

A Compendium Platform for Reproducible, R-based Research with a focus on Statistics Education

Introduction

- Acknowledgments
- Motivation (based on frustration)
- Reproducible Research and the Compendium:
 - Literature
 - The compendium redefined
 - Proposed solution
- Screenshots
- Conclusions & Future work

<http://www.freestatistics.org> >> Publications

<http://www.wessa.net/download/user2008.pdf>

Acknowledgments

- **Funding** (we accept money):
 - K.U.Leuven Association, OOF 2007/13
 - Donations from private companies
- **Contributors:**

Bart Baesens, Eric Bloemen, Eddy Borghers,
Christophe Croux, Claude Doom, Dirk Janssens,
Christine Lourdon, Koen Milis, Stephan Poelmans,
Riko van Dijk, Guido Van Rompuy, Ed van Stee,
Larry Weldon, Patrick Wessa
(www.freestatistics.org)

My frustration

- Teaching Time Series Analysis
- Exam question:
Compute $(1-B) Y[t]$ if you know that
 $Y[t] = \{5, 8, 2, 3, 7, 1, 4\}$
 $BY[t] = Y[t-1]$

My frustration

- Teaching Time Series Analysis

- Exam question:

Compute $(1-B) Y[t]$ if you know that

$$Y[t] = \{5, 8, 2, 3, 7, 1, 4\}$$

$$BY[t] = Y[t-1]$$

- Result

- Less than 8% of students got it right.
- More than 90% of students could prove Wold's decomposition theorem!

Conclusion?

- I am an extremely bad educator.
- I shouldn't have asked that silly question: Students can only reproduce theories – they are not required to understand them!
- ...
- Or maybe there is something wrong with our approach towards statistics education?

A new approach is needed

- Within the pedagogical paradigm of (social) constructivism:
 - Interaction & collaboration (peer review)
 - Experimentation
 - Responsibility (social control)
- => learning & computing technology
- => we need to Free Statistics of irreproducible research
- => www.FreeStatistics.org

Computing

**Reproducible ~~Research~~ and the
Compendium**

Green's comment

- Now the methodology is often so complicated and computationally intensive that the standard dissemination vehicle of the 16-page refereed learned journal paper is no longer adequate. ...

Most statistics papers, as published, no longer satisfy the conventional scientific criterion of reproducibility: could a reasonably competent and adequately equipped reader obtain equivalent results if the experiment or analysis were repeated?

Claerbout's principle*

- An article about computational science in a scientific publication is not the scholarship itself, it is merely **advertising** of the scholarship. The actual scholarship is the complete software development environment and that complete set of instructions that generated the figures.

Jan de Leeuw's comments*

- First, there is **no reason to single out figures**. The same "Principle" obviously applies to tables, standard errors, and so on. The fact that figures often happen to be easier to reproduce, does not preclude that we should apply the same rule to any form of computer-generated output.
- Second, there is **no reason to limit the Claerbout's Principle to published articles**. We can make exactly the same statement about our lectures and teaching, certainly in the context of graduate teaching. We must be able to give our students our code and our graphics files, so that they can display and study them on their own computers (and not only on our workstations, or in crowded university labs).
- And third, and perhaps most importantly, it is not clearly defined what a "software environment" is. Buckheit and Donoho apply the principle in such a way that everybody who wants to check their results is forced to buy MatLab(R). Not Mathematica(R), Macsyma(R), or S-plus(R). Those you may need to buy for other articles. This violates the **Freeware Principle**...

