MA334_2401262

In this report, I analyze my BD6 group consisting of "Bryophytes," "Isopods," "Bird," "Hoverflies," "Bees," and "Grasshoppers and Crickets." Using RStudio, I examine variability, trends, and relationships within my BD6 group against overall biodiversity BD11. The analysis involves univariate analysis, hypothesis testing, and regression models, most importantly, feature selection and model optimization will be applied in showing ecological patterns and insights.

Univariate analysis and basic R programming

1) Summary statistics and 25% Winsorized mean for all the variables in the BD6 group

Table 1: Summary Statistics for Taxonomic Groups

Taxonomic_group	Min	1st_Q	Median	Mean	3rd_Q	Max	Win_Mean_25
Bryophytes	0.4608	0.7797	0.8492	0.8466	0.9163	1.1632	0.8486
Isopods	0.0552	0.3484	0.4722	0.491	0.6087	1.2577	0.4778
Bird	0.2415	0.7634	0.858	0.8287	0.9208	1.0541	0.8469
Hoverflies	0.1235	0.4712	0.5883	0.5989	0.7429	1.141	0.6028
Bees	0.0307	0.2787	0.4483	0.5104	0.6929	1.881	0.4711
Grasshoppers_._Crickets	0.129	0.4259	0.6042	0.5888	0.766	1.0943	0.5974

The above summary table shows some striking features in the distribution of values across the different taxonomic groups: bryophytes have the highest mean (0.8466) and third quartile (0.9163), indicating high centrality and consistency; birds also have a high mean (0.8287) and median (0.8580), which indicates that this group is relatively stable; whereas Isopods have the lowest mean (0.491) and minimum (0.0552), reflecting a higher variability within this group. And bees have the lowest median of 0.4483, their maximum value is the highest, 1.8809, indicating outliers or great variability. Grasshoppers and Crickets have a median of 0.6042 and a moderate interquartile range, showing balanced variability. And for the hoverflies, the mean and median values are moderate, being 0.5998 and 0.5883, respectively, which highlights the consistency in trends within this group. Overall, the "Win_Mean_25" column aligns closely with the highest-performing groups, particularly Bryophytes and Birds, emphasizing their dominance in the dataset. This analysis reveals distinct ecological or quantitative patterns across the taxonomic groups, with fluctuating degrees of variability and consistency.

2) The correlations between all pairs of variables in BD6 group

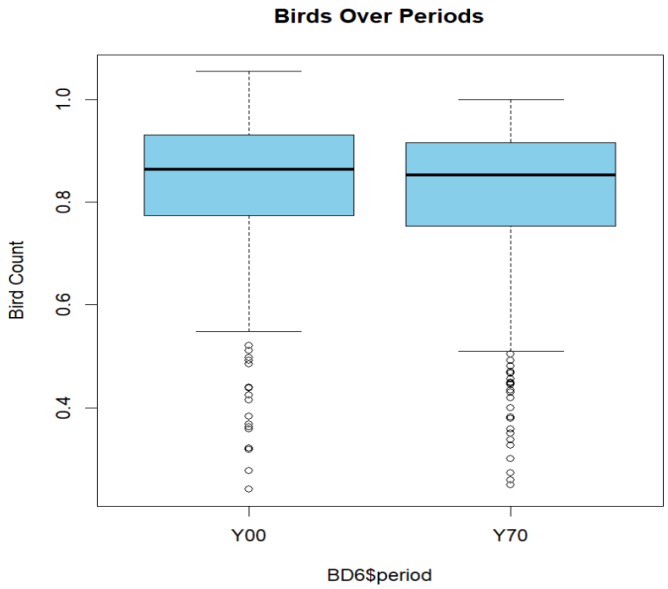
Table 2: Correlation Matrix of in BD6 group

