ISSN: 1859-2171
e-ISSN: 2615-9562
TNU Journal of Science and Technology 225(04): 107 - 112

KHAI THÁC ỨNG DỤNG CỦA PHẦN MỀM R TRONG GIẢNG DẠY ƯỚC LƯỢNG VÀ KIỂM ĐỊNH TRONG Y, DƯỢC HỌC TẠI TRƯỜNG ĐẠI HỌC Y DƯỢC – ĐẠI HỌC THÁI NGUYÊN

Đỗ Thị Phương Quỳnh*, Nguyễn Thị Tân Tiến, Lê Thị Oanh
Trường Đại học Y Dược - ĐH Thái Nguyên

TÓM TẮT

Phần mềm R là một phần mềm mã nguồn mở, có nhiều ứng dụng tốt cần được khai thác. Hiện tại phần mềm này đang được rất nhiều người quan tâm tìm hiểu. Vì vậy bằng cách nghiên cứu và khai thác sâu các ứng dụng của phần mềm R, đặc biệt khai thác nhiều ứng dụng vẽ biểu đồ minh họa. Bài báo đã giới thiệu một cách ngắn gọn cách cài đặt phần mềm R với mã nguồn mở, đồng thời thiết kế một mẫu các câu lệnh liên tiếp của phần mềm R tạo thành chuỗi tư duy logic của việc sử dụng phần mềm R trong việc giảng dạy ước lượng và kiểm định, tại trường Đại học Y Dược – Đại học Thái Nguyên.

Từ khóa: *Phần mềm R; cài đặt R; ứng dụng R trong ước lượng; ứng dụng R trong kiểm định, biểu đồ.*

Ngày nhận bài: 23/10/2019; Ngày hoàn thiện: 28/4/2020; Ngày đăng: 28/4/2020

APPLICATION OF SOFTWARE R IN TEACHING ESTIMATION AND HYPOTHESIS TESTING IN MEDICINE AND PHARMACY AT UNIVERSITY OF MEDICINE AND PHARMACY - TNU

Do Thi Phuong Quynh*, Nguyen Thi Tan Tien, Le Thi Oanh
TNU – University of Medicine and Pharmacy

ABSTRACT

R Software is open source software, there are many good applications that need to be exploited. Today, this software is interested by many people. So by studing and exploiting deeply applications of R software, specially exploiting many application of illustrating chart. The article introduced how to install the software with open source and designed a sequence of successive statements of the software and formed a logical thinking sequence in teaching estimation and hypothesis testing in Thai Nguyen university of medicine and pharmacy.

Keywords: *Software R; settings R; applications in estimation; hypothesis testing, chart.*

Received: 23/10/2020; Revised: 28/4/2020; Published: 28/4/2020

* Corresponding author. *Email: phuongquynhtn@gmail.com*

1. Giới thiệu tổng quan

Phân tích số liệu và biểu đồ thường được tiến hành bằng các phần mềm thông dụng như SAS, SPSS, *Stata, Statistica* và *S-Plus*. Đây là những phần mềm được các công ty phần mềm phát triển và giới thiệu trên thị trường khoảng ba thập niên qua, và đã được các trường đại học, các trung tâm nghiên cứu và công ty kỹ nghệ trên toàn thế giới sử dụng cho giảng dạy và nghiên cứu. Nhưng vì chi phí để sử dụng các phần mềm này tương đối đắt tiền (có khi lên đến hàng trăm ngàn đô-la mỗi năm), một số trường đại học ở các nước đang phát triển (và ngay cả ở một số nước đã phát triển) không có khả năng tài chính để sử dụng chúng một cách lâu dài. Do đó, các nhà nghiên cứu thống kê trên thế giới đã hợp tác với nhau để phát triển một phần mềm mới, với chủ trương mã nguồn mở, sao cho tất cả các thành viên trong ngành thống kê học và toán học trên thế giới có thể sử dụng một cách thống nhất và hoàn toàn miễn phí.

Năm 1996, trong một bài báo quan trọng về tính toán thống kê, hai nhà thống kê học Ross Ihaka và Robert Gentleman thuộc trường đại học Auckland, New Zealand phác hoạ một ngôn ngữ mới cho phân tích thống kê mà họ đặt tên là R. Sáng kiến này được rất nhiều nhà thống kê học trên thế giới tán thành và tham gia vào việc phát triển R. Cho đến năm 2006, qua chưa đầy 10 năm phát triển, càng ngày càng có nhiều nhà thống kê học, toán học, nghiên cứu trong mọi lĩnh vực đã chuyển sang sử dụng R để phân tích dữ liệu khoa học. Trên toàn cầu, đã có một mạng lưới hơn một triệu người sử dụng R [1].

