MA334 AU 2023 Individual Assignment

Reg no: 2316207

Outline:

I have received the "proportional_species_richness_NAs_removed.csv" dataset. The dataset contains 7 taxonomic groups and two time periods, Y00 and Y70. It also contains locations based on the UK national grid system, including dominant land classification and northing and easting values.

I have been assigned to analyse five taxonomic groups, termed BD5. My allocated 5 taxonomic groups include Vascular plants, Macromoths, Hoverflies, Isopods, and Bees. My analysis will focus on identifying how BD5 differs from the mean of all 11 taxonomic group proportional species values named BD11, and how BD5 changes between the two time periods.

Univariate Analysis and Basic R Programming:

Summary statistics table for each of the five variables in my BD5 group, including the 20% Winsorized mean:

Taxonomic Group	Min	1st Quarter	Median	Mean	3rd Quarter	Max	Winsorized Mean (20%)
Vascular Plants	0.42	0.72	0.79	0.79	0.86	1.2	0.79
Macromoths	0.09	0.79	0.88	0.85	0.94	1.26	0.86
Hoverflies	0.12	0.57	0.7	0.68	0.81	1.15	0.69
Isopods	0.05	0.39	0.54	0.55	0.72	1.26	0.55
Bees	0.03	0.35	0.59	0.61	0.82	3.31	0.59

Figure 1

From Figure 1 it can be observed that the BD5 taxonomic groups have symmetrical distributions, as both the median and mean values are very close.

Compared to other taxonomic groups Vascular plants and Macromoths have relatively consistent values across the first and third quartiles. However, Bees show a wider range when comparing the first and third quartiles.

Additionally, the winsorized mean shares the same actual mean for Vascular plants and Isopods, and only a small difference between the actual mean for the other taxonomic groups. This indicates that there are no extreme outliers in the data, which could have significantly impacted the mean.

Correlations between all pairs of variables in BD5:

	Bees	Hoverflies	Isopods	Macromoths	Vascular Plants
Bees	1	0.36	0.05	0.47	0.17
Hoverflies	0.36	1	0.38	0.39	0.34
Isopods	0.05	0.38	1	0.07	0.34
Macromoths	0.47	0.39	0.07	1	0.13
Vascular Plants	0.17	0.34	0.34	0.13	1

Figure 1.1

In Figure 1.1, I have compared all variables within my BD5 group. The correlation table indicates that there are no negative correlations between variables. The strength of the correlation varies across different pairs. For instance, Bees and Macromoths have a relatively strong positive relationship with the strongest correlation coefficient of 0.47. In contrast, Isopods and Macromoths have a weaker positive relationship with the weakest correlation coefficient of 0.07.

Boxplot for Vascular plants:

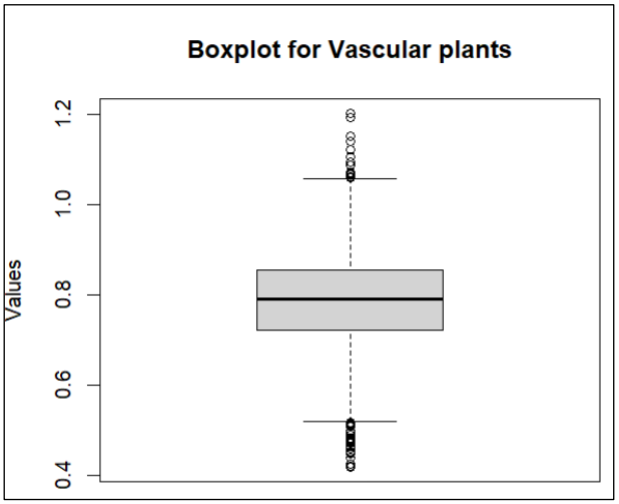

Figure 1.2

In Figure 1.2, I have created a boxplot for Vascular plants. This boxplot consists of many outliers clustered at the bottom of the whisker. This suggests that there is a concentration of values in the lower range. Yet, there are fewer outliers above the top whisker, indicating that there is a smaller number of values in the upper range. Therefore, the data on Vascular plants is skewed to the left.

