UseR! for Teaching

Sanford Weisberg
University of Minnesota
Minneapolis, MN USA

June 16, 2006

Framework

Audience: (Post)-graduates, both in statistics and particularly in other areas.

Non-stats student goals: Leave the class able to apply what they have learned to what they really care about.

Stats student goals: The material in the course, including the computing, is the end in itself.

Instructor's goals: Provide transferable knowledge, and keep computing from getting in the way (for non-stats students).

Three types of courses

- Teaching about R.
- Teaching *Analyzing Survey Data Using R*. This can imply teaching what the program can do under the general rubric of survey analysis.
- Using R in a course about sample surveys. This implies R is a tool that could be replaced by other tools.

Teaching about R

- R provides a high-level language for research statisticians
- R is great for exploration of new ideas; packages.
- How to. . . courses, for example, graphics using R.
- Guru creation.

Analyzing [your choice here] Using R

- Tailor the course to match what the program does. This often requires *compromise*.
- *Often, this is just what students want!*

“The University of Minnesota is not a technical or trade school.”

. . . Tom Burk, Forestry Prof.

Methods primary, R incidental

- The program should enable, not hinder, learning methods. **Easy to say, hard to do.**
- Common metaphors for working with the computer are: browsers, iTunes, and possibly Excel. . . R is nothing like any of these and *therefore is not obvious to students*.
- Students get stuck on HOW rather than WHY; memorization (is it `header` or `col.names` or `colnames`?) and inconsistency are a hinderance.
- Irregular users *forget* — no visual cues: a blank screen is intimidating.“
- Documentation is oriented toward the expert, not the novice (what is an S3 and why do I care?)

Textbooks

1999: Applied Regression Including Computing and Graphics

- Based on Arc and XLispStat: Book and program are strongly linked: book and program inseparable: an intellectual success, but an overall failure.

2005: Applied Linear Regression, 3rd Ed

- Synthesis of last edition (1985), some graphics from 1999 book, and some new stuff
- Little mention of computing in the text.
- Web supplements for *ALR* using R, S-Plus, SAS, SPSS and JMP. (google `applied linear regression`).

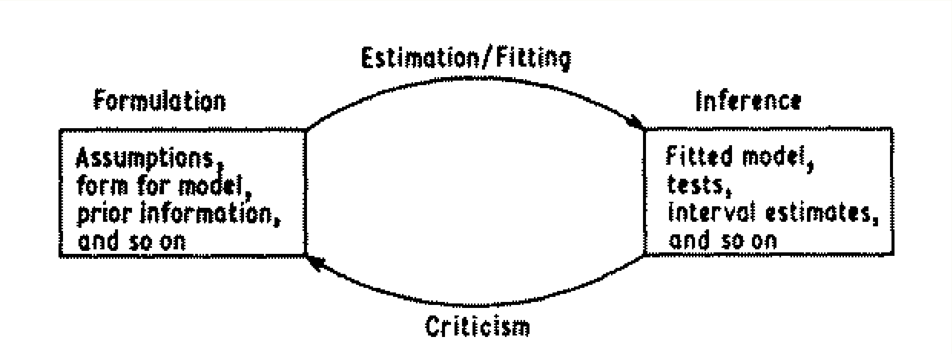
Primer download statistics

For January 1 – May 28, 2006, 11,000 web vists:

SPSS Primer	319	19%
SAS Primer	361	22%
JMP Primer	261	16%
R/SPlus Primer	725	44%

No program was adequate. R/S-Plus was closest with added package.

Does R encourage good data analysis?



