Using R/R Studio In An Introductory Statistics Course

Nkechi Agwu and Piotr Bialas

nagwu@bmcc.cuny.edu; pbialas@bmcc.cuny.edu

Borough of Manhattan Community College (BMCC)

City University of New York (CUNY)

Abstract: *In this paper we will share our experiences and approaches, based on* ***GAISE***[1] *and* ***Project Mosaic***[2] *to teaching Introductory Statistics class to students with no programming background. We think that R, open-source, and free programming language can be used as a "calculator" in teaching/discovering various statistical concepts in the realm of descriptive statistics and simple linear regression. Several examples of simple homework assignments will be presented; in addition, exemplary student report will be displayed.*

INTRODUCTION

In July 2014 one of the authors attended International Conference on Teaching Statistics, (ICOTS 9), held in Flagstaff, Arizona and enrolled in workshop *Statistics Using R and RStudio.* The workshop was created and conducted by Randall j. Purim, Nicholas J. Horton, and Daniel T. Kaplan and included introduction to R Studio, Mosaic, and R Markdown.

Two half-day sessions convinced attendant about necessary changes in ways of teaching an Introductory Statistics course. Main change involved using different kind of technology, different from Texas Instrument graphing calculators, Excel, SPSS, and even SAS packages.

R and R Studio computer application programs were on desk to be tried.

WHY R AND RSTUDIO …?

Both computer application programs are **free** and open source entities; so, no more complaints from students about additional monetary expenditure for technology which in some cases was thought to be useless after completing the course.

[1] College Report 2016, (GAISE), Guidelines for Assessment and Instruction in Statistics Education

[2] *Randall J. Pruim; Nicholas J. Horton; Daniel T. Kaplan;* Project MOSAIC, *Start Teaching with R; Preliminary Edition*

R and R Studio were considered suitable for use because R seems to be intuitive and efficient tool in conjunction with R Studio when used in teaching the Introductory Statistics Course.

Moreover, R as an open source program can be used along with various packages to manipulate data and construct graphical displays that were not easily or ever obtainable using other technologies.

R/RStudio IN *Calculator* MODE

After numerous meetings and discussions authors agreed on trying R and RStudio in teaching an introductory course using mentioned above in a "calculator" mode.

Students were to use handout with assignment and R/R Studio package for calculations, then they were instructed to copy obtained results to their handout, notebook, homework assignment or special project, (typically in MS. Word format). Constructed graphical displays were to be saved in appropriate format to a desktop in order to be copied later. No working files in R/R Studio format such as ***filename*.r**, ***syntax1.txt***, or work directory were to be saved. Students who intended to replicate procedures executed by R were suggested to **redo** original handout procedures for a particular assignment. The list of commands was frequently expanded from the original one submitted to students by the authors on day one.

Authors emphasized importance of working in groups of two, (voluntary grouping), use of help resource internal to R, and Internet resources such as YouTube, PDF's, .html files, and many others. The authors agreed on the following purposes.

LEARNING OBJECTIVES

The authors wanted to introduce R and RStudio software packages to Mat 150, Introductory Statistics students that would function primarily as a "calculator", as well as, introduce the concept of density, and inform students how to read R output.

They intended to have students use these packages to produce and to interpret summary statistics for a single-variable or/and multiple-variable data set, as well as, to use these packages when working on simple regression topics to produce and analyze summary statistics related to a particular linear regression model and to graph/analyze a scatterplot including regression line.

In addition, they wanted students to produce simple graphical displays of one-variable, two-variable data set (e.g. bar plots, various histograms, different frequency polygons, boxplots, and multi-graphs presented in a single frame or multiple graphs in multiple frames presented at once for comparison purposes

Objectives that are NOT intended while teaching statistics with R and RStudio:

The authors did not intend to teach programming/programming techniques for programing purposes or for graphics only.

Neither they intend to teach proper saving of files including workspaces, complex coding, and complex graphics.

Expected benefits while teaching statistics with R or/and RStudio

Authors expected that students using R/RStudio as

a "calculator" in the introductory course would benefit as follows.

Students will be able to work/analyze a multiple-variable data set next to a single-variable/two-variable data set.

They would be able to expand their set of skills that are needed when comparing various data sets/variables' distribution within data frame.

In addition, students will save time from number-crunching activities, and using various tables, (e.g. z-tables, t-tables).

Lastly, students will be able to work simultaneously with two programs: R/R Studio and MS. Word when preparing assignments or/and term papers and exams.