*Source: Jan de Leeuw, Reproducible Research: the Bottom Line, 2001, online

Sweave package

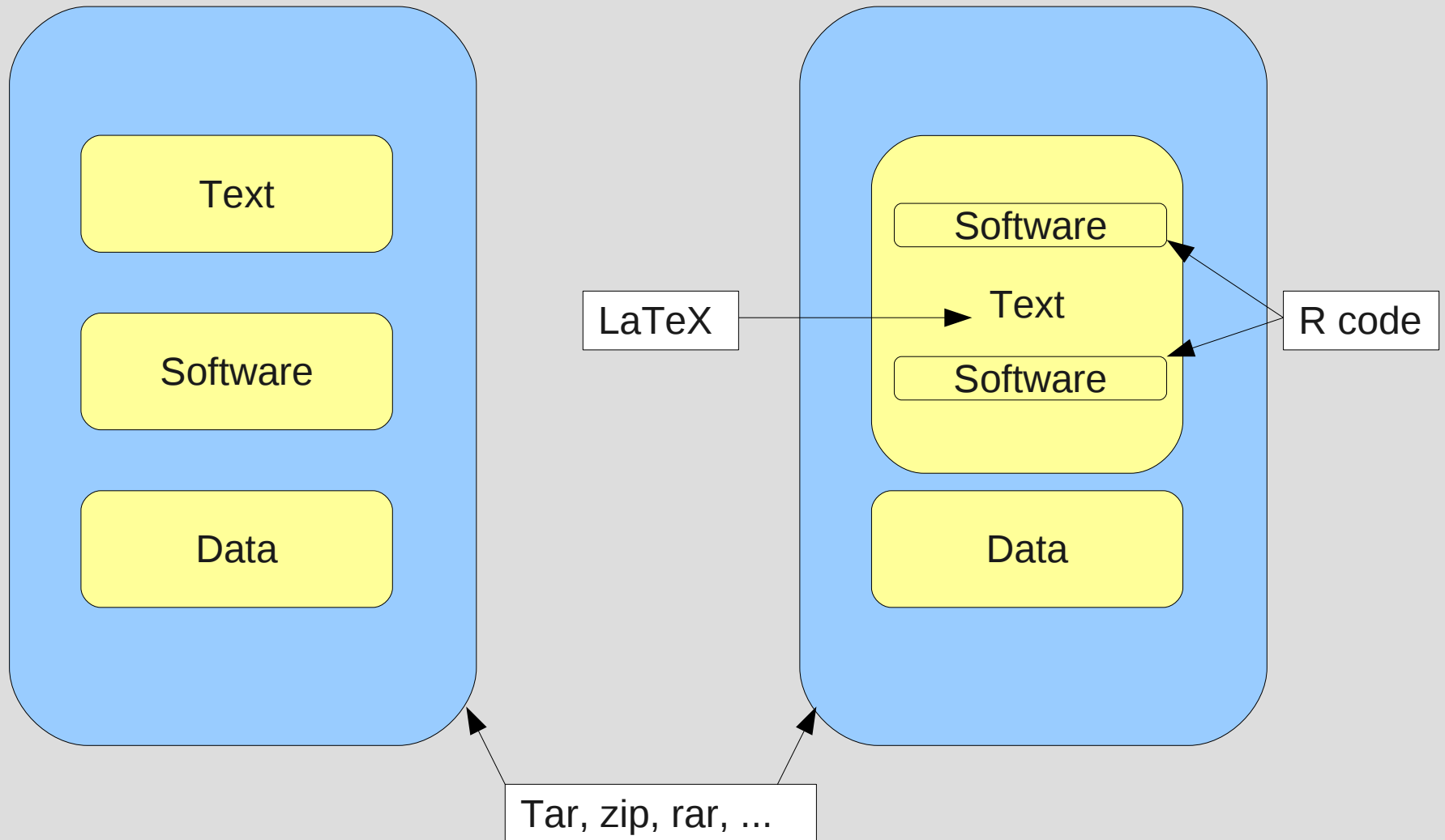
- Excellent solution (in general)
- Somewhat impractical for education because the student:
 - is required to DIE (Download, Install, Execute)
 - must have a working knowledge of LaTeX and R
 - must recreate a working compendium (for each submission)
- Not designed with educational research in mind: there is no way to monitor/measure the actual learning activities

