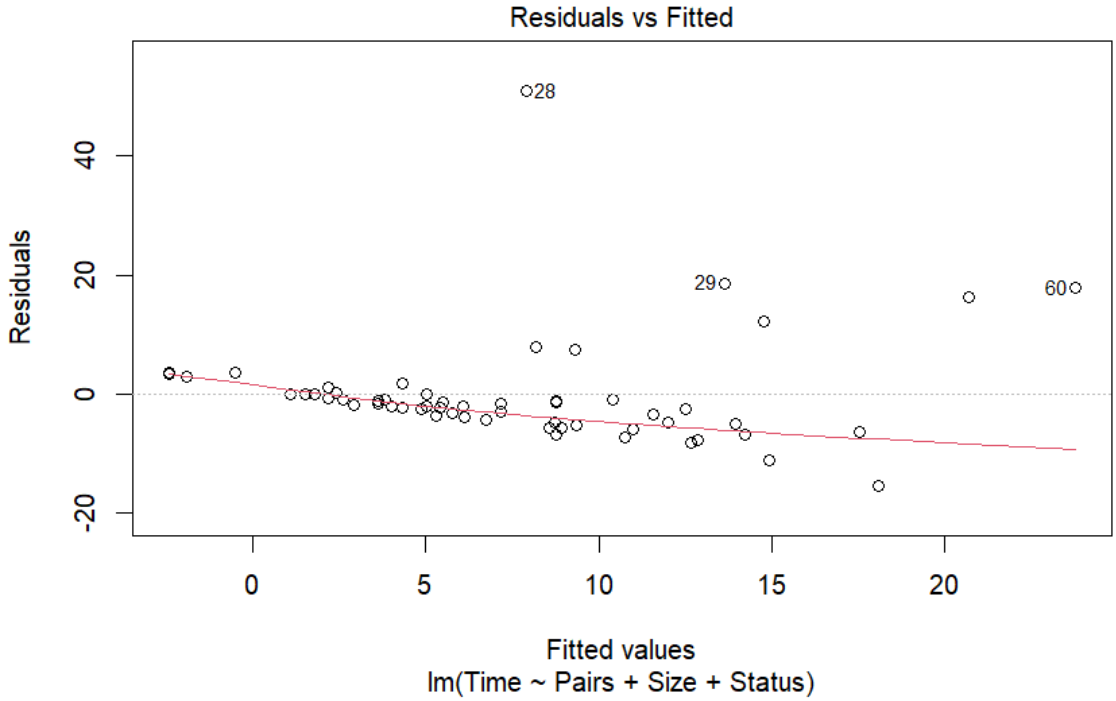


Figure 2: Residual vs Fitted plot of the first model

The Q-Q plot, Figure 3, displays whether the two subsets of data sets are from populations with common distributions, and if so, then the line in this graphic should be almost 45°. The Q-Q plot of the model is far from the 45-degree line, and it can be used as another argument against the original model. The conclusion of all these factors is that the model is unstable, and it should be transformed for efficiency.

