BA 464 – Week 5B

Using Regression to Predict Categorical Output

Categorical output

- An output is categorical if it takes values (also called clases) from a finite set.
- If the values also have an ordering, it is called an “ordered categorical variable” or “rank variable”

Binary Output

- Binary output: that takes one of two values.
- Example: TitanicSurvival data set from “effects” library

(Binomial) Logistic Regression

- Linear regression is not designed to predict discontinuous outcome variables.
- Logistic regression relies on a trick: it is linear regression that predicts log(arithm of)-likelihood (probability) of being in one of the two classes in the case of a binary output. When probability is greater than 0.5 it chooses that class.

Quality of model

- Quality of a binomial (or multinomial, as we will see) models are evaluated using a confusion matrix

Proper testing of models

- Using a model to predict for inputs it has already seen is “not fair”. In the real world applications, the model has to be proven by predicting on previously unseen inputs
- Standard practice is to split the data in hand to train, validate, and test sets, usally with a 70/15/15 percent split standard practice.
- Validate set is used to select models trained with the train set. And test set is used to assess errors for the final model.

Multinomial Logistic Regression

- Is an extension of logistic regression idea to predict a categorical output variable that can take more than two values/classes.
- Example: “wine” dataset from “rattle.data” library