Data Analysis Tools in an Age of Computer Technology

John Maindonald

Centre for Mathematics and Its Applications,
Australian National University

May 28, 2013

The Two R's of R

Ross Ihaka and Robert Gentleman developed the initial version of R at Auckland University in NZ. They set out to create an open source version of the earlier AT&T S language.

From Auckland to Canberra

Auckland University and Mt Albert Research Centre

Computing for Statistics — Then & Now

1970s	2013
As pencil & paper, but faster	Improve on pencil & paper
Rely on statistical theory	Theory, plus computation
(eg, Central Limit Theorem)	(simulation, bootstrap, . . .)
Data was similar to 1930s	New types of data; 'big' data
Many incompatible programs	Incompatible programs + R

Why R? — Data Analysis

- Designed to encourage good statistical practice
  - analyses and graphics go hand in hand
- User can do most things without writing new code
  - but can use new code where necessary
- New methods appear first in R, later elsewhere
- R is much more than a language and a system
  - It is a community of users and of expertise
  - Working co-operatively makes great things possible
- More reasons below (and some negatives!)

Why R? – Two Opinons

- I use R exclusively to perform data analysis, and Python for more generic programming tasks (e.g. workflow control of a computer model).
  [Paul Hemstra: stackexchange]
- Since the language has been around for ever, lots of people have done things that you're likely to want to do. . . . , when faced with a hard problem, . . . download the package and get to work. And R "just works": you give it a dataset, and it knows what summary statistics are useful. You give it some results, and it knows what plots . . . . [Ben Dundee: stackexchange]

http://programmers.stackexchange.com/questions/181342/r-vs-python-for-data-analysis

R: Language++ & Library++

- The R System is a Library of packages
- Each package has access, in principle, to all others
- New packages leverage off existing code –
  - I/O, data manipulation, graphics, statistics,
  - spatial interpolation, ...
- A package management system handles dependencies
- Minimal consistency is enforced across packages
  - Documentation, Coding standards, Directory tree

R's Role in the CoheRent Whole

The R project is

- an international community
  - statistical experts, expert users,
  - software development experts
- a repository for software and documentation
  - statistics, graphics, . . . , and much else
- an information resource
  - again, for statistics and much else
- a marketplace for ideas
  - It brings together diverse skills & insights
  - Ideas reflect diverse scientific backgrounds.

UmbRella R – Examples

- Spatial Analysis
  - Initially, interfaces to GRASS, SAGA, etc.
  - Steadily, abilities have moved to R
- Motion Charts
  - For now, *googleVis* interfaces to Google apps
  - Work at Auckland Univ will make these part of R.
- Commentary
  - Interfacing is a good way to start
  - Gradually, transfer functionality to R

What is Wrong with R?

- The language model is dated
- R is powerful, but not powerful enough
  - Users now want even more powerful tools
- R code is slow
  - hence call C code if faster execution is needed
- R has been a first draft of a rewrite of S
  - Much has been learned as the draft has evolved
- Some specific areas require attention
  - e.g., the graphics for lm (regression) models.

On Balance, Virtues vs Defects

- R is a Mature System with a Dated Language Model
  - Maturity wins!
- Slow R code is mainly a problem for power users
- The Package Management System is a big plus
- An evolving R will be with us for a while yet!
- Challenge: Use what we now know to improve R.
  - Easiest if changes to base R can be avoided
  - *knitr* is a good example of what is possible

Rethinking Statistical Analysis

- Simulation is the basic tool for estimating sampling distributions
- Sometimes we are lucky, and mathematical results make the recourse to simulation unnecessary
- Simulation does make theoretical assumptions
  - Theoretical assumptions are inescapable!
- Alternatives – the bootstrap
  - Here the data determines the distribution
  - This introduces data sampling error