Compendium

- Original definition:

An electronic collection of Text, Data and Software that allows the reader to reproduce the research that is presented in the document

Compendium



File Edit Action Settings Help



Search:

Filename	Permissions	Owner	Group	Size	Timestamp	Link
GOstats	drwxr-xr-x	madman	bioc	0 B	06/01/04 08:30 pm	
└─ data	drwxr-xr-x	madman	bioc	0 B	06/01/04 08:30 pm	
└─ ALL.rda	-rw-r--r--	madman	bioc	12.6 MB	04/05/04 02:00 pm	
└─ Bdists.rda	-rw-r--r--	madman	bioc	171.3 KB	04/05/04 02:00 pm	
└─ Ndists.rda	-rw-r--r--	madman	bioc	171.3 KB	04/05/04 02:00 pm	
└─ inst	drwxr-xr-x	madman	bioc	0 B	06/01/04 08:30 pm	
└─ doc	drwxr-xr-x	madman	bioc	0 B	06/01/04 08:30 pm	
└─ GOstats.bib	-rw-r--r--	madman	bioc	61.7 KB	04/11/04 11:33 pm	
└─ GOstats.Rnw	-rw-r--r--	madman	bioc	48.7 KB	06/01/04 08:30 pm	
└─ GOstats.tex	-rw-r--r--	madman	bioc	6.2 KB	06/01/04 08:30 pm	
└─ GOvis.Rnw	-rw-r--r--	madman	bioc	14.7 KB	05/14/04 06:13 pm	
└─ Scripts	drwxr-xr-x	madman	bioc	0 B	06/01/04 08:30 pm	
└─ distance.R	-rw-r--r--	madman	bioc	4.1 KB	05/13/04 07:47 pm	
└─ TESTP.R	-rw-r--r--	madman	bioc	2.1 KB	04/05/04 02:00 pm	
└─ man	drwxr-xr-x	madman	bioc	0 B	06/01/04 08:30 pm	
└─ DESCRIPTION	-rw-r--r--	madman	bioc	497 B	06/01/04 08:30 pm	
└─ install.R	-rw-r--r--	madman	bioc	78 B	04/05/04 02:00 pm	
└─ R	drwxr-xr-x	madman	bioc	0 B	06/01/04 08:30 pm	
└─ GOgraph.R	-rw-r--r--	madman	bioc	8.1 KB	06/01/04 08:30 pm	
└─ GOhpytest.R	-rw-r--r--	madman	bioc	2.9 KB	04/05/04 02:00 pm	
└─ shortestPath.R	-rw-r--r--	madman	bioc	3.1 KB	06/01/04 08:30 pm	
└─ triad.R	-rw-r--r--	madman	bioc	1.4 KB	04/05/04 02:00 pm	
└─ zzz.R	-rw-r--r--	madman	bioc	411 B	05/05/04 10:25 pm	
└─ R_PROFILE.R	-rw-r--r--	madman	bioc	78 B	04/05/04 02:00 pm	

