

Rozy Bathrick  
Week 9  
I worked alone

**Q1 (1 pt.):** Briefly (1 - 2 short paragraphs) describe at least two tradeoffs between the customized ML methods and the canned methods.

Customized = multi-linear models  
Canned = Least squares

The customized multi-linear models are more complex and less prescriptive than the “canned” simple linear models. The benefit to a “canned” model is that they are simple – much of the work has been done for you and has convenient computational speed. The tradeoff is that the fit may not be as good as it could be – which the customized model would give you. Another tradeoff is the level of explanation needed in the methods section – with a canned approach, there is more of a shared language between authors and readers that doesn’t require lengthy explanation of model building.

**Q2 (1 pt.):** Briefly (1 - 2 sentences) describe each of the four key assumptions of the general linear modeling approach.

- Independent observations in the sample – they are random, non-hierarchical, aren’t reliant on each other somehow
- Constant variance – the stochastic model is normal (NOT mega-phone shaped around the mean, i.e. the variance doesn’t increase)
- Normalcy – the residuals (difference between expected value of a model and observed value) are normally distributed
- Fixed x – we have “perfect” accuracy in the sampling parameters

**Q3 (1 pt.):** Explain how the normality assumption can be met in a general linear model, even if the response variable is not normally-distributed. (1 - 2 paragraphs)

The Shapiro test for bill length confirmed the distribution of frequency of bill lengths is not normal with a low p-value. The actual variables do not need to be normally distributed in order for Normalcy in a linear model. The linear model poses the question: how much is bill length explained by body mass and species? A model is generated to predict bill length, given penguin sex and body size. The Shapiro test in the linear model describes the distribution of the residuals – the difference between the expected values of bill length and the actual, observed values. It passes the test with a high p-value, which means we do NOT reject the null – the residuals are normally distributed, thus the normality assumption is not violated.