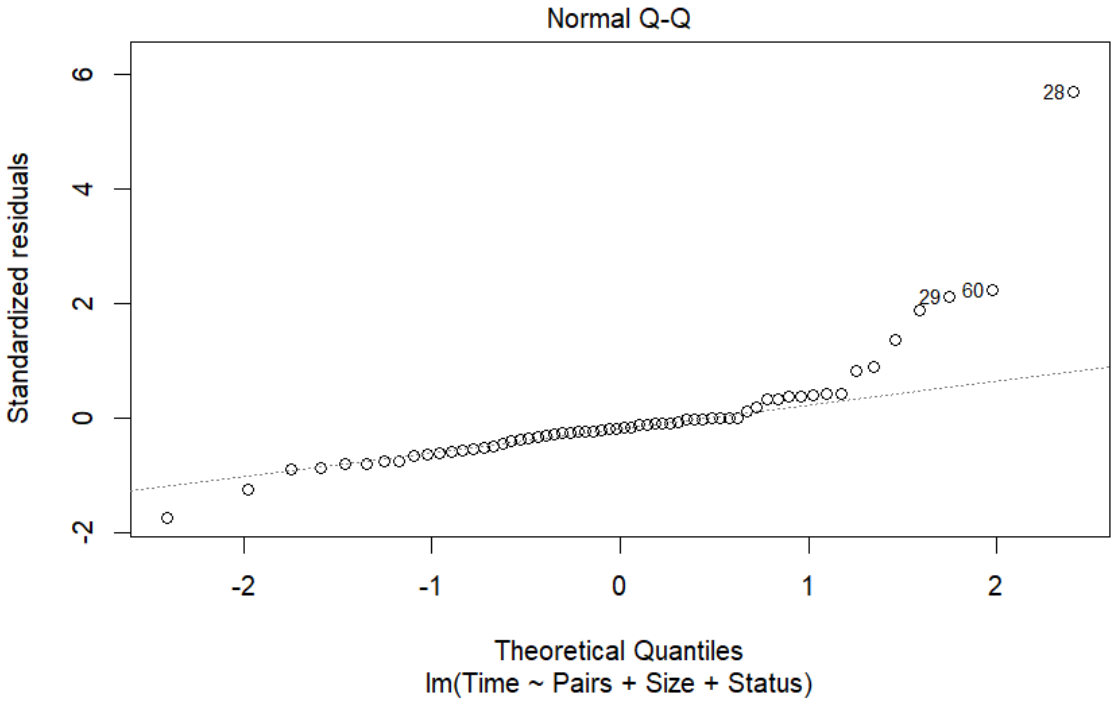


Figure 3: Q-Qplot of the first model

The following transformations were experimented with to find a better model: log(time), sqrt(time), 1/(time). Each of the values was fitted to the model, and the summary of each transformation was plotted, Figures 4 - 6. The best Adjusted R-squared were given by the transformations log(time) and 1/(time).

```
Residual standard error: 0.6523 on 58 degrees of freedom
Multiple R-squared:  0.5984,    Adjusted R-squared:  0.5776
F-statistic: 28.81 on 3 and 58 DF,  p-value: 1.557e-11
```

Figure 4: Summary of the log(time) model

```
Residual standard error: 1.069 on 58 degrees of freedom
Multiple R-squared:  0.454,     Adjusted R-squared:  0.4257
F-statistic: 16.07 on 3 and 58 DF,  p-value: 1.012e-07
```

Figure 5: Summary of the sqrt(time) model

```
Residual standard error: 0.1941 on 58 degrees of freedom
Multiple R-squared:  0.6038,    Adjusted R-squared:  0.5833
F-statistic: 29.46 on 3 and 58 DF,  p-value: 1.059e-11
```

Figure 6: Summary of the 1/time model

As the Adjusted R-squared values for both of these transformations are similar, the Residuals vs. Fitted plot determined the model that represented a better linear relationship, Figures 7 and 8.