Estimation/Fitting

Formulation

Inference

Assumptions, form for model, prior information, and so on

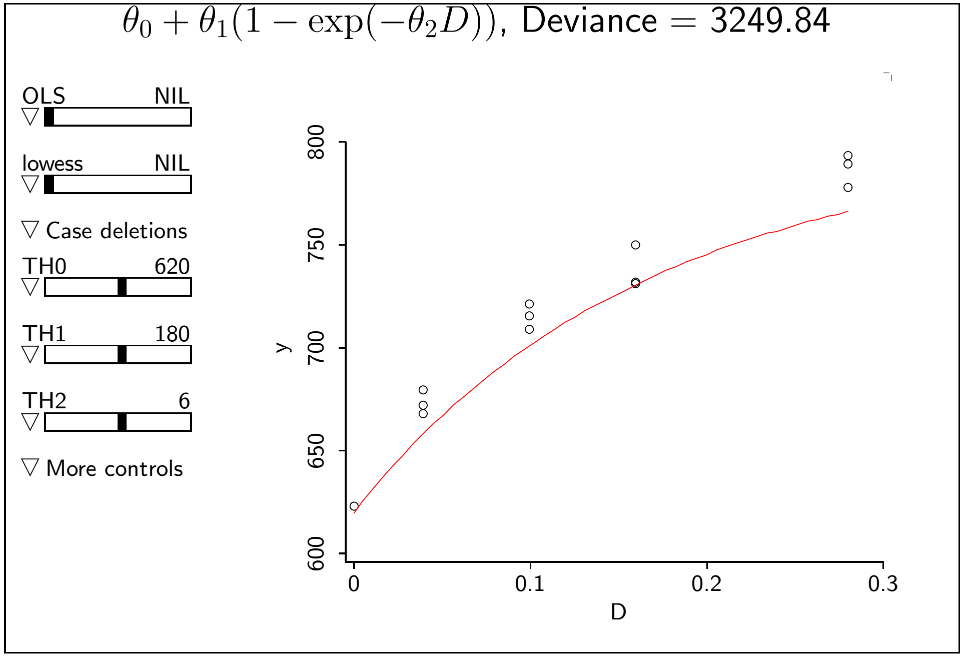
Fitted model, tests, interval estimates, and so on

Criticism

If

```
> m1<-nls( y˜th0+th1*(1-exp(-th2*x)),start=start)
```

How do you find `start`? How to chose the formula? What next? Or, what before? How do you find out? No visual cues.



$\theta_0 + \theta_1(1 - \exp(-\theta_2 D))$, Deviance = 3249.84

OLS NIL

lowess NIL

▽ Case deletions

TH0 620

TH1 180

TH2 6

▽ More controls

Summary

- R works differently for different students, and R is unlikely to work for everyone.
- To help students:
  - Continued work on GUIs.
  - Improved, accessible documentation (Wiki).
  - Continued efforts to promote consistency that might be impossible with a commercial program but can be done in R.
  - Visual cues:

```
> library(alr3)
> m1 <- lm(LBM ˜ Ht + Wt + RCC, data=ais)

> hints(m1)

Methods that understand lm objects:
  conf.intervals     confidence intervals
  inf.index          influence index plots
  mmp                marginal model plots
  pod                partial one-dimensional models
  pure.error.anova   pure error analysis of variance
  anova              analysis of variance
  hatvalues          hat values
  residual.plot      residual plotting methods
  predict            predictions/fitted values
  residuals          residuals of various types
  inv.res.plot       inverse response plot
  delta.method       estimate/se for nonlinear fns
```

Teaching Social-Science Statistics Courses with R

useR! Vienna

John Fox
McMaster University

Canada

June 2006

Course Objectives

- My central pedagogical objectives are to teach
  - statistical concepts (at the introductory level);
  - the application of statistical methods to data (at and beyond the introductory level).
- A statistical "package" (in the broad sense) is a means to an end.
  - Teaching the package is not an end in itself.
  - The package must therefore support the central course objective.
  - The "best" package for a research statistician (surely R at present) may not be best for a social-science student.

Characteristics of the Students

- Most social science students whom I encounter
  - are taking the course because it is required (the case for three of the four courses I'll describe);
  - are math-phobic;
  - have difficulty installing, maintaining, and using computer software;
  - are MS/Windows users.
- Sociology students may be a relatively extreme case, but these kinds of issues are, I suspect, more general.
- It is necessary to meet the students where they are.

