

Presentation at JSM03 in San Francisco:  
R vs. Commercial Software?  
Some New Developments in SAS v.9

Wolfgang M. Hartmann

August 11, 2003

## 1 Comparison between R and Commercial Stat Packages

This is somehow limited and maybe a little bit biased:

<b>Advantage</b>	<b>Disadvantage</b>
No fee	none (maybe no tax writeoff :-)
Open source code is: easy addition of code very educational, useful for teaching researchers can be active themselves serves fast communication of new ideas  users can tailor their own needs	algorithms are visible to everyone: difficult cleanup of outdated ideas can only be second best in long run since some people may lookup and improve may become bulky by overlapping packages may lack common look and feel of input options and output maybe subject to fashionable trends
Bugs and testing: users will find bugs and hopefully report them developer will fix bugs for free	accepted is compilable version uncontrolled testing AFTER release some users maybe just turned off hopefully
Tech support: free by newsgroup or forum developer may answer question and improve his implementation	no phone number to call maybe nobody knows maybe conflicting replies nobody to insult :-)

## 2 Software Updates

- much of good software needs updating
- adding new features instead of adding new packages
- it needs rewriting, restructuring from time to time
- update will be difficult if developer is no longer active in the specific area
- sometimes its easier to write a new package than to add features to an older one who's author is no longer interested in the subject
- multiple packages with much overlap and with personal flavors, e.g. seven (7!) packages for Correspondence Analysis
- documented are only separat packages, no overview doc for competing packages outlining differences and helping users to decide inbetween

### 3 Including Features Common to all Packages

Not all researchers who want to add a package are good programmers and will see the need of complying to coding standards. General and efficient subroutine systems should be used by almost all packages for:

- some input options should be the same for a number of packages
- something like SAS' *Output Delivery System* (ODS)
- standardized input (use of XML?)
- use of common graphical output (at SAS v.9 connected to ODS)
- file in- and output
- parallel processing
- sparse matrix algebra (does R know about sparsity?)
- implement central message file system
- PMML (frame computing environment) similar to DCOM

A message file system makes translation of message output possible without having a foreign translator touching the code.

## 4 Competition among Commercial Stat Software Providers

- Labor intensive software development takes large initial investment for successful startup
- Competition between Stat software packages is very strong and not much profit can be made
- SAS and SPSS changed successfully to selling "Business" software
- Now also S-plus and Statistica are trying to enter this much more profitable market
- Some smaller stat software sellers have left the market or merged with larger ones (BMDP and SYSTAT with SPSS, Statview with JMP)

## **5 New PROCs in SAS Stat v. 9.0,9.1**

**MI** (prod.) missing value imputation

**MIANALYZE** (prod.) analysis of missing value imputation

**TPHREG** (exp.) adds CLASS statement to PHREG

**ROBUSTREG** (prod.) M-, S-, MM-, LTS estimation

**POWER** (prod.) power and sample size

**GLMPOWER** (prod.) power and sample size

**SURVEYREG** (prod.) finite population

**SURVEYLOGISTIC** (prod.) finite population

**DISTANCE** (exp.) various distance measures

## **6 New PROCs in SAS ETS v. 9.0,9.1**

**VARMAX** (prod.) multivariate time series

**UCM**

**TIMESERIES**

## **7 New PROCs in Enterprise Miner v. 9.0,9.1**

**EMCLUS** (prod.) model based clustering

**SVM** (exp.) support vector machines

## 8 New Features of old PROCs in SAS v. 9.0,9.1

- FREQ**
- exact CI's for 2 by 2 tables
  - exact CI's for common odds ratios
  - BDT option (Tarone's adjustment)
  - ZERO option in WEIGHT statement

- GLM**
- exact  $p$  values for multivariate tests

- MIXED**
- various residual diagnostics added

- LOGISTIC**
- exact logistic regression
  - SCORE statement
  - new CLASS parametrizations

- PHREG**
- NORMALIZE option
  - WEIGHT statement

- GENMOD**
- new CLASS parametrizations

- LIFETEST**
- tests for comparing multiple samples
  - improved CI's for survivor function

## 9 Stat Graphics in ODS: SAS v. 9.0,9.1

ANOVA	CORRESP	GAM
GENMOD	GLM	KDE
LIFETEST	LOESS	LOGISTIC
MI	MIXED	PHREG
PRINCMP	PRINQUAL	REG
ROBUSTREG	TPSPLINE	