Real-World Data Analysis

Liew How Hui

January 2016

1 Objective and Scope

This project plans to understand how to analyse real-world data or problems using statistical software.

By perform a Google search on "real-world data analysis", a lot of "real-world data" related to clinical data, big data pop up. This project is not going to explore those data but is going to limit the scope to a UTAR data obtain during consultation and data organise from workload problem.

It is important in an information society to be able to process data efficiently with the right software tools. This project is going to contribute to the understanding of real-world data and problems that one will encounter as well as to master the right software tools to process real-world data.

2 Literature Review

A Google search on "real-world data analysis" returned something many links on health study and big data as well as links to Coursera's "Data Science in Real Life by Johns Hopkins University", "Python for Data Analysis", "Practical Data Science Cookbook". We are interested in the later, so we will limit the search to "books on data analysis", which returns a link on "free books for learning data mining and data analysis":

1. Data Jujitsu: The Art of Turning Data into Product (O'Reilly)
2. Data Mining Algorithms in R (Wikibook)
3. Data Mining and Analysis: Fundamental Concepts and Algorithms
4. Introduction to Data Science (Jeffrey Stanton, using R, Creative Common License)
5. Mining of Massive Datasets (www.mdds.org)
6. School of Data Handbook

Note that a book is free does not mean that we can copy and paste from it without permission but we can download and read the electronic copy.

There are also so-called "must-read" (popular) books based on the recommendation of some newspaper comments or Amazon users:

1. Big Data: A Revolution That Will Transform How We Live, Work, and Think
2. The Signal and the Noise: Why So Many Predictions Fail — but Some Don't
3. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die
4. Web Scraping with Python: Collecting Data from the Modern Web

3 Data Analysis Procedure

How could we achieve our objective. We already know that Excel is a basic tool in data analysis. However, Excel is not powerful enough to handle large data. We need more powerful software. An Internet search gave a list of data analysis software below:

- ELKI (Java)
- Weka (Java)
- DataMelt (Java, previously SCaVis)
- ROOT, PAW (Physics Analysis WOrkStation) from CERN, written in C++, Fortran
- R
- Encog Machine Learning Framework
- SPSS (IBM)
- SAS
- Stata

UTAR does have the license for the popular commercial software SAS which can solve statistical and data analysis problems. However, the installation of the software take up a lot of space and the knowledge gained from using it is only good when working in a large corporation.

A better choice is probably a mixed of Excel and open source software for data analytics. Python and R are two very popular data analysis software. Syntax-wise, Python is a better choice.

4 Mathematical Formulas in LaTeX

Linear regression formula Let $\mathbf{y}$ be the respondent variable and $\mathbf{x}$ be the factor. Then ...

$$\mathbf{y} = (A^T A)^{-1} A^T \mathbf{x}$$

The matrix

$$\begin{bmatrix} a_{21} & a_{22} & a_{23} & \\ a_{31} & a_{32} & a_{33} & \\ b_{31} & b_{32} & b_{33} & b_{34} \end{bmatrix} \quad \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

Integration and Differentiation

$$F'(x) = \frac{dF}{dx} = f(x)$$
$$\int_a^x f(x)dx = F(x) - F(a)$$

or special function

$$y = \begin{cases} f_1(x) & a \leq x < c \\ f_2(x) & c \leq x \leq b \end{cases}$$

5 Project Planning

Week	Plans	Complete?
2	• Check out "United Nations Statistics Division" at `http://unstats.un.org/unsd/`	
3	Learn software for gathering and reading data	
4	Use software to collect and gather real-word data	
5	Typing the review of literature and data into the proposal	
6	Clean up and print out the proposal. After discussion with supervisor, start to transfer the results into the interim report LaTeX file.	
7–9	Learn how to use software to clean and format data	
10–11	Type the above research result into the interim report and make a plan for the coming Project II. Discuss with supervisor to correct some mistakes in the report for the submission in Week 12.	