JMSCR Vol||07||Issue||03||Page 1194-1197||March 2019

www.jmscr.igmpublication.org
Index Copernicus Value: 79.54
ISSN (e)-2347-176x ISSN (p) 2455-0450
crossrefDOI: https://dx.doi.org/10.18535/jmscr/v7i3.202

IGM Publication
Journal Of Medical Science And Clinical Research
An Official Publication Of IGM Publication

Application of R Software in Life Sciences

Authors
Immad A Shah, Shakeel A Mir, Imran Khan, Nageena Nazir, Owais Bhat, Shakeel Bhat
Sher-e-Kashmir University of Agricultural Sciences and Technology-Kashmir India
Email: *Immad11w@gmail.com*

Abstract
R programming a perfect choice to execute and analyze life science data, as there are vast developers working and coming up with new packages of R programming. R software is worth its popularity worldwide and it is going to scale further. R software allows a wide variety of statistical techniques like classical statistical tests, modeling (linear and nonlinear), classification, time series analysis, cluster analysis, as well as the graphical visualization of data. Besides, R software is highly extensible and an easy to learn language making this software an ideal choice for manipulating big data and life science data.

Introduction

Life sciences and healthcare data are expected to grow exponentially in the future. Life science data sources provide a wealth of information for life science analysts. This gigantic amount of information underpins a wide extent of medical and healthcare capacities, which include disease surveillance, clinical analytics, and clinical decision support. Thus, it is fundamentally vital to procure accessible tools, infrastructures, and techniques so as to use this endless sum of information successfully. To meet these targets R program has been progressively utilized as a device to attain scientific objectives.

A whole list of the most widely used statistical analysis software is provided on several internet sites (Wegman & Solka, 2005). However, R is the only programming language that permits analysts and scientists to perform the foremost complicated analysis. R is free as it is an open source programming language. R was introduced to the world by Ross Ihaka and Robert Gentleman (1996) and was later developed and managed by R-core group members and other scientific contributors. R software syntax codes can be used across all the available platforms viz. Windows, Linux, and Mac. It has all the standard data analysis tools to access data in varied formats, for several data manipulation operations and also includes both conventional as well as modern tools for statistical modeling including ANOVA, Regression, GLM etc. R has some great data visualization tools as well to create graphs, bar charts, multi-panel lattice charts, scatter plots and new custom-designed graphics. Unparallel charting and graphics offered by R language are highly influenced by data visualization experts. Some of the best choices for an introduction to the R software include A Handbook of Statistical Analysis Using R (Everitt and Hothorn, 2006) and

Introduction to Statistics through Resampling Methods and R/S-plus (Good, 2005).

Applications of R Software

1. Use by Researchers and Scientists

R is an extremely prominent language in academia. Various researchers and analysts utilize R for investigating distinctive avenues regarding life science data. R codes used for analysis have been presented in Modern Applied Statistics with S (Venables and Ripley, 2002). Being a favored computer program, R software has advanced a huge pool of individuals who have sound information of R programming, thus leading to skilled statisticians.

2. Use in Data Management and Wrangling:

Data Management is the process of cleaning complex and messy data sets to enable easy and convenient handling of data for further analysis. R has an extensive library of tools for data manipulation and wrangling. Some of the commonly used packages for manipulation of life science data in R include:

- **dplyr ( ):** The dplyr package was created by Hadley Wickham. This package is best known for data transformation and data exploration. Other than, the package encompasses a profoundly adaptive chaining syntax.
- **data.table ( ):** This package of R software allows faster data manipulation and simplifies data aggregation.
- **readr ( ):** The readr package helps in reading various forms of data into R software. It performs a given task at a 10x faster speed by not converting characters into factors.

3. Data visualization and graphics

Data visualization is defined as the visual representation of data in graphical form. The graphical visualization gives an overview of the data structure. R has numerous tools that can help in data visualization and graphics. The R packages "**ggplot2**" and "**ggedit**" are the two standard data visualization packages. The ggplot2 package is centered on visualizing data, while the ggedit package helps users to bridge the gap between making a plot and getting the plot aesthetics precisely correct.

