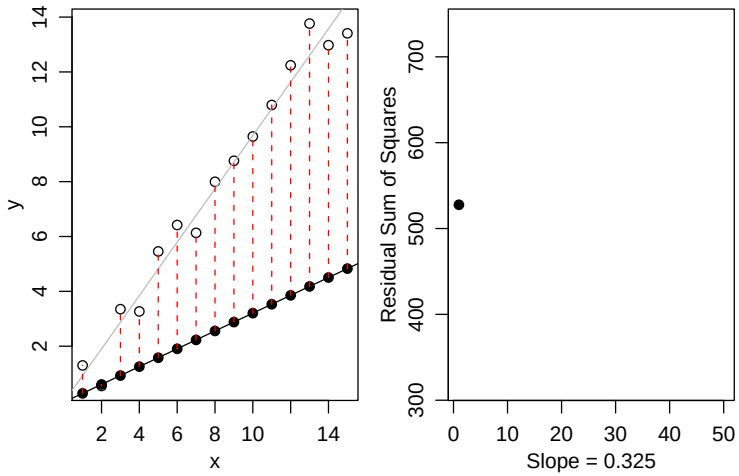
R in Stat500: Demonstration, Computing, and Graphics

with An Introduction to the R Package **ISU.Stat500**

Yihui Xie

Department of Statistics, Iowa State University

October 22, 2009



Abstract

In this talk I'll mainly intRoduce how to use R to fuRtheR examine the pRoblems and datasets we met in the couRse Stat500 based on what we leaRned from Stat579. First, I'll show the possibility of doing RepRoducible homework in a all-in-one manneR (data + R + Sweave + LaTeX $\Rightarrow$ PDF output) and making our TA Chuanlong's life pRobably easieR in this way; then I'll bRiefly Review the mateRials taught in Stat579 and Stat500, and focus on selected topics using R as a complement to Dr Dixon's class (tentatively include demos for bootstRapping and CI, sample size computation, power, contRasts in R, QQ plots, etc); in the end I'll give an incomplete compaRison of SAS and R. To sum these pieces of thoughts up, Hui and I decided to contRibute an R package named **ISU.Stat500** which will hopefully serve as a useful collection of R code for Stat500.

A self-intRoduction

- speaks two types of English (new to US)
- 4-year R user (SAS?)
- use R twice a day
- three types of statisticians (I want to be...?)

Outline

Demo of R and Sweave (homework 6)

- Q1: need a formula sheet beside you, then arithmetic calculations with R
  - whenever you need p-values, use the probability distribution functions `p*(...)` (e.g. t dist'n: `pt(q, df)`; Normal: `pnorm(q, mean, sd)`)
  - whenever you need model-based CI, use quantile functions `q*(...)`
- Q2: `lm()` is your friend; fit a linear model first (e.g. `fit = lm(y~x)`) and extract *exactly* what you want – Chuanlong does not like numbers flooding everywhere in your homework
  - want coefficients? `coef(fit)`
  - want t test on slope? `summary(fit)` and extract the single p-value from the coefficient matrix
  - want CI of $\hat{Y}$? `predict(fit, newdata, interval = 'confidence')`
  - prediction interval? `predict(..., interval = 'prediction')`

Demo cont'd (homework 6)

- Q3: apply `lm()` to each data set (use `by()` to avoid loops over data sets)
  - want regression summary? again, extract whatever you want (R square? p-values?) to show to Chuanlong
  - want picutres? `plot(x,y)` (with regression line, `abline(fit)`, or LOWESS line if you like)
- Q4: still use the object returned by `lm()`
  - plot of residuals vs predicted values? `plot(fit, which = 1)`
  - box-cox transformation? get $\hat{\beta}_1$ from `lm(log(sd(y))~log(mean(y)))`
  - lack of fit test? something like `sum((yi-yhat)^2)` and `pf()` to get the p-value
  - inverse prediction? arithmetic

What's the point of Sweave?

- reproducible
  - anything in a scientific paper should be reproducible (what about homework? my terrible experience)
  - even your grandson can reproduce your homework if interested 50 years later (if data, R and LaTeX still exist on this planet)
- LaTeX
  - free (yeah!)
  - plain text (Micro$oft word? oh no)
  - can produce high-quality (PDF) output (yeah!)
- R
  - free (yeah!)
  - let you know the computation details whenever you want to know
  - richness of (statistics-/data-related) functions (>2500 in base)
  - highly extensible (more than 2000 add-on packages; thus drowned us)

The R package exams

- Chuanlong is extremely good at programming
- If all of us could use R, ...
- R is object-oriented; everything is an object
- just send the Sweave document to Chuanlong and use R to compare our answers to the correct answer; why bother using eyeballs?
- an implementation of using R to generating exams: Grün and Zeileis (2009)
- let R grade our homework? (difficulties, of course)

Outline

Another big big world

- base R is a big world (I'm only familiar with a small portion of functions, 100? 200?)
- probably should not learn too much about base R, because you'll get addicted and unwilling to use add-on packages
- Stat579 is more focused on manipulating messy data (`reshape` and `plyr`) and drawing elegant graphics (`ggplot2`)
- that's only another big world of R (can we appreciate these packages without having been tortured by really messy data from our clients? poor Hadley...)
- unfortunately, I have already been addicted to base R (so congratulations to 579 students!)

