MA334 SP Individual Assignment 2023 (100% of module marks)
Deadline 27th April 1200

Materials provided and general instructions

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

proportional_species_richness.csv

Or a variant suffixed "_V2" which has added two extra variables.
Or a variant suffixed "_V3" which has no NAs (see below)

Missing Values (NAs)

You are not required to consider the treatment of missing values but may choose to do so within your written assignment. It is acceptable to use a version of the data which has been cleaned of NAs and selected for squares which have data for the same taxonomic groups for both periods to facilitate individual square comparisons between the two periods. This cleaned up version of the data is suffixed "_V3" and is available on Moodle.

As described in the module, you will be randomly allocated an individual selection of seven 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 seven taxonomic groups(hereafter termed BD7). Questions of interest are how BD7 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 BD7, 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.

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. Please follow these basic rules. The written assignment itself must be simply a single PDF. There is no 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).

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 can 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.

In summary, there are two files, to be submitted separately (not zipped together):

1. The written assignment in a format like “MA334_reg_no.pdf”

2. The code in a format like “MA334_reg_no.R”.

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.

Specific instructions for the written assignment

Data Exploration

For this section you should explore your seven variables as single variables (univariate) and also look at the correlations between your seven variables. Choose some features to concentrate upon so that this section is not excessively long. If you wish, you may also consider the other variables, period, land classification, Easting or Northing. Aim to write a report which might only highlight some interesting aspects of some of your seven values for proportional species richness.
Marks: 15/100 broken down into accuracy (10/15) and quality of writing (5/15)

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)

Simple linear regression

Report back using simple linear regression on how BD7 matches BD11. Also consider this simple linear regression for each period separately.
Marks: 15/100 broken down into accuracy (10/15) and quality of writing (5/15)

Multiple linear regression

Removing your seven variables (the mean of which is BD7) from the eleven in the data set(the mean of which is BD11) leaves four taxonomic groups. Calculate the mean of the proportional species richness values for these remaining four taxonomic groups (hereafter termed BD4). BD4 is itself a measure of the biodiversity excluded from BD7. Perform a multiple linear regression of BD4 against all seven of your proportional species values. Thus, you will initially have seven predictors and BD4 as the response variable. Then perform any feature selection based on p values from the regression and also use AIC to justify the removal or otherwise of variables as predictors. Finally write a report on your final model including an in interpretation of the regression coefficients.
Marks: 25/100 broken down into accuracy (20/25) and quality of writing (5/25)

Open analysis

Using any of the variables and the proportional species richness values, in any way you choose, write a report. The natural primary focus is change 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 these data. Use methods presented in the module. This section is the chance to demonstrate your skill in telling a story about some chosen aspect of these data.
Marks: 30/100 broken down into accuracy (20/30) and quality of writing (10/30)

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.