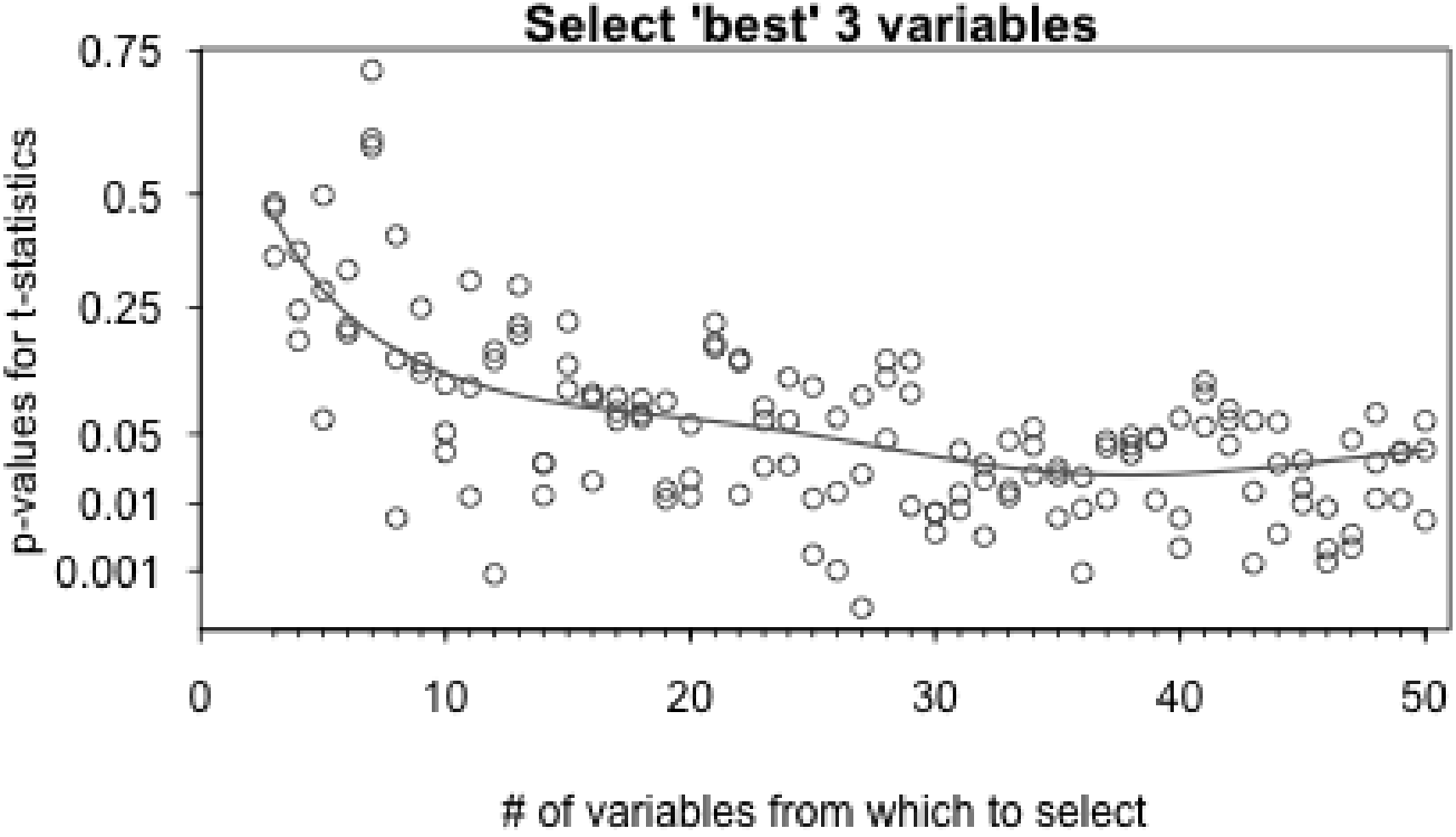
Simulation – Variable Selection

y variable	x1	x2	x2	. . .
Noise	Noise	Noise	Noise	. . .
rnorm(100)	rnorm(100)	rnorm(100)	rnorm(100)	. . .

- With no variable selection (3 explanatory variables x1, x2 and x3), p-values have an expected value of 0.5
- As the number of variables available for selection increases, selecting the best 3 gives ever smaller notional *p*-values.

Simulation – Variable Selection

Here variable values are random normal noise



Message: Stepwise regression has serious traps!