Although there are more outliers falling below the lower whisker of the boxplot, the median value is situated in the centre of the interquartile range. This suggests that Vascular plants is symmetrical in the central 50% of BD5. The skewness of the Vascular plants is -0.13, indicating a minor deviation from symmetry, which results in a slight leftward skew.

Hypothesis tests:

Kolmogorov-Smirnov test:

I performed a Kolmogorov-Smirnov test to compare the distributions of eco_status_5 and ecologicalStatus using the Empirical Cumulative Distribution Function (ecdf) in R. The null hypothesis (H0) being tested is whether the distributions are the same. In the KS-test from Figure 2, the red line represents the mean for all variables in ecologicalStatus, and the green line represents the mean for the five assigned taxonomic groups in eco_status_5.

The test statistic (D) was calculated to be 0.1070 and the p-value was found to be <2.2e-16. Since the p-value is lower than the significance level of 0.05, the null hypothesis is rejected. Rejecting the null hypothesis for the KS test suggests that there is a significant difference between the eco_status_5 and ecologicalStatus distributions.

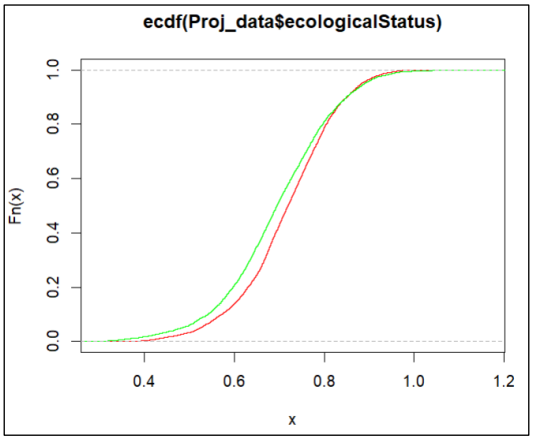

Figure 2

One-sample T-test:

I conducted a second Hypothesis test - the one-sample t-test. The Figures 2.1 and 2.2 depict the distribution of the change in BD5 and BD11 for two periods, Y00 and Y70. In the t-test, the null hypothesis (H0) is that mu=0, which tests whether the population mean μ is equal to 0. Meanwhile, the alternative hypothesis (H1) tests whether the true mean is not equal to 0.

