MA334 AU 2023 Individual Assignment (100% of module marks)
Deadline: 1200 15th January 2024

The project folder on Moodle has all the materials and data you will need. Follow guidelines given in the module for their use, specifically use only the data from the csv file named (*which does not include any missing values*):

proportional_species_richness_NAs_removed.csv

As described in the module, you will be randomly allocated an individual selection of five of the eleven taxonomic groups. Your individual biodiversity measure, following the authors of the paper from which this data is derived, is simply the mean of the proportional species values for your allocated five taxonomic groups (hereafter termed BD5). Questions of interest are how BD5 differs from the mean of all 11 taxonomic group proportional species values (hereafter termed BD11). The authors in their paper report some numerical results based on BD11 and compare BD11 between two time periods, you should aim to report back in a similar way based on BD5, as far as your analysis allows. Much of the scientific paper on which this assignment is based is outside the scope of this assignment. For example, we do not have the biodiversity measure which is based on priority species, and you should ignore that aspect and others such as the "Fescalo" method for the purposes of this assessment.

Be sure that you follow the following basic rules:

- Compile your final assignment as single PDF, no other format, do not zip the files and do not submit in WORD. If the Turnitin plagiarism detector software is unable to read your file for any reason, for example if you compress it, then it cannot be marked. The written assignment itself must be simply a single PDF.

- There is no set limit to the number of pages but an assignment with more than about seven pages is likely to be poorly executed (See the list of warnings below) and may be penalised if it is too verbose or repetitive.

- This is an individual assignment, and you must use R. In addition to the single pdf described above you must submit your final R code in one separate file as a "*.R" file which could be run directly within the markers RStudio with only a change in working directory for reading in the data file. You must add your own comments to any parts of code developed by yourself.

- Note it's the registration and no other number that's needed in these formats. Do not write your name or any other ID on the pdf or R script and submit on Faser only before the deadline. Uploading multiple versions will likely result in any one of them being marked so delete old versions from Faser if you make a mistake uploading.

In summary, submit two separate files, a written assignment in a format like "MA334_reg_no.pdf" and code in a format like "MA334_reg_no. R" (Do not Zip !)

Specific instructions for the written assignment

Univariate analysis and basic R programming

Provide a concise and well written univariate report (in PDF form) on your five values for proportional species richness. This report must include all the following:

1) Present a table which provides the following summary statistics for each of the five variables in your BD5 group. This table must provide the 7 statistics:
   a) The six statistics which are commonly found using the summary() command, namely Min, 1st Quarter, Median, Mean, 3rd Quarter, Max.
   b) In addition, add a column to this table which provides a new statistic, that is the 20% Winsorized mean for all the variables in the BD5 group. In your R code file add comments to the relevant lines. (Mark: 4/100).

2) Estimate the correlations between all pairs of variables in BD5 and put them into a table which consists of 5 rows and 5 columns where each row/column represents a BD5 variable (Mark: 4/100).

3) Perform the boxplot for only one variable in BD5 (Mark: 1/100).

4) Using the results which you have obtained above and using your table or boxplot, write some conclusions. (Mark: 3/100).

Total Marks: 15/100 broken down into accuracy (12/15) and quality of writing (3/15)

Hypothesis tests

Perform two distinct types of hypothesis test aside from the linear regression results. You may choose any tests within the scope of the module using the given data set. You should precisely report the p values and also an interpretation of the results.

Marks: 15/100 broken down into accuracy (10/15) and quality of writing (5/15)

Contingency table/comparing categorical variables

If BD5 is sufficiently like BD11 then where there is an increase of BD11 there will also be an increase in BD5, likewise with decreases. Perform a test of independence between BD5 increases/decreases and BD11 increases/decreases between two periods. That is, you need to define two binary categorical variables which signify either an increase or a decrease in the biodiversity measure under consideration. Term these new binary categorical variables BD11up and BD5up.

