Experiences in Teaching S-PLUS

Frank E Harrell Jr
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
University of Virginia School of Medicine
Box 600 Charlottesville VA 22908 USA
fharrell@virginia.edu
hesweb1.med.virginia.edu/biostat

1999 S-PLUS INTERNATIONAL USER CONFERENCE
NEW ORLEANS, LA
22 OCTOBER 1999

Outline

1. Target audience of course
2. The biggest hurdle
3. Course description
4. Texts
5. Commands vs. GUI
6. Most difficult ideas to get across
7. Can one be a good data analyst without being a good programmer?

Target Audience

- Graduate students in intensive 1–year M.S. in Health Evaluation Sciences program
- Undergraduates
- Have had or be taking intro. statistics course
- General computer skills but no programming
- Jobs: entry–level analyst, epidemiologist, pharmaceutical research, . . .

The Biggest Hurdle

- Convincing curriculum committee to use S-PLUS as primary statistical software in masters program
- Job market for graduates
- Time horizon for use of statistical computing knowledge
- Instructor extremely familiar with SAS, so knew its limitations
- S-PLUS required for later courses, esp. biostatistical modeling

Fortunately . . .

- Bugs in S-PLUS 4.0r1 were not known
- Problems with S-PLUS 5 were not known

Statistical Computing & Graphics: Objectives

To be able to use a high-level object–oriented statistical computing language (Windows/NT S-PLUS) to

1. Input data into a computer
2. Transform, recode, and manage data
3. Perform calculations
4. Examine data
5. Compute and display descriptive statistics
6. Compute standard probabilities, power, and confidence limits
7. Understand how commonly used graphical techniques can result in optical illusions and poor communication of information
8. Learn how to construct graphics that can accurately display information in multiple variables

Based on premise that graphical presentation of information is fundamental to understanding data and presenting research findings

Course Description

1. S-PLUS introduction
   (a) Overview of GUI
   (b) Basic commands
   (c) Entering and saving commands
2. Objects, Getting Help, Functions, Attributes, and Libraries
3. Data in S-PLUS
   (a) Importing data
   (b) Adjustments to variables after input
   (c) Inspecting data
4. Operating in S-PLUS
   (a) Reading and writing data frames and variables
   (b) Functions for manipulating and summarizing data

   `tapply, by, aggregate, summary.formula, summarize`

(c) Recoding variables and creating derived variables

(d) Re–shaping datasets

5. Review of data frame creation, annotation, and analysis
6. Probability and statistical functions
7. Making tables
8. Graphics horror stories
9. Elements of graphical perception
10. Presentation of students' graphics examples
11. Some Useful Traditional Graphics and the S-PLUS GUI

(a) More about the GUI

(b) 1–d scatterplots, histograms, density plots, CDFs, box plots, scatterplots

12. Newer graphical techniques

(a) Dot plots

(b) Multi–panel displays

(c) 3-D plots

(d) Interactive and dynamic graphics

(e) Nonparametric trend lines

13. Using S-PLUS Functions (Commands) for Graphing Data

(a) Basic plotting commands

(b) Converting tables into plots

(c) Trellis graphics

(d) Plotting summary statistics

Other Aspects of Class

- 20 fast NT stations with large screens for 26 students
- Projector for instructor station
- Class E–mail list
- Anonymous E–mail
- Assignments on web page
- Instructor's commands put on web page after class
- Datasets, links on web page
- MathSoft has been extremely helpful in getting labs equipped with S-PLUS

Texts

- Alzola CF, Harrell FE: *An Introduction to S-PLUS and to the Hmisc and Design Libraries*, 1999.
- Cleveland W: *The Elements of Graphing Data*, Summit NJ: Hobart Press, 1994.
- *S-PLUS 4.5 Student Edition User's Guide*, Pacific Grove CA: Duxbury Press, 1999.
- References: Cleveland, Tufte, V&R, Wilkinson, Wallgren *et al.*

Commands vs. GUI

- Learning curve for GUI: fast then levels off
- Learning curve for commands: slow then faster, resulting in higher analytic capabilities
- Which to emphasize?
- Order: GUI (brief), commands (long), GUI, mix
- Students have very little trouble with GUI for graphics
- Script editor and report window are nice
- Students see advantages of commands for reproducible research [1], ultimate savings of time
- Half–life of knowledge about a command: 1 week

Most Difficult Ideas to Get Across

- Search list and masked objects
- `attach`
- When to `attach` in search position 1
- Present both `attach` and $ (for a few variables)
- What words to quote
- Subsetting data
- Example templates do help

Can One be a Good Data Analyst without being a Half–Good Programmer?

- NO
- GUI is useful mainly for overcoming fear of system and for interactive graphics (especially multi–panel)
- GUI for data management & recoding variables is of unknown value
- Data updates happen!
- Audit trail

Summary

- Convincing other faculty to emphasize S-PLUS was difficult
- Need to teach both GUI and commands
- Commands emphasized for data management & manipulation, computing summary statistics for tables and plots
- Accessing and storing objects in multiple places remains confusing to students
- Course has gotten good reviews although students think it's a lot of work
- Combination of texts is becoming useful
- Availability of S-PLUS Student Edition for students to be able to work at home has been a major plus
- Previous programming experience (e.g., Basic) would have helped

- Students have gone on to use S-PLUS effectively in later projects
- Students have gotten jobs directly because of S-PLUS skills
- Students appreciate the advantages of S-PLUS , especially for graphics and statistical modeling

References

[1] R. Koenker. Reproducible econometric research. Technical report, University of Illinois, `www.econ.uiuc.edu`, 1996.

Abstract

This talk covers experiences in teaching S-PLUS to graduate students who have not previously studied statistical languages or packages. First, the course, “Statistical Computing and Graphics” will be overviewed (a detailed syllabus and course notes are available at `hesweb1.med.virginia.edu/biostat` under Teaching Materials). This course is based on S-PLUS for Windows/NT and uses a 20–station computer lab with a video projector. A brief discussion of teaching using GUI menus vs. commands will be given. The relative ease of teaching graphical methods vs. data manipulation and management will be covered and a source of data manipulation and management examples will be given. Students’ experiences with the Student Edition of S-PLUS and the book that accompanies it will be discussed. Some of the controveries that will be addressed are attaching data frames vs. operating on `dataframe$variable`, and whether students can utilize reproducible and efficient research practices without mastering the command language to some extent.