BIOL343 – Assignment #7

Mytilus educate us about polynomial regression

Assigned Sunday 8 March 2020
Due Saturday 14 March 1159pm

Here's an assignment that will put your new regression, model selection and data transformation skills to work. No big preamble for this assignment, except to say that for as long as biologists have been able to assay genetic variation, they have been interested in whether it exhibited any geographic pattern or correlation with environmental variables. A non-random pattern indicates a potential role for natural selection or some other interesting process in structing genetic variation.

The classic dataset (**mytilus.csv**) that you're going to work with here comes from the golden era of allozyme studies when biologists routinely assayed structural variation in major protein enzymes, which they assumed (usually correctly) had a basis in DNA sequence variation.

The **mytilus.csv** dataset contains frequencies of the "94y" allele at the leucine aminopeptidase locus for 17 populations of the blue mussel (*Mytilus edulis*) scattered along the eastern seaboard of North America. The first column is the population frequency of the *94y* allele, the second column is the geographic position of the population measured as miles east of Southport CT (red marker in the map below) along a gradient of increasing ocean salinity.

The question here is simple: does the frequency of *94y* vary geographically and what is the form of the geographical "cline" in allele frequency? To address this question fit simple, quadratic, cubic and quartic linear regressions (along with a null model) and determine the minimum adequate model (MAM) through stepwise selection using likelihood ratio tests and comparison of model AICc.

Ah…but here's the hitch. The response variable is a proportion that varies 0–1. By definition, proportions violate the assumptions of linear models because the theoretical variance of a proportion varies across its range, with a maximum at 0.5 and minima at 0 and 1. So biologists routinely

transform proportional response variables. Unfortunately, the transformation traditionally used is wrong. There is a better transformation. You must find it and use it. Include in your R notebook text a reference to the refereed literature that justifies your use of your chosen transformation.

Of course, just because scientists routinely use a certain data transformation, doesn't mean that it actually improves the fit of the data to model assumptions. So, you'll have to check this and indicate in your R notebook whether transformation has improved adherence to regression assumptions and/or changed your conclusions from the analysis in any important way.

One you've found the MAM, you will, as per usual, evaluate outliers and produce a publication-quality **ggplot** figure of the pattern of geographic variation in *94y* (transformed?) with the appropriate regression line. For educational purposes, it will be informative to overlay on the same graph the other regression models that you evaluated and emphasize (using a colour scheme of your design) the model that your analysis has indicated is the MAM.

Here's a summary of what Hana and I are expecting from you:

1. Code importing and checking the dataframe (no marks for this, you should be pros at this).
2. Report the reference of a reliable source indicating how variables like allele frequencies should be transformed, and apply this transformation to your data (using **dplyr**).
3. Use `lm()` to fit linear, quadratic, cubic, quartic and null models to variation in transformed allele frequencies.
4. Backwards selection to determine the MAM, plus interpretation of the results.
5. AICc-based model selection to determine the top model. Be careful because not all possible combinations of model terms are legitimate models on their own. You'll have to do some research to figure out how to use AICc to evaluate different polynomial regression models. Make sure you calculate and interpret the Akaike weights and evidence ratios for all the models in the appropriate model set.
6. Did likelihood-ratio tests and AICc lead you to the same MAM?
7. Publication-quality graph + caption illustrating geographic variation in allele frequency with the various regression models superimposed. Use colours to highlight the models indicated as the MAM by the two approaches to model selection.
8. Comparison of how transformed vs. untransformed data meet model assumptions and how transformation influences statistical results. You'll want to see how transformation affected your choice of MAM, how it affected the analysis of residuals and whether it tamed particularly influential data points. In the end you'll have to state whether the data transformation is appropriate.
9. If you find an influential data point, then identify it and explore how its inclusion in the data affects the results and conclusions.
10. Figuring out points 8 and 9 might require a bit of supplemental graphical analysis. If so, you don't have to provide formal captions for these graphs but the graphs do have to appear in your R Notebook.

You should upload your file to the Assignment #7 OnQ dropbox by Saturday 14 March 2020 at 1159pm. As before, please submit a **PDF version** of your .html R notebook document called "StudentNumber_A7.pdf", where the file name starts with your student number.

Before you upload your assignment, please make sure that your PDF file is **complete** and as nicely formatted and organized as possible. Write **concise** code with **no redundancy**. Pay close attention to all the formatting guidelines and tips provided for the previous assignments.

As always, Hana and I will be at tutorial to help. Good luck!!