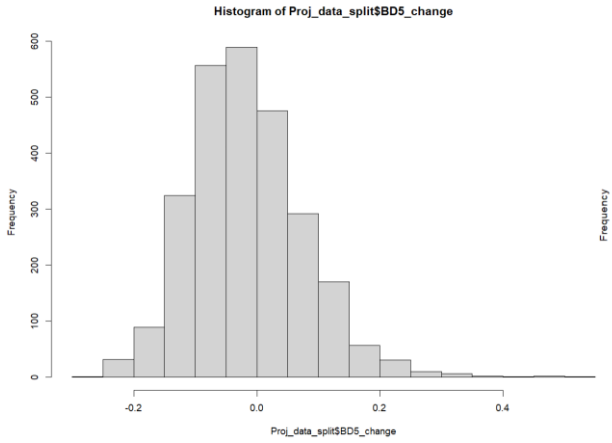

Figure 2.1

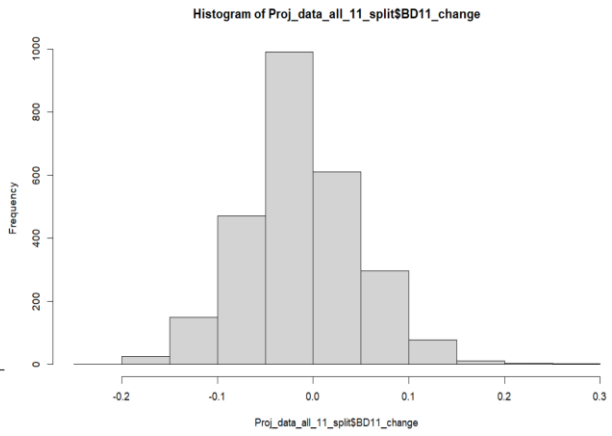

Figure 2.2

The p-values for BD5_change and BD11_change are both very low both <2.2e-16 which is less than the significance level of 0.05. This means that we reject the null hypothesis and accept the alternative hypothesis, which suggests that the population is not equal to 0. The negative sample mean for BD5_change (-0.01521558) implies that there has been a decrease in eco_status_5 from Y00 to Y70. Similarly, BD11_change also has a negative sample mean, indicating a decrease in ecologicalStatus.

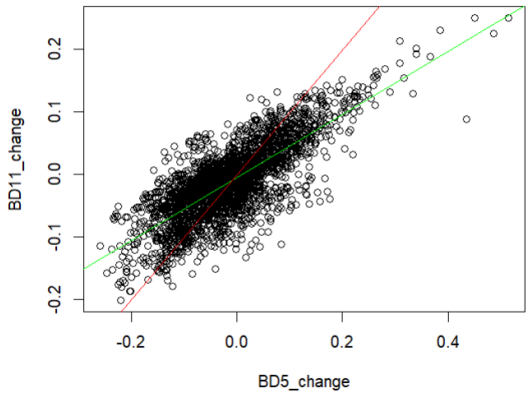

Figure 2.3

Figure 2.3 illustrates a scatter plot displaying a concentration of data points near (0,0), where the changes in both BD11_change and BD5_change are zero. The green line in the plot represents the best-fit line and indicates a positive relationship. Therefore, on average, as the changes in BD5_change increase, changes in BD11_change also increase.

The correlation coefficient between BD11_change and BD5_change is 0.7768772. This indicates that there is a strong positive correlation between the two variables. This positive correlation aligns with the positive slope observed from the scatter plot. This reinforces the fact that when BD5_change increases, BD11_change also increases.

Contingency table/comparing categorical variables:

I have created a contingency table for the mean difference in both periods for BD5 and BD11. BD5Down/BD11Down is for instances where there is a decrease, and BD5Up/BD11Up is for instance where there is an increase. I have also created the corresponding independence model for Table_up_down.

Table_up_down:

	BD5 Down	BD5 Up	Row Total
BD11 Down	1349	289	1638
BD11 Up	243	759	1002
Col Total	1592	1048	2640

Figure 3

Corresponding independence model:

	BD5 Down	BD5 Up	Row Total
BD11 Down	988	650	1638
BD11 Up	604	398	1002
Col Total	1592	1048	2640

Figure 5.1

Using the log likelihood ratio test, I compared the proportions of increase for Table_up_down and the corresponding independent model. In this test , H0 assumes that there is no difference in the proportions of increase in BD5 and BD11, while H1 assumes that there is a difference.

For Table_up_down the corresponding p-value was <2.2e-16, which is below the chosen confidence level of 5%. Therefore, rejecting the null hypothesis. BD5 and BD11 have different increase proportions, indicating they are not increasing at the same rate.

For the independence model the p-value is equal to 0.9845, which is higher than the chosen confidence level of 5%. Accepting the null hypothesis indicates no difference in the increase proportions of BD5 and BD11, implying that they would increase at the same rate.

I estimated the Odds Ratio, Sensitivity, Specificity, and Youden's Index for the contingency table Table_up_down. The Odds Ratio was estimated to be 14.57973, an odds ratio of 1 would indicate no difference in proportions of increase in BD5 and BD11. Therefore, this high odds ratio would indicate a high proportional increase between BD5 and BD11.