Four Courses at the McMaster Sociology Department

- **Sociology 3H06**: A two-semester introductory-statistics course required of Sociology honours majors.
- **Sociology 6Z03**: The same course, but taught in one semester, for Sociology PhD students with a weak background in statistics.
- **Sociology 740**: A one-semester introduction to data analysis, applied regression, linear models, and generalized linear models, required of Sociology PhD students.
- **Sociology 761**: A one-semester selected-topics course for interested graduate students.
  - Recent content: Introductions to matrices, linear algebra, calculus; structural-equation modeling; survival analysis; mixed-effects models for hierarchical and longitudinal data.
- For details: `http://socserv.mcmaster.ca/jfox`

Desirable Features of a Statistical Package

For Basic Statistics Taught to Social-Science Students

- Easy to use.
  - Probably requires a point-and-click interface.
- Easy to install.
  - Permitting the students to work on their own computers.
- Appropriate coverage.
  - In the case of Sociology 3H06/6Z03, corresponding at minimum to Moore's *Basic Practice of Statistics* (the course text).
- "Low threshold/high (or no) ceiling" (borrowing LOGO's motto).
  - Package should not be a dead-end.
- Inexpensive
  - Depends on the institutional context (e.g., availability of site licenses).

Desirable Features of a Statistical Package

For More Advanced Social-Statistics Courses

- Ease of use and installation are less important issues (but not entirely absent).
- Coverage appropriate to the course, use of the package beyond the course, and expense are still important.
- The ability to tailor the package to the course can be important, particularly if certain features are absent (as they nearly inevitably are).

Strengths of R

Extensibility: The Key Strength

- Lisp-like structure enables bottom-up programming:
  - functional programming language;
  - lexical scoping.
- Object orientation facilitates building onto what is already there.
  - Contrast, e.g., the way that statistical models are handled in SAS.
- Package system facilitates organizing, distributing, and using relatively ambitious extensions (and sharing them on CRAN!).
- These characteristics encourage "building the language towards the course" (adapting Graham's approach to Lisp programming).

Strengths of R

Illustrative Course-Related Extensions

- Sociology 3H06 and 6Z03: The **Rcmdr** package, which provides a basic-statistics GUI for R.
- Sociology 740: Diagnostics and other facilities for linear and generalized linear models provided by the **car** package.
  - E.g., added-variable plots via `av.plots()`, component-plus-residual plots via `cv.plots()`, non-sequential ANOVA and analysis-of-deviance tables via `Anova()`.
  - Everything in the course is supported by the **Rcmdr**, though students at this level are better served by learning to write commands.
- Sociology 761:
  - Simple didactic functions for matrix operations–e.g., `GaussianElimination()`.
  - Simple function for constructing a life table, `lifeTable()`.
  - The **sem** package for structural-equation modeling.
  - Survival analysis and mixed-effects models are already handled by recommended and contributed packages (**survival**, **nlme**, **lme4**).

Strengths of R

Other Strengths

- Simple surface syntax (e.g., relative to Lisp)
  - makes it easy to compose commands;
  - makes simple programs intelligible even to novice users.
- Cross-platform availability.
  - For my audience, availability on the Windows platform is key.
- Relatively simple installation, maintenance, and extension.
  - The package system is important here as well.
  - Distribution on CDs or via the Internet is convenient.
- Consistency of use despite the wide diversity of available applications (more than 700 contributed packages and counting on CRAN). For example:
  - the formula interface for linear-like models;
  - the help system;
  - organization of rectangular data sets as data frames.
- Cost can't be beat.

Limitations of R

- Relative difficulty of building cross-platform, easy-to-install GUIs.
  - The **Rcmdr** GUI, for example, is based on a very limited set of widgets (but **tcltk2** may solve this problem).
  - Though extensible, extension of the **Rcmdr** requires at least some **tcltk** programming.
- Lack of high-interaction graphics.
  - Compare what one can do with Lisp-Stat (e.g., in Weisberg and Cook's Arc software).
  - Linking to other software (e.g., GGobi) is *not* what I have in mind (though the ability to link software such as GGobi to R is useful in other contexts).
  - There are some promising developments: clever use of **tcltk**; **tkrplot**; the **rgl** package; **iplots**/JGR.

Limitations of R

- Relative inconvenience of handling very large data sets (but more convenient access to data sets stored in DBMSs may be on the way).
- The S language is not (yet) seen as standard among social scientists.
  - Students may be expected in other contexts to know how to use SPSS or SAS.
  - One shouldn't exaggerate, however, how difficult it is to acquire that knowledge.