Research and Statistical Support
University of North Texas

RSS Matters

Link to the last RSS article here: Equivalence Tests - Ed.

Using Statistical Software in Classroom Instruction: S-Plus/R, An Accessible, Low Cost Alternative

By Dr. Rich Herrington, Research and Statistical Support Services Manager

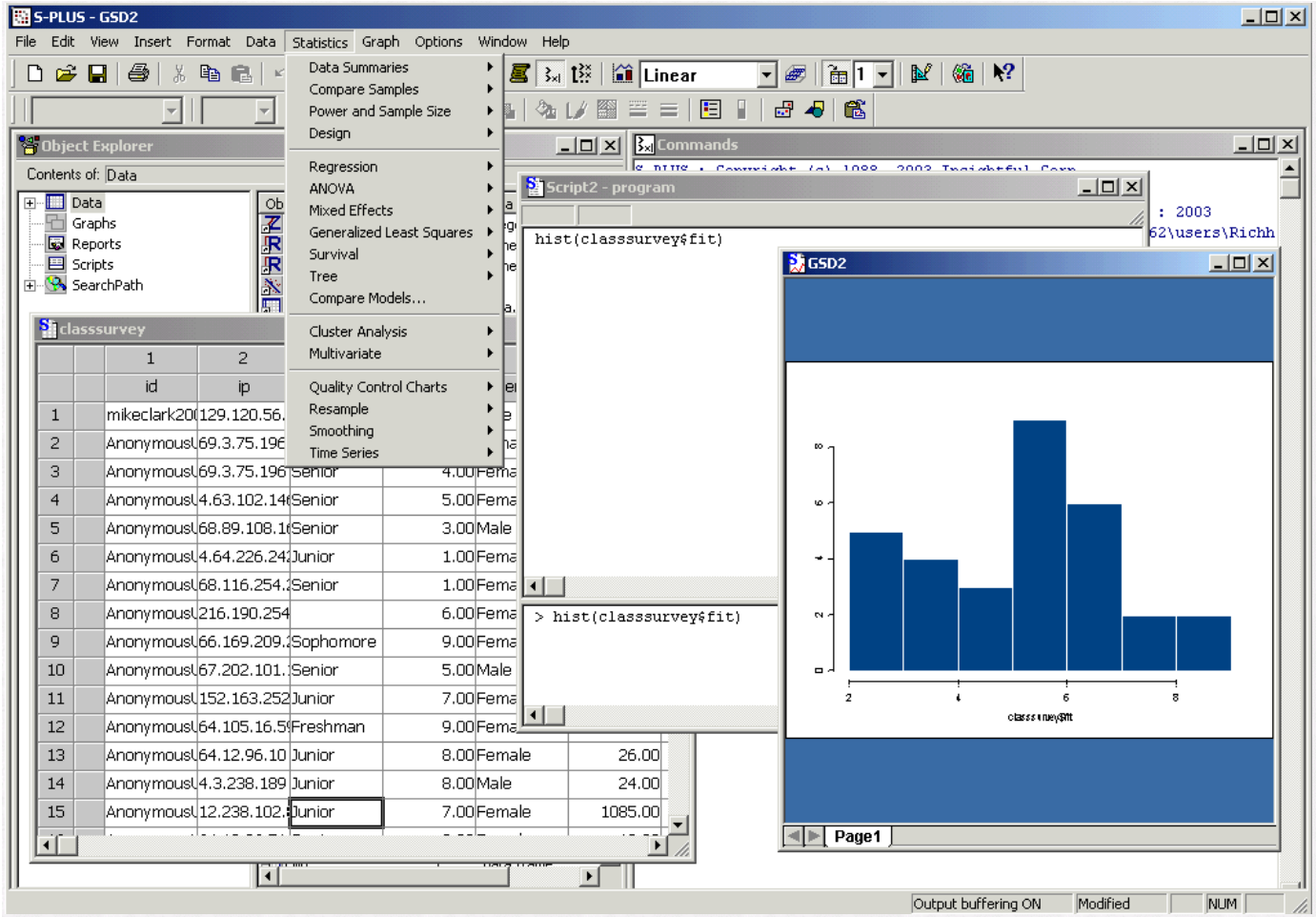
The choice of which statistical package to use in an introductory statistics or advanced statistics course can be determined by a number of considerations:

1. Which statistics package is the instructor most comfortable with?
2. Popularity of the statistics package
3. Goals of the intended student user - will the student be doing more involved research and development, or will they be engaging in intermittent cursory usage?
4. Ease of use - are there drop down menus? How easy is the syntax/language to learn?,
5. Flexibility
6. Cost for the student
7. Will the student be using modern, advanced statistical technologies, or will they be relying mostly on well known classical methods?
8. How important is high quality, publication ready graphics (both exploratory and classical)?
9. Availability of the software during course work, and after the student leaves the academic institution
10. Is there an active, supportive community of users?
11. How available are documentation, tutorials, and books?
12. Are there statistics textbooks that cover software usage along with theory?

