IT UNIVERSITY OF CPH

Factors Affecting Extinction

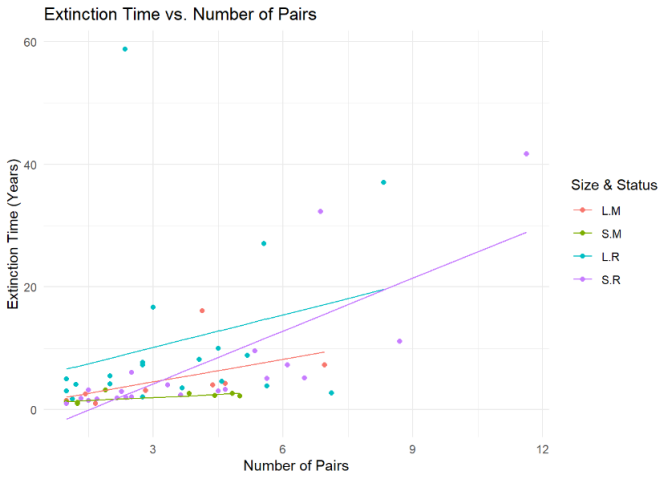
March 16, 2023

Rinalds Lipenitis Jakub Mráz Adam Rosenørn Costel Gutu Richard Kentoš

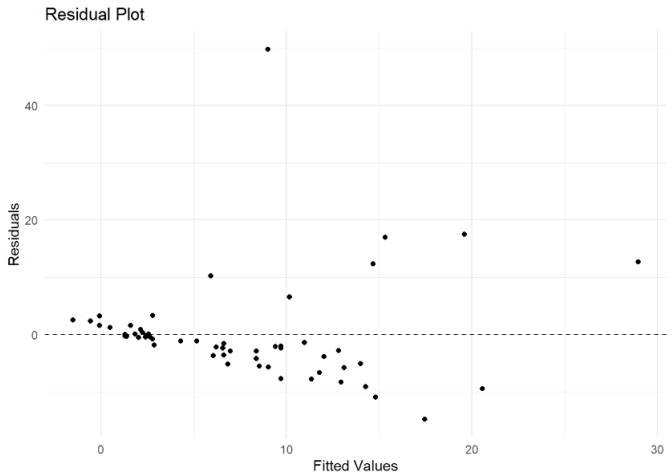
March 2023

1 Exercise 1

```
##
## Call:
## lm(formula = Time ~ Pairs * Size * Status, data = bird_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.769  -3.644  -0.652   1.128  49.831
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.7775     6.6831   0.116    0.908
## Pairs                 1.2397     1.7232   0.719    0.475
## SizeS                 0.1759     8.3168   0.021    0.983
## StatusR               4.0489     7.8704   0.514    0.609
## Pairs:SizeS          -0.9058     2.4206  -0.374    0.710
## Pairs:StatusR         0.5317     2.0037   0.265    0.792
## SizeS:StatusR        -9.4013     9.8736  -0.952    0.345
## Pairs:SizeS:StatusR   2.0047     2.7218   0.737    0.465
##
## Residual standard error: 9.237 on 54 degrees of freedom
## Multiple R-squared:  0.3398, Adjusted R-squared:  0.2542
## F-statistic:  3.97 on 7 and 54 DF,  p-value: 0.001446
```



2 Exercise 2



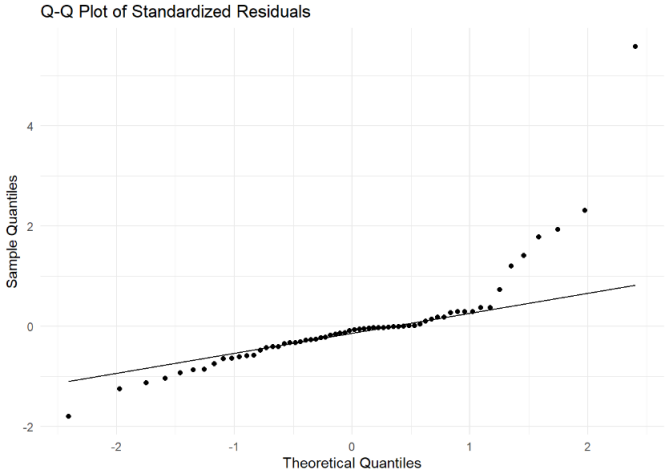
When examining the residuals, we look for:

Homoscedasticity: The residuals should have constant variance across the range of fitted values. If the spread of residuals seems to change across the fitted values, a transformation may be needed. The variance of the residuals becomes bigger and bigger as the value being fitted (number of pairs + extinction time) goes up.

Independence: There should be no patterns or trends in the residuals. If you notice any patterns or trends, it might indicate that a variable is missing from the model, or a transformation is needed. There appears to be a slight downward slope pattern.