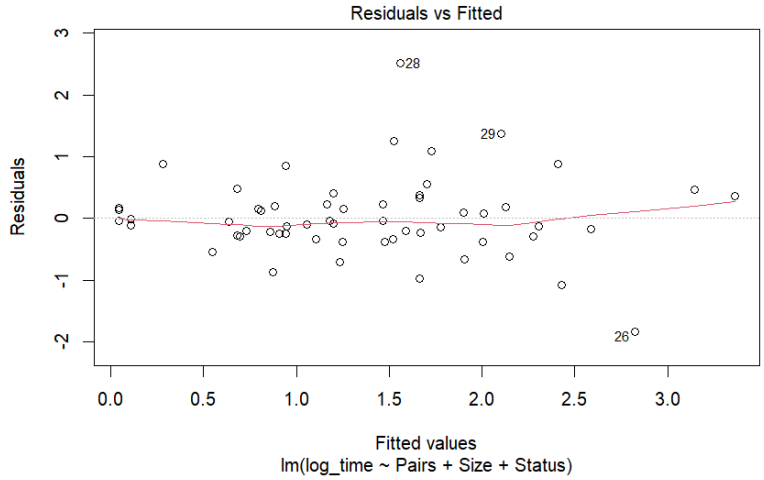


Figure 7: Residuals vs Fitted plot of the log_time model

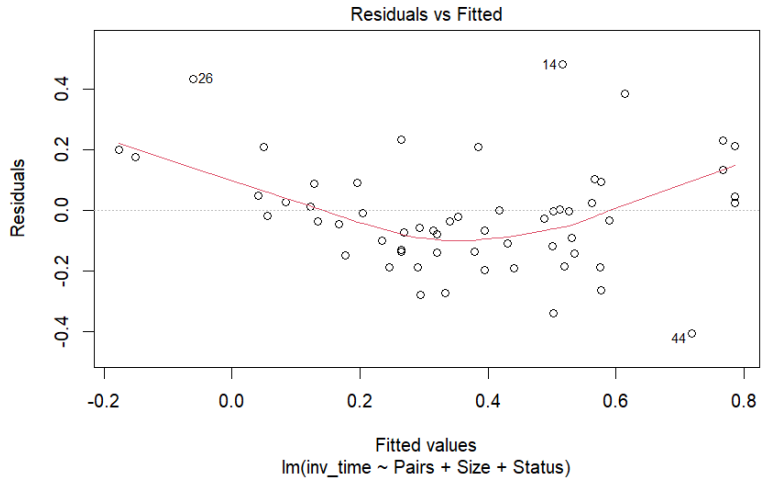


Figure 8: Residuals vs Fitted plot of the 1/time model

After observing the plots, it is obvious that the log(time) transformation gives a better linear relationship (the red line of the log(time) transformation is approximately horizontal compared to the one for 1/(time)). The Residuals plot also shows three outliers (26, 28, 29) that should be further investigated and some of them (or even all) to be removed to create a more sustainable model.

5 - Outliers and Further Transformations in order to create the reduced model

In order to identify the most relevant outliers on each end of the scale, a Cook's distance plot of the new model was generated along with using the the built-in R function "identify" to find influential outliers.

The Cook's distance plot, Figure 9, suggests that the birds with indices 26, 28, 29 and 18 are the most influential outliers.

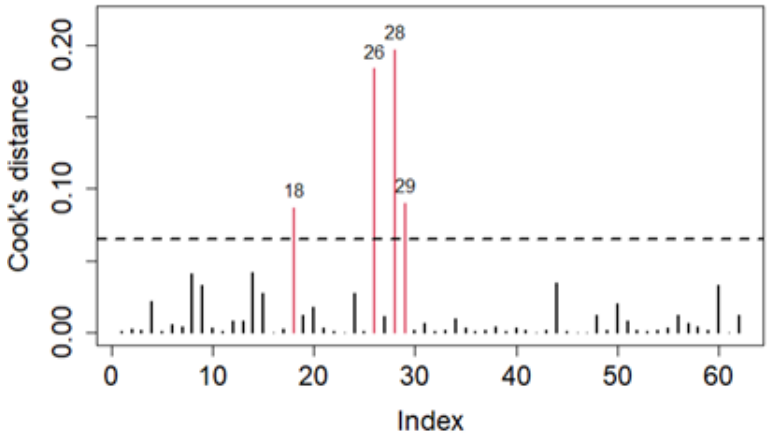


Figure 9: Cook's distance plot of our current model

Inspecting the data set showed that the bird with index, 28, the Raven, is the outlier at the high end of the scale, having the longest extinction time in the entire data set, and the bird with index 26, the Jackdaw, is the outlier at the lower end of the scale, with one of the shortest extinction times in the data set. This is confirmed by the model's Q-Q plot, where the outliers 28, 29 and 26 are visible, with 28 and 26 being at the upper and the lower ends respectively. (see Figure 10)