These are only a few of the considerations involved in selecting a statistics package for a statistics course. In this article, we bring two data analysis/statistical systems to the attention of educators: "S-Plus" (the commercial version of the "S" language) and the public domain "R" (free version of the "S" language). We discuss the cost and availability of S-Plus and R to the community of UNT researchers, instructors, and students.

S-Plus

S-Plus incorporates the object-oriented language S, developed at AT&T Bell Labs statistics research group (Lucent Technologies). Marketed by Insightful Corp., S-Plus fits statistical models as "objects", making data analysis much more flexible than the older, procedural language approach (e.g. SPSS, SAS). S-Plus incorporates a highly useable graphical user interface (see this online tutorial for examples), along with the capability of script based processing. Additionally, S-Plus allows the user to "interact" with data and graphics through a command line interface. The figure below provides an example of the S-Plus GUI interface:



S-Plus has an active world-wide user community - S-NEWS. Additionally, Insightful Corp. provides online versions of all S-Plus documentation (this documentation is also installed locally upon software installation). Students, instructors and researchers will be glad to know that many books and tutorials have been published on the S-Plus system. Advanced researchers should be excited about the continuing expansion of the S-Plus system with the newest statistical technologies available. **Insightful Corp. provides numerous "experimental" research libraries at no-charge for download.** Currently, these libraries include: S+CorrelatedData (mixed effects generalized linear models), S+Best (B-Spline methods), S+Resample (bootstrap library), S+Bayes (bayesian analysis), S+FDA (functional data analysis). Many of the libraries utilize both a "drop-down" GUI menu system and a command line interface approach. One particular library that could be particularly useful to introductory statistics instructors is the S+Resample library. A current trend in statistics education is to use resampling methods (e.g. bootstrap & permutation methods) to illustrate empirical sampling distributions and non-parametric confidence intervals based on the empirical sampling distribution. One notable example: Tim Hesterberg and co-authors have teamed up with the authors of the highly acclaimed *"Introduction to the Practice of Statistics, Fifth Edition"* by David Moore and George McCabe, to produce a book chapter that integrates the bootstrap into the statistics curriculum at an elementary level. This book chapter utilizes the S+Resample library to provide easy accessibility to resampling methods at an introductory statistics level. Tim Hesterberg has also written about using resampling and simulation methods in teaching statistics. Researchers who are interested in "data-mining" methodologies can use S-Plus in conjunction with Insightful Corp.'s "Insightful Miner" product to explore undetected patterns in massive datasets. A quick search on Google search engine demonstrates that S-Plus is a popular system for research and instruction (e.g. a search on "S-Plus" returned 482,000 hits).

Pricing and Availability of S-Plus at the University of North Texas

Students can purchase an "Academic" version of S-Plus at the UNT University Bookstore for $25. This is a specially licensed copy of S-Plus (for UNT campus) that expires one year after installation (MicroSoft Windows version). This academic version has all the features of S-Plus "Professional", except that it expires one year after installation. **Insightful Corp. also provides a "Student" version of S-Plus that is freely available** at http://elms03.e-academy.com/splus/ This version of S-Plus is free, and has full statistical functionality of the academic version, but: 1) Has a 20,000 cell or 1,000 row limitation; 2) Is only for educational use; 3) Expires after one year; 4) Has a large download (more than 100 meg). Students register at the website, download the software, and are given a license code that enables the software. The "Student" version of S-Plus is an attractive alternative to the "Academic" version of S-Plus for those instructors teaching a "long distance" learning course where students are incapable of purchasing S-Plus from the bookstore. For full-time faculty, S-Plus can be obtained at no cost from the Research and Statistical Support Office (RSS) at UNT. S-Plus is gaining in popularity (it is already a favorite amongst professional statisticians); S-Plus excels in incorporating modern statistical methodology while maintaining a large inventory of

classical statistical methodologies; There are many tutorials, advanced methodology books, and introductory statistics textbooks that incorporate S-Plus. S-Plus compares favorably on the all software-choice considerations enumerated above. That is, S-Plus can accommodate both novice users and heavily research oriented practitioners of statistics.