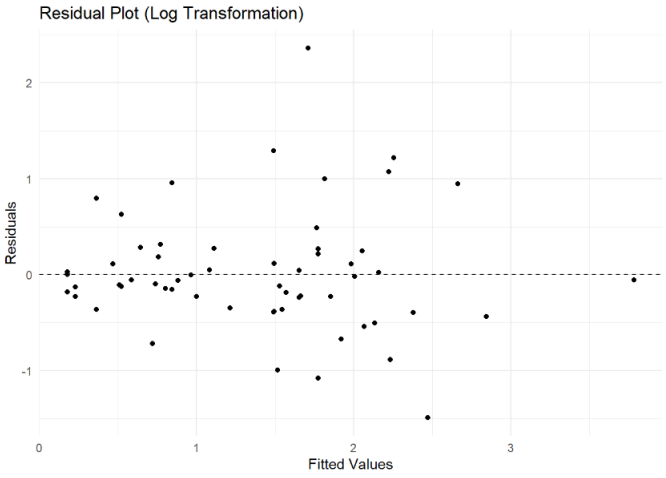
Outliers: Look for points that stand out from the rest, as they may be outliers. Investigate these points further to determine if they are errors or if they represent genuine observations. You may consider removing outliers if they are affecting the model's performance.
Biggest outlier:
Species: Raven
Time: 58.82
Pairs: 2.35
Size: L
Status: R

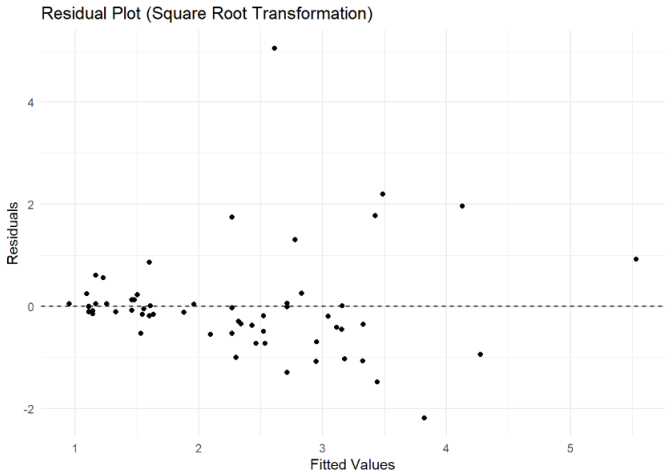
Normality: Ideally, the residuals should be approximately normally distributed. If the residuals exhibit a non-normal distribution, you may need to consider a transformation of the dependent variable or use a different modeling approach.

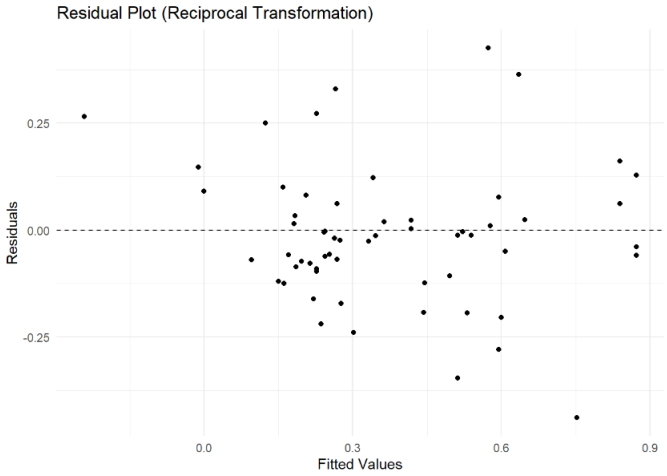


The residuals appear to be normally distributed except for the lower and upper ends, where they become more skewed, this suggests that a data transformation may be necessarry.

3 Exercise 3







Immediately, we can tell that the square transformation does not improve the model. To compare the log and reciprocal transformation, we will use Q-Q plots again.

Based on the Q-Q plots, the log transformation seems to be the best fit for the majority of the fitted data points.

4 Exercise 4

Results can be significantly impacted by the outlier in each end of the scale. Outliers can also affect the measures of the central tendency, more specifically mean, median and mode. Moreover, they can also affect standard deviation and range - the spread of the data.

We believe that is is vital for us to keep the outliers in the dataset for various reasons:

1. Outliers provide important information about the data distribution. To be more specific, they can indicate the presence of extreme or unusual values, which might be beneficial for understanding the range and variability of our data. For instance, the outlier for the Peregrine species in this dataset indicates that it has a much lower average time until it becomes extinct than the other species. This could also be because of the specific environmental or ecological factors.

