BIO782P - Week 1 Assessment

Joe Parker

10th November 2017

BIO782P Statistics and Bioinformatics Assignment 2017

Due: 17:00 Friday 1 December 2017

Introduction

You have been allocated two unique datasets. To find yours, look at the table at the end of this document. Once you know the numbers of the datasets to use look in the folder called `Student_data` (see class GitHub). If your UID listed below is (for example) 17, then you should look in folder `Student_data/17` for input files `17_part1.tdf`, `17_part2.tdf` and `17_part3.tdf`. All datasets are supplied as tab-delimited text files.

Having read the description of the way the data were collected, you need to import each dataset into R. Remember to check the datasets when you import them for things like proper allocation of each variable as a factor or a numeric variable etc, and to carry out proper exploratory data analysis. Once you're happy that you've got the data into a condition where it can be analysed, fit a model and use that to try to answer the questions given below. Don't forget to check the model assumptions!

Report format

For each dataset, your writeup should consist of a properly annotated R script that will allow me to exactly replicate your analysis, and the results section of a paper describing your findings (please collate the reports into a single document). NOTE: Marks will be deducted for scripts that do not run, while bonus marks (max 5%) are available for scripts submitted as pull requests to the class GitHub.

The report should be no more than 1000 words for all three parts (and can be a lot less…) and should not contain any unnecessary figures or tables. The word count excludes code, figure/table legends, and any references you want to include. If in doubt have a look at some papers in (for example) PLoS Biology or Genome Biology & Evolution to see how it's done. Please don't paste output from R directly into your report, by which I mean don't just paste a summary table or an ANOVA table in (not that you shouldn't use graphics from R).

Dataset 1: Marine microbial diversity

In two successive research cruises (January and August), the RRS Discovery survey ship has sampled 20l of seawater from 10 separate locations within approximately equatorial (0-8 degrees N) and equatorial (48-55 degrees N) waters of the North Atlantic (e.g., the Azores and British Isles, respectively). Each sample was filtered with a coarse filter to remove macroinvertebrates, then filtered using a 50uM mesh to collect microbiota. DNA was extracted, WGS sequenced with an Oxford Nanopore MinION, and sequence reads identified using BLASTN. Microbial diversity was assessed using UniFrac, and is expressed relative to a reference sample sequenced at MBL, Plymouth.

For the report, please use an appropriate analysis to answer the following questions (***Note:*** ignore variables X and Y in the datafile):

1. How does microbial diversity change with latitude?
2. How does microbial diversity change with time of year?
3. Is there an interaction between the season, and location?

Dataset 2: Pairwise nucleotide substitutions and RNA expression levels

A particular protein family, luciferase (P08659; http://www.uniprot.org/uniprot/P08659 is responsible for bioluminescence. As part of an RNA-seq project we have sequenced over 900 coding loci from 208 species of plants within the Brassicaceae (a large flowering plant family), and are surprised to discover 30 genes with homology to luciferase, and measure their exprssion levels in each species. We build a phylogenetic tree using the sequences and compute the pairwise genetic distances (averaged number of amino-acid substitutions) between a 31st copy of the 'putative' luciferase gene, in the closely-related, high-quality genome sequence of ***Arabidopsis thaliana***, and the copy in each species.

For your report, please use an appropriate analysis to answer the following questions:

1. How does the putative 'luciferase' homologue expression change with genetic distance (amino acid substitutions)?
2. Comment on whether the model assumptions are valid.
3. Assuming your model is ***statistically*** valid, can you guess what effect is responsible for the relationship you've found?

Dataset 3: HIV viral load and within-patient population dynamics

Populations of the HIV virus are able to evade drug therapy and persist in certain immunoprotected tissues in the body, principally areas of the central nervous system (CNS) where CD4+ T-cells are prevented from entering. An HIV+ patient has enrolled in a study and consented to viral load sampling over 40 weeks in a year. Each week a sample was taken from the brain or spinal cord, and the viral population size ('load', here expressed as log10(number of viral particles per ml)) measured using a chip assay. Average Shannon population diversity, and mean pairwise genetic distance from individual viruses present in each weekly sample to a reference sequence was assessed using single-copy PCR amplification of the viral Env gene, and we hope to find some relationship between population size, diversity, evolutionary distance and tissue.

The sampling (involing an epidural needle inserted deep into the CNS) is painful and carries a high risk of complications, so we need to make sure we make the most of this data. Do your best to fit a model explaining viral load in terms of the other variables, using any combination of variables you see fit, and any model selection procedure. Write up your results accordingly, with any figures which are appropriate for the research paper we plan to write. Since we're unsure how likely the host immune system is to penetrate the CNS, CD4+ cell counts in the patient's general circulation were assessed using flow cytometry, and categorized as 'low' or 'high' - you may wish to consider these in your analysis. The rows in the data table are in chronological order.

General feedback from previous years.

Overall, we have been impressed with the standard of previous reports. Most of you manage to get the analyses more or less right, and most of you produce good- quality scripts. None of them had too much annotation - remember that when you're annotating a script you're really writing a guide to yourself describing what it does. Think about the information that might be useful if you come back to the analysis a year or two in the future.

Regarding the "results section" of the reports, these are a lot more variable. Many of you put too much analysis into these, and in particular include material that would not normally go in a journal results section. Preliminary and exploratory analysis, diagnostic plots and the like are not generally included in journal results sections. Many of you have redundant graphs – you shouldn't show the reader the same set of data twice. A common problem is too much focus on the statistics and not enough on what the statistical results mean in terms of the biology of the system, or in terms of effect sizes: don't just tell me that there's an effect, tell me how big the effect is, and put it in meaningful biological terms.

Figure captions are something that most of you need to work on. When you're writing a figure caption, it's a good idea to try to write it so that a casual reader who is skimming through the paper can look at the figure, have a look at the caption and have at least a rough idea about what the figure is showing. That doesn't mean that each figure should have the whole methods section reproduced, but you can include a sentence or two that gives the casual reader a basic idea of what's going on.

Other common problems include:

- Including p-values without test statistics or degrees of freedom
- Including multiple tests of the same thing (e.g. using a post-hoc test on a fitted model, collapsing two factor levels and then comparing models with a partial F- test)
- Carrying out tests for normality on data prior to analysis (skipping data exploration/eyeballing and)
- Mixing p-values and AIC as criteria for model selection - the philosophy behind these is very different and you should use one or the other but not both.
- Giving significance levels for main effects when they are also included in higher- order interaction terms
- Including graphs showing no effect. There are circumstances when you might want to do this, if you're reporting a negative result and you think it's necessary to make a point about how little relationship there is, but in general we wouldn't put this sort of thing in.
- Including code, function names etc. from R. Results sections from journals wouldn't usually include this kind of material unless you're describing an esoteric analysis that the readers will not be familiar with, in which case you might say "We fitted a generalised additive model to the data using the gam() function as implemented in the mcgv package (Wood 2014)" but usually you would just tell the reader the type of analysis used.
- Referring to "non-significant" results as "insignificant". Don't do this - have a think about why.
- No references! Not a single person put a reference in their results section.

(Student IDs overleaf)

Student names and datasets:

student_UID	Surname
student_993	Akkari
student_190	Baskan
student_75	Chowdhury
student_403	Gomez
student_322	Grant
student_137	Grigoriadis
student_273	Jackson
student_572	Kale
student_3	Khan
student_235	Matthews
student_596	Miller
student_101	Nicholson
student_61	Owosu
student_611	Patel
student_930	Pink
student_896	Selvachandrarajah
student_897	Soormally
student_322	Turner
student_909	Zafar
student_918	(spare)