Both authors were present during instruction related to introduction to R and RStudio, facilitated group work and independent work of students. Discussions and *live* R computer outputs generated from instructor station presented by different students followed.

SETTINGS FOR MAT150 – 171 W CLASS

There were 22 mostly non-traditional students with gender-split almost 1:1.

The instruction took place in a computer laboratory, N553 where

each station included machine with R and RStudio installed and active Internet connection.

Instructor's laptop was used by student-volunteer for demonstration of various activities.

Modality of Instruction in Mat 150-17W Class

The Mat 150-171W class met once a week every Sunday between the hours of 5:00 p.m.

and 8:30 p.m. in computer lab for five sessions over duration of the fall 2017 semester.

The class Mat150-171W is described as writing intensive section that meets the CUNY writing intensive graduation requirement.

Professor Nkechi Agwu was principal instructor of Introductory Statistics course. Topics included in R/R Studio students' activities were already taught to students and needed no reteaching by Professor Agwu or Professor Bialas.

Professor Piotr Bialas was visiting the course five times each for about 90 minutes in order to instruct students to use R/R Studio in their class.

Session 1 of Introduction to R/RStudio

During session 1 student's tasks consumed approximately 60 to 75 minutes of instruction.

Activity 1 lasted for 15 minutes, its intended goal was to reinforce students' knowledge about R/RStudio and possibility of uploading R/RStudio to personal computer's desktop or laptop.

Students used Internet in order to provide answers two questions: What is R /RStudio? And, how to download R and RStudio to a laptop/desktop computer? Finally, they shared they answers.

Activity 2 also lasted 15 minutes, when students had to determine how to get resources related to downloading R and R Studio. YouTube resources prevailed, and authors recommended Tutorial 1.1, Tutorial 1.2, Tutorial 1.3 from Series1 of *MarinStatsLectures about R,* (https://www.youtube.com/watch?v=cX532N_XLIs).

Activity 3 was allocated 30 minutes and intended to inform students about simple syntax needed to perform calculations selected by the authors.

Students obtained handout related to use of R/RStudio to complete series of calculations like that on (http://www.pbialas.com/chapter-1.html) They were to work in groups of two. Finally, self-selected students would provide answers using instructor laptop to project the syntax and answers on white board. Analyses of errors would follow.

Authors observed numerous instances of *aha*-moments students expressed during this activity with respect to application of parentheses, order of operation, and justifications for their use.

In addition, demonstration of calculations of an individual standardized *z-score* for a value in the data set, lead students to understanding that calculations of *z-scores* can be made with respect to all data values at once, students found those useful in detection of *unusual* data set values and also helpful in explanation of calculations related to the *Pearson's Linear Correlation Coefficient*.

Homework assignment for Session 1 involved uploading both packages, (R and R Studio), exploration of www.pbialas.com site, and practice of calculations like in CHAPTER 1 of the website.

Session 2 of Introduction to R/RStudio

Session 2 begun with homework review when the authors observed, that large majority of students completed homework Assignment 1, and just very few were unable to download R/RStudio. Those who did not, were able to download the programs in class.

During Session 2 students' tasks consumed 45 to 75 minutes of instructional time. Students were instructed to visit www.pbialas.com website and review Assignment 0 as well as Model Solution to Assignment 0.

Assignment 0 was related to Simple Linear Regression topic involving bivariate data where students had to enter data by hand into R/R Studio file, follow instructions related to calculations, graphing, and write short interpretations of obtained results. The authors assisted students during Session 2 and noticed that some students were having *technical* difficulties related to copy-and-paste of the results obtained from R to MS. Word file.

Homework assignment for Session 2 was placed on www.pbialas.com and was similar to activities students worked on in class time. This website was developed for students' homework assignments and model problems. Students were expected to create a report in MS Word format by following detailed instructions provided online by the authors, (Assignment 1 Report).

Session 3 of Introduction to R/RStudio

Session 3 begun with collection of Assignment 1 and homework review. Homework review included the list of issues students encountered while working on Assignment 1. Two students did not submit report on time and were given extension for completion of their homework.

During Session 3 students' assignment consumed between 45 and 75 minutes of instructional time. Students' activities were related to uploading a .csv file into R Studio. The authors provided short demonstration of the procedure along with explanation of each step. After the authors' presentation, students watched the YouTube video about importing data from MS. Excel to R Studio. Finally, they worked in self-selected groups or individually on a *small* data set in MS. Excel format in order to import it into R Studio.