Chính vì tính ưu việt của phần mềm R nên chúng ta cần nghiên cứu ứng dụng của phần mềm trong giảng dạy thống kê y sinh học nói chung và việc dạy học phần ước lượng và kiểm định tại trường Đại học Y Dược Thái Nguyên nói riêng.

2. Phương pháp nghiên cứu

Để giải quyết được vấn đề trên, tác giả đã nghiên cứu chi tiết từ cách cài đặt đến cách sử dụng từng câu lệnh của phần mềm R, đặc biệt quan tâm đến phần biểu đồ. Sau đó tác giả kết hợp các câu lệnh phù để tạo ra các mã mới sử dụng trong việc giảng dạy phần ước lượng và kiểm định tại trường Đại học Y Dược – Đại học Thái Nguyên (ĐHTN).

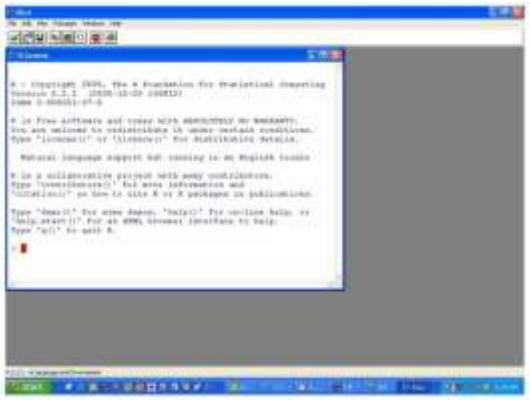
2.1. Cài đặt phần mềm R

Bước 1: Truy cập vào trang chủ: https://cran.r-project.org/, click tiếp vào một trong 3 dòng sao cho phù hợp với máy tính:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

Tiếp tục chọn install R for the first time.

Sau khi cài đặt (set up), tạo icon R để chạy phần mềm. Mở phần mềm R xuất hiện cửa sổ lệnh (hình 1). Ngoài ra chúng ta có thể cài đặt thêm phần mềm R studio có thêm tính năng xuất bản.



Hình 1. *Cửa sổ lệnh R*

Khi cài đặt xong ta có thể tính toán một số hàm đơn giản trong R ví dụ như hàm tính giá trung bình, tính phương sai…

Bước 2: Để phục vụ cho thống kê (cụ thể ước lượng và kiểm định) ta cài thêm gói hỗ trợ *BSDA* dùng để tính ước lượng và kiểm định giả thuyết và gói lm cho phần phân tích hồi quy tuyến tính.

Để cài được hai gói này các bạn [2] chọn **Install packages trên thanh công cụ** trong **packages** của R. Chọn *BSDA* và *lmodel2* (viết tắt từ **l**inear **m**odel) cho phân tích hồi quy tuyến tính. Hoặc đánh trực tiếp lệnh

>install.packages("BSDA");
>install.packages("lmodel2").

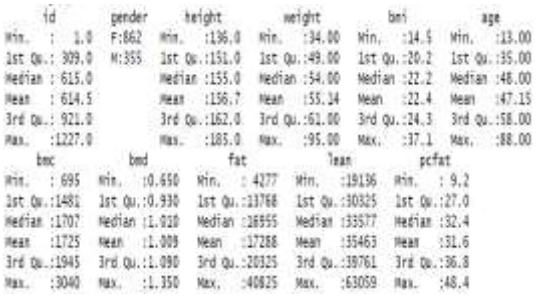
2.2. Khai thác các ưu điểm của phần mềm R trong giảng dạy phần ước lượng và kiểm định tại trường Đại học Y Dược - ĐHTN

Đối với người làm thống kê, chúng ta thường xử lý dữ liệu theo hai bước, bước 1 là thống kê mô tả, bước 2 thống kê suy diễn. Việc sử dụng phần mềm R cho hai bước này rất hữu hiệu vì ngoài các ưu điểm đã kể trên như phần mềm có mã nguồn mở và cách sử dụng khá thân thiện chúng ta còn thấy ưu điểm khác như:

+ Cách nhập dữ liệu trong R có thể nhập trực tiếp và ưu việt hơn cả là R có thể đọc được các dữ liệu từ Excel, từ Stata. Trong phạm vi bài báo này tác giả sẽ sử dụng nguồn dữ liệu đáng tin cậy bằng phương pháp điều tra [3].