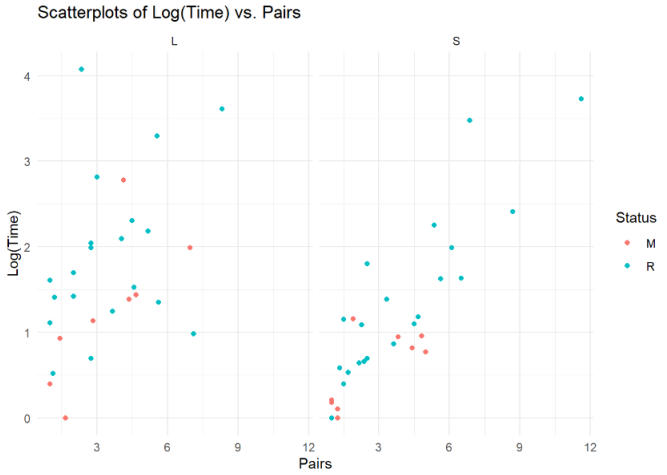
2. Preventing loss of information and biased analysis. Outliers may be representing rare but important occurrences that we should not leave out. Removing them might result in a distorted view of the true data distribution and lead to incorrect conclusions. For instance, if we decided to remove the outlier for the Raven species in this dataset, we would miss the fact that it has much longer average extinction time than the other species.

3. Allowing more robust statistical analysis. Methods such as robust regression or non-parametric tests are able to handle outliers and provide more accurate results.

We therefore think that is is inevitable to keep outliers in our dataset and carefully consider how they impact the results. Rather than removing them, it may be better to investigate why they exist in the first place and how they may affect the analysis.

5 Exercise 5

Firstly we need to understand the motivation behind transforming a variable. Transformation of a variable is often done to linearize a relationship between two variables which is not linear in its original form. However, in this question we are specifically asked to assess whether there are linear relationships between log('time') and 'pairs' in all possible combinations of 'size' and 'migratory status'.



The relationship between log(Time) and Pairs does appear to be a positive linear relationship, thus indicating that a transformation of Pairs is not needed.

The relationship also appears to be more linear than with log-transformed Pairs, as shown below, where a slight curve pattern emerges.

6 Exercise 6

6.1

```
##
## Call:
## lm(formula = log(Time) ~ Pairs * Size * Status, data = bird_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4887 -0.3577 -0.1033  0.2065  2.3643
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.1943     0.4716   0.412   0.6819
## Pairs                 0.3141     0.1216   2.583   0.0125 *
## SizeS                -0.2207     0.5869  -0.376   0.7084
## StatusR               1.1412     0.5554   2.055   0.0448 *
## Pairs:SizeS          -0.1089     0.1708  -0.637   0.5266
## Pairs:StatusR        -0.1546     0.1414  -1.093   0.2790
## SizeS:StatusR        -1.0754     0.6967  -1.544   0.1285
## Pairs:SizeS:StatusR   0.2717     0.1921   1.415   0.1629
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6518 on 54 degrees of freedom
## Multiple R-squared:  0.6267, Adjusted R-squared:  0.5783
## F-statistic: 12.95 on 7 and 54 DF,  p-value: 1.219e-09
```

After running this code, you will see the coefficients for each term in the model, including the interactions. To determine whether the slopes are equal for all combinations of "size" and "migratory status," examine the coefficients

for the interaction terms Pairs:Size, Pairs:Status, and Pairs:Size:Status. If the interaction terms are not statistically significant (i.e., their p-values are larger than a chosen significance level, typically 0.05), it suggests that the slopes are not significantly different between the four combinations of "size" and "migratory status."

Pairs:SizeS: The coefficient is -0.1089 with a p-value of 0.5266, which is not statistically significant at a 0.05 significance level.

Pairs:StatusR: The coefficient is -0.1546 with a p-value of 0.2790, which is not statistically significant at a 0.05 significance level.

Pairs:SizeS:StatusR: The coefficient is 0.2717 with a p-value of 0.1629, which is not statistically significant at a 0.05 significance level.

Since none of the interaction terms are statistically significant, there is not enough evidence to conclude that the slopes for all four combinations of "size" and "migratory status" are different. This suggests that the relationship between "Pairs" and log("Time") may not be significantly different among the four groups based on "size" and "migratory status". However, it is essential to note that a lack of statistical significance does not necessarily mean the slopes are equal; it indicates that there isn't enough evidence to reject the null hypothesis that the slopes are equal.

