Week 9 Reading Questions

Mercy Melo (no collaborators)

Question 1: Briefly (1 - 2 short paragraphs) describe at least two tradeoffs between the customized ML methods and the canned methods.

While custom-built procedures can allow statisticians to reveal novel interactions in biological data, standardized procedures are sometimes favored for their apparent simplicity and common perception. Of course there are tradeoffs when choosing between these two types of procedures. One of the main reasons researchers choose to use standardized methods when approaching biological data is that the greater scientific community is more likely to have an understanding of the standardized statistical approach, leading to less explanation being required in communication. When writing up a manuscript's data analysis portion of the methods section, a researcher using standard methods would simply have to state the name of the statistical procedure ("we used linear regression…", for example) and most readers would be able to understand the exact steps that were taken to analyze the data. If a custom-built procedure was used, however, this methods section would have to include much more detail about each step in the analysis, any parameters specified in the model, and why the analysis was formatted in such a way. Besides making communicating methods simpler, using standardized methods can also help other scientists trust your data as they have at least a broad understanding of the analysis method from previous experience. Whether they have ran an analysis of your type previously or just heard the name of the analysis from reading scientific literature, people are more likely to trust results from analyses they have prior experience with. Further, using standardized methods can also help other researchers to compare your results to their results using the same standardized method. Straightforward comparison of data in this way is helpful for different researchers to compare their studies as well as multiple researchers on the same long-term study to have comparable data analysis tools.

Question 2: Briefly (1 - 2 sentences) describe each of the four key assumptions of the general linear modeling approach.

There are four key assumptions of using general linear models. First, the data must be collected from independent observations, meaning that all data points cannot be grouped in any way. For instance, if 10 observations were collected from 2 different sample sites (with each site's datapoints being more closely related to each other than to the other site), these datapoints would violate the independent observation assumption. Second, the residuals of the data must be normally distributed, entailing that the model is fitted to the data in such a way that the separation between the points and the model fit follows a normality pattern. The third assumption is that the variance must be constant within the data, meaning that data must be clustered just as tightly at low x values as it is at high x values. Finally, there must be no error in the measurement of continuous predictor variables. This final assumption is typically violated and/or ignored.

Question 3: Explain how the normality assumption can be met in a general linear model, even if the response variable is not normally-distributed.

The normality assumption in general linear models refers to the residuals of the model being normally-distributed, not the data itself. The residuals of the model are based on the fit of the data to the model line, with a closer fit yielding smaller residuals and a distanced fit yielding larger residuals. The residuals themselves are therefore not based on the data alone, but instead are curated by the data's relationship to the model fit. Any distribution of data can have any distribution of its residuals and the residuals can therefore remain normally-distributed even when the response variable itself is not normally-distributed. By altering the fit of the model to the distribution of the data, the residuals can be made normal with any distribution of the data itself.