MA334 FINAL PROJECT REPORT

BY MARYAM TAJ

DATED: 27TH APRIL,2023

Introduction

The aim of this report is to explore seven variables of proportional species richness and investigate their correlation, perform hypothesis testing, multiple linear regression, and open analysis. The analysis has been done by using R software. The data of proportional species richness is used for this analysis. The seven variables; Bees, Bird, Hoverflies, Ladybirds, Macromoths, Grasshopper, and Vascular_plants were merged using row mean score method (taking average) into a single univariate variable named as BD7.

Objective:

Our main questions of interest are how BD7 differs from the mean of all 11 taxonomic group proportional species values (which is termed as BD11). The BD7 variables allocated to me were:

Bees, Birds, Hoverflies, Ladybirds, Macromoths, Grasshoppers, Vascular plants.

Descriptive Analysis

The summary statistics including mean and standard deviation was used to check the central tendency and variation among the variables.

Table 1

Descriptive Statistics

Variables	Mean	SD
Bees	0.60502	0.3105858
Birds	0.8872	0.1065161
Hoverflies	0.6795	0.1771395
Ladybirds	0.6140	0.2660338
Macromoths	0.84927	0.1406266
Grasshoppers	0.6289	0.2093839
Vascular_plants	0.7869	0.1010546

A descriptive statistics summary presented in above table (1). The analysis was carried out to check the average and standard deviations of the seven main variables: bees, birds, hoverflies, ladybirds, macromoths, grasshoppers, and vascular plants in a certain study area. The mean abundance of bees was 0.61 (SD = 0.31), while the mean abundance of birds was 0.89 (SD = 0.11). Hoverflies had a mean abundance of 0.68 (SD = 0.18), ladybirds had a mean abundance of 0.61 (SD = 0.27), and macromoths had a mean abundance of 0.85 (SD = 0.14). Finally, grasshoppers had a mean abundance of 0.63 (SD = 0.21), and vascular plants had a mean abundance of 0.79 (SD = 0.10). These findings suggest that the study area has varying levels of abundance for each of these species.

Correlation Analysis

Variables	Bees	Birds	Hoverflies	Ladybirds	Macromoths	Grasshoppers	Vascular_plants
Bees	1.00						
Birds	0.38	1.00					
Hoverflies	0.36	0.44	1.00				
Ladybirds	0.51	0.55	0.40	1.00			
Macromoths	0.47	0.59	0.39	0.52	1.00		
Grasshoppers	0.35	0.24	0.42	0.31	0.29	1.00	
Vascular_plants	0.17	0.22	0.34	0.18	0.13	0.37	1.00

Table (2) presents the correlation matrix between the seven variables of interest: bees, birds, hoverflies, ladybirds, macromoths, grasshoppers, and vascular plants. The correlation significance level was tested using the P-value less than 0.01. A significant correlation is reported for all pairs of variables ($p < .01$). Specifically, the correlation coefficients between bees and birds ($r=0.38$, $P<0.01$), hoverflies and bees ($r=0.36$, $p<0.01$). Bess and ladybirds ($r=0.51$, $p<0.01$). The correlation between Macromoths and Bess is reported to be positive ($r=0.47$, $P<0.01$). Similarly, Bees and Grasshoppers ($r=0.35$, $P<0.01$), and Vascular_Plants and bees also indicated a positive and significant correlation ($r=0.17$, $P<0.01$). On the other hand, Bird with Hoverflies ($r=0.44$, $P<0.01$) also indicated positive and significant correlation. Bird and Lady Birds showed significant and positive correlation ($r=0.55$, $P<0.01$). Bird and macromoths pairs also indicated positive and significant correlation ($r=0.59$, $P<0.01$). The correlation between pairs like Bird and Grasshoppers ($r=0.24$, $P<0.01$). The Vascular_Plant with birds indicated a significant correlation ($r=0.22$, $P<0.01$). The other pairs of Hoverflies and Ladybirds showed significant and positive correlation ($r=0.40$, $P<0.01$). Hoverflies and Macromaths ($r=0.39$, $P<0.01$), with Grasshoppers ($r=0.42$, $P<0.01$). Vascular_Plant showed positive and significant correlation ($r=0.34$, $P<0.01$).