Rethinking Regression

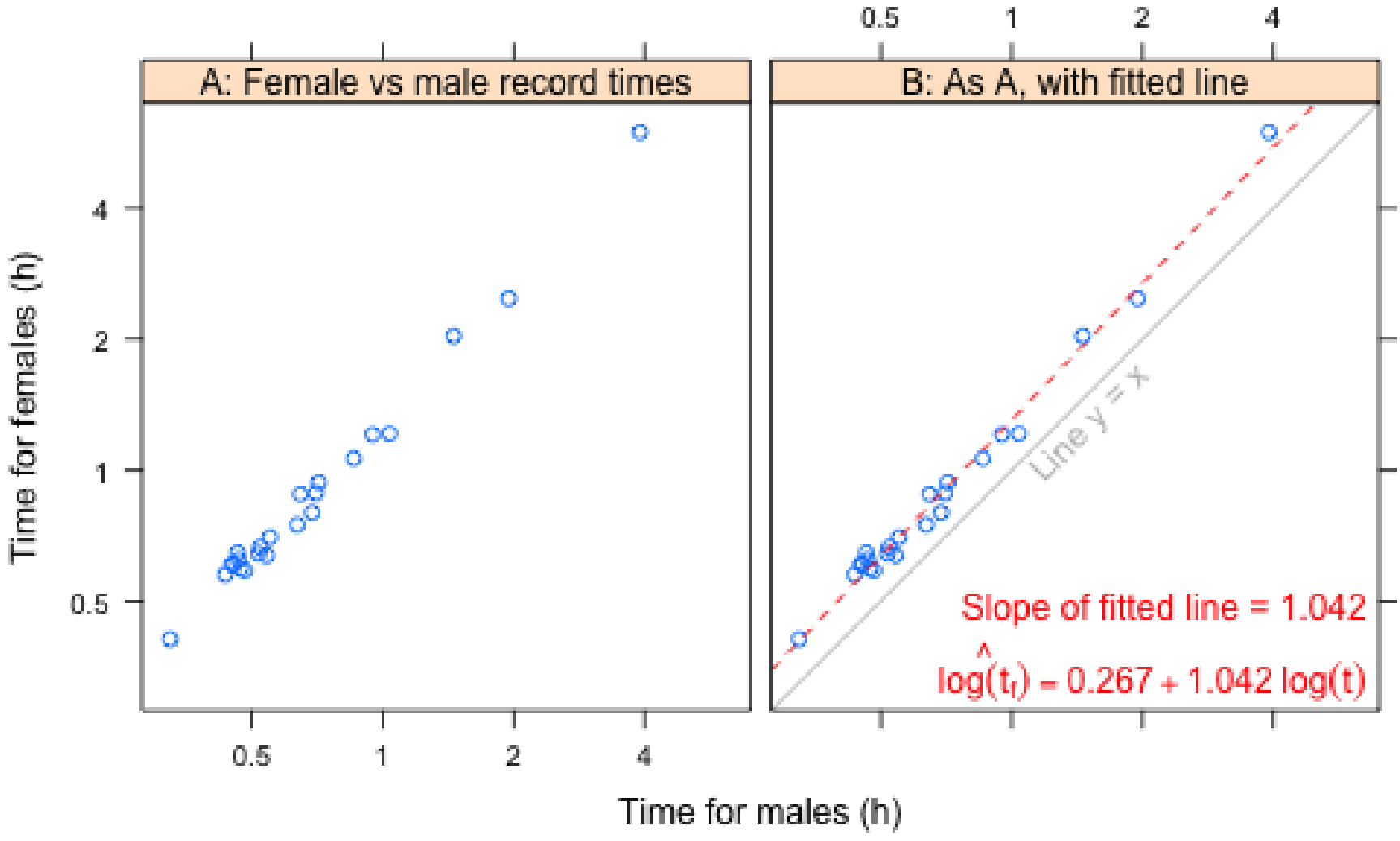
- Most R model objects can both be printed and plotted
  - *plot()*, with an lm object, gives diagnostic plots
- In small datasets, some features are just noise
  - Answer – do such plots for simulated data
  - Hence plotSimDiags() in DAAG (>= 1.16)