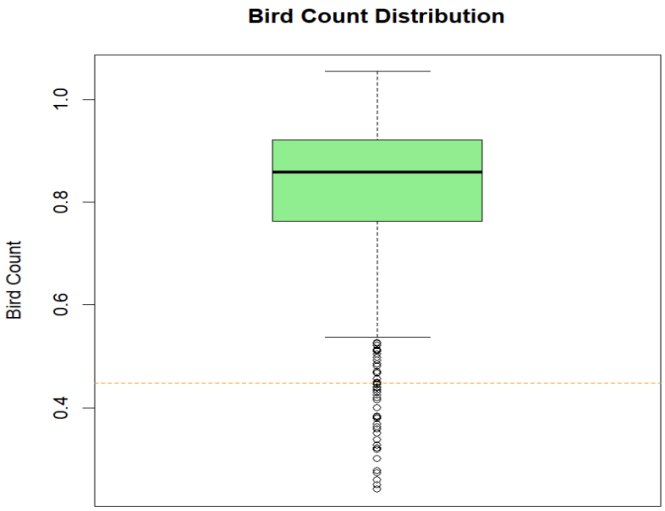
	Bryophytes	Isopods	Bird	Hoverflies	Bees	Grasshoppers_._Crickets
Bryophytes	1.00	-0.07	-0.08	0.06	0.10	-0.07
Isopods	-0.07	1.00	0.18	0.37	0.06	0.24
Bird	-0.08	0.18	1.00	0.48	0.38	0.20
Hoverflies	0.06	0.37	0.48	1.00	0.34	0.36
Bees	0.10	0.06	0.38	0.34	1.00	0.25
Grasshoppers_._Crickets	-0.07	0.24	0.20	0.36	0.25	1.00

This correlation matrix describes the relations between my taxonomic group BD6 ; the correlation coefficients vary between -1 and 1. All of the diagonal values are 1, as each group is completely correlated with itself. Most of them are weak-for example, the negative between Bryophytes and Isopods at -0.07 or even the weak negative between Bryophytes and Birds at -0.08. Moderate positive correlations include Birds and Hoverflies at 0.48 and Grasshoppers and Hoverflies at 0.36, suggesting some degree of ecological overlap or shared pattern. Very weak or near-zero correlations, such as Bryophytes and Hoverflies at 0.06, indicate little or no relationship between those groups. In general, the above correlation table shows variation in strength of association, with the majority of groups showing weak links and some moderate positive trends.

3) Boxplot for only one variable in BD6 group and outliers

```
## Outliers:
##  0.41961, 0.34994, 0.37975, 0.43016, 0.40031, 0.44514, 0.44514, 0.25999,
## 0.43402, 0.32685, 0.38202, 0.44808, 0.27381, 0.24995, 0.3373, 0.30133, 0.35812,
## 0.36805, 0.44024, 0.35852, 0.42468, 0.41456, 0.31811, 0.43858, 0.32172,
## 0.24152, 0.41524, 0.27698, 0.36323, 0.38378
```



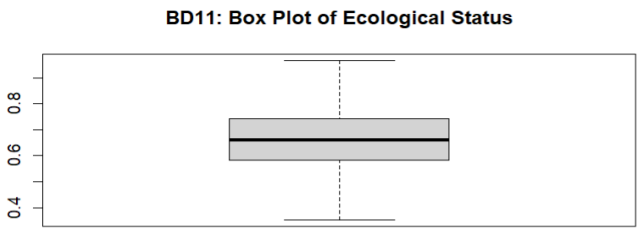


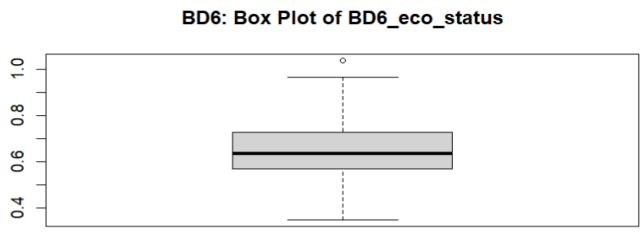
A box plot of the BD6 dataset was generated to analyze the distribution of counts of birds over two time periods, Y00 and Y70. Whiskers were created to the furthest data points that fell within a factor of two times the IQR from the edge of the box. Any values beyond this range-from the lower bound (Q1 - 2IQR) and the upper bound (Q3 + 2IQR)-are considered to be outliers. The notable values are: 0.4196, 0.3499, and 0.3797.