Moreover, the pairs of Macromaths and Grasshoppers indicated significant correlation ($r=0.29$, $P<0.01$), and with Vascular_Plant also showed significant positive correlation ($r=0.13$, $P<0.01$). The last pair of variables Grasshoppers and Vascular_Plant indicated significant and positive correlation ($r=0.37$, $P<0.01$).

Summary Statistics based on Periods:

In this part, the analysis provides a useful framework for exploring and analyzing summary statistics for a set of selected variables, and for comparing the statistics between two periods of interest.

It provides an overview of the descriptive statistics for the selected variables, including their mean, median, minimum, maximum, standard deviation, IQR, and skewness. This information can be useful in identifying any outliers or trends in the data.

	period	variable	mean	median	sd	skewness	min	max
Bees1	Y00	Bees	0.71	0.70	0.32	1.21	0.10	3.31
Bird1	Y00	Bird	0.90	0.92	0.11	-1.35	0.24	1.17
Grasshoppers_._Crickets1	Y00	Grasshoppers_._Crickets	0.60	0.58	0.23	0.07	0.07	1.59
Hoverflies1	Y00	Hoverflies	0.64	0.65	0.17	-0.40	0.12	1.15
Ladybirds1	Y00	Ladybirds	0.63	0.66	0.29	0.22	0.06	1.84
Macromoths1	Y00	Macromoths	0.90	0.92	0.13	-1.36	0.25	1.26
Vascular_plants1	Y00	Vascular_plants	0.76	0.76	0.10	0.04	0.42	1.20
Bees	Y70	Bees	0.50	0.47	0.26	0.30	0.03	1.00
Bird	Y70	Bird	0.87	0.89	0.10	-1.90	0.25	1.00
Grasshoppers_._Crickets	Y70	Grasshoppers_._Crickets	0.66	0.67	0.18	-0.13	0.16	1.00
Hoverflies	Y70	Hoverflies	0.72	0.74	0.17	-0.63	0.14	1.00
Ladybirds	Y70	Ladybirds	0.59	0.60	0.24	-0.40	0.06	1.00
Macromoths	Y70	Macromoths	0.80	0.83	0.14	-1.27	0.09	1.00
Vascular_plants	Y70	Vascular_plants	0.82	0.82	0.09	-0.24	0.56	1.00

Hypothesis Testing

Ho: The mean of BD7 does not significantly differs from the mean of BD11

H1: The mean of BD7 significantly differs from the mean of BD11

Table 3

Mean Values BD7	Mean of BD11	T	DF	P	Decision
0.72	0.71	2.6533	10210	0.007983	Significant

Table (3) shows t-test results conducted to check the mean level significant difference. BD7 differs significantly from BD11 (t_{10210}=2.6533, P<0.05). The mean difference is significant. This indicates that on the average BD7 differs significantly from BD11.

Ho: The mean of BD7 does not significantly differs from the mean of BD4

H1: The mean of BD7 significantly differs from the mean of BD4

Table 4

Mean Values BD7	Mean of BD4	T	DF	P	Decision
0.72	0.70	7.5212	9896	<0.05	Significant

Table (4) shows t-test results conducted to check the mean level significant difference. BD7 differs significantly from BD4 (t_{9896}=7.5212, P<0.05). The mean difference is significant. This indicates that on the average BD7 differs significantly from BD4.

Simple Linear Regression

In this section, we are implementing simple linear regression between two variables BD7 and BD11. The histogram of variable BD7 is plotted to check its normality. The lm() function is used to fit a linear regression model between BD7 and BD11, and the summary() function is used to get the model summary. The plot() function is used to plot the model to check its linearity and normality.