The Sensitivity was estimated using BD5Down/BD11Down which resulted in a value of 0.8473618 which is the true positive rate. The high rate of 0.8473618 indicates that 85% of the cases in both BD11 and BD5 are decreasing. Specificity was estimated using BD5Up/BD11Up which resulted in a value of 0.7242366, this is the true negative rate. The high rate of 0.7242366 indicates that 72% of the cases in both BD11 and BD5 are increasing.

Using the Sensitivity and Specificity, I Performed the Youden's Index, which shows the diagnosis correctness for capturing increasing and decreasing cases for both periods. This resulted in 0.5715985, this value is not particularly high in terms of capturing the increasing and decreasing cases. However, I would say that it performs reasonably well.

Simple linear regression:

I conducted a simple linear regression with BD1 (Birds) as the response variable and BD5 (eco_status_5) as the predictor variable.

Summary of Simple linear regression:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.574593	0.007077	81.19	<2e-16 ***
Proj_data$eco_status_5	0.450323	0.010037	44.87	<2e-16 ***

Figure 4

---Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Slope and intercept:

(Intercept)	Eco_status_5
0.574593	0.450323

The positive coefficient 0.45032 for the predictor variable indicates a positive association between BD5 and BD1. The significance level is 0.001, which I have identified from the significance codes. The p-value for BD5 was <2e-16 therefore, rejecting the null hypothesis, concluding that the slope is significant.

Multiple linear regression:

Using the same response variable I chose for my simple linear regression, I performed a multiple linear regression (MLR) with all five of my BD5 variables as the predictors. My Initial MLR model is as follows: lm(Bird ~ Bees + Isopods + Macromoths + Vascular_plants + Hoverflies).

I then calculated the Akaike information criterion (AIC) for my initial MLR model, this resulted in the value -11437.65, which displays the goodness-of-fit for the MLR model.

I then performed a feature selection based on the p-values and the AIC on my Initial MLR model. I removed the taxonomic group with the highest p-value, which was Isopods with a p-value equal to 0.24, exceeding the significance level of 0.05. After removing Isopods from the MLR model, the AIC was reduced to -11438.27. As the AIC has decreased this represents a better-fitting model, therefore, removing Isopods has improved the MLR model.

I tried to find a linear regression model with a lower AIC by creating an interaction term between two predictor variables in my BD5 group. The interaction term I created was between the variables Bees and Macromoths. By including an interaction term, I reduced the AIC to -11505.43, which improved the MLR model's fit.

I split the dataset into two subsets the first being train_set for the period Y70 and the second being test_set for the period Y00. I then performed the mean square error (MSE) on the test set and on the training set.