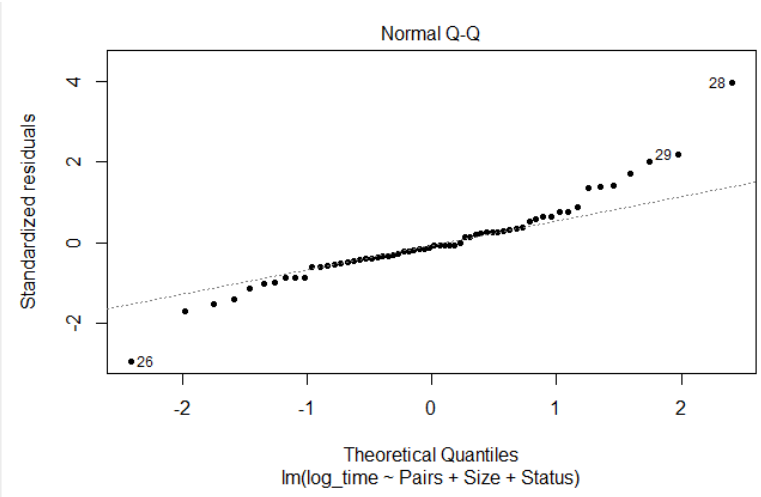


Figure 10: Q-Q plot of our current model

The built-in R function "identify" reports the outliers 26, 28 and 29, which yet again confirms our previous results. Since the outliers 26, 28 and 29 are confirmed using 3 different methods, the Cook's distance plot, the Q-Q plot and the "identify" function, the decision was made to remove them to

ascertain if it would improve the model. Before removing the outliers, the adjusted R^2 of the model was 0.5776. and the p-value was 1.557e-11 . (see Figure 11)

```
Call:
lm(formula = log_time ~ Pairs + Size + Status, data = birds)

Residuals:
     Min       1Q   Median       3Q      Max
-1.83997 -0.29458 -0.07187  0.21712  2.51691

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.43056    0.20706   2.079 0.042011 *
Pairs           0.26509    0.03679   7.206 1.32e-09 ***
SizeSmall      -0.65237    0.16665  -3.915 0.000241 ***
StatusResident  0.50406    0.18261   2.760 0.007717 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6523 on 58 degrees of freedom
Multiple R-squared:  0.5984,    Adjusted R-squared:  0.5776
F-statistic: 28.81 on 3 and 58 DF,  p-value: 1.557e-11
```

Figure 11: Summary of model with outliers

After removing the outliers, the adjusted R^2 of the model is 0.7217. It is also notable that the model obtains a very low p-value, of only 6.339e-16. (see Figure 12)

```
Call:
lm(formula = log_time ~ Pairs + Size + Status, data = birds_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-1.07187 -0.28701 -0.05488  0.26100  1.22410

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.39042    0.15645   2.496  0.01560 *
Pairs           0.28236    0.02843   9.932 7.11e-14 ***
SizeSmall      -0.66713    0.12703  -5.252 2.51e-06 ***
StatusResident  0.43983    0.13650   3.222  0.00214 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4841 on 55 degrees of freedom
Multiple R-squared:  0.7361,    Adjusted R-squared:  0.7217
F-statistic: 51.15 on 3 and 55 DF,  p-value: 6.339e-16
```

Figure 12: Summary of model after removing outliers described above

The R^2 and p-value of the model with the removed outliers are greatly improved from those of the model before eliminating the outliers, which argues for our cleaned data set, without the outliers 26, 28 and 29.

However, the working data set is very small, with only 62 observations. Removing too many outliers could make the model over-fitted, and cause poor performance regarding new predictions. The 3 chosen outliers represent approximately 5 percent of the data set. Removing further outliers in order to achieve a model with a higher scoring R^2 can have an adverse affect on the model because of the risk of over-fitting and the high percentage of data this would exclude.

