Journal homepage:
www.jmisci.com
Contact:
jmiseditor@gmail.com

Journal of Military and Information Science
Institute of Science

Editorial

De Facto Language of Data Science: The R Project

Kerim Goztepe

Dear Readers,

This is the last issue of the year 2016. Another year has come to an end. I want to confess that this year has not been as successful as we expected.

I always support science for humanity and environment. Besides, I dream the world to be a more livable place. But my dreams may be another topic for another paper. Before you get bored, I want to talk about the R Project, whose popularity is increasing every year, in this letter.

As you know, data science has taken the world by storm. Firms, people, states produce a great amount of data every day (Berry and Linoff, 1999). Every field of science and business realizes the value of the incredible quantities of data (Lumley et al., 2002; Chen et al., 2012). The matter is not to produce data, it is to extract value from those data. A data analysis must have data science abilities in order to cope with validation of data. Here, the importance of R appears. Most of the statisticians, optimizers, and data analyzer are well aware of the R programming language (Muenchen, 2012). They probably confess, it has become the de facto programming language for data science. Because of its flexibility, it is sophisticated, highly configurable, and has no cost (open source). Plus, many scientists not only use it but also, contribute to amazing R environment for better problem solutions. With R you can It is possible to write functions, do sophisticated calculations, create simple or complicated graphs, use almost any available statistical techniques and even write your own scripts for a purpose. A great number of researchers supports R and many research institutes, companies, and universities have migrated to R.

It is not easy for me to mention or explain the R project in a letter. But I can refer a fast introduction to the R and give you some references about it. Below you can follow a basic introduction to R step by step.

Step 1. Do not hurry

R is not so easy to learn. But once you learn, it opens a new horizon to you. Do not hurry. Be patient and learn every day some issue about it. Many researchers see the R as a one of the main language of science.

Step 2. Visit R website

If you are interested in the R project, you should visit the official website of R first and examine the eco-system of R. Official website of R is https://www.r-project.org/about.html. R is not only a software for statistics. It provides a platform for many different implementations, it has fuzzy logic, neural network, data mining, image processing tools for instance. R can be extendable with different packages.

Step 3. Install R and read basic documents

You may download R for Windows, Linux and Mac OS X operating systems using https://cran.r-project.org/mirrors.html website. Open this site and select a mirror for downloading R. Read basic R documents after installation. Do not give up reading and understand R idea. Some of my literature recommendations are given below(Table.1).

Table.1. Some references for beginners

Author	Reference	Level
Paradis, E. (2002)	R for Beginners.	Beginner
Matloff, N. (2009)	The art of R programming.	Beginner, Intermediate

Chang, W. (2012)	*R graphics cookbook*. " O'Reilly Media, Inc.".	Intermediate, Advanced
Stowell, S. (2014)	*Using R for statistics*. Apress.	Intermediate, Advanced
Venables, W. N., Smith, D. M., & R Development Core Team. (2004)	An introduction to R.	Beginner
Gentleman, R., Hornik, K., & Parmigiani, G. (2009)	Use R!.	Beginner, Intermediate

Web sites	**Info**
https://www.r-project.org/	Official R Project site
https://www.rdocumentation.org/	Documents about R usage, manuals etc.
https://www.r-bloggers.com/	Best blogger site for R
http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html	Supports data for R
https://journal.r-project.org/	R Journal
http://adv-r.had.co.nz/	For advanced R users.

Step 4. Do not hesitate to use R

You can start using soon after installation R. Read some “immediate starting references” and use examples given in them. There are many examples in related books from simple to advanced. I tried to give some basic samples below.

Sum of two plus two,

```
> 2+2

[1] 4
```

Suppose you want to calculate the logarithm of 3 with base 10. You may type

```
> log(3)

[1] 1.098612
```

Assign a value to q,

```
> q <- 12*5+3

> q
[1] 63
```

Combine values into a vector x,

```
> x <- c(1,2,3,4)
> x
[1] 1 2 3 4
> help(c) #if you want to learn to combine command
(c) use help
```

Let say y= x^2 and we request graphic of it,

```
> y <- x^2

> y
[1]  1  4  9 16
> plot(y, col=34, lwd=2, pch=10) #this command
gives you graphics of y= 1,4,9,16 (Fig.1)
```

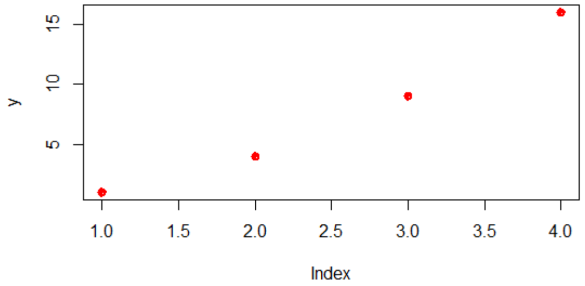

Fig.1. Plot of y.