+ Sử dụng biểu đồ mà R biểu diễn rất rõ, chính xác và đẹp mắt để chúng ta dễ dàng suy diễn được tổng thể nghiên cứu [4].

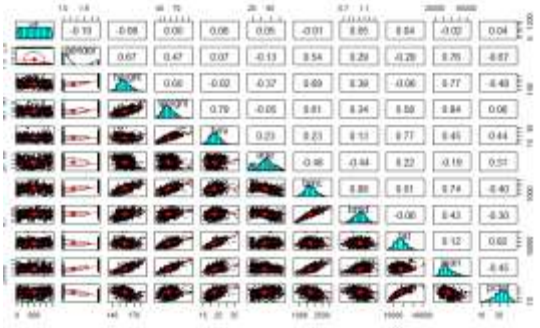
Thông qua dữ liệu trên tác giả đã sử dụng nhiều ứng dụng của phần mềm R trong giảng dạy phần ước lượng và kiểm định. Trước tiên dùng lệnh: >dim(Q) để thấy dữ liệu gồm 1217 hàng và 11 cột. Sau đó dùng lệnh >View(Q); >summary(Q) với 2 câu lệnh này cho ta cái nhìn tổng quan về toàn bộ dữ liệu như hình 2.



Hình 2. *Tổng quan dữ liệu*

Với kết quả này cho ta toàn cảnh về dữ liệu, đây là điều rất cần thiết cho người học: Tổng điều tra có 862 nữ và 355 nam. Dữ liệu cho biết các số đo của height (chiều cao); weight (trọng lượng); bmi (tỷ trọng cơ thể); age (tuổi); bmc (khối lượng xương); bmd (mật độ chất khoáng trong xương); fat (khối lượng mỡ); lean (lượng cơ); pcfat (tỷ trọng mỡ toàn thân). Min là giá trị nhỏ nhất, Max là giá trị lớn nhất, 1st Qu= Q1 là giá trị mà 25% số liệu nhỏ hơn Q1; 3st Qu= Q3 là giá trị mà 75% số liệu nhỏ hơn Q3; Từ đó ta có thể suy ra 50% số liệu sẽ nằm trong khoảng tứ phân vị (Q1;Q3).

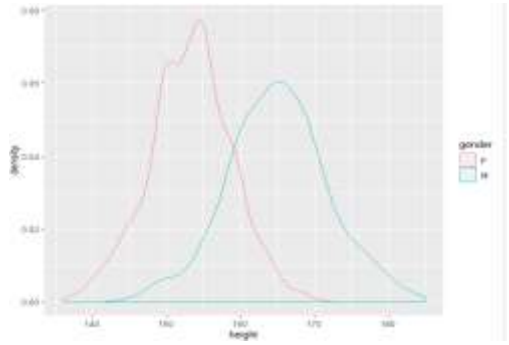
Để tiếp tục khai thác được ứng dụng R chúng ta cài đặt gói lệnh psych (sử dụng lệnh >install.packages("psych")), sau đó gọi gói lệnh đã cài (sử dụng lệnh >library(psych)), dùng lệnh: >pairs.panels(Q) (ta được hình 3), hình này cho ta cách nhìn tường minh về dữ liệu mình đã thu thập được: Các biến heigh, weight, bmi, bmc, bmd, fat có biểu đồ hình chuông cân đối vậy ta dự đoán các biến này tuân theo quy luật phân bố chuẩn. Các biến còn lại có dáng biểu đồ không giống hình chuông nên có thể không tuân theo quy luật phân bố chuẩn. Để khẳng định chính xác xem biến ngẫu nhiên có tuân theo quy luật phân bố chuẩn hay không ta có thể dùng lệnh >Shapiro.test, với lệnh này p – value >0,05 thì biến đó được coi là có phân bố chuẩn [5].