Assumption of Normality

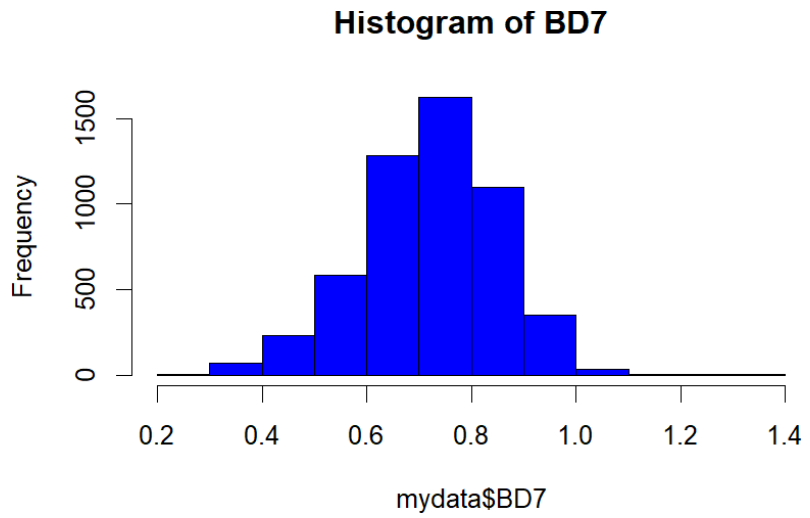


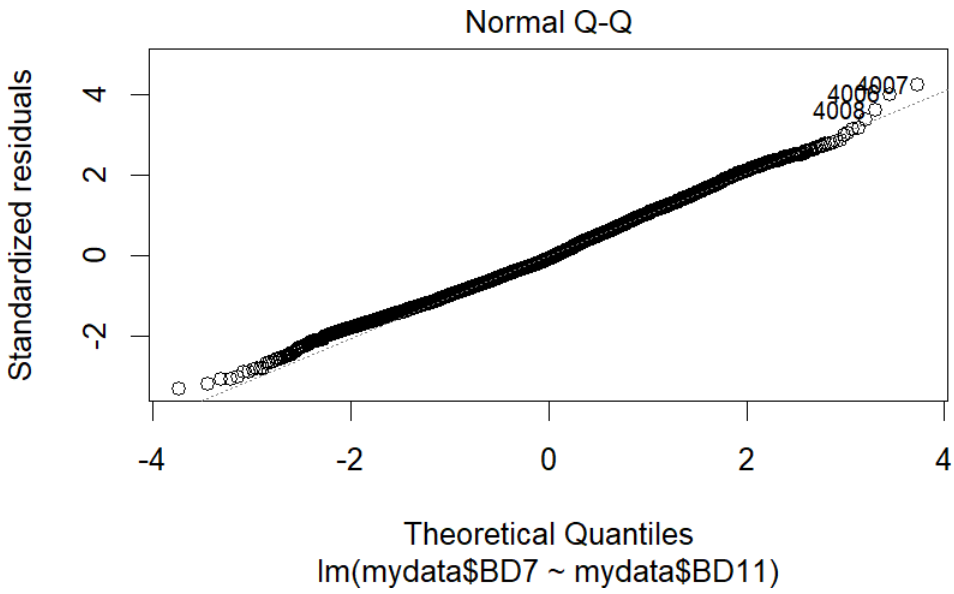
Figure 1

Figure (1) above shows symmetrical shape for the dependent variable BD7. This is an indication that BD7 follows normal distribution.

Figure 2

QQ plot for normality test.
The QQ plot shown below indicates that the data follows normality.

The QQ-plot shown in figure (2) shows that all the points lies around the central line. This is an indication that there is no violation of the assumption of linearity. Additionally, this is proofed using



the plot (3) below.

Model Explanation

R-squared: This is a measure of how well the model fits the data. In this case, the multiple R-squared is 0.9247, which means that the model explains 92.47% of the variation in the response variable. Adjusted R-squared: This is a modified version of the R-squared that considers the number of predictors in the model. In this case, the adjusted R-squared is also 0.9247, which means that adding more predictors to the model is not likely to improve its performance.

F-statistic: This is a test of whether the regression model is significant. The F-statistic is 6.483e+04, which is very large, and the p-value is less than 2.2e-16, which is much smaller than the typical significance level of 0.05. This means that the model is highly significant. Overall, these results suggest that the regression model is a good fit for the data, and that it explains a large amount of the variation in the response variable.