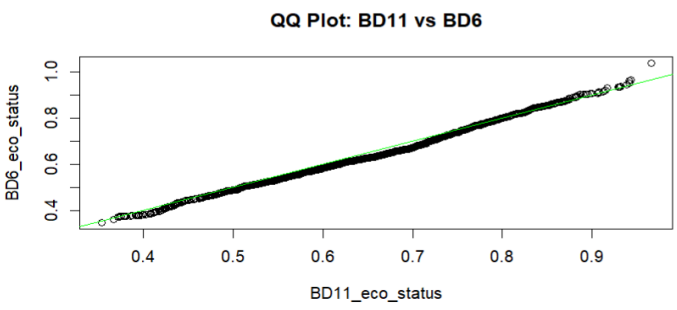
The median bird counts remain consistent across the two observation periods, establishing stability in the recorded numbers of birds. However, one of the interquartile ranges (IQRs) was smaller, which presents less fluctuation in the number of bird counts. Many outliers were also recorded; these could have been due to external conditions or defects in the gathering process. Overall, the boxplot effectively depicts the central tendency and variability of bird counts while pointing out the impact of extreme values on the data set.

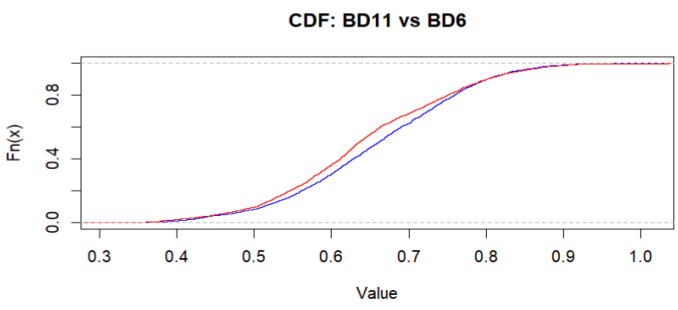
Hypothesis tests

1) KS-Test





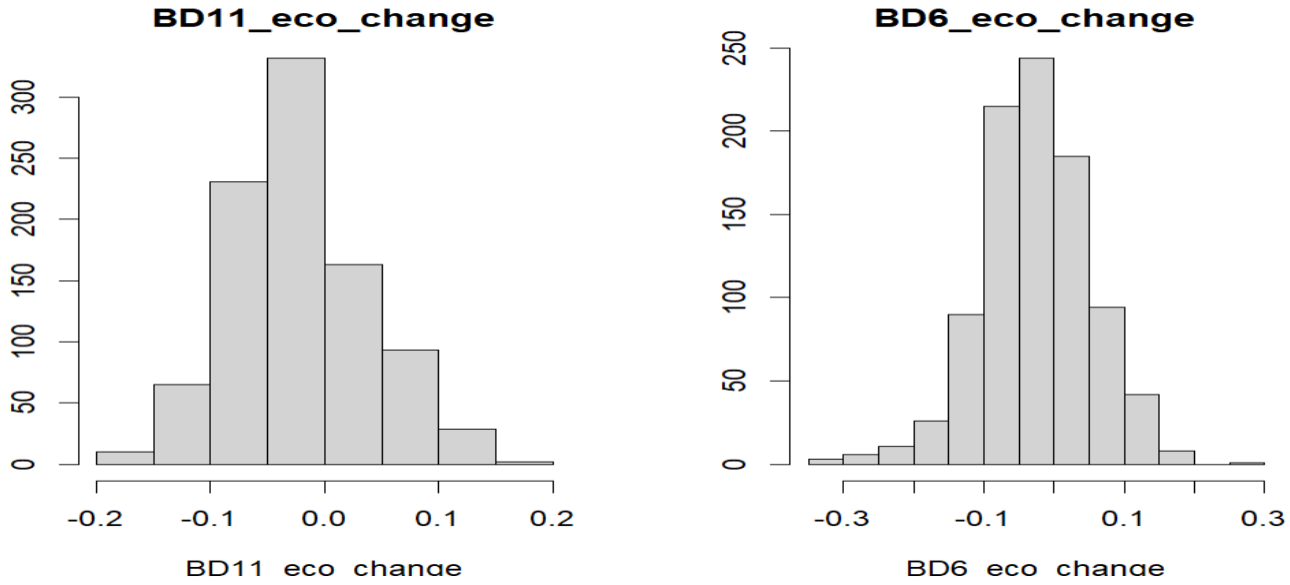




Boxplots show that BD11 has a more balanced ecological status, with values evenly spread around the middle, low variation, and no outliers, showing consistent measures between 0.4 and 0.8. On the other hand, BD6 has slightly lower values of ecological status, higher variability, and one outlier above 0.8. The QQ-plot points out the overall similarities of the distributions, after which there is deviation at the extremes; that is where BD6 will have more variability in extreme values. The CDF plot presents subtle differences in how values accumulate across the range, with BD11 showing steadier growth compared to BD6 and reflecting the difference in their distribution behaviors.