The current model, using the cleaned data set with the removed outliers described above, and a logarithmic transformation of the extinction times, **(model ← - lm(log(Time) ~ Pairs + Size + Status, data = cleaned_bird_data)**, relies on an already linear relationship between the predicted variable, Time and the predictor variables Pairs, Size and Status. It is plain to see that this is the case from the residual plot of the current model, which shows no pattern in the residuals, and an approximately horizontal line at 0. (see Figure 13)

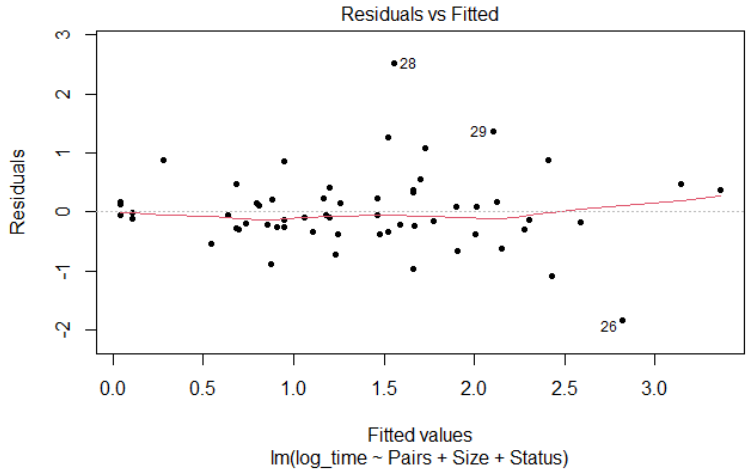
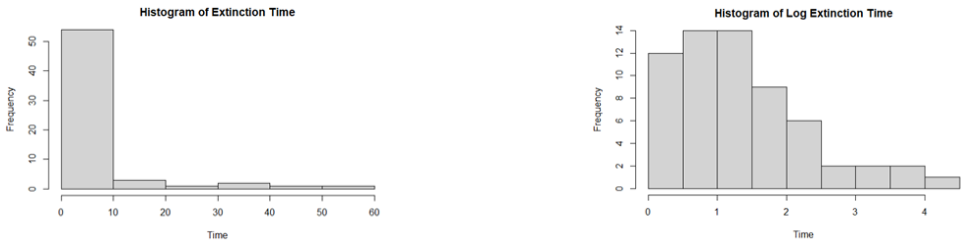


Figure 13: Residual plot of our current model

Since there is already a linear relationship, there is no reason to transform the Pairs variable. We decided to transform the Time variable, because the original data did not show a linear relationship between Time and the predictor variables. It was therefore not possible to construct a valid model. Performing the transformation was necessary to create said linear relationship. It also made sense to use a logarithmic transformation on Time based on its histogram plot. The histogram plot below shows that the original values of Time were not normally distributed, which is another assumption of linear regression that was not met by our original full model. (see Figure 14a)



(a) before log transformation

(b) after log transformation

Figure 14: Histogram of extinction time with different time transformation

This issue and the linearity issue discussed above were accounted for by logarithmically transforming the time variable. (see Figure 14b)However, considering the last changes in the model, the assumptions of linear regression are already met and the model now has a fairly satisfying R^2 value. It is neither required to further transform any variables, nor does it make logical sense in the context of what the data represents.

There is no reason to further transform any variables, as this would lead to merely using transformations to obtain better metrics, at the cost of said transformations eradicating all logical context to our data and losing much of the original meaning of the data. Such a model may obtain a higher R^2 value or a lower p-value, but it could not be used for accurate predictions for any new data, and thus would not be useful. There would be no further transformations on this model.

6 - Testing and the Reduced Model

To test whether the slopes for all four combinations of "size" and "migratory status" are equal, one can create interaction terms between these variables and include them in the model (Figure 15). This allows for testing differences in the slopes between the different groups. The lm() function in R can be used to perform this analysis, with the natural log of extinction time as the dependent variable.

```
{r}
model_logclean_int <- lm(log_time ~ Pairs + Size + Status + Size:Status, data =
birds_clean)
summary(model_logclean_int)
```

```
Call:
lm(formula = log_time ~ Pairs + Size + Status + Size:Status,
    data = birds_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11837 -0.24665 -0.06583  0.24818  1.31059

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.29097    0.19730   1.475   0.1461
Pairs                      0.28550    0.02876   9.928 8.83e-14 ***
SizeSmall                 -0.51064    0.22731  -2.246   0.0288 *
StatusResident             0.56810    0.20627   2.754   0.0080 **
SizeSmall:StatusResident -0.22916    0.27567  -0.831   0.4095
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4854 on 54 degrees of freedom
Multiple R-squared:  0.7395,    Adjusted R-squared:  0.7202
F-statistic: 38.32 on 4 and 54 DF,  p-value: 3.545e-15
```