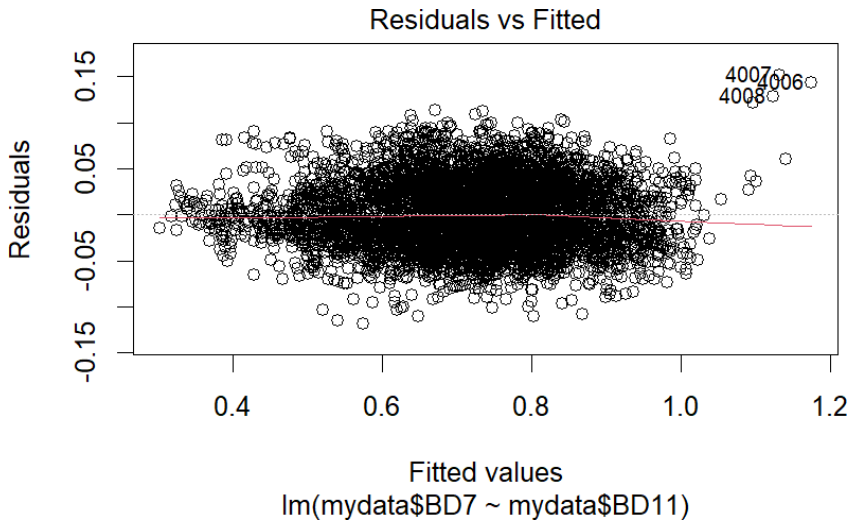


Table 6 Regression

outputs

Description	Estimates	Std.Error	t-value	P-value
Intercept	-0.107607	0.003293	-32.67	<0.05
BD11	1.159063	0.004552	254.61	
Dependent Variable BD7				

A simple linear regression was conducted to examine the relationship between BD11 and BD7. BD11 was found to have a statistically significant effect on BD7 ($\beta = 1.16$, $t(5278) = 254.61$, $p < .001$). The positive coefficient estimate indicates that as BD11 increases, BD7 also increases. Therefore, the results suggest that BD11 has a significant positive effect on BD7.

Simple Linear Regression separately for each Period

Table 6.1

Regression outputs

Description	Estimates	Std.Error	t-value	P-value
Intercept	-0.175385	0.004302	-40.77	<0.05
BD11	1.235024	0.005946	207.70	<0.05
Dependent Variable BD7, separated for the period Y70				

Table 6.2

Regression outputs

Description	Estimates	Std.Error	t-value	P-value
Intercept	-0.175385	0.004302	-40.77	<0.05
BD11	1.235024	0.005946	207.70	<0.05
Dependent Variable BD7, separated for the period Y00				

The R-square value for the periods is almost same, like it is around 0.89%. The F-test predictsthe significant result for both periods. The regression outputs are predicting significant results for both periods.

Test Model

Model Explanation

The test model was run using the training data. This was consisting of the 80% of the sample. The correlation between the variable in the test data is around 0.69. Which shows a strong correlation.

The F test indicated overall significance impact of the seven variables on the combined variable BD4 ($F_{7,4216}$=568.8, P<0.05). The R-square in this model explains around 48% of variation.

The regression outputs are as follow:

Table 7

Regression outputs

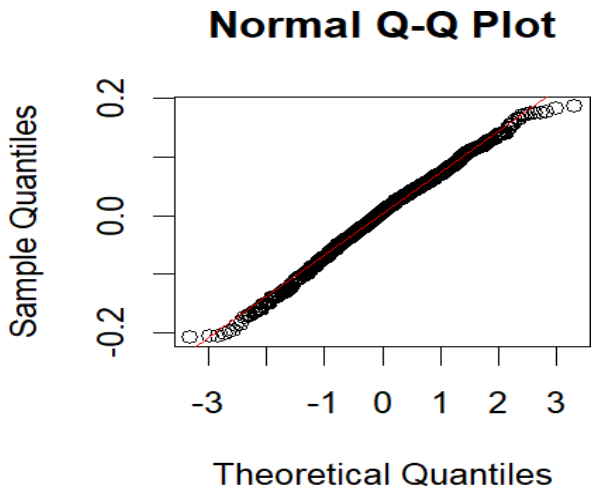
Description	Estimates	Std. Error	t-value	P-value
Intercept	0.267372	0.012447	21.480	<0.05
Bees	0.003364	0.004372	0.769	0.441
Bird	-0.043352	0.014244	-3.044	0.002
Hoverflies	0.234689	0.007637	30.729	<0.05
Ladybirds	0.030083	0.005494	5.476	<0.05
Macromoths	0.053086	0.010569	5.023	<0.05
Grasshoppers	0.066406	0.006208	10.697	<0.05
Vascular_plants	0.265139	0.011993	22.109	<0.05
Dependent Variable BD4				