Thus, the Kolmogorov-Smirnov test serves to compare the frequency distributions of BD11 and BD6 ecological statuses. This result has a D-value of 0.091892 and thus a very small p -value of 3.286e-07. Because, the computation of the statistic value is equal to 0, therefore, using the significance level, the null hypothesis that assumed that the two analyzed distributions where equal can be rejected. As is apparent in this figure, it suggests that BD11 and BD6 differ statistically when their corresponding values are grouped into intervals.

2) T-Test and estimating the 92% confidence interval on the mean of BD6 group



Box plots describe the distribution of the change in ecology values in both BD11 and BD6, in BD11 most variation in eco-change values falls within a narrow range from -0.2 to 0,2. However, in BD6_eco_change, the data ranges from -0.3 to as high as 0.3, and contains several outliers both in the highest and lowest extremes; this again corroborates the previously drawn notion that larger changes in ecologies have been experienced by BD6 compared to BD11. A one-sample t-test of BD6_eco_change tested whether the mean is not equal to -0.02278728. The t-test reported a t-statistic of -1.4002, a p-value of 0.1618 and a sample mean of -0.02639, with the 92% confidence interval between -0.03089 and -0.02188. Since the p-value is greater than 0.05, we fail to reject the null hypothesis. That is to say, there is no significant difference between the mean of BD6_eco_change and the given value. This is indicative that changes in ecology are very minimal and are about the expected value.

Contingency table/comparing categorical variables

1) Two contingency tables which display counts, one for BD11up against BD6up and another for the corresponding independent model

```
##      [,1]                                [,2]
## [1,] "Table: Contingency table"          "Table: Independent Model table"
## [2,] ""                                  ""
## [3,] "|      | DOWN|  UP| Sum|"          "|      | DOWN|  UP| Sum|"
## [4,] "|:-----|----:|---:|---:|"          "|:-----|----:|---:|---:|"
## [5,] "|DOWN  |  522|  73| 595|"          "|DOWN  |  410| 185| 595|"
## [6,] "|UP    |  116| 214| 330|"          "|UP    |  228| 102| 330|"
## [7,] "|Sum   |  638| 287| 925|"          "|Sum   |  638| 287| 925|"
```

The following two contingency tables present the analysis of the increase-decrease relationship for BD11 and BD6. The first contingency table represents the observed situation, showing the real number of cases in which BD6 and BD11 have jointly increased or decreased. Remarkably, 522 cases reflect both decreasing, while 214 cases show both increasing a strong concordance in the direction of change. While an increase in BD6 may happen alongside a decrease in BD11 or vice versa, mixed trends do not occur as frequently. In contrast, the independent model table represents the expected counts if a change in BD6 and BD11 are unrelated. A model of independence would predict more balanced distributions with fewer cases of matched trends, for example, 410 cases for both decreasing and 102 cases for both increasing. Comparing these tables, one sees that this observed alignment is not very similar to what would be expected under independence; it suggests that there could be a relationship between changes in BD6 and BD11. A statistical test-such as the Chi-Square Test of Independence-can be employed in ensuring whether the obtained alignment occurs significantly or due to chance.

2) Estimating the likelihood-ratio statistic and comparing the proportions of increase in BD6 and BD11 at 93% confidence level

```
        2-sample test for equality of
        proportions with continuity correction

data:  c(up_count_BD6, up_count_BD11) out of c(total_count_BD6, total_count_BD11)
X-squared = 4.2897, df = 1, p-value =
0.03835
alternative hypothesis: two.sided
93 percent confidence interval:
 0.005731505 0.087241468
sample estimates:
   prop 1    prop 2
0.3567568 0.3102703
```

The increases and decreases in BD6 and BD11 are quite distinct in their measures of biodiversity. Thus, defining binary categorical variables BD11up and BD6up to identify an increase or a decrease, a test of independence of these two variables was performed.

