Sean Cordrey

October 31, 2014

Stat 3654

Roberts

Weekly Paper 10

Chapter 7 in our R textbook discusses two modelling methods, linear and logistic regression. Linear regression is quite known by anyone who has done a moderate amount of statistics and knows the power behind it. While it is called linear regression, it can be applied to nonlinear relationships and while the model itself may not be usable, we can still get insight from the linear regression to influence what our next steps may be. With the model we can use it to predict values which will assist us in answering questions swayed by what numbers we want. A few ways of validating a linear model involves the determination coefficient R^2 which looks into what percentage of y’s variation is explained by our x factors. We can put into practice our model’s predictive power. Or we can look at the residuals. The residuals not only give us an idea of how good our predictions were but it may also show us some patterns that will help us determine what form the data actually takes.

Logistic regression is also a common method that deals with values between 0 and 1, most notably probabilities. Logistic regression predicts the probability that an instance of a variable belongs to a specific category. An important argument in the R command glm() when constructing our logistic regression is the family argument. Here we specify the distribution our y variable will take. To assess our model we examine the scores given to the different instances where we want high scores for positives and low scores for negatives. This is easy to check by plotting them and looking at the distributions. Ideally there would be two groupings with positives together and negative together at high and low ends respectively. To make this model into a classifier, thresholds need to be set. These thresholds should be made while trying to balance precision and recall.

Both model methods return coefficients that are easily interpreted that help us make decisions in answering our questions of interest. Also worth noting is that these are both methods of prediction and are useful for answering questions that involve future concerns. Another good little thing, is that both of these models are extremely easy to carry out in R so much so that there is no reason not to at least try them even if they don’t fit into the data well.