Figure 15: Summary of the model_logclean

The output indicates that both Size and Status are significantly related to the log-transformed extinction time ($p < 0.001$). However, the model assumes that the slopes for all four combinations of Size and Status are equal. To test this assumption, a new model can be fitted that includes an interaction term between Size and Status (Figure 16):

```
{r}
model_interaction <- lm(log_time ~ Pairs + Size*Status, data = birds_clean)
summary(model_interaction)
```

```
Call:
lm(formula = log_time ~ Pairs + Size * Status, data = birds_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11837 -0.24665 -0.06583  0.24818  1.31059

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.29097    0.19730   1.475   0.1461
Pairs                      0.28550    0.02876   9.928 8.83e-14 ***
SizeSmall                 -0.51064    0.22731  -2.246   0.0288 *
StatusResident             0.56810    0.20627   2.754   0.0080 **
SizeSmall:StatusResident -0.22916    0.27567  -0.831   0.4095
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4854 on 54 degrees of freedom
Multiple R-squared:  0.7395,    Adjusted R-squared:  0.7202
F-statistic: 38.32 on 4 and 54 DF,  p-value: 3.545e-15
```

Figure 16: Summary of the model_interaction

The "*" between Size and Status specifies the inclusion of an interaction term in the model. The model will estimate separate slopes for each combination of Size and Status, rather than assuming they are all equal. Comparing the fit of this model to the previous model **(model_logclean)** will determine if the new model fits significantly better, indicating that the slopes for at least one combination of Size and Status are not equal (Figure 17).

```r
model_logclean_int <- lm(log_time ~ Pairs + Size + Status + Size:Status, data =
birds_clean)
summary(model_logclean_int)
```

```
Call:
lm(formula = log_time ~ Pairs + Size + Status + Size:Status,
    data = birds_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11837 -0.24665 -0.06583  0.24818  1.31059

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.29097    0.19730   1.475   0.1461
Pairs                     0.28550    0.02876   9.928 8.83e-14 ***
SizeSmall                -0.51064    0.22731  -2.246   0.0288 *
StatusResident            0.56810    0.20627   2.754   0.0080 **
SizeSmall:StatusResident -0.22916    0.27567  -0.831   0.4095
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4854 on 54 degrees of freedom
Multiple R-squared:  0.7395,    Adjusted R-squared:  0.7202
F-statistic: 38.32 on 4 and 54 DF,  p-value: 3.545e-15
```

Figure 17: Summary of the model_logclean_int

The output above shows the results of fitting a new model that includes an interaction term between Size and Status. The model suggests that the effect of Size on log-transformed extinction time depends on the migratory Status of the bird species. The coefficient for the interaction term "SizeSmall:StatusResident" is negative and not statistically significant (p = 0.4095), which suggests that there is no significant difference in the slopes of the relationship between log-transformed extinction time and Size for resident and migratory birds. However, both Size and Status still have significant main effects on the log-transformed extinction time (Size: p = 0.0288; Status: p = 0.0080).

The new model has a slightly lower Adjusted R-squared value (0.7202) compared to the previous model (0.7217). This suggests that the interaction term does not improve the model fit significantly. Therefore, the reduced model based on the findings would remain as the first complete model. This model suggests that the log-transformed extinction time is significantly influenced by the number of breeding pairs, as well as the bird species' size and migratory status. Based on the findings from the previous items, we can make a reduced model to simplify our analysis. The **model_logclean ←- lm(log_time Pairs + Size + Status, data = birds_clean)** (Figure 18) is a suitable candidate for a reduced model. This model includes only the variables that were found to be significantly related to the log-transformed extinction time: Pairs, Size, and Status.

```r
model_logclean <- lm(log_time ~ Pairs + Size + Status, data = birds_clean)
```