The authors observed that those students who completed work early offered their assistance to other students. Finally, self-selected students would provide answers using the instructor laptop to project the syntax and answers on white board. Analyses of errors followed.

Homework assignment for Session 3, Assignment 2 was placed on www.pbialas.com website. It involved uploading a .csv data set file, creation of a report in (MS. Word format) related to histograms and boxplots comparison topics using R Studio as "calculator" for calculations and graph(s). Students were directed to http://www.pbialas.com/sunday-m1502.html for instructions related to Assignment 2 and were given extended time to complete this assignment.

Session 4 of Introduction to R/RStudio

Session 4 begun with collection of Assignment 2 and homework review. Homework review included the list of issues students encountered while working on Assignment 2. Some students did not submit report on time.

During Session 4 Students continued to work on Assignment 2 Report in class in self-selected groups of two or individually for 45 to 75 minutes. Some students who completed their tasks offered again their assistance to other students. Students' completed assignment was to be re-submitted to the authors upon arrival to the next class session. (place the discussed issues)

Session 5

Session 5 begun with collection of Assignment 2 and homework review. Homework review included the list of issues students encountered while working on Assignment 2.

During Session 5 the authors reviewed with students four previous sessions related to R and R Studio packages and discussed issues students had while working on their respective assignments. One of the issues involved *to save or not to save* particular workspace while learning simple R syntax. In addition, one of the authors provided demonstrations of saving files in R and R Studio. Students agreed that at their stage of learning R/R Studio simplicity would prevail Further the authors discussed students' creativity with respect to their completed assignments.

ASSESSMENT OF STUDENTS CREATIVITY BY SESSION

The authors believe that in Introduction to R/R Studio in Introductory Statistics Course process of assessment of students' creativity is related to the following: originality of topics selected for the process, method of delivery (e.g. lecture, group-work, technology uses in the process-YouTube and more), originality of student's activities and homework assignments, and finally to use of Depth of Knowledge, (DOK) questioning levels.

In Session 1 of students' activities the authors observed various reactions of students to error messages produced by R/R Studio. Explanation of error messages by students and authors created numerous *I got it* … moments to correctly executed commands, (e.g. need of parenthesis to preserve Order of Operation Rule and *what-if- the need* is violated …). Students were expected to create their own problems; many did and shared their work using instructor's laptop via LCD projector.

The authors observed that *copy-and-paste and edit* idea when writing a syntax was immediately accepted and widely used by students through the reminder of all activities, (e.g. student's comment: *What a useful way to correct errors!*).

In Session2 students' activities the authors offered students *Assignment 0* and presented them with *Model Solution to Assignment 0* link. Students were expected to work with two programs at once;

perform computations using R Studio, then copy those into MS. Word file and provide narrative/interpretation of the computations in MS. Word file for each task in *Assignment 0*.

Homework for Session 2 involved creation of report, Assignment 1 Report. The authors observed that some students questioned the use of MS. Word format requested by the authors and were able to modify it successfully to their own format.

In addition, those students who wanted to replicate their calculations used in Assignment 1 Report found questionable the authors' suggestion: *do-not-save-file in-RStudio rule*. In class discussions and presentations convinced some students about usefulness of the *rule*, (*do-not-save rule*). This rule allowed students to avoid errors related to variable names, incorrect workspace location, and other aspects such as saving of graphical displays.

There were several students who decided to save their calculations to RStudio file and who attempted to use saved file to subsequent calculations. Later, they reported multiple error messages regarding syntax and obtained output including incorrect calculation(s). Among those students, one decided to save her calculations to *filename.r* format, she was able to access it, and use it multiple times for subsequent calculations correctly after consultations with one of the authors and her independent research online.

In Session 3 and Session 4 students' activities the authors noticed that students in general understood the need of uploading a *.csv-type* of file into R/RStudio and some students were successful in this activity at their first attempt, (*aha moment:* no more typing multiple-variable data set(s) with *large* number of cases!). The authors observed that some students had to revisit topics involved in this activity related to histogram(s) and boxplot(s). They did it on their own, (extension was granted). It was their first time when they were exposed to analyses of a multi-variable data set including comparisons of distributions for selected variables.

Assignment 2 varied in quality and format. Some students followed directions with respect to format and presented their work early, no resubmission was necessary for those students. There were students who created their own format, (viz. copied console outputs and graphs into MS. Word formatted file, then answered questions). Some students requested more time for completion of Assignment 2 since quality of submitted homework was in question.