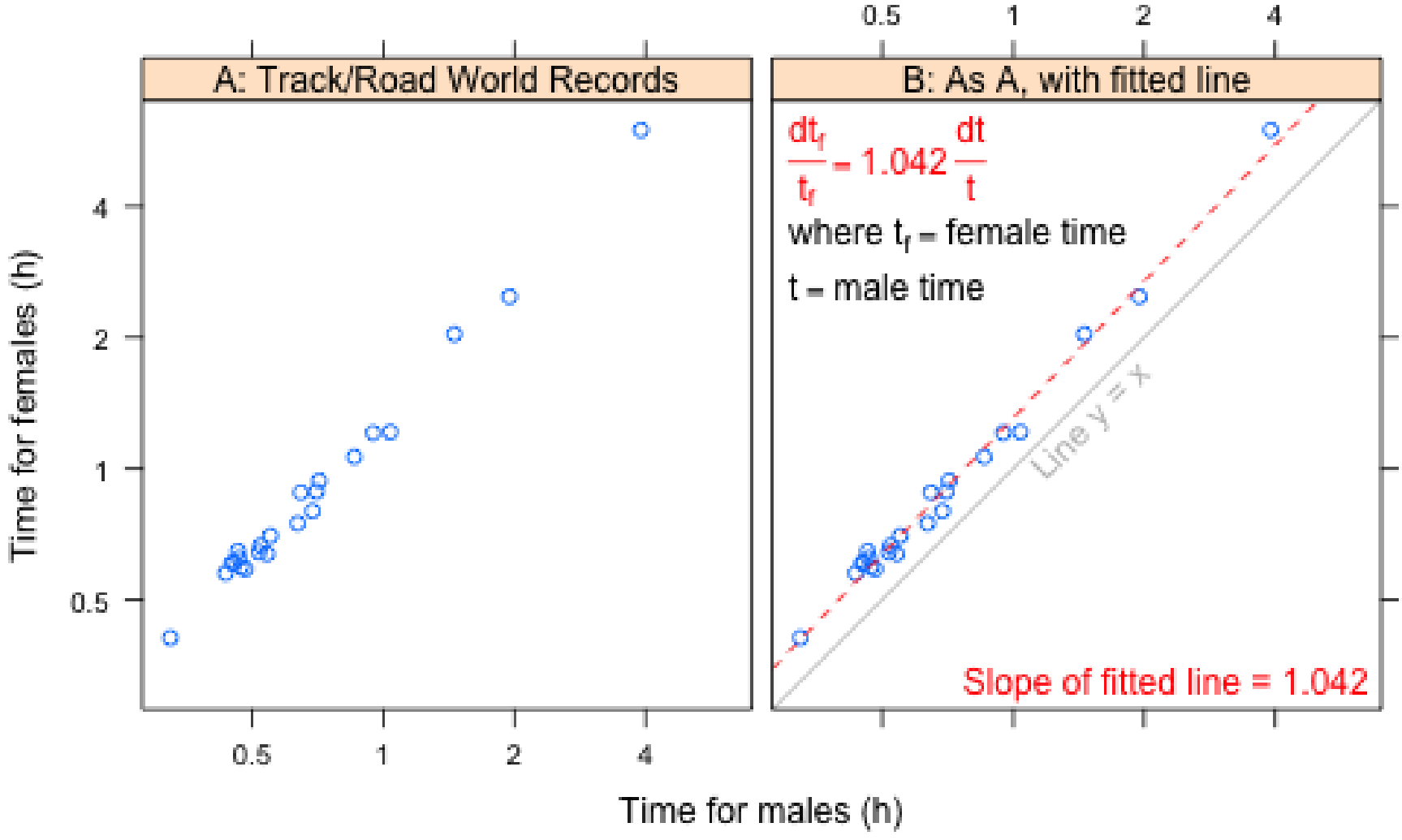
Rethinking Regression – Data

The following are male and female record times for hill races in Northern Ireland. Race distances are in miles, extent of climb is in feet, and times are in hours.

