Week 9 Reading Questions

Stella Dee

I did not work with any other students.

Q1: Briefly describe at least two tradeoffs between the customized ML methods and the canned methods.

Customized ML methods are sometimes faster from a computational perspective. However, given the amount of computing power in the present day, this is unlikely to be of a concern except for very large datasets. Standard models can be easier to choose parameters for. Also, readers and reviewers may have an easier time accepting canned methods, whereas customized ML methods require very clear explanation and justification, because the reader or reviewer cannot be assumed to have prior knowledge of the methodology.

Q2 : Briefly describe each of the four key assumptions of the general linear modeling approach.

The four assumptions are that the observations are independent, there is constant variance, there is fixed x, or no measurement error in the predictor variables, and that the residuals are normal. The observations being independent means that they are not rendered dependent by some variable such as geographic proximity. Constant variance means that the variability does not depend on the x-value, but rather the stochastic model follows a normal distribution. Fixed x means that there is accuracy in the predictor variables. For example, if your predictor variable is tree dbh, a general linear model assumes that you correctly measure dbh—that your tape is not drooping or incorrectly used, but is taking perfectly accurate measurements every time. The model residuals being normally distributed means that the difference between predicted and observed values follows a normal distribution.

Q3 : 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 refers to the model residuals, not to the distribution of data itself. The model residuals are the differences between the predicted and measured data, and as long as these take a normal distribution, the response variable itself can follow any distribution. In the example used in class, Salamander dispersal rates can perhaps be modeled with a Ricker curve, which is not a normal distribution. However, as long as the model residuals follow a normal distribution, the normality assumption can still be met. This is true of any linear model that appropriately models the data.