1 file selected 8.1 KB

33 files 13.1 MB

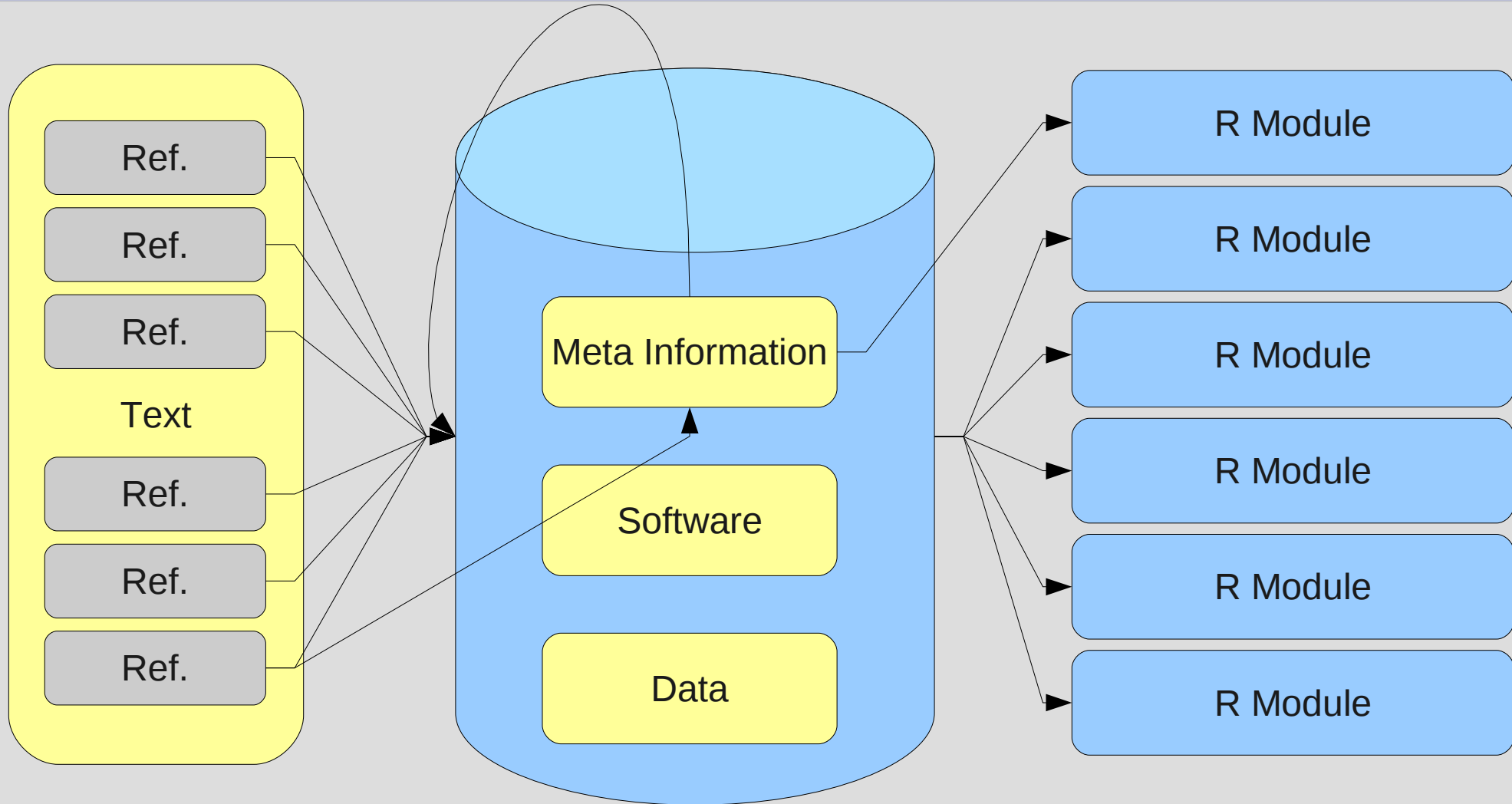
Compendium redefined

- New definition:

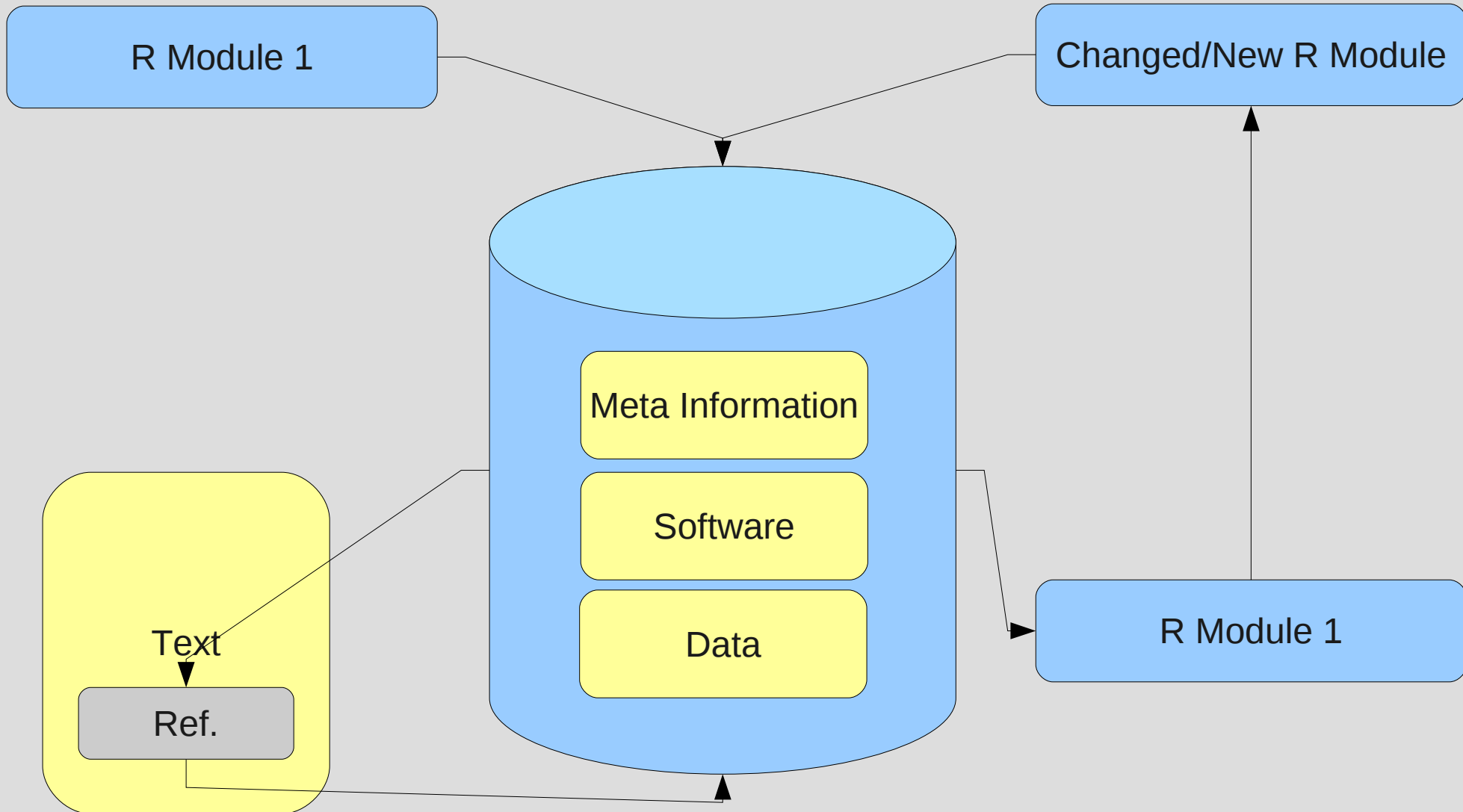
A document with (open-access) references to (remotely) archived Computations (including Data, Meta-data, and Software) that allow us to reproduce, and reuse the underlying analysis

- Complete separation of:
 - text and computing
 - computational result and computing infrastructure
- => the compendium platform is a tool for collaboration, dissemination, and monitoring.