6.2

Nested models are a series of models where each model is a subset of the previous one. This approach helps in assessing the contribution of variables and their interactions to the overall model fit. To create nested models, start with the simplest model and gradually add variables and interaction terms to evaluate their contributions.

```
## Analysis of Variance Table
##
## Model 1: log(Time) ~ Pairs
## Model 2: log(Time) ~ Pairs + Size
## Model 3: log(Time) ~ Pairs + Size + Status
## Model 4: log(Time) ~ Pairs * Size + Status
## Model 5: log(Time) ~ Pairs * Size * Status
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     60 34.795
## 2     59 27.924  1    6.8710 16.1719 0.000181 ***
## 3     58 24.682  1    3.2424  7.6315 0.007825 **
## 4     57 23.976  1    0.7056  1.6608 0.202995
## 5     54 22.943  3    1.0332  0.8106 0.493550
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When we move from Model 1 to Model 2 (adding the Size variable), the p-value is 0.000181 (significant at the 0.05 level). This indicates that adding the Size variable significantly improves the model fit.

When we move from Model 2 to Model 3 (adding the Status variable), the p-value is 0.007825 (significant at the 0.05 level). This suggests that adding the

Status variable significantly improves the model fit.

When we move from Model 3 to Model 4 (adding the Pairs * Size interaction term), the p-value is 0.202995 (not significant at the 0.05 level). This suggests that adding the Pairs * Size interaction term does not significantly improve the model fit.

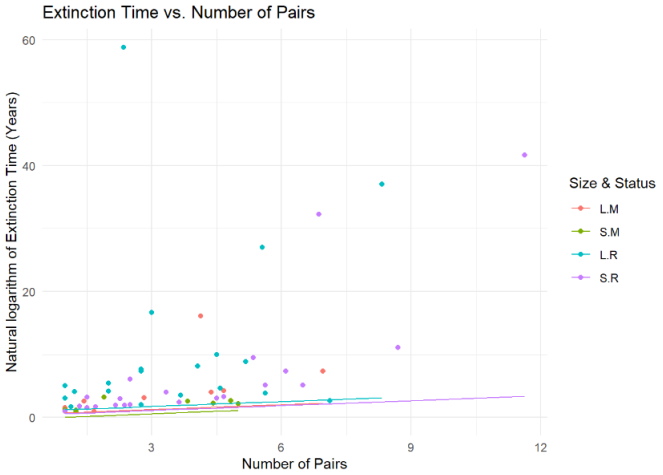
When we move from Model 4 to Model 5 (adding the Pairs * Size * Status interaction term), the p-value is 0.493550 (not significant at the 0.05 level). This indicates that adding the Pairs * Size * Status interaction term does not significantly improve the model fit.

Based on the ANOVA results, it seems that Model 3 (Time Pairs + Size + Status) is the most appropriate model for this dataset, as adding interaction terms does not significantly improve the model fit.

7 Exercise 7

Based on the findings from previous items, we can create a reduced model using just the Pairs, Size, and Status variables, without interaction terms.

```
##
## Call:
## lm(formula = log(Time) ~ Pairs + Size + Status, data = bird_data)
##
## 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 ***
## SizeS       -0.65237    0.16665  -3.915 0.000241 ***
## StatusR      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
```



8 Exercise 8

After analyzing the data and accounting for the number of nesting pairs, we can conclude that the number of nesting pairs has a significant impact on the time it takes for a species to become extinct. Larger numbers of nesting pairs tend to result in longer times before extinction.

However, when considering the size and migratory status of species, their individual effects on extinction time are less clear. We could not establish a strong relationship between these factors and the time to extinction, indicating that they may not be as influential as the number of nesting pairs.

It is also worth noting that there were a few outliers with unusually large extinction times compared to other species with similar explanatory variable values. These outliers should be further investigated to understand the underlying factors contributing to their atypical extinction times.

Biggest outlier:
- Species: Raven
- Time: 58.82
- Pairs: 2.35
- Size: L
- Status: R

Based on the reduced model, the "theoretical" regression formula for the logarithm of extinction time can be constructed using the coefficients from the summary. The formula would look like this:

$$ln(Time) = 0.43056 + 0.26509 * Pairs - 0.65237 * SizeS + 0.50406 * StatusR$$

Here,

ln(Time) represents the natural logarithm of the extinction time (note that log() in R defaults to ln);

Pairs is the number of nesting pairs (as a continuous variable);

SizeS is a binary variable indicating the bird size, with 1 for small-sized birds and 0 for large-sized birds;

StatusR is a binary variable indicating the migratory status, with 1 for resident birds and 0 for migratory birds.

The coefficients in the formula represent the effects of each variable on the logarithm of the extinction time while keeping the other variables constant.

For example, an increase in the number of nesting pairs by one unit is associated with an increase in the logarithm of extinction time by 0.26509 units, holding size and migratory status constant. Similarly, small-sized birds have a 0.65237 units lower logarithm of extinction time compared to large-sized birds, holding the number of nesting pairs and migratory status constant.

This formula can be used as a conclusion to describe the relationship between the extinction time and the variables of interest (number of nesting pairs, size, and migratory status) based on the reduced model.