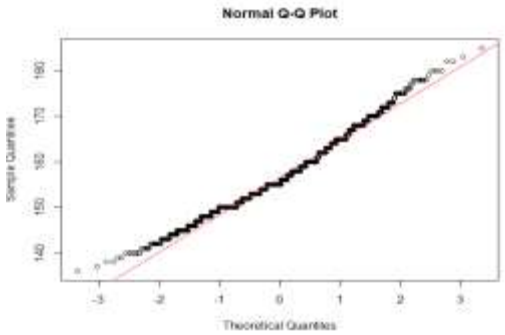
Hình 3. *Mối tương quan giữa các đại lượng nghiên cứu*

Qua hình 3 chúng ta cũng thấy các đám mây dữ liệu gần giống hình chữ nhật dẹt, đồng nghĩa là các biến tương ứng sẽ có mối tương quan tuyến tính với nhau. Trong phạm vi bài báo này đã phân tích chủ yếu 3 sau: biến chiều cao (height), cân nặng (weight), tỷ trọng cơ thể (bmi). Để tách dữ liệu làm hai nhóm nam và nữa riêng chúng ta sử dụng lệnh [5]: >nam<-subset(Q, Q$gender=='M') và >nu<-subset(Q, Q$gender=='F'). Để vẽ được biểu đồ thể hiện dữ liệu về chiều cao cho cả 2

nhóm nam và nữ trên cùng biểu đồ ta sử dụng tổ hợp lệnh: >library(ggplot2);>DT<-ggplot(data = Q, aes(x = height, color = gender)) + geom_density();DT.



Hình 4. *Biểu đồ mô tả phân phối chiều cao của nam và nữ*



Hình 5. *Chiều cao của nam*

Nhìn hình 4 cho chúng ta thấy chiều cao của nam và nữ tuân theo quy luật phân phối chuẩn, chiều cao trung bình của nữ rơi quanh giá trị 155 cm và chiều cao trung bình của nam rơi quanh khoảng 165 cm.

Hoặc thông qua hàm hàm lệnh > qqnorm(Q$height); >qqline(Q$height,col=2) (ta được hình 5) cho thấy giá trị quan sát về chiều cao của mẫu trên (các điểm trên biểu đồ) rất gần với giá trị kỳ vọng của quy luật phân bố chuẩn (là đường màu đỏ). Tương tự như vậy chúng ta cũng có thể kiểm định các biến khác xem có tuân theo quy luật phân phối chuẩn hay không?

Tiếp tục nghiên cứu về số lượng nam nữ thừa cân, béo phì chúng ta dùng tổ hợp lệnh:

>table(cut(nam$bmi,breaks = c(0,18.5,25,40), include.lowertail=TRUE))

>plbmi<-cut(nam$bmi,breaks = c(0,18.5,25,40),include.lowertail=TRUE,labels = c("nam thieu can"," nam binh thuong","nam beo phi"))

>pie(x1)

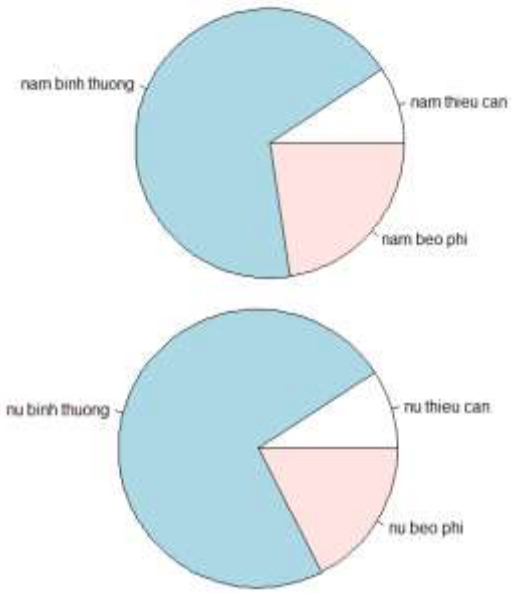
>table(cut(nu$bmi,breaks= c(0,18.5,25,40),include.lowertail=TRUE))

>plbminu<-cut(nu$bmi,breaks = c(0,18.5,25,40),include.lowertail=TRUE,labels = c("nu thieu can","nu binh thuong","nu beo phi"))

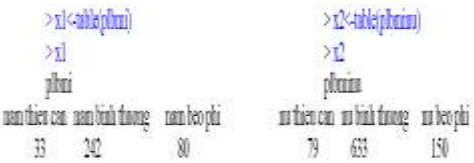
>summary(plbminu)