In Session 5, the authors summarized their experiences and pointed to numerous instances of students' creativity related to completion of assignments. For example, some students who did not want to replicate an assignment or part of it just printed R console content for themselves for future reference. Another student was able to learn on her own saving workspace procedure and demonstrated her ability to access saved file again. In addition, students were able to save graphical displays created in R/R Studio in various formats outside of R/ R Studio using copy-and-paste procedure.

WHAT ARE POSSIBLE IMPROVEMENTS TO USING R/RStudio IN AN INTRODUCTORY STATISTICS COURSE AS A "*Calculator*" APPROACH

The authors believe that introduction to R/RStudio in an introductory statistics course can take place in a computer laboratory classroom where all machines have *R* and *RStudio* programs uploaded or/and regular classroom equipped with access to desktop/laptop connection to LCD projector. The use of R/RStudio does not have to consume excessive amount of lecture time and can be left to discretion of instructor. Although, instruction related to R/RStudio should start early in the course and last through duration of the semester.

The authors find that students should have these programs installed on campus computer labs available for all after instruction time and access to trained in basic R/RStudio tutors in these settings. Newly minted tutors may be made of students who successfully completed the course, are interested in learning more about R and willing perform tutor function.

The authors believe that *"Less Volume, More Creativity"*[3] approach is a key to successful introduction of R/RStudio to students. According to Randal Randal J. Pruim et.al ideas initial set of commands should be relatively small, coherent and powerful. In addition, students should be able to use R/RStudio to execute numerical summaries, graphical summaries, and create linear models.

The authors found that students liked to use Internet resources such as YouTube videos for instruction in R/RStudio as well as instructional tools in learning about selected statistical topics. Some students claimed that abundance of Internet resources may be a reason for not purchasing expensive textbook(s).

WHAT DOES THE FUTURE HOLD …?

The authors believe that recent technological developments in computer science and data science cause already changes in ways introductory statistics courses are being taught. Such changes may include ability to work with *medium-size* and *large-size* data sets and ability to manipulate these sets in order to create graphical multivariable summaries, summaries that were not available in recent past few years.

The authors think that future challenges in teaching various statistics courses could be met with using appropriate teaching tools, one of them may be R with abundant number of packages and R Studio being used as computing device in a *calculator mode*.

The authors believe, that sooner or later some instructors may be tempted to utilize In-Cloud computing using R/R Studio in order to be independent of installation of R and RStudio packages.

[3] *Randall J. Pruim; Nicholas J. Horton; Daniel T. Kaplan;* Project MOSAIC, *Start Teaching with R; Preliminary Edition*

Example of Student's Work-Assignment 1 Report

MAT 150-171W
Assignment 1 Student Worksheet

Open R Studio, write a command in console pane, execute it by pressing *ENTER* and then, copy-and-paste obtained answer to a box below.

REMEMBER to copy-and-paste command from console to this worksheet; see your model solution to As signment 0 worksheet!!!

Simple Linear Regression Assignment 1

The data set below contains two variables: Study Time in minutes students spend on preparation for Quiz 2 and Grade in percent obtained on Quiz 2 (Assume that requirements for using linear regression model are satisfied.) Follow directions for the following tasks: #1-#17.

1. StudyT

```
studyT<-c(45,78,80,83,55,45,30,25,90,25,30,15,40,25,30)
```

2. Enter your data into R Studio console quiz 2

```
quiz2<-c(55,70,85,78,65,70,75,85,95,46,48,75,70,80,85)
```

3. Determine Summary Statistics for StudyT variable by typing into console

```
summary(studyT)
```

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
15.0    27.5    40.0    46.4    66.5    90.0
```

4. Determine Summary Statistics for SQuiz2 variable. Summary

```
summary(quiz2)
```

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
46.00   67.50   75.00   72.13   82.50   95.00
```