1) Create two contingency tables which display counts, one for BD11up against BD5up and another for the corresponding independent model. (Mark: 2/10).

2) Using R, estimate the likelihood-ratio statistic from the two above contingency tables. Based on this statistic, compare the proportions of increase in BD5 and BD11 at 5% confidence level (Mark: 3/10).

3) Estimate odds-ratio, sensitivity, specificity, and Youden's index. Draw some conclusions from these quantities. (Mark: 3/10).

Total Marks: 10/100 broken down into accuracy (8/10) and quality of writing (2/10).

Simple linear regression

Choose only one of the taxonomic groups which is not one of your BD5 group, hereafter termed "BD1".

1) Perform a simple linear regression with BD1 as the response variable and BD5 as the predictor variable. (Mark: 5/10).

2) Interpret the slope and determine from your output if the estimated slope is significant (Mark: 5/10).

Marks: 15/100 broken down into accuracy (10/15) and quality of writing (5/15)

Multiple linear regression

Again, using same BD1 chosen above. Perform a multiple linear regression of BD1 against all five of your five proportional species values. Thus, you will initially have five predictors and BD1 as the response variable. We call this regression model "initial MLR". You are not required to split the data into a training and a testing subset except in section (4) below, but you may wish to do so. The tasks are as follows:

1) Estimate the AIC for the initial MLR model (Mark: 5/100).

2) Perform any feature selection based on p values and AIC from the regression. Justify the removal or otherwise of variables as predictors (Mark: 5/100).

3) Assume that you can use one interaction term between any two predictor variables in the BD5 group. Find a linear regression model which has a lower AIC, if possible (Mark: 10/100).

4) Divide the dataset into two sub-datasets where each subset is for one period (Y00 or Y70). Use the earlier period Y70 subset as a training set and the later period Y00 subset as the test set (also known as the validation set). Then perform the usual diagnostics, compute the mean square error on the test set and on the training set and discus these results (Mark: 5/100).

Marks: 30/100 broken down into accuracy (25/30) and quality of writing (5/30)

Open analysis

Using any of the variables provided, including land classification, period for the data, eastings and northings, write a report on the dependence of BD5 on these variables. A primary focus would be the changes in BD5 between the two periods, but you can

choose another aspect. It is essential that you write this section in a clear and precise way, avoid speculation but rather base statements on the data. Use only methods presented in the module set book.

Marks: 15/100 broken down into accuracy (10/15) and quality of writing (5/15)

Plagiarism and other more minor bad practice; avoiding low marks

With apologies to the great majority of students on MA334, it has become necessary to be explicit. The following mistakes will result is a lowering of the total score, depending on the individual cases, of course, so no numbers cannot be offered. Note that this assignment is worth 100% of module marks, hence it is important to be aware of the following…

- The Turnitin software automatically provides a report which details text copied from anywhere including all other Essex submissions. Plagiarism will result in an academic offence hearing and very likely a penalty total score for the assignment. Plagiarism includes collusion as this is an individual assignment. Any copying of text is easily picked up by Turnitin, that and other collusion such as obviously recycled figures or tables cannot be ignored. Clumsy manipulations of borrowed wording do not escape the software's eagle eye. Finally, be aware that all suspected plagiarism cases cause considerable problems for everyone involved in their processing as well as the suspect. It is easy to avoid any such problems by simply working alone and being careful to collaborate on stats principles but not on the presentation of your analysis.

- There are some relatively minor bad practices which are still bad practice. These are obvious as they usually denote a lazy approach. For example, the pasting of computer output or screen plots into a written assignment, this is very poor practice. Plots must be useful to the argument and discussed in the main text and avoid plots which are either excessive or very meagre in information content. Output from R could be compiled into a table or presented as a well labelled plot which is referenced and used in the main text. Do not cut and paste raw computer output into a report. The use of R markdown may be useful but not if code and computer output is provided as some kind of substitute for analytical argument.