Sean Cordrey

September 19, 2014

Stat 3654

Roberts

Weekly Paper 4

Chapter 4 is concerned with cleaning and managing data to put the data into a usable form for modeling purposes. A big issue that can arise in datasets is missing values, denoted NA in R. The first thing is to discern what is actually meant by these missing values. Is it a lack of data or is there a meaning to why they are “missing”. It could simply also be an issue with compiling or importing the data and some information was lost in the process. Regardless of the case, what is actually intended needs to be figured out so that the missing values can be treated properly in the modeling. Another thing to do with data is transformed it into more meaningful forms. This could mean turning continuous variables into discrete ones, or applying logarithms to get a better curve. What transformations one may want to apply depends on what kind of information we want to get across in our modeling. The last little section in chapter 4 is about sampling. Although we could work with the entire dataset, this is usually inefficient and often the conclusions drawn from a sample through inference is a very good indicator of the population. A note though is using a method so that it is not a random sample that is drawn every time. We want to draw a sample and then be able to refer to that same sample again. This helps in testing new models. If we know the old model gave us the correct answer for the sample, we can check if the new model does as well. Random samples also can cause issues with coding since the results are random and not reproducible.