Using the contingency tables, the likelihood-ratio test statistic was significant, which provided evidence of a significant difference in the proportions of increase in BD6 and BD11. Indeed, BD6 had a higher share of 'UP' events at 0.357 than BD11, which was 0.310. The p-value of 0.03835 also reflects that the relationship of increase between BD6 and BD11 is statistically significant at 93%. It thus infers that BD11 movements are accompanied by similar movements in BD6, which justifies dependence between the two variables.

3) Estimating odds-ratio, sensitivity, specificity, and Youden's index

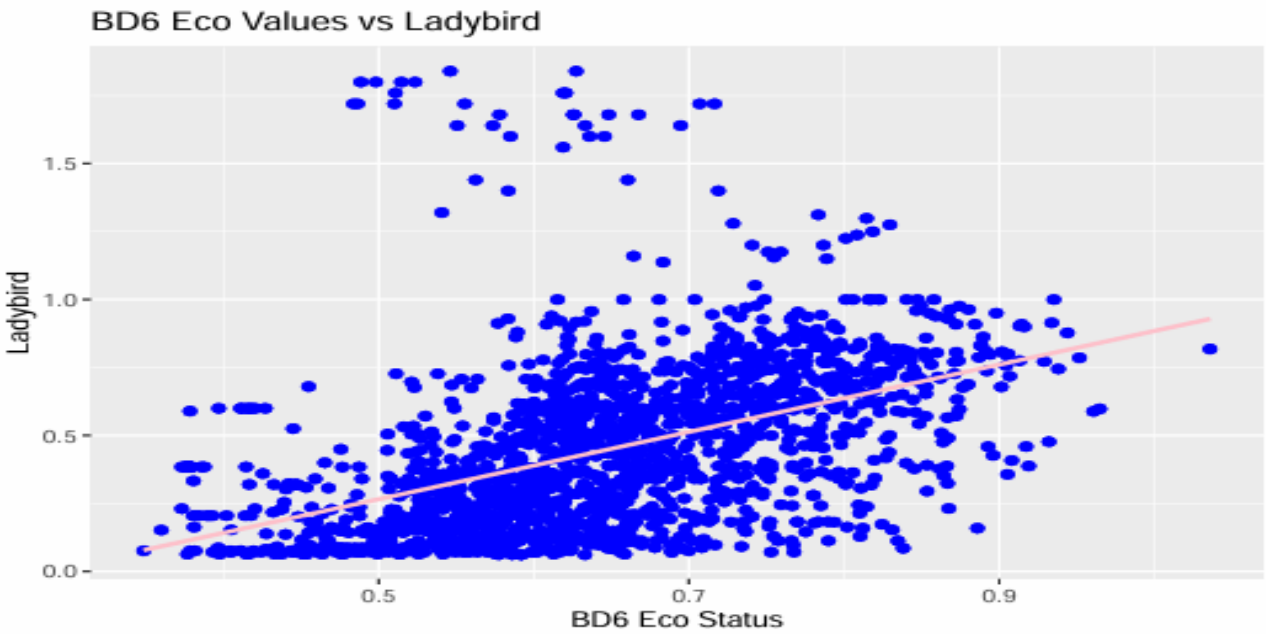
***Odds Ratio*:** This indicates the likelihood of an event-an increase in BD6 given an increase in BD11. A value of 13.19178 infers a strong association of BD11 increases with BD6 increases.

