Business Analytics in R: Introduction to Statistical Programming

Timothy Wong
James Gammerman

June 30, 2018

1 Overview

`R` is a very popular computer language for statistical programming. It is widely used in academia and data-centric industries such as medicine, biostatistics, quantitative finance and insurance. In the business analytics community, open-sourced programming languages such as `R` are increasingly accepted as the primary analytical tool.

2 Course Structure

This course begins by introducing the modern `R` ecosystem. This covers the software toolsets, fundamental programming concepts, data types and basic operations. Participants will learn and practice essential data wrangling techniques such as filtering, aggregation and table joining.

In later part of the course, participants learn about how to perform traditional statistical procedures such as Ordinary Least Squares regression in `R`. More advanced statistics and machine learning topics are included, such as tree-based mathods, time series analysis and clustering techniques, etc.

Each module of the course will begin with a short lecture, followed by a series of practical tasks. Participants will spend approximately 50% of their time on practical programming.

3 Learning Outcomes

Upon course completion, all participants will have a good understanding of the `R` ecosystem and be able to perform basic data wrangling tasks. Participants will also be able to build statistical models in `R` using various techniques. Besides, participants will learn about how to compare modelling techniques. This includes understanding the pros and cons of various modelling approaches.

4 Prerequisites

This is a statistical programming course designed for non-`R` users. Participants do not need any prior experience in the `R` language but they are expected to have substantial experience in computer programming. Proficiency in at least one computer programming language is required. In addition, knowledge in statistics is highly beneficial.

5 Preparation

5.1 IT Equipment

Course participants are welcomed to bring along their private laptop. Alternatively, participants can use their corporate device as well. Participants must make sure they have access to the `R` programming environment. Access to corporate systems is ***not*** provisioned as part of course enrolment.

5.2 Training Material

Training materials including the book, slides and datasets are open-sourced. These can be digitally accessed through the URL: `https://git.io/fzxW9`. There is no need to read the materials in advanced. Participants are welcomed to bring along a printed copy of the slides for easier note-taking.

6 Syllabus

R Ecosystem (1 hour)

- Open Source R
- RStudio IDE
- Packages
- Repositories
- User Communities

Programming Concepts (1 hour)

- Data types
- Logical operators
- Control statements
- Code vectorisation
- Subsetting
- Functions

Data Transformation (3 hours)

- The `tidyverse` package
- Filtering
- Aggregation
- Comupting new variables
- Grouping by
- Table joining

Regression Methods (2 hours)

- Linear regression
  - Polynomial regression model
  - Interaction terms
  - Regression table
  - Model diagnostics
  - Overfitting
- Poisson regression
  - Poisson distribution
  - Goodness-of-fit test
- Logistic regression
  - Binomial distribution

  - Odds-ratio

Tree-based Methods (1.5 hours)

- Decision tree
  - Recursive partitioning
  - Tree pruning
  - Visualisation
- Random forest

Neural Networks (1.5 hours)

- Neural activation
- Non-linear activation functions
- Gradient descent methods
- Topologies
- Multi-layer perceptron (MLP)

Time Series Analysis (3 hours)

- Temporal correlation (Correlogram)
  - Auto-correlation function (ACF)
  - Partial auto-correlation function (PACF)
  - Cross correlation function (CCF)
- Decomposition
  - Additive time series
  - Multiplicative time series
  - Time series regression using decomposed components
- Auto-regressive integrative moving average (ARIMA) model
  - Autoregressive (AR) model
  - Moving average (MA) model
  - Stationarity
  - Seasonal ARIMA

Survival Analysis (1.5 hours)

- Kaplan-Meier estimator
- Cox proportional harzard model

Unsupervised Learning (1 hours)

- K-means Clustering
- Agglomerative hierarchical clustering

7 Assessment

This is a non-assessed course but participants should respect each other by paying attention.