4. Packages useful for health and life science data:

- **gsub ( ):** The gsub package makes cleaning data much simpler when working with vectors and strings.
- **melt ( ):** The melt is a part function of the reshape2 package. It allows pivot-table style options to restructure data without losing values.
- **Sqldf ( ):** The sqldf package makes data analysis convenient. For instance, the sqldf package allows getting the average estimate for heart attack patients by state.
- **MILC ( ):** The MICL stands for "Microsimulation Lung Cancer" model. It is used for the prediction of patient outcomes and describing the natural history of lung cancer.
- **Epi Model ( )**: The EpiModel is a package for mathematical modeling of infectious diseases, which includes stochastic agent-based models, stochastic network models, and deterministic compartmental models.
- **Survival ( )**: This package of R programming is applicable in survival analysis and comprises survival analysis functions, including Kaplan-Meier, multi-state curves and Cox models.
- **Flexsurv ( )**: The Flexurv package is used for Flexible Parametric Survival and Multi-State Models for time-to-event data, including spline models.
- **Mstate ( ):** This package is used in data preparation, estimation, and prediction in Multi-State Models. The package includes functions for data preparation, descriptives, and hazard prediction.
- **TP msm ( ):** The TPmsm package is used in the estimation of transition probabilities in Multistate Models. It also allows for the estimation of transition probabilities for

illness-death or three-state progressive disease model.

- **Ipw ( ):** The Ipw package allows to estimate Inverse Probability Weights which includes functions to estimate the probability to receive the observed treatment, based on individual characteristics.
- **RIS med ( ):** The R package comprises of a set of tools to extract bibliographic content from the National Center for Biotechnology Information (NCBI) database directly.
- **HMDHFD Plus ( )**: This package allows to read the data from the Human Fertility Database (HFD) and Human Fertility Database (HFD) into an R session as data. frame ( ) objects.
- **Demography ( ):** The demography package in R allows forecasting mortality, fertility, migration and population data Human Mortality Database (HMD). It also comprises of the functions for demographic analysis including, life table calculations.
- **SEER2R ( ):** This package of R allows reading and writing of surveillance, epidemiology, and end results (SEER) datasets.
- **Genalg ( )**: This Genalg is an R Based Genetic Algorithm optimization and is of great used in genetic optimization studies.

5. Execution in R: To start with a CRAN mirror is specified. CRAN is mirrored on more than 80 registered servers.

To install packages in R, the following syntax is used:

> install. packages ("package name")

> install. packages ("fortunes")

R gives some information on the installation of the package:

```
Installing package(s) into 'D:/R/library'(as 'lib' is unspecified)
....
opened URL
downloaded 165 Kb
package 'fortunes' successfully unpacked and MD5 sums checked
....
```

The library is the directory where the packages are installed. After the packages are installed they are loaded into R console using the load syntax as:

> load ("package name")

To unload a package the detach() function is used, but the package to be detached needs to be specified.

> detach ("package name")

Conclusion

A plethora of uses for the software package R, and specifically on its helpful applications in life science data analyses are presented. The packages applicable in the field of life sciences are given along with their applications. The great advantage of the R software package is its capacity to adjust to the ever-changing needs of the software client. Through the coordinated effort of programming developers and freeware, R is equipped for addressing the necessities and filling the specialties of several separate software packages while remaining exceptionally cost-effective.

References

1. R. Ihaka and R. Gentleman.1996.R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5:299–314.
2. Wegman EJ, Solka JL.2005.Statistical software for today and tomorrow.

X[http://binf.gmu.edu/~jsolka/PAPERS/ess2542_rev1.pdf, Accessed on November 11, 2010.

3. Everitt, B.S. and Hothorn, T.2006. A Handbook of Statistical Analysis Using R. Chapman and Hall, Boca Raton, FL.
4. Good, P.I.2005. Introduction to Statistics through Resampling Methods and R/S-PLUS. John Wiley and Sons, Inc., Hoboken, NJ.
5. Venables, W.N. and Ripley, B.D. 2002. Modern Applied Statistics with S. Springer, New York, NY.