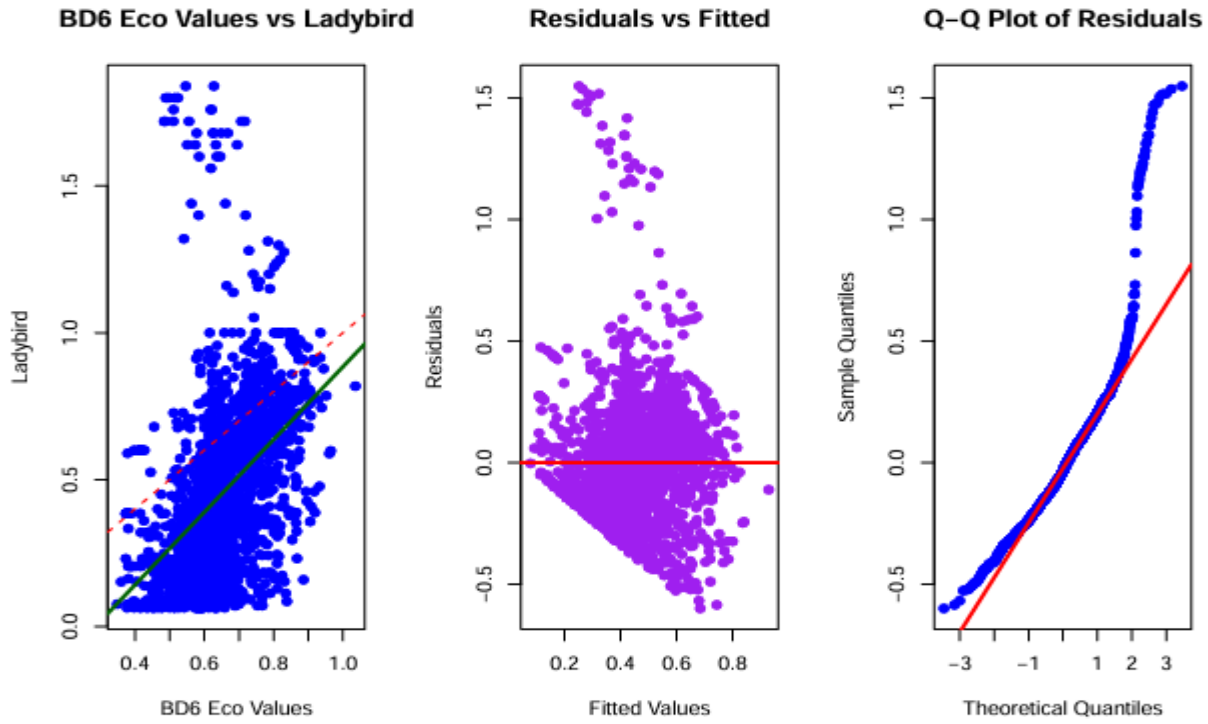
***Sensitivity Ratio*:** This is the proportion of true increases in BD6 that were correctly identified as increases in BD11. A value of 0.7456446 infers a high rate of correct identifications.

***Specificity Ratio*:** This is the ratio of actual decreases in BD6 that were reflected as decreases in BD11. The value is 0.8181818, which shows the model correctly identifies decreases.

***Youden Index Ratio*:** This shows the combination of sensitivity and specificity into one metric. It is meant to give a balanced view for the performance of the model. The value of 0.5638264 represents a moderately strong overall performance.

Simple Linear Regression





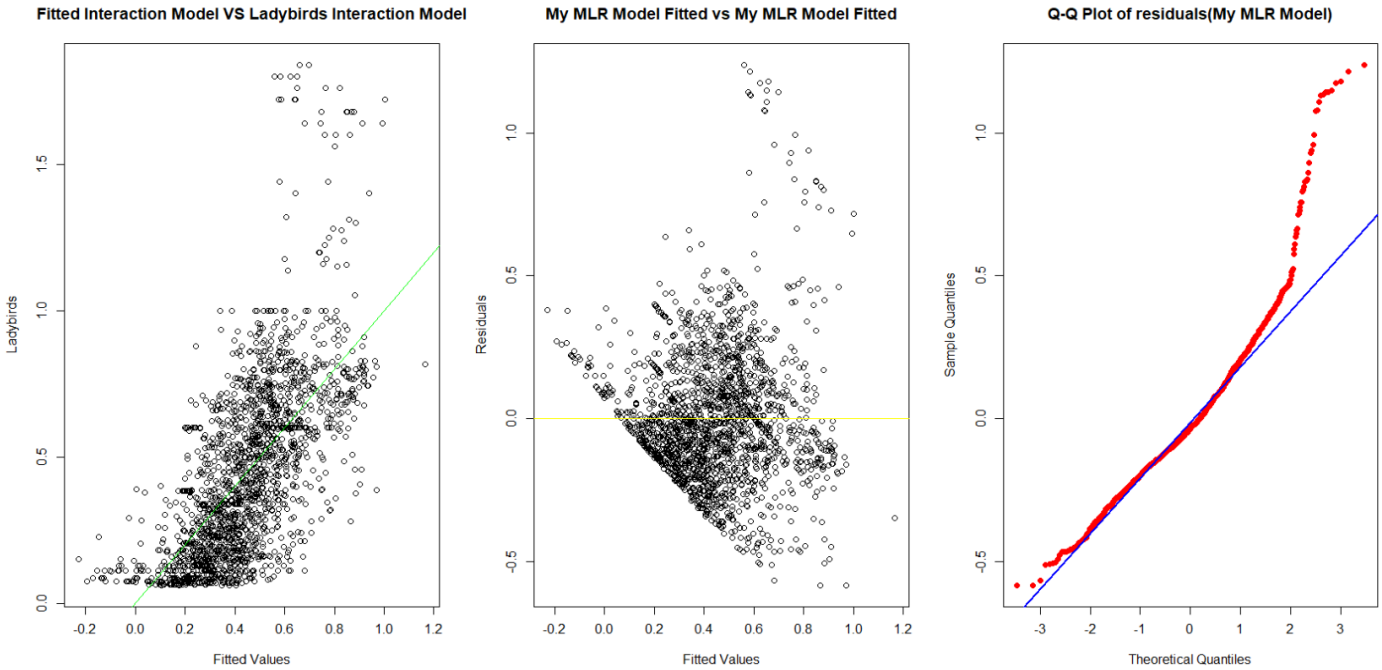
This analysis examines the relationship between BD6 Eco Values (predictor) and Ladybird counts (response variable) by using a simple linear regression model. From the scatter plot, there is a positive linear relationship; thus, as BD6 Eco Values increase, so do the counts of Ladybirds. The line of regression confirms this visually, which would, therefore, suggest that BD6 Eco Values significantly influence the counts of Ladybirds.