5. Determine standard deviation of study time variable

```
sd(studyT)
[1] 24.8619
```

6. Determine standard deviation of Quiz 2 variable

```
sd(quiz2)
[1] 14.0299
```

7. Determine correlation coefficient of Quiz 2 and study time variable

```
cor(studyT,quiz2)
[1] 0.3520541
```

8. Determine r-squared value

```
cor(studyT,quiz2)*cor(studyT,quiz2)
[1] 0.1239421
```

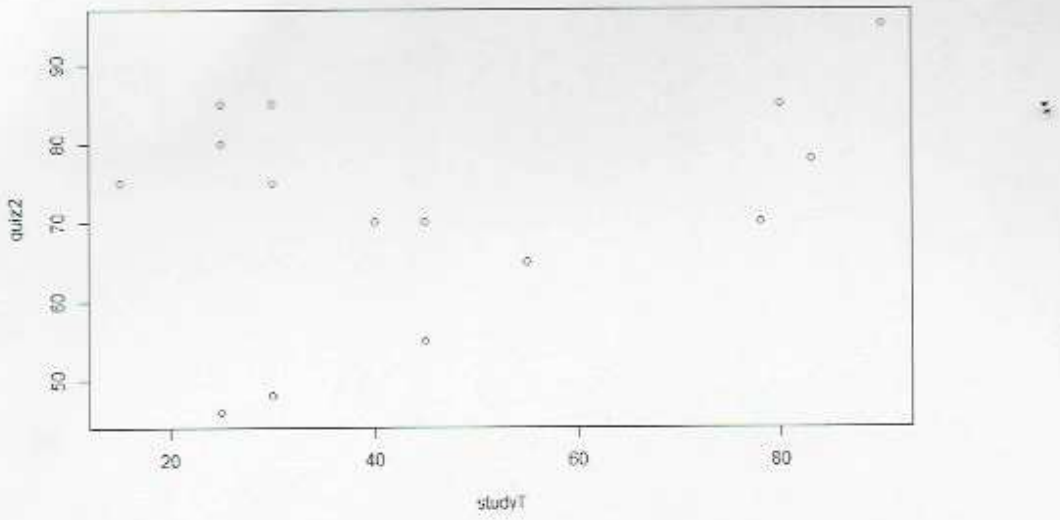
9. Explain what it (r-squared) means in context

```
[1] "just about 12.4% of variation in grade can be explained by variation in study time
."
```

10. Graph scatterplot of Grade 2 vs. StudyT plot

Source: homework assignment from →http://www.pbialas.com/sunday-m150.html

MAT 150-171W



11. Determine linear regression model equation lm1

```
lm1<-lm(quiz2~studyT);lm1
Call:
lm(formula = quiz2 ~ studyT)

Coefficients:
(Intercept)       studyT
    62.9151       0.1987
```

12. Write the equation of the linear model; y-hat = b0+b1x

```
[1] "y-hat=62.9151+0.1987*x"
```

13. Write the equation of the linear model in context

```
[1] "averagequiz2grade=62.9151+0.1987*studytime"
```

14. Explain what the slope means in this context

```
[1] "for every minute change in study time average grade on quiz2 will change by 0.1987
%(just about 0.2%).  OR for every 10 minutes increase in study time it is expected that
the grade will increase by 1.987%"
```

15. Explain what the y-intercept means in this context

```
[1] "average quiz2 grade for a student whose study time is ZERO minute, (e.g.Average qui
z2 grade=62.9151+0.1987*0, average quiz2 grade= 62.9151+0, average quiz2 grade=62.9151"
```

16. Estimate Quiz2 grade for a student who studied 30 minutes

```
"average quiz2 grade=62.9151+0.1987*study time, average quiz2 grade=62.9151+0.1987*30,
average quiz2 grade=68.8761"
```

17. Estimate Quiz2 grade for a student who studied 330 minutes and comment on it

```
[1] "average quiz2 grade=62.9151+0.1987*study time, average quiz2 grade=62.9151+0.1987*
330, average quiz2 grade=128.4861" [1] "comment: 330 minutes is far from range of value
s for studyT variable, therefor the grade of 128.4861% is not realistic by common sense
measure"
```

Source: homework assignment from →http://www.pbialas.com/sunday-m150.html

Selected Bibliography

1. College Report 2016, (GAISE), Guidelines for Assessment and Instruction in Statistics Education
2. mosaic-web.org. Project Mosaic | Modeling Statistics, Calculus, and Computation;
3. Randall J. Pruim; Nicholas J. Horton; Daniel T. Kaplan; *Start Teaching with R*; Preliminary Edition; July 2014
4. J. Pruim; Nicholas J. Horton; Daniel T. Kaplan, *Compendium of Commands to Teach Statistics with R*
5. MarinStatsLectures – YouTube channel, More than 50 YouTube videos related to R/RStudio use in teaching introductory statistics course
6. www.pbialas.com; 2016