**Session 9 DBs**

Differentiate between logistic regression and linear regression. What are the similarities and differences? When is it appropriate to use each type?

**Similarities:**

* Both are forms of supervised learning.
* Both are trying to predict the outcome of a variable based on other related variables.
* Collinearity is an issue for both models.
* Metrics like ANOVA and the F-test are utilized to assess goodness-of-fit.
* Both models can use continuous and discrete variables on the predictor side of the equation.

**Differences:**

* Linear Regression utilizes Least Squares to predict future values.
* Logistic Regression utilizes Maximum Likelihood to predict future values.
* Linear Regression requires that 5 major assumptions be met in order for the model to be valid.
* Logistic Regression is not constrained by linearity for the response variable.
* Logistic Regression predicts log-odds/probability of an outcome.

**Appropriate Usage:**

* When the response variable is not linear (binary), logistic regression is the right model.
* When the 5 major assumptions cannot be assuaged, use logistic regression.
* When there is a continuous response variable, continue the process of building a regression model.
* If the errors are not correlated, continue to use linear regression.

Differentiate between “logit” and “probit.” What are the differences between these two measures? Why are there differences? In what instances do we use each?

Probit and Logit models are very similar. The logit model was developed as a result of the heave calculations required for probit. The difference between these two measures comes down to ease of interpretation as well as the tails of both models. As probability increases, the tail from 99.9% to 99.99% percent becomes rather obscured. In the logit model, the same tail is easier to understand because odds in not bound by a range of 0-1. From researching this topic, I believe it is a personal preference on when to use on model verses the other. I prefer to use probit when the data is normally distributed and my results are not on either tail of the curve. Plus, probability is easier to understand and explain than odds. In the rare case that my data is in the range of 99% and up on the distribution curve, I would use the logit model because it is not bound by a specific range.