Statistics 135 - Statistical Computing Software

Mark E. Irwin
Department of Statistics
Harvard University

Autumn Term

Monday, September 19, 2005 -
January 2006

Personnel

Instructor:	Mark Irwin
Office:	611 Science Center
Phone:	617-495-5617
E-mail:	irwin@stat.harvard.edu
Web-site:	<http://www.courses.fas.harvard.edu/~stat135/>
Lectures:	MWF 10:00 - 11:00, Sever 102
Office Hours:	Monday 1:00 - 2:00, Thursday 2:00 - 3:00, or by appointment
Teaching Fellow:	Darin McGill
E-mail:	dmcgill@stat.harvard.edu

Syllabus

- S Language (S-Plus / R)
- SAS
- LaTeX
- Implementing standard analyses and graphics
- Exploratory data analysis
- Computationally intensive procedures
  - Clustering and classification
  - Non-parametric regression
- Simulation (Monte Carlo) Methods

- Programming new methods.
- Other packages as time permits (Matlab, Stata, etc)

References

S-Plus / R:

- Venables WN and Ripley BD (2002). Modern Applied Statistics with S (4th edition), Springer. (Ordered by COOP)

  Often referred to as MASS.

- Krause A and Olson M (2005). Basics of S-PLUS (4th edition), Springer. (Ordered by COOP) An online version of the 3rd edition is available through the HOLLIS online catalog and the course web site.

- Venables WN and Ripley BD (2002). S Programming, Springer.

SAS:

- Delwiche, L.D. and Slaughter, S.J (2003). The Little SAS Book - A Primer (3rd edition), The SAS Insitute. (Ordered by COOP)

- Cody R and Pass R (1995). SAS Programming by Example, The SAS Insitute.

LaTeX:

- Mittelbach F, Goossens M, with Braams J, Carlisle D, and Rowley C (2004). The LaTeX Companion, 2nd edition. Addison Wesley, Reading, MA. (1st edition was by Goossens M, Mittelbach F, and Samarin A)

- Lamport L (1994). LaTeX: A document preparation system, 2nd edition. Addison Wesley, Reading, MA.

Plus much more on the course web site.

Grading

- Homework (50%): 6 or 7 during the term.
- Midterm (20%): Take home exam
- Final (30%): Take home exam

S

An **S** environment is an integrated suite of software facilities for data analysis and graphical display. Among other things it offers

- an extensive and coherent collection of tools for statistics and data analysis,
- a language for expressing statistical models and tools for using linear and non-linear models,
- graphical facilities for data analysis and display either at a workstation or as hardcopy,
- an effective object-oriented programming language that can easily be extended by the user community.

(from Venables and Ripley (MASS), page 1)

There are two major implementations of **S**

- **S-Plus**: Developed at Bell Laboratories (Lucent Technologies, formerly AT&T). Currently exclusively licensed to Insightful Corporation, which distributes an enhanced version.

  S-Plus is currently available for Unix/Linux and Windows. There is no Macintosh version available.

  In **S-Plus**, many analyses can be run from a menu based system, but the full programing language is always available and is needed for some analyzes.

- **R**: An open source (GNU) project originally started by Ross Ihaka (Univeristy of Auckland) and Robert Gentleman (Harvard Biostat).

  R is currently available for Unix/Linux, Windows and Macintosh from <`http://www.r-project.org/`>

  R does not currently allow analyzes to be performed with a menu based system. But if you want to write one, it would be appreciated.

Much of the course discussing the **S** environment will be presented in **R** for the following reasons

- Most things discussed will work in either environment
- Tends to be faster and has better memory management
- Faster development of add on libraries (and possibly more developement)
- Runs on just about anything, as opposed to **S-Plus**
- Easier to get up to date versions here at Harvard

Both packages are extendable, with the built-in object oriented programming language and the abilities to incorporate C, C++, and Fortran routines.

SAS

Probably the most popular Statistical software worldwide.

SAS claims that its products are used at over 40,000 sites, including at 90% of the Fortune 500.