R

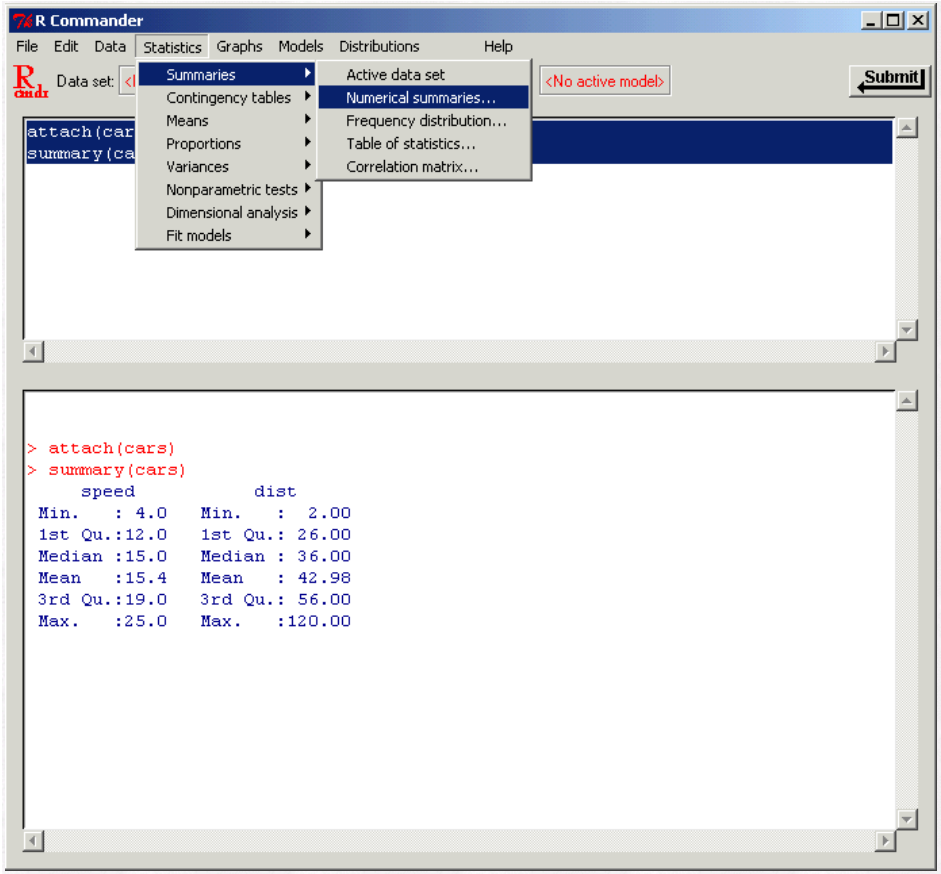
R is an open-source initiative whose aim is to create and distribute the same high quality, "cutting-edge" statistical technology that S-Plus is known for (see the R homepage). Quoting from the R homepage:

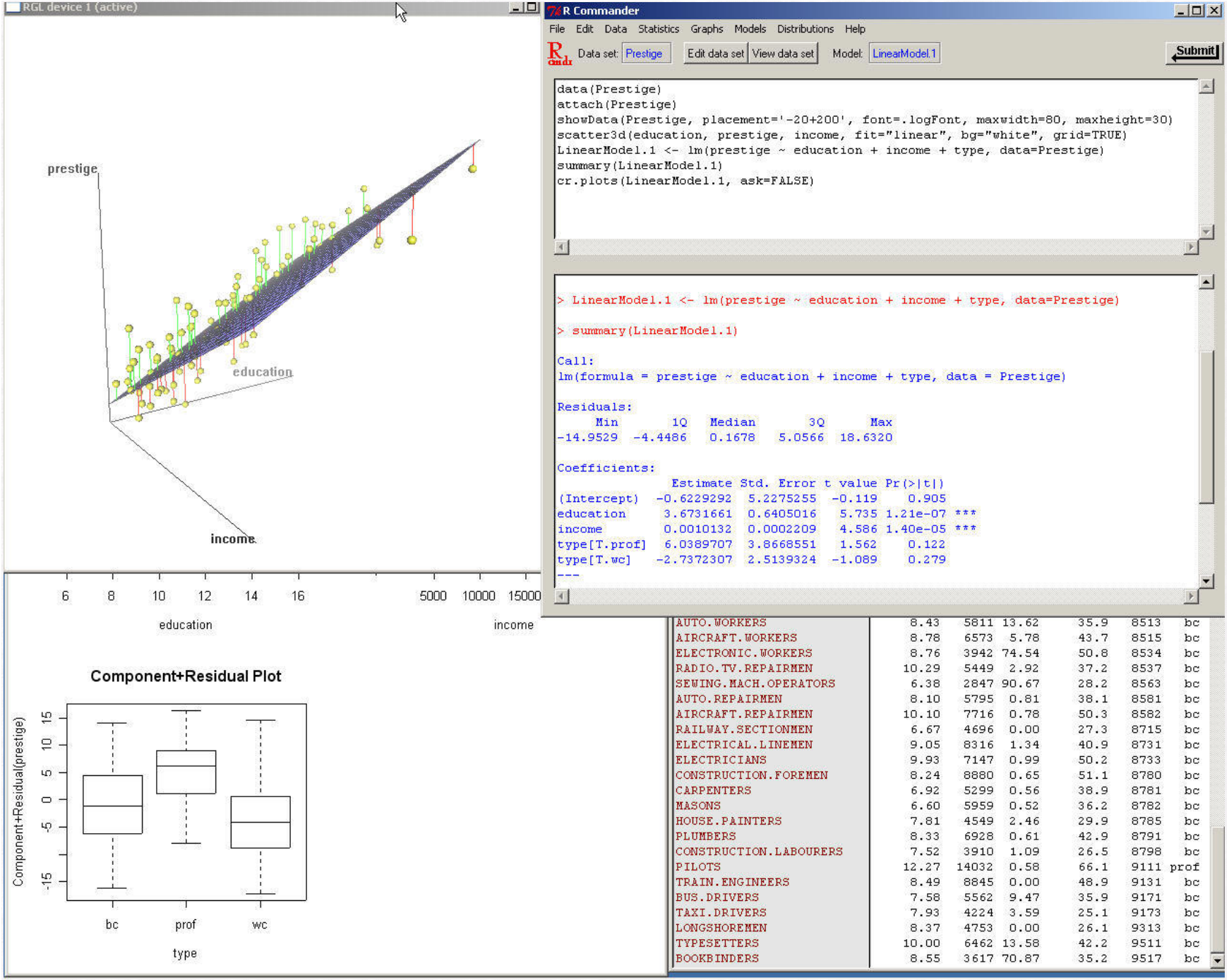
> *R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.*
>
> *R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.*
>
> *One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.*
>
> *R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.*

As a free alternative to S-Plus, R cannot be beat. Available to the R system are hundreds of user contributed libraries that cover large areas of both classical and modern statistics (see UNT's R server help page on installed packages). While S-Plus excels at providing advanced functionality through a menu system, R excels in providing breadth in statistical functionality (e.g. our own RSS R Server has 587 libraries installed). Much of this statistical functionality is not duplicated for the S-Plus environment. Partly, this is a result of the R system being an open-source project. Since the R source code is available to developers of statistical technology, much integration of R with existing statistical tools, databases, and operating systems has occurred. The "Omegahat" project being the prime example of such efforts. From the Omegahat website:

> *Omega is a joint project with the goal of providing a variety of open-source software for statistical applications. The Omega project began in July, 1998, with discussions among designers responsible for three current statistical languages (S, R, and Lisp-Stat), with the idea of working together on new directions with special emphasis on web-based software, Java, the Java virtual machine, and distributed computing. We encourage participation by anyone wanting to extend computing capabilities in one of the existing languages, to those interested in distributed or web-based statistical software, and to those interested in the design of new statistical languages.*