Step 5. Use datasets during learning phase

You can apply ready to use preinstalled datasets in R. Especially in the learning phase, you do not need to find data. I used “women” data here.

```
> data("women")

> women
  height weight
1     58    115
2     59    117
3     60    120
4     61    123
5     62    126
---------------# there are 15 data in this sample.
```

Lets' use data “women” for linear regression.

```
>names(women)   #use it to get or set the names of
an object.
[1] "height" "weight"
> dim(women)   #gives dimension of data. There are
15 rows, 2 columns in data “women”
[1] 15  2
> x←women$height #height values copied in x
> y←women$weight #weight values copied in y
> lm(x~y) # create linear model for x and y
# The simple linear regression model is
```

$y=\beta_0+\beta_1 x_1+\varepsilon_i$ where β_0 is the intercept and β_1 is the slope of the linear relationship assumed between the response and explanatory variables and ε_i is an error term.

```
Call:
lm(formula = x ~ y)

Coefficients:
(Intercept)          y
   25.7235      0.2872
>mymodel←lm(x~y) # output of lm is copied to
"mymodel"
> summary(mymodel) # this gives you summary of
"mymodel"

Call:
lm(formula = x ~ y)
Residuals:
     Min       1Q   Median       3Q      Max
-0.83233 -0.26249  0.08314  0.34353  0.49790

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.723456   1.043746   24.64 2.68e-12 ***
y            0.287249   0.007588   37.85 1.09e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.44 on 13 degrees of freedom
Multiple R-squared:  0.991,	Adjusted R-squared:  0.9903
F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

Our model explains if there is a relation between women height and weight. We can obtain some plot of data as seen below (Fig.2,Fig.3, Fig.4, Fig.5)

```
> plot(height ~ weight, data = women)  # This
produces a scatterplot of velocity and distance
as seen Fig.2
```

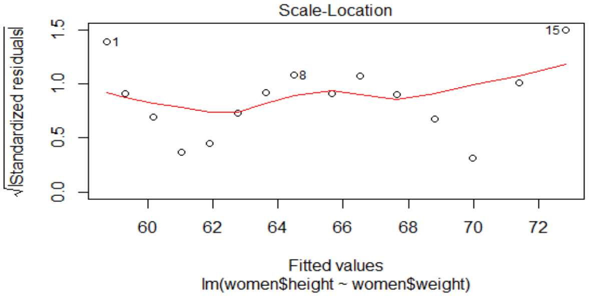

Fig.2. Scatterplot of fitted values.

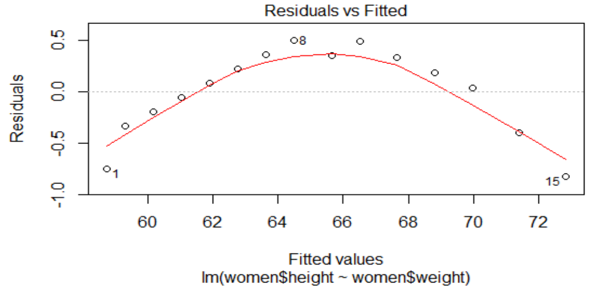

Fig.3. Plot of residuals against fitted values.

You can try other ways to obtain plot of your study.

```
>plot(height, weight, main="Scatterplot Example",
+     xlab="Weight ", ylab="Height", col=34, pch=19)
> abline(lm(height~weight), col="red") # regression line (height-weight) data is women.
> lines(lowess(height, weight), col="blue") # lowess line (height-weight)
```

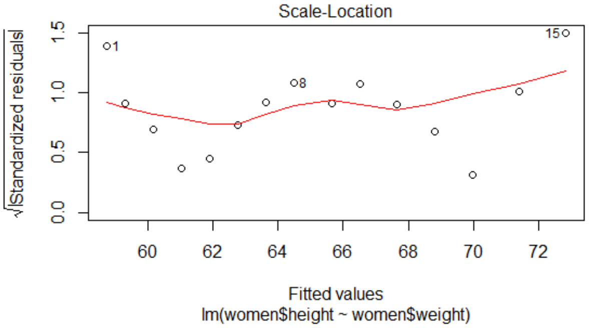

Fig.4. Fitted values.

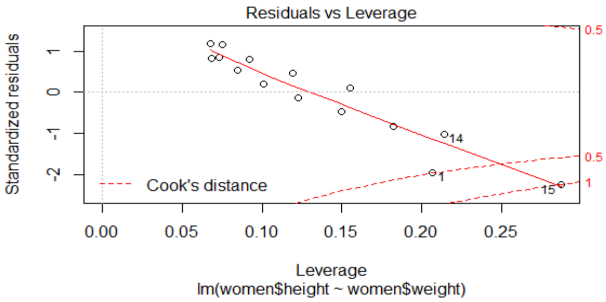

Fig5. Index plot of Cook's distances for data.

I tried to write a fast beginning letter for the people who are interested in R language.

I would like to thank our authors and reviewers. I believe that none of our successes would have been possible if their efforts had not been of the authors who submitted their quality papers.

I thank our reviewers who tirelessly supervised the review process and, on occasions, provided me with great suggestions and advice. Prof.Dr. Cengiz Kahraman, Prof.Dr. Şeref Sağıroğlu, Prof.Dr. Orhan Torkul, Dr. Alper Kayaalp and Serhan Ateş deserve special thank for their valuable support.

I look forward to another successful year; meanwhile, please feel free to contact me with your suggestions and comments.

Sincerely,

Kerim Goztepe, IE, Ph.D
Editor-in-Chief
Journal of Military and Information Science

References

Books&Articles

Berry, M., & Linoff, G. (1999). Mastering data mining: The art and science of customer relationship management. John Wiley & Sons, Inc..

Chang, W. (2012), R graphics cookbook. " O'Reilly Media, Inc.".

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS quarterly, 36(4), 1165-1188.

Matloff, N. (2009), The art of R programming.

Muenchen, R. A. (2012). The popularity of data analysis software. UR L http://r4stats. com/popularity.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. Annual review of public health, 23(1), 151-169.

Paradis, E. (2002), R for Beginners.

Stowell, S. (2014), Using R for statistics. Apress.

Venables, W. N., Smith, D. M., & R Development Core Team. (2004), An introduction to R.

Web sites

Best blogger site for R, https://www.r-bloggers.com/

Documents about R usage, manuals etc.,https://www.rdocumentation.org/

Official R Project site, https://www.r-project.org/

R Journal, https://journal.r-project.org/

Supports data for R, http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html

For advanced R users., http://adv-r.had.co.nz/