Table (7) above indicates overall significant results for all predictors except BEES (β= 0.003364, $t_{(5271)}$ = 0.769, p =.44). The predictor Bird showed significant but negative influence on BD4 (β=-0.043352, t=-3.044, P<0.05). The remaining predictor variables (Bees, Hoverflies, Ladybirds, Macromoths, Grasshoppers & Crickets, and Vascular_plants) had statistically significant positive effects on BD4 (all p < .05). Therefore, the results suggest that the seven predictor variables collectively have a significant positive effect on BD4, while Bird has a significant negative effect.

The AIC for this model is around - -10317.54.

Figure 4 Normal

QQ plot



The normal QQ plot shown in above indicated that all the points are around the center line and there is no violation of the assumption of normality.

Multiple Linear Regression

In this section, we perform a multiple linear regression of BD4 against all seven of your proportional species values. Thus, initially we have seven predictors and BD4 as the response variable.

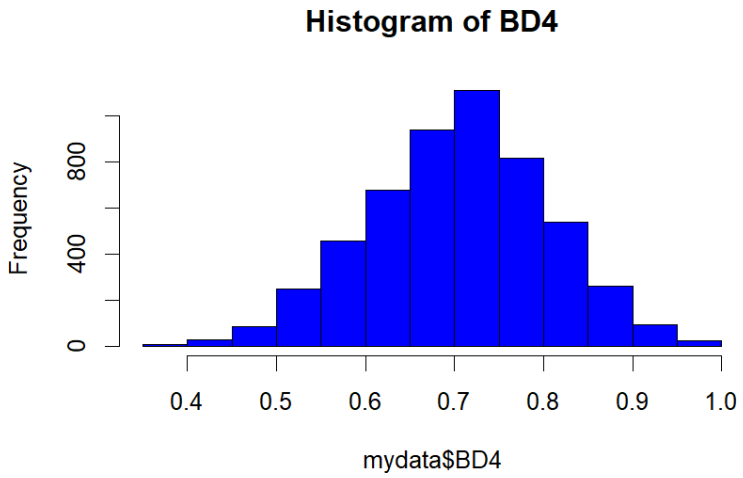


Figure 5: Histogram shows symmetrical shape. An indication of normality

Model Explanation

R-square value for this model is 0.49, this indicates that around 49% of variation is explained in the model through the predictors. The overall ANOVA result was tested using F test. The F test indicated significant result ($F_{7,5272}$=725.3, P<0.05).

Table 8

Regression output of multiple regression

Description	Estimates	Std.Error	t-value	P-value
Intercept	0.261460	0.011073	23.613	
Bees	0.002107	0.003921	0.537	0.5911
Bird	-0.029854	0.012583	-2.373	0.0177
Hoverflies	0.235622	0.006863	34.331	<0.05
Ladybirds	0.030042	0.004927	6.097	<0.05
Macromoths	0.048492	0.009477	5.117	<0.05
Grasshoppers_._	0.067481	0.005565	12.127	<0.05
Vascular_plants	0.261948	0.010800	24.254	<0.05
Dependent Variable BD4				

Among the predictor variables, only Bird had a statistically significant negative effect on BD4 (B = -0.03, t(5271) = -2.37, p = .018). The remaining predictor variables (Bees, Hoverflies, Ladybirds, Macromoths, Grasshoppers & Crickets, and Vascular_plants) had statistically significant positive effects on BD4 (all p < .05). Therefore, the results suggest that the seven predictor variables collectively have a significant positive effect on BD4, while Bird has a significant negative effect.

Whereas only Bees showed insignificant effect on the dependent variable BD4 (β=0.002107, t=0.537, P=0.5911).

The AIC for this model is around -12880.12.

Multiple Regression with 6 predictors

A new regression is suggested following the P-value criteria. The predictor BEES indicated an insignificant result with a P-value greater than 0.05. Therefore, as model selection criterion, we removed the predictor BEEs and re-run the analysis. The results are as follow.