R's integration with web servers should be of particular interest to instructors who are interested in web-based statistics courses. For a number of years now, I have been using a modified version of Rcgi to create online, interactive tutorials for *Benchmarks* articles and introductory statistics courses. Our *RSS Matters* column has a number of examples of using R to create interactive tutorials: robust statistics, kernel density estimation, false detection rate, robust correlation, bootstrap, too name a few. If, as an instructor, you are concerned about the lack of a default drop-down menu system for R, some efforts have gone toward developing a GUI system for the R system. The most notable of these efforts is John Fox's *R Commander* (see our past Benchmarks articles on this GUI - Article1; Article 2; Article 3 - these articles are somewhat dated). See the main R Commander website for the most recent updates. R Commander uses both a drop down menu system and a script window. Similar to other statistical packages, R Commander pastes syntax into a syntax editor whenever the contents of a menu system window have been submitted. This allows easy access to default syntax (via a GUI) , but allows the user to see the syntax, change the syntax, and save the syntax, for later submission. This facilitates learning to program in the "S" language. A couple of examples of R Commander's interface is presented below:


R Commander
File Edit Data Statistics Graphs Models Distributions Help
Data set:
Summaries
Contingency tables
Means
Proportions
Variances
Nonparametric tests
Dimensional analysis
Fit models
Active data set
Numerical summaries...
Frequency distribution...
Table of statistics...
Correlation matrix...
<No active model>
Submit
attach(car
summary(ca
> attach(cars)
> summary(cars)
speed dist
Min. : 4.0 Min. : 2.00
1st Qu.:12.0 1st Qu.: 26.00
Median :15.0 Median : 36.00
Mean :15.4 Mean : 42.98
3rd Qu.:19.0 3rd Qu.: 56.00
Max. :25.0 Max. :120.00



Like the S-Plus user community, the R user community is highly active as well - R-HELP. In addition, the R developers publish a high quality, edited newsletter that covers software development news, R package development and usage, as well as the usual tips and hints about using R. The user community is also quite generous in providing free tutorials, books, and documents on R. R's documentation is very high quality as well. The basic R language is well documented with examples that can be executed as is, then modified as the user needs. For example, fitting a regression, ANOVA, or ANCOVA model can be fit with the "lm" function. The help function for *lm* gives the user an example that can be executed by pasting the text into the R console, then altered as needed. The "foreign" package gives users the ability to import other file formats: SAS, SPSS, Stata, Minitab, SYSTAT, to mention some of the more common formats available. R's base language

is mostly compatible with the S-Plus base language (greater than 95%?). That is, most code written with the base R language will run unaltered in S-Plus and vice-versa. It is not inconceivable that a student or researcher would use both R and S-Plus in conjunction with one another.

Conclusion

In summary, R compares favorably with S-Plus (and is arguably superior in some ways). In regards to some of the statistical-software choices enumerated at the beginning of this article: 1) Both S-Plus and R are readily available and inexpensive to the student and instructor; 2) Both S-Plus and R are readily available to instructor and student; 3) Both S-Plus and R are inexpensive alternatives to more popular statistical packages (e.g. SAS, SPSS, Stata); 4) Both S-Plus and R excel at providing a broad range of classical and modern statistical methodologies; 5) S-Plus utilizes an advanced menu system that is more accessible to students, however, R is gaining some ground on that issue; 6) Both S-Plus and R can accommodate a range of users from novice to advanced, that is, both cursory users and researchers; 7) Both S-Plus and R have high quality documentation and textbook usage; 8) The user communities of both S-Plus and R are highly active and accessible to both student and researcher; 9) S-Plus and R are already favorites amongst theoretical and applied statisticians, and both of these systems are becoming increasingly important in the environmental, biological, medical, and social sciences, as evidenced by the increase in classes being taught utilizing these environments and the increase in statistical texts being published; **10) And most importantly - THE PRICE IS RIGHT!**

Resources

Everitt, Brian S. (2005). An R and S-Plus Companion to Multivariate Analysis, Springer.

Faraway, Julian (2005). Linear Models with R, CRC Press.

Good, Phillip (2005). Introduction to Statistics Through Resampling Methods R/S-Plus, Wiley.

Heiberger, R.M. & Holland, Burt (2004). Statistical Analysis and Data Display: An intermediate Course with Examples in S-Plus, R and SAS, Springer.

Verzani, John (2005). Using R for Introductory Statistics, CRC Press.

Crawley, Michael (2002). Statistical Computing: An Introduction to Data Analysis Using S-Plus, Springer.

Dalgaard, Peter (2002). Introductory Statistics with R, Springer.

Kraus, A & Olson, M. (2002). The Basics of S-Plus, Third Edition, Springer.

Venables, W.N. & Ripley, B.D. (2002). Modern Applied Statistics with S, Fourth Edition, Springer.

Pinheiro, J.C. & Bates, D.M. (2000). Mixed-Effects Models in S-Plus, Springer.

Return to top