>pie(x2)

Kết quả thu được:



Hình 6. *Biểu đồ thể hiện tỷ lệ phân loại của nam, nữ theo chỉ số BMI*



Hình 7. *Bảng phân loại tần số của nam, nữ theo chỉ số BMI*

Qua hình 6 và hình 7 bước đầu chúng ta nhận định được có thể tỷ lệ nữ béo phì hoặc thừa cân là thấp hơn nam, để khẳng định được chắc chắn vấn đề đó, chúng ta sẽ ước lượng

khoảng tin cậy 95% cho tỷ lệ béo phì và thừa cân của 2 nhóm này bằng lệnh [6],[7]: >install.packages("rms"); >require(rms);

>binconf(x=80,n=355,method = "all"); >binconf(x=150,n=862,method = "all")

Ta sẽ thu được kết quả sau

> binconf(x=150,n=862,method = "all")

Kết quả ước lượng khoảng tin cậy 95% của tỷ lệ béo phì của nam, nữ như sau

	PointEst	Lower	Upper
Exact	0.1740139	0.1492739	0.2010049
Wilson	0.1740139	0.1501662	0.2007543
Asymptotic	0.17401	0.1487050	0.1993228

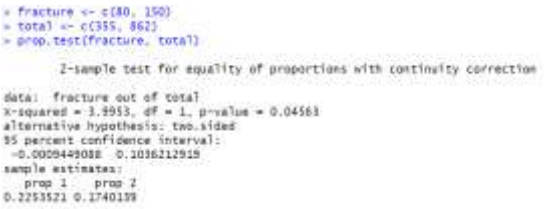
> binconf(x=80,n=355,method = "all")

	PointEst	Lower	Upper
Exact	0.2253521	0.1829302	0.2724222
Wilson	0.2253521	0.1849629	0.2716216
Asymptotic	0.2253521	0.1818894	0.2688148

Hình 8. *Kết quả ước lượng khoảng của tỷ lệ béo phì*

Chúng ta thấy kết quả phân tích theo phương pháp chính xác thì khoảng ước lượng cho tỷ lệ béo phì của nữ là: (0,1493;0,2010) và của nam là (0,1829;0,2724). Vậy câu trả lời tương đối rõ ràng rằng tỷ lệ béo phì của nam cao hơn nữ, với tình trạng thiếu cân ta cũng sử dụng lệnh tương tự.

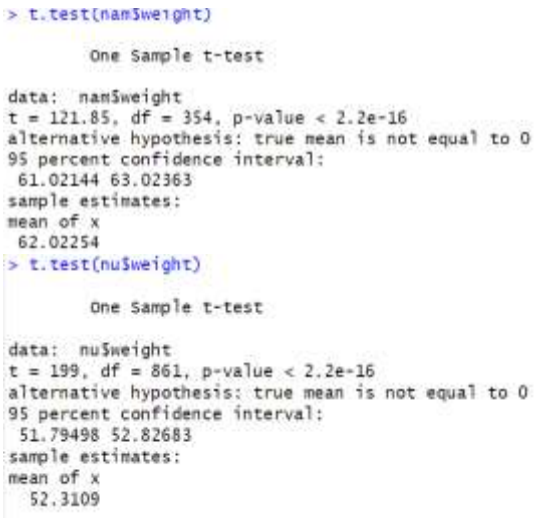
Qua hình 8 chúng ta thấy 2 khoảng ước lượng cho tỷ lệ béo phì của nữ là: (0,1493;0,2010) và của nam là (0,1829;0,2724), hai khoảng này không có sự khác biệt lắm, vậy một câu hỏi đặt ra là tỷ lệ béo phì của nam và nữ có khác biệt không với mức ý nghĩa 95% cho biết ý kiến trên có chấp nhận được không? Để trả lời cho câu hỏi này chúng ta dùng tổ hợp lệnh sau:



Hình 9. *Kết quả của sự khác biệt tỷ lệ béo phì ở nam và nữ*

Kết quả phân tích ở hình 8 cho chúng ta thấy tỷ lệ béo phì của nam là 0,2254, tỷ lệ béo phì ở nữ là 0,174. Phân tích trên cho thấy với độ tin cậy 95% độ khác biệt giữa nam và nữ là 0,0009 đến 0,1 (tức 0,09% đến 10%), với trị số p-value = 0,04<0,05 ta có thể nói tỷ lệ béo phì của nữ thấp hơn tỷ lệ béo phì của nam. Với tỷ lệ bình thường hoặc thiếu cân ở nam và nữ ta có thể kiểm định tương tự.