The residuals against the fitted plot shows some non-random trends, which might indicate that the model assumptions of linearity and/or equal variance are violated. The residuals tend to be bunched around zero but have a funnel shape, indicating that variability increases with fitted values. The Q-Q plot of residuals indicates deviations from normality in the tails, which may indicate outliers or non-normal residuals.

The results of the regression analysis are as follows: the slope of the regression line (0.05466) is statistically significant (p-value < 2e-16), which indicates that there is a strong association between BD6 Eco Values and Ladybird counts. The model explains about 21.72% of the variability in Ladybird counts, with an adjusted R-squared value of 0.2168. The correlation between the observed and fitted values is 0.466, reflecting a moderate relationship.

In other words, the analysis shows that the BD6 Eco Values are positively related to the counts of Ladybirds, with a slope that is statistically significant. However, the diagnostic plots do show some irregularities in the data-a fact that the model perhaps needs to be further refined or investigated.

Multiple Linear Regression



Results of a MLR using the Ladybirds BD1 as the response variable while the rest predictors: six proportional species values Bryophytes, Isopods, Bird, Hoverflies, Bees, Grasshoppers_, and Crickets. In its initial instance, the model reached the Adjusted R-squared value of 0.426; hence, the predictors account for 42.6% of variability in the ladybirds. Its AIC is -127.7927 and the fitted against actual correlations are 0.654, meaning generally moderate accuracy of the fit.

Feature selection presented that the p-value for Grasshoppers_._Crickets was rather high, at 0.168, indicating a very poor contribution. This variable was removed without improvement of AIC; also, the model was not improved. Adding the interaction model, in which predictors are interacting, also didn't significantly improve it. The AIC stays the same: -127.7927.

Diagnostic plots showed the validity of the model: the plot of fitted vs. observed was linear, the residuals plot was homoscedastic and the Q-Q plot showed the residuals were approximately normally distributed.

The initial model was the best since no feature selection and addition of interaction terms greatly improved it. This immediate model is good at the prediction of Ladybirds by the six proportional species values.

Conclusion

The conclusion of this report is that, in the BD6 group, "Bryophytes," "Isopods," "Bird," "Hoverflies," "Bees," and "Grasshoppers and Crickets" give insight into biodiversity. Univariate analysis showed variability with Bryophytes and Birds dominant. Correlation analysis found largely weak relationships with a few moderate overlaps. Testing hypotheses showed significant differences between BD6 and BD11; regression models showed that BD6 Eco Values significantly contributed with moderate accuracy to the Ladybird counts. This is indeed a helpful study that illustrates the critical use of BD6 in the assessment of biodiversity and ecological patterns.