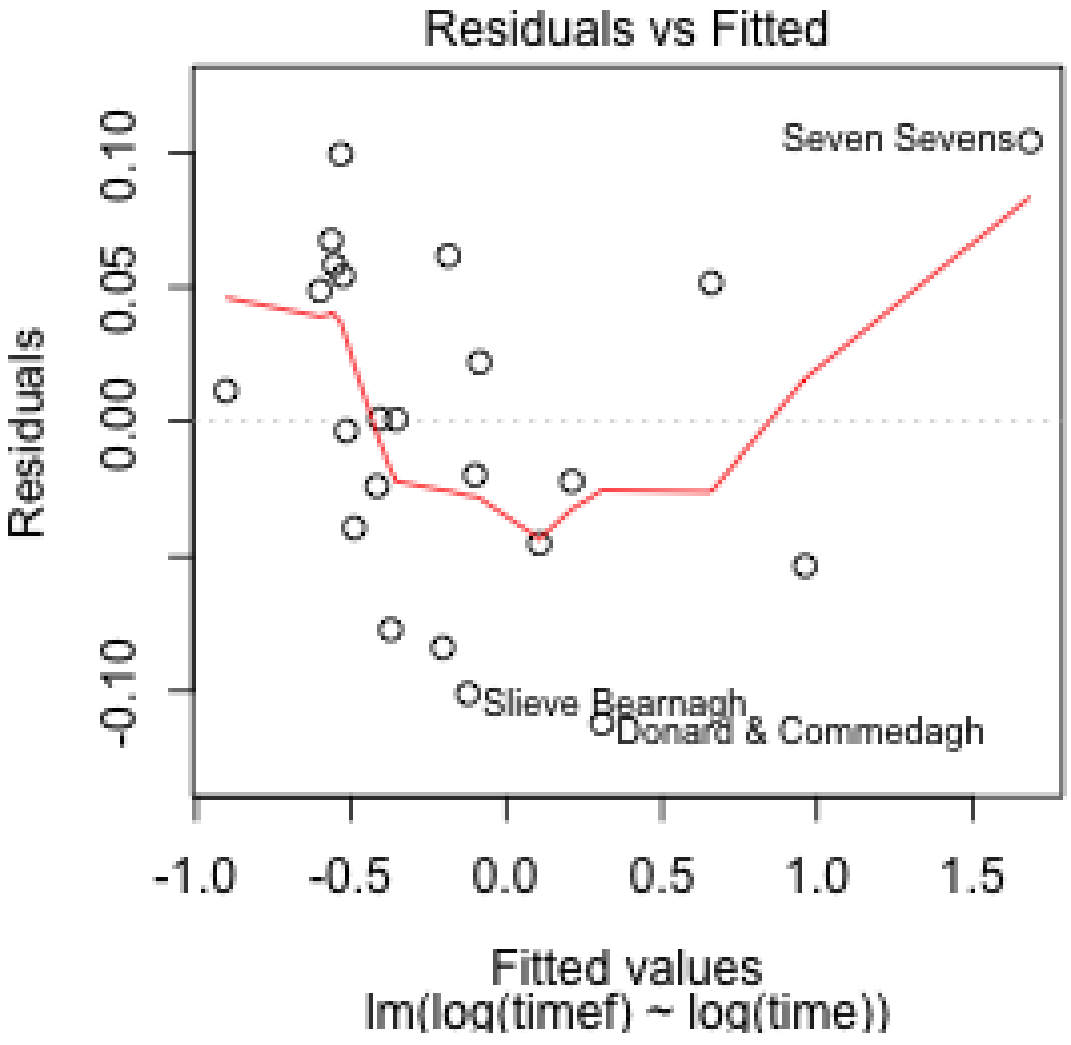
	dist	climb	time	timef
Binevenagh	7.5	1740	0.8583	1.0644
Slieve Gullion	4.2	1110	0.4667	0.6231
Glenariff Mountain	5.9	1210	0.7031	0.8869