My concerns

- we don't have messy data in Stat500 (thanks, Dr Dixon)
- `plot()`, `hist()`, `boxplot()` will do in most cases
- we play with `t.test()`, `aov()` and `lm()` more

Outline

Dealing with data sources

- with small data sets, `read.table()` suffices (what about large data?)
- can read the data from Dr Dixon's website directly (e.g. `read.table('http://www.public.iastate.edu/~pdixon/stat500/data/creativity.txt', header = TRUE)`)
- can download the data files using R (see `?download.file`)
- can share the data with our friends through the intranet in R (see `?socketConnection`)
- can read data from (g)zipped files (see `?gzip`)
- even can read/share data with Google Docs in R! (see Temple Lang (2009))
- imagination exhausted?

Demo: boostrapping, CI and QQ plot

- R package `animation` by Xie (2009) (hey! we hate advertising!!)
- `library(animation)`
- `?boot.iid`
- `?conf.int` (change α)
- `?sim.qqplot` (don't overinterpret QQ plots; this function is available on R-Forge now; to appear in `animation v1.0-6`)

Computing: sample size, power, contrasts

- when you feel "R cannot do this, this and this", it is time to learn more about R or just contribute something to R
- sample size computation based on power:
  $\delta = (t_{2(n-1),1-\alpha/2} + t_{2(n-1),1-\beta}) s_p \sqrt{2/n}$
  - what we learned in class: iterate explicitly from an initial n
  - what I do: treat the equation as $f(n) = 0$ and find its root by `uniroot()` (e.g. homework 5 Q3)
  - can contribute a simple function with one line of code to compute sample size based on power, effect size, etc
- relationship of power and distribution assumption
  - a problem raised in lab before mid-term exam but not easy to answer (no definite solution)
  - can get power β from simulation: given an effect size, compute the proportion of cases not rejecting null hypothesis (i.e. type II error rate)

Computing: sample size, power, contrasts (cont'd)

two sample t-test on U(0, $\sqrt{12}$) and U(.5, .5+$\sqrt{12}$)

```
# real power
compute.power = function(n = 30, conf.level = 0.95,
    delta = 0.5, sigma = 1) {
    pt(delta/(sigma * sqrt(2/n)) - qt(1 - (1 - conf.level)/2,
         2 * (n - 1)), 2 * (n - 1))
}
# simulated power from Uniform dist'n
sim.power = function(n = 30, conf.level = 0.95, B = 10000) {
    mean(replicate(B, t.test(runif(n, 0, sqrt(12)), runif(n,
        0.5, 0.5 + sqrt(12)))$p.value) < 1 - conf.level)
}
> compute.power()
[1] 0.4741093
> sim.power()
[1] 0.4689
> compute.power(50)
[1] 0.696329
> sim.power(50)
[1] 0.6979
```

Computing: sample size, power, contrasts (cont'd)

- contrasts in `lm()` does not look straightforward, but it does have an argument `contrasts`
- need further reading on Venables and Ripley (2002) Section 6.2
- for impatient guys, just use `make.contrasts()` in Warnes (2009) to convert contrasts like `c(1, -1/2, -1/2)` to what R needs

ISU.Stat500

- goals
  - save our time (e.g. use `library(ISU.Stat500);data(creativity)` instead of `read.table()` for ever)
  - examples for further exploring our data sets
  - provide functions or guides for statistical analysis which do not exist in base R (e.g. Benjamini and Hochberg procedure for controlling FDR; `levene.test()` already exists in `car`)
  - simulations and demos helping us gain insight into methods
- on R-forge now
  - https://r-forge.r-project.org/R/?group_id=208
  - `install.packages('ISU.Stat500', repos = 'http://r-forge.r-project.org')`
  - only a sketch there
  - demo

Outline

Curious about "source code"?

Figure 1: We want to see the source code; freedom of obtaining source code is important

Open source – share with other people

Figure 2: Open source software community (e.g. R-help@r-project.org)

Learning R is like...

Figure 3: It hurts to learn R? But it's rewarding!

Comparing SAS with R is like...

Figure 4: R is too young compared to SAS (and I'm too young to talk about it)

So take a look at what they said

Frank Harrell (SAS User, 1969-1991) R-help (September 2003)

Overall, SAS is about 11 years behind R and S-Plus in statistical capabilities (last year it was about 10 years behind) in my estimation.

Frank Harrell R-help (November 2003)

I quit using SAS in 1991 because my productivity jumped at least 20% within one month of using S-Plus.

