Sean Cordrey

September 12, 2014

Roberts

Stat 3654

Weekly Paper 3

Chapter 2 is concerned with importing data, particularly in R. R easily imports data with the read.table() command. If the data is well-structured, variables are clearly defined and values throughout the data hold meaning. The summary() command gives a full overview of the variables and the values taken within them. While well-structured data is nice, the data may not be in such a form and so R has a number of ways of refining the data. Critical to this though is documentation that gives information on what the otherwise meaningless values are, so that we can apply R commands to refine and transform the data. Another issue that may come up is dealing with large amounts of data, too much to be put simply into something like a spreadsheet. In these situations, databases built to hold vast quantities of data are necessary to employ. Dealing with big data is best done in databases and then samples can be drawn from the database into software like R. It’s important throughout the process of working with data that documentation is made of the procedures taken so that they are repeatable and show how one came to the answers they got.

Chapter 3 starts to apply methods of exploring the data now that we have it imported and visualizing it. Like stated above, the summary() commands gives summary statistics for every variable in the data frame. This quick command is important because we can get a good overview of the data and see if there are any values that may cause issues. For example NAs which correspond to missing values and values that are negative when we would expect them to be at least greater than or equal to zero or outliers. Noticing these issues with the summary() command helps us, early on, consider what to do with these values in our modeling. Some pointers for good graphics come from William Cleveland. A couple being that the graphic should try to give as much information as possible with as little strain on the viewer and make the graphic clear so that the data stands out and important trends can be seen. The chapter then goes on to give a number of forms of visualization for one and multiple variables like histograms, bar charts, scatterplots and number of others each with their own pros and cons.