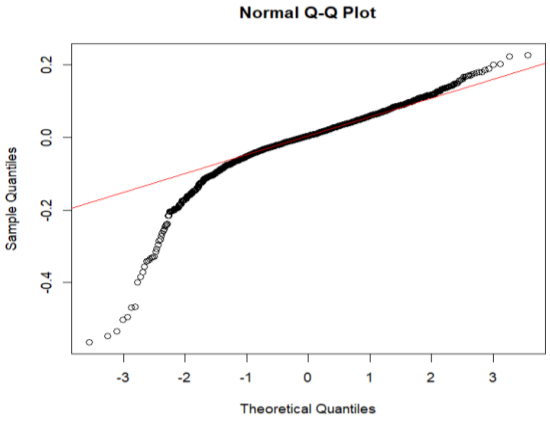

Figure 5

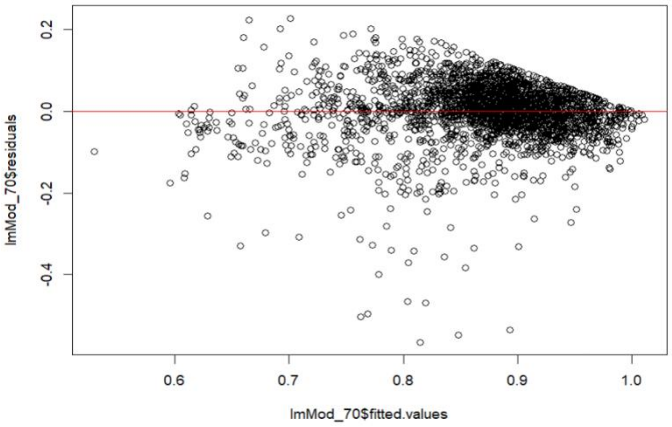

Figure 5.1

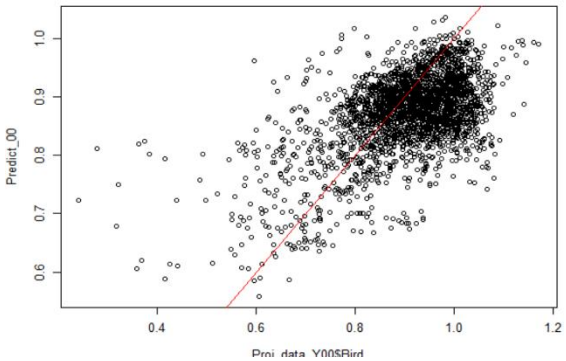

Figure 5.2

In Figure 5, we can see a Q-Q Plot that demonstrates the normality of the residuals. The plot indicates that normality is present, however, there is a deviation from normality on both tails, particularly on the left-hand tail indicating more outliers in this area.

Figure 5.1 displays the relationship between the fitted values and the residuals. The absence of any observed pattern in Figure 5.1 indicates that the residuals have a normal distribution with constant variance. Figure 5.2 compares the predicted bird results from lmMod_70 with the ecological status for Bird in Proj_data_Y00.

The MSE for the training set estimates the difference between the actual bird values and the fitted values of lmMod_70 is low and measures at 0.00516861. This indicates that the model is a good fit for the training data.

However, the MSE for the test set, which estimates the difference between the actual bird values and the lmMod_70 predicted values for Y00, is higher and measures at 0.00939536. Suggesting that the model overfits since it does not fit the test set as well as it fits the training set.

Open Analysis:

For my Open analysis, I will compare the mean values of eco_status_5 for the most frequently occurring land classification in Wales between the periods Y70 and Y00.

I performed a frequency count on the eight land classifications in Wales. I referred to the 2017 Land Classification codes to identify all eight land classifications associated with Wales.

I found that "17w2" (Rounded mountains/scarps/upper valleys, mid/S Wales) was the most frequent land classification for Wales.

Figure 6

Figure 6 displays a histogram that shows the land class frequency on the y-axis and the mean values for BD5 associated with the land class "17w2" on the x-axis for both periods Y70 and Y00.

The green area represents the period Y00, while the red area represents the period Y70. Y70 has three peaks, all with a frequency of 5. The mean values for these peaks range from 0.6 to just below 0.8.

However, in Y00 there are three peaks, all with a frequency of 4. The mean values for these peaks range from 0.7 to 0.9. Therefore, indicating that there is a shift in distribution towards lower mean values for the BD5 taxonomic groups from Y00 to Y70.

To determine if the observed differences in mean between the two periods are statistically significant, I performed a two-sample t-test using the mean values for BD5 grouped with the land classification 17w2 for both time periods.

The resulting p-value was equal to 1.751e-07, which is less than the significance level of 0.05. As a result, I reject the null hypothesis. Therefore, there is a significant difference in the mean between the two periods for BD5, as conveyed by the histogram.

Additionally, based on the difference in sample means 0.105754, it does not appear that the difference between the two variables is relatively large. Again, regarding the sample means, the mean has increased from Y70 to Y00, which suggests that the species richness has increased between periods, concluding that BD5 is dependent on the variables I have chosen for this open analysis.