Thêm một ứng dụng nữa của R khi giảng dạy ước lượng kiểm định là sau khi chúng ta nhận định được các biến nghiên cứu trong mẫu, biến nào có quy luật phân phối chuẩn rồi, chúng ta sẽ sử dụng lệnh [8] >t.test() để tính khoảng tin cậy và kiểm định. Ví dụ chúng ta muốn tính khoảng tin cậy 95% cho giá trị trung bình trọng lượng của nam trong dữ liệu trên ta thu được kết quả như hình 10.



Hình 10. *Kết quả ước lượng khoảng tin cậy 95% cho trọng lượng trung bình của nam, nữ*

Vậy với độ tin cậy 95% trọng lượng trung bình của nam rơi vào khoảng [9]: (61,0214;63,0236) và trọng lượng trung bình của nữ rơi vào khoảng (51,795; 52,8268).

Trong phạm vi bài báo này đã khai thác được một số ứng dụng của phần mềm R trong quá trình giảng dạy phần ước lượng và kiểm định tại trường đại học Y Dược - ĐHTN, và còn rất nhiều ứng dụng khác của phần mềm R có

thể làm cho bài giảng của chúng ta sinh động hơn, và người học dễ hình dung, tiếp cận cũng như xử lý dữ liệu một cách tường minh. Tác giả cũng rất mong sự đóng góp ý kiến của bạn đọc để có thể khai thác được nhiều hơn ứng dụng của phần mềm R trong quá trình giảng dạy học phần này.

3. Kết quả và luận bàn

Như vậy qua nghiên cứu tác giả đã sử dụng biểu đồ để minh họa được một đại lượng ngẫu nhiên tuân theo quy luật phân phối chuẩn và không chuẩn cũng như sử dụng các câu lệnh kiểm tra quy luật phân phối của một đại lượng ngẫu nhiên bất kỳ. Ứng dụng này đã giúp cho người học không thấy mơ hồ về phân phối của đại lượng ngẫu nhiên. Bằng việc sử dụng nhiều tổ hợp lệnh khác nhau đã cho tác giả một mã mới trong việc phân tích dữ liệu sử dụng cho giảng dạy phần kiểm định và ước lượng trong thống kê y sinh học.

Trong phạm vi của bài báo này tác giả chỉ nói đến phần ước lượng và kiểm định, hy vọng trong thời gian tới tác giả có thể xây dựng được các mã mới để phục vụ cho việc giảng dạy phần tương quan hồi quy trong giảng dạy phần thống kê y học.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1]. V. T. Nguyen, *Data analyze with question and answers*. Publishing company Ho Chi Minh city, 2018.

[2]. T. H. Dang, *Statistics for social sciences and life sciences with R software*. Publishing company Ha Noi University, 2019.

[3]. V. T. Nguyen, "Data analysis and application," University pharmacy Hanoi, 2019. [Online] Available: http://www.hup.edu.vn/cpbdv/pcntt/noidung/SiteAssets/Lists/huongdanvecntt/NewForm/Datasets%20for%20practice.zip. [Accessed Jan. 2020].

[4]. V. T. Nguyen, *Data analysis and chart R*. Garvan Institute of Medical ResearchSydney, Australia, 2103.

[5]. J. Veani, "Simple R-Using R for Introductory Statistics," 2001. [Online]. Available: https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf. [Accessed Jan. 2020].

[6]. E. Paradis, "R for Beginners," 2005. [Online]. Available: https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf. [Accessed Jan 2020].

[7]. J. H. Maindonald, "Using R for Data Analysis and Graphics," Australian National University, 2008. [Online]. Available: https://cran.rproject.org/doc/contrib/usingR.pdf. [Accessed Feb. 2020].

[8]. M. Staniak, and P. Biecek, "The landscape of R packages for automated exploratory dât Analysis," *The R Journal*, vol. 11, no. 2, pp. 347-369, 2019.

[9]. W. Djatschenko, "An R package for fixed Coupon Bond Analysis," *The R Journal*, vol. 11, no. 2, p.124, 2019.