R-square value for this model is 0.49, this indicates that around 49% of variation is explained in the model through the predictors. The overall ANOVA result was tested using F test. The F test indicated significant result ($F_{6,5273}$= 846.3, P<0.05).

Table 9

Regression output of multiple regression after removal of BEES as predictor

Description	Estimates	Std.Error	t-value	P-value
Intercept	0.260864	0.011016	23.680	
Bird	-0.029889	0.012582	-2.376	0.0177
Hoverflies	0.235947	0.006836	34.515	<0.05
Ladybirds	0.030784	0.004729	6.510	<0.05
Macro moths	0.049604	0.009248	5.364	<0.05
Grasshoppers.Crickets	0.067963	0.005491	12.377	<0.05
Vascular_plants	0.261918	0.010799	24.253	<0.05
Dependent Variable BD4				

Among the predictor variables, only Bird had a statistically significant negative effect on BD4 (β= -0.03, t (5271) = -2.37, p = .018). The remaining predictor variables (Hoverflies, Ladybirds, Macromoths, Grasshoppers & Crickets, and Vascular_plants) had statistically significant positive effects on BD4 (all p < .05). Therefore, the results suggest that the sixth predictor variables collectively have a significant positive effect on BD4, while Bird has a significant negative effect.

There is not any predictor whose P-value is greater than 0.05. This shows that the predictors are valid and significant. No need to further change the mode.

The AIC value for the new model shows -12881.83. The AIC value for the new model is less than in the previous one. Therefore, it can be stated that the revised model is better.

Open Analysis

Methodology

An open analysis was carried out to check the significant difference of the combined variable BD7 on location variable and time variable. It is believed that the average responses vary with respect to time and location.

In this regard an independent sample t-test is used to test the change in average biodiversity score in between the two periods.

Further, an ANOVA test was also used to check the biodiversity in between different states, i.e, England, Wales, and Scotland.

Results:

Table 10

Two sample t test result

Mean Values BD7	T	DF	P	Decision
0.72	402.56	5279	<0.05	Significant

Our analysis showed a significant increase in biodiversity between the two periods. The mean BD7 was 0.72 differs significantly between the two time periods ($t = 402.56$, $p < 0.05$). This indicates that there has been an overall change in biodiversity over time.

It is of great interest that we there the variable of interest change in biodiversity score varies for the three states or not. To investigate this ANOVA analysis was run.

Table 11

One-way ANOVA result

Mean Values BD7	DF	Sum of square	Mean Square	F	P-value
Location	402.56	5279	0.030434	8.703	
Residuals	2640	9.23	0.003497		

The ANOVA result was tested using F test. The F test indicated significant result ($F= 8.703$, $P<0.05$).

This shows that BD7 averages values varies with respect to location as well.

Conclusion:

In this report, we explored 11 variables that measure biodiversity and analyzed them individually and in relation to each other.

In this report correlation analysis was carried between the main variable of interest. From the results reported in table (2) above, it can be concluded that all the seven variables of interest are correlated with each other. Interesting findings regarding correlation is the positive correlation shown by these variables with each other. It is pertinent to mentioned that all these variables are internally linked. Changing in any one variable will affect the other variables. Therefore, any intervention to disturb the biodiversity will have long lasting effect.

The hypothesis testing in the first section illustrates mean different between the composite variable of BD7 and BD11. The first hypothesis result shown in table (3) indicates significant result. This shows that average of the seven variables differs significantly from the 11 variables. Similarly, the remaining four variables composite average also differs significantly from the average of the composite variable of seven variables. This analysis provides a base for the future analysis that on average the combine effect of the seven variables (Bees, Bird, Hoverflies, Ladybirds, Macromoths, Grasshopper, and Vascular_plants) differs significantly from any other combination of variables.

In the later section the effect of BD11 on BD7 was tested using simple regression, and effect of all seven variables on BD4 using multiple linear regression.

It can be concluded that all eleven variables are interrelated. Changes in the behavior or dimension of one variable will affect positively or negatively the BD7 or BD4.

Lastly, changes in the combined effect of seven variables BD7 is affected by time and region. This was proofed using independent sample t-test and one-way ANOVA