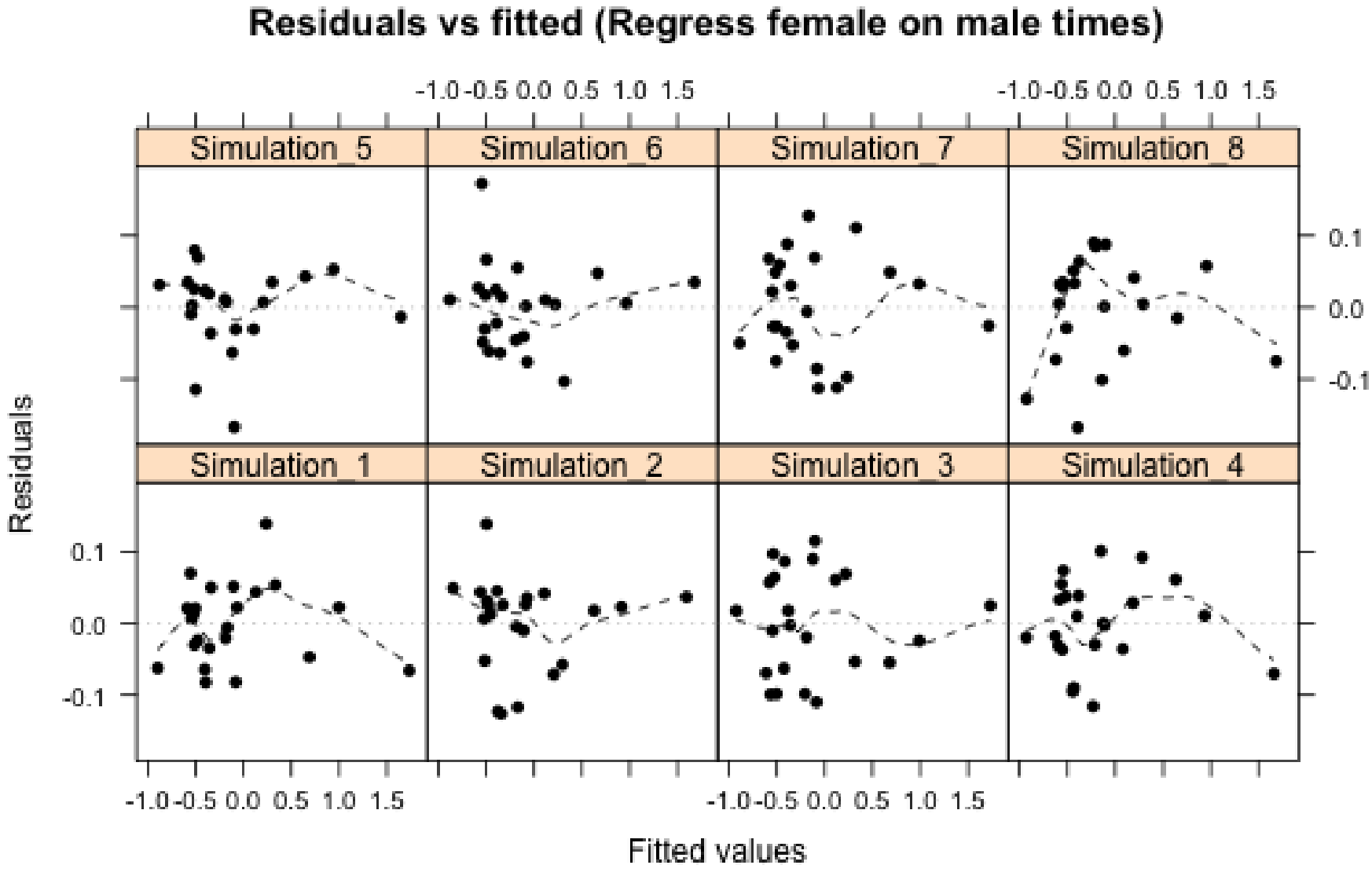
... 22 further rows of data


A: Female vs male record times
B: As A, with fitted line
Time for females (h)
Time for males (h)
0.5
1
2
4
Line y = x
Slope of fitted line = 1.042
$\hat{\log(t_f)} = 0.267 + 1.042 \log(t)$


A: Track/Road World Records
B: As A, with fitted line
$\frac{dt_f}{t_f} = 1.042 \frac{dt}{t}$
where t_f = female time
t = male time
Line y = x
Slope of fitted line = 1.042
Time for females (h)
Time for males (h)
0.5
1
2
4

Diagnostic plot 1





These simulate the sampling distribution of diagnostic plot 1.

Fixing termplot()

- Currently, useful for a limited range of models
  - For these, it gives highly helpful insight
- Output is not ideal for transformation terms
  [e.g., log(x) terms]
- Shows partial regression effects
  - currently, for main effects terms only

Rethinking the Textbook

Make the web an adjunct to the printed page
The same issues arise for course notes.

- Texts commonly have their own R package
- Provide data sets (obviously) & functions
- Use functions to package any major code chunks
  - For ephemeral fn's, create by sourcing scripts – e.g., *gamclass* has functions like fig2.1()
- Vignettes, R markdown, ...
- There is much more to be done and said ...
  - This is, after all, a time of rapid computing related technological change!

Acknowledgements

- Friedrich Leisch and R-core, for *Sweave()*
- Yihui Xie for *knitr*, which improves on *Sweave()*
  _ Slides were created usingf *knitr*.