Figure 18: The final chosen model: model_logclean

Although adding an interaction term between Size and Status did improve the model fit, the change in performance was not substantial. Additionally, by keeping the original model without the interaction term, the risk of over-fitting the data is reduced. Therefore, it is reasonable to choose to keep the **model_logclean ¡- lm(log_time Pairs + Size + Status, data = birds_clean)** as a reduced model. This model performs well, as indicated by its high multiple R-squared value (0.7361), which suggests that the model explains a substantial proportion of the variance in the log-transformed extinction time. The p-values for all the coefficients in the model are significant, indicating that the variables are all important predictors of extinction time (Figure 19).

```r
{r}
model_logclean <- lm(log_time ~ Pairs + Size + Status, data = birds_clean)

summary(model_logclean)
```

```
Call:
lm(formula = log_time ~ Pairs + Size + Status, data = birds_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-1.07187 -0.28701 -0.05488  0.26100  1.22410

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.39042    0.15645   2.496  0.01560 *
Pairs           0.28236    0.02843   9.932 7.11e-14 ***
SizeSmall      -0.66713    0.12703  -5.252 2.51e-06 ***
StatusResident  0.43983    0.13650   3.222  0.00214 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4841 on 55 degrees of freedom
Multiple R-squared:  0.7361,    Adjusted R-squared:  0.7217
F-statistic: 51.15 on 3 and 55 DF,  p-value: 6.339e-16
```

Figure 19: Summary of the model_logclean

7 - Conclusion

In this project, a data-set containing measurements of breeding pairs of land-bird species collected from 16 islands around Britain over several decades was analysed. The aim was to investigate the factors influencing extinction in bird populations, particularly the effect of nesting pairs, size, and migratory status.

The preceding analysis revealed that nesting pairs, size, and status have a significant effect on the extinction times of bird species. Species with larger numbers of nesting pairs, resident rather than migratory, and large in size tend to remain longer before becoming extinct.

To conduct the analysis, several statistical tests were performed, including residual plots and model fitting. The data-set was also examined for possible transformations and outliers by plotting Q-Q plots. It was discovered that a logarithmic transformation of the extinction times provided the best fit for the data-set compared to squared and inverse times.

This investigation also identified a few species with unusually small or large extinction times compared to other species with similar values of the explanatory variables. These were distinguished as outliers and were not include in the final model since they could skew the model's representation of the majority of bird species.

Using the data from the analysis, a model was developed that could be used to predict a bird species' extinction period based on the "Number of pairs," "Size," and "Status" parameters in relation to the extinction time of analyzed birds (Figure 20).

```r
{r}
# Suppose we want to predict Time for a bird with the following characteristics:
new_bird <- data.frame(Pairs = 1.2, Size = "Large", Status = "Resident")

# Make the prediction using the model
prediction <- predict(model_logclean, newdata = new_bird)

# View the predicted Time value
prediction
```

```
      1
1.16908
```

Figure 20: Example of a prediction generated by the new model

For the future research, the selection of analysed data could be expanded to increase accuracy of prediction of the model.

Overall, this analysis contributes to the understanding of factors affecting extinction in bird populations. The findings of this project could be useful in developing strategies to protect disappearing and endangered bird species and their habitats.

8 - Bibliography

Carlsen, L. M. (2023). https://learnit.itu.dk/mod/folder/view.php?id=171649. Retrieved from Project Factors Affecting Extinction.

Dr. Iain Pardoe, D. L. (n.d.). Stat 501. Retrieved from Regression Methods: https://online.stat.psu.edu/stat501/lesson/4/4.2

Stuart L. Pimm, H. L. (1988). On the Risk of Extinction. The American Naturalist, Number 6.

Virginia Humanities, B. (2020). 1980S ENVIRONMENTALISM AND HOW THE REAGAN-ERA SHAPED THE NATURAL WORLD [Recorded by B. B. Ayers]. Virginia, United States of America.