Brian D. Ripley s-news (May 1999)

Some of us feel that type III sum of squares and so-called ls-means are statistical nonsense which should have been left in SAS.[a]

[a] note here a story of 50000 SVN commits

So take a look at what they said (cont'd)

Bill Venables R-help (November 2000)

I'm really curious to know why the "two types" of sum of squares are called "Type I" and "Type III"! This is a very common misconception, particularly among SAS users who have been fed this nonsense quite often for all their professional lives. Fortunately the reality is much simpler. There is, by any sensible reckoning, only ONE type of sum of squares, and it always represents an improvement sum of squares of the outer (or alternative) model over the inner (or null hypothesis) model. What the SAS highly dubious classification of sums of squares does is to encourage users to concentrate on the null hypothesis model and to forget about the alternative. This is always a very bad idea and not surprisingly it can lead to nonsensical tests, as in the test it provides for main effects "even in the presence of interactions", something which beggars definition, let alone belief.

So take a look at what they said (cont'd)

Bill Venables 'Exegeses on Linear Models' paper (May 2000)

I was profoundly disappointed when I saw that S-PLUS 4.5 now provides "Type III" sums of squares as a routine option for the summary method for aov objects. I note that it is not yet available for multistratum models, although this has all the hallmarks of an oversight (that is, a bug) rather than common sense seeing the light of day. When the decision was being taken of whether to include this feature, "because the FDA requires it" a few of my colleagues and I were consulted and our reply was unhesitatingly a clear and unequivocal "No", but it seems the FDA and SAS speak louder and we were clearly outvoted.

So take a look at what they said (cont'd)

Frank Harrell R-help (November 2004)

There are companies whose yearly license fees to SAS total millions of dollars. Then those companies hire armies of SAS programmers to program an archaic macro language using old statistical methods to produce ugly tables and the worst graphics in the statistical software world.

David Kane R-SIG-Finance (December 2004)

I have never heard anyone (knowledgable or otherwise) claim that, in the absence of transition costs, SAS is better than R for equity modeling. If you come across any such claim, I would be happy to refute it.

So take a look at what they said (cont'd)

Rene M. Raupp and Kjetil Brinchmann Halvorsen R-help (May 2005)

Rene M. Raupp: Does anybody know any work comparing R with other (charged) statistical software (like Minitab, SPSS, SAS)? [...] I have to show it's as good as the others.
Kjetil Brinchmann Halvorsen: Sorry. That will be difficult. Couldn't it do to prove it is better?

David Brahm (announcing the sudoku package) R-packages (January 2006)

Any doubts about R's big-league status should be put to rest, now that we have a Sudoku Puzzle Solver. Take that, SAS!

So take a look at what they said (cont'd)

Douglas Bates R-help (March 2007)

You must realize that R is written by experts in statistics and statistical computing who, despite popular opinion, do not believe that everything in SAS and SPSS is worth copying. Some things done in such packages, which trace their roots back to the days of punched cards and magnetic tape when fitting a single linear model may take several days because your first 5 attempts failed due to syntax errors in the JCL or the SAS code, still reflect the approach of "give me every possible statistic that could be calculated from this model, whether or not it makes sense". The approach taken in R is different. The underlying assumption is that the useR is thinking about the analysis while doing it. – (in reply to the suggestion to include type III sums of squares and lsmeans in base R to make it more similar to SAS or SPSS)

On Stat500...

- "... (Wilcoxon rank sum test) algorithm used by R does not work if there are ties in the data..." (see `?wilcox.test` and will get the answer)
- "... Here's where R is not nearly as flexible as SAS..." (agree if "*giving everything no matter you want or don't want*" means "*flexible*")
- "... and from here on out, you're stuck, at least without hand coding things... there may be appropriate functions in add on packages, but I haven't found them..."[1] (as we have seen (1) R does not want to regard SAS as a industry-standard (2) R is flexible to be extended)

[1] `http://www.public.iastate.edu/~pdixon/stat500/R/judges.r`

News

- Norman Nie, founder and former CEO of SPSS, has moved to REvolution (a commercial company based on R)
- I'm not sure this is good or bad news

Materials used in this talk

- homework 6 for Stat500: http://www.public.iastate.edu/~pdixon/stat500/homework/hw6.pdf
- datasets can be found at: http://www.public.iastate.edu/~pdixon/stat500/datasets.html
- quotes are from the R package `fortunes`
- what? picutres?...

Bibliography

Grün B, Zeileis A (2009). "Automatic Generation of Exams in R." *Journal of Statistical Software*, **29**(10), 1–14. ISSN 1548-7660. URL http://www.jstatsoft.org/v29/i10.

Temple Lang D (2009). *RGoogleDocs: Primitive interface to Google Documents from R*. R package version 0.2-2, URL http://www.omegahat.org/RGoogleTrends/.

Venables WN, Ripley BD (2002). *Modern applied statistics with S*. Springer verlag.

Warnes GR (2009). *gmodels: Various R programming tools for model fitting*. R package version 2.15.0, URL http://CRAN.R-project.org/package=gmodels.

Xie Y (2009). *animation: Demonstrate Animations in Statistics*. R package version 1.0-6, URL http://animation.yihui.name.

Thanks!

- Questions and comments?
- Email: `sprintf("%s@%s", "xie", "yihui.name")`
- Slides will be available online later:
  `browseURL("http://yihui.name/en/vitae")`