This will not be all SAS as they make other products, such as JMP (a menu and dialogue based stat package)

Huge in the pharmaceutical industry. There is a belief, though 100% FALSE, that the FDA requires the analysis of all clinical trials to be done in SAS. You can do it in anything, you just need to document it as part of your approval submission.

SAS, the company, has the reputation of being a fantastic place to work as well.

Extremely powerful package. You name it, it probably does it. If it doesn't, they are probably working on it.

It is available for many platforms including Windows, Unix, Macintosh, mainframes (z/OS, CMS, VSE, VMS, MVS). However they don't keep all versions updated at the same rate (e.g. Mac version only goes to version 6.12, OS/2 is at 8.2, whereas Windows is at 9.1).

It's a program based package. You need to write a program for your analysis. There is no menu based approach like Stata, Minitab, or SPSS have available.

These programs will have a block structure, with each block corresponding to a different part of the analysis.

Each block will usually start with a different PROC statement, such as PROC REGRESS, PROC SORT, PROC LOGISTIC, etc. Within each block, commands will be given, options set, etc.

It is also extendable with the built-in Macro language.

What is TEX?

> TEX represents the state-of-the-art in computer typesetting. It is particularly valuable where the document, article, or book to be produced contains a lot of mathematics, and where the user is concerned about typographic quality. TEX software offers both writers and publishers the opportunity to produce technical text, in an attractive form, with the speed and efficiency of a computer system.
>
> (from the back cover of The TEXbook by Donald E. Knuth, the initial developer of TEX)

Most of what is done today is not in plain TEX, but with addon macro packages. Parts of TEX tend to be cryptic, so these addons have been created to make things easier. The most popular of these addons is LATEX. Others you might come across are AMS-TEXand AMS-LATEX.

What is LaTeX?

LaTeX is a document preparation system based on the TeX formatter. It is a set of macros which simplifies much of plain TeX. LaTeX is the most popular approach to writing papers in Science, particularly in Mathematics and the Physical Sciences. Much publishing today is done in LaTeX. I believe everything published by Springer-Verlag, is prepared in LaTeX. Many journals have also gone the LaTeX route as well, with packages prepared to assist with matching the journal format.

The most current version commonly available is LaTeX 2_ε, though LaTeX3 is currently in development. The previous version was LaTeX 2.09, which is quite a bit different. While files written in this version will run in LaTeX 2_ε (usually), you want to stick with LaTeX 2_ε for your writing.

LaTeX, and all TeX derivatives are page markup languages, with formatting commands mixed in with the text, similar to HTML. It is not a WYSIWYG approach, like Microsoft Word, though there are products that try to bring this to the LaTeX world.

A LaTeX file, is a plain text file which is then processed to give the desired output. The output is usually in a `.dvi` file, which then can be converted to other formats, such as postscript or pdf, if desired.

This approach to preparing documents allows for great flexibility and many things to be automated. These include

- Pagination
- Equation numbering
- Creation of tables of contents, list of figures, etc
- Bibliography creation and proper citation in the text
- Changing fonts and font sizes

TEXis available for just about any platform available and is free. It is usually available in standard Linux implementations (you might need to add it to the installation) and is easily available for Windows and Mac (see course web site).

These slides are produced in LATEX with the `foils` package.

Statistics/Mathematics Packages Available at Harvard

- R: nice, lab PCs, lab Macs?
- S-Plus: ice (versions 3.4, 5.1), nice (version 6.2?), lab PCs (version 4). Version 6.0 for Windows can be checked out from Cabot Library.
- SAS: lab PCs (version 8), available for download for Windows (details to be discussed later).
- Stata: lab PCs and Macs (version 9), available for download from FASCS software download site (Keyserved)
- SPSS: Windows (version 13.0), Mac (version 11.0.3), available for download from FASCS software download site (Keyserved)
- Minitab: Windows only (version 14)

- Matlab: ice, nice, Windows, Macs (version 6.5, release 13), available for download from FASCS software download site (must be on 140.247 subnet to run).