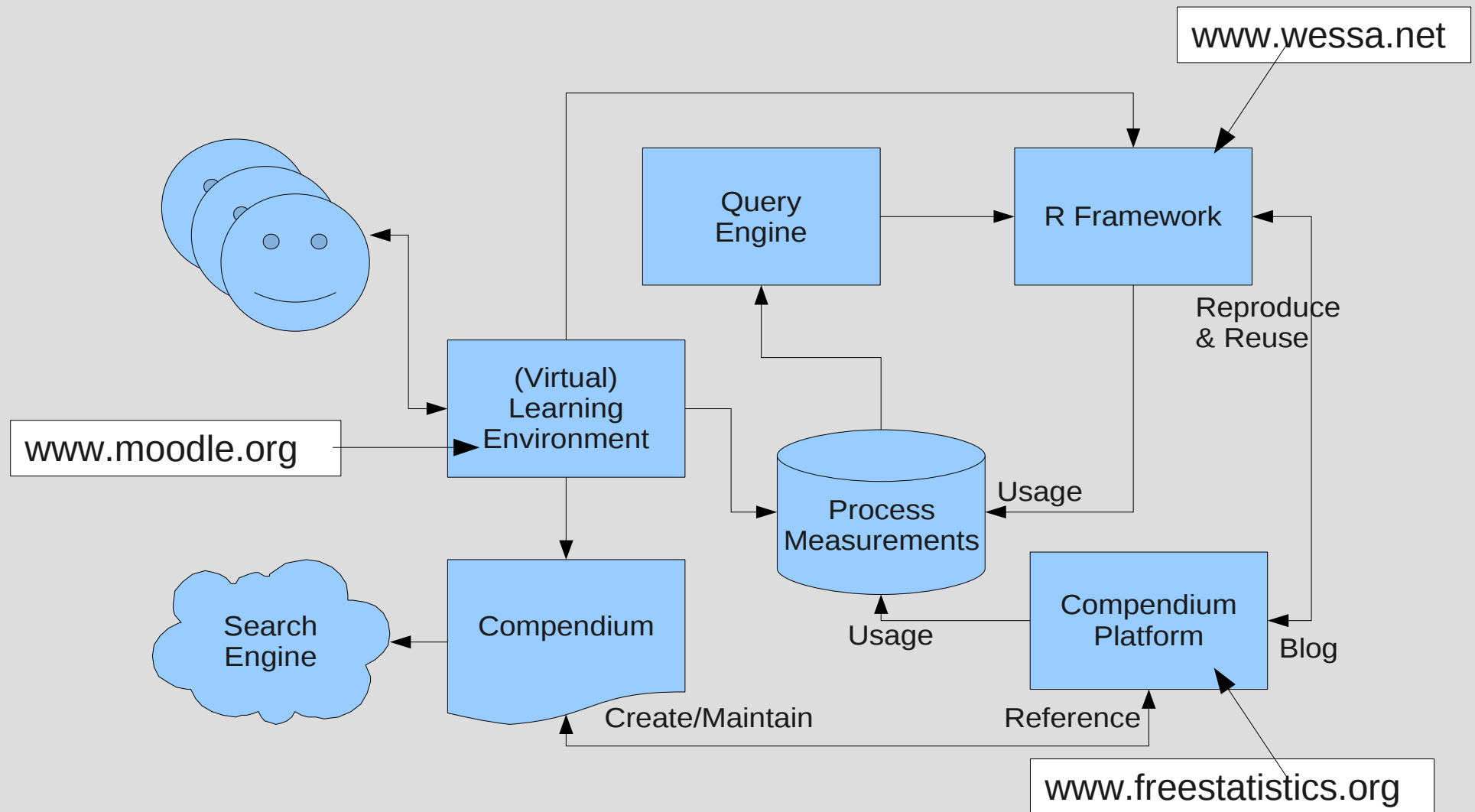
Computations Database



Compendium Dynamics



Learning System or Educational Laboratory?



Examples of Compendia

<http://www.wessa.net/download/tutorial.pdf>

(Descriptive Statistics – Central Tendency)

<http://www.wessa.net/download/tutorial1.pdf>

(Time Series Analysis - Introduction)

Note: both documents are “work in progress”
Please, send corrections & suggestions to
patrick@wessa.net

Screenshots

Fixed Seasonal Effects
 Include Monthly Dummies

Type of Equation
 Linear Trend

Chart options
 Width: 600
 Height: 400

R Code

```

library(lattice)
par1 <- as.numeric(par1)
x <- t(y)
k <- length(x[,1])
n <- length(x[,1])
x1 <- cbind(x[,par1], x[,1:kt=par1])
mycolnames <- c(colnames(x)[par1], colnames(x)[1:kt=par1])
colnames(x1) <- mycolnames #colnames(x)[par1]
x <- x1
if (par3 == 'First Differences'){
x2 <- array(0, dim=c(n-1,k), dimnames=list(1:(n-1), paste('1-
B)', colnames(x), sep='')))
for (i in 1:n-1) {
for (j in 1:k) {
x2[i,j] <- x[i+1,j] - x[i,j]
}
}
x <- x2
}
if (par2 == 'Include Monthly Dummies'){
x2 <- array(0, dim=c(n,11), dimnames=list(1:n, paste('M', seq(1:11), sep='')))
for (i in 1:11){
x2[seq(i,n,12),i] <- 1
}
x <- cbind(x, x2)
}
    
```

Compute

Summary of computational transaction

Raw Input	view raw input (R code)
Raw Output	view raw output of R engine
Computing time	6 seconds
R Server	'Sir Ronald Aylmer Fisher' @ 193.190.124.24

Multiple Linear Regression - Estimated Regression Equation

$ongevallen[t] = + 2324.06337310277 - 226.386033602688[t] - 451.374973296311M1[t] - 636.461063323769M2[t] - 563.133697991392M3[t] - 694.56342689015M4[t] - 556.476867326636M5[t] - 609.464131994261M6[t] - 532.074276661864M7[t] - 515.434421329607M8[t] - 460.86700697131M9[t] - 319.71721066475M10[t] - 116.386866332377M11[t] - 1.76486633237686t + e[t]$

Multiple Linear Regression - Ordinary Least Squares

Variable	Parameter	S.D.	T-STAT	H0: parameter = 0	2-tail p-value	1-tail p-value
(Intercept)	2324.06337310277	44.029839	52.7837		0	0
x	-226.386033602688	41.037226	-55166		0	0

A framework for statistical software development, maintenance, and publishing within an open-access business model, 2008, Computational Statistics

Computations are “blogged” (not archived)

Blog & Share - Free Statistics and Forecasting Software (Calculators) v.1.1.23-r1 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.wessa.net/blogshare.wasp?outtype=&id=8&command=blog&check=Sun, 03 Aug 2008 04:33

Server Status ... Gmail - Inbox (... Blog & Share S... Course: Applie... Blog & Share S... International C... Blog & Share - ...

http://www.freestats.org/) where it is permanently archived for reference purposes. In addition visitors of the Blog can Discuss, Reproduce, and Reuse all Statistical Computations in the archive.

Submit your Statistical Computation to the FreeStatistics.org Archive

Field	Value
Title (optional, meaningful title)	this is my title
Keywords (optional, comma-delimited list)	statistics, assignment 5, hypothesis testing, any other keyword
Your Comments (optional, any meaningful text)	I computed this hypothesis test to answer question 3 in assignment 5.
E-mail (optional, private - this is required if you want to edit/delete the post at a later time)	patrick@wessa.net
Type of Access (optional, do you want to grant everyone access to your archived computation?)	Public (anybody can access my computation)
Moratorium date (enter the moratorium date - only needed if 'Moratorium' is selected in 'Type of Access')	YYYY-MM-DD
Captcha	

Done

Multiple Regression (old)
Descriptive Statistics
Statistical Distributions
Hypothesis Testing
Statistics Education

Academic citations
Computations Archive
Search Computations
R Project
FAQ
About Wessa.net
Powered by Linux

Server status page
History list

Google

Web wessa.net

Search

workshop - OpenOffice.org Writer

File Edit View Insert Format Table Tools Window Help

Equation Expand Config

Default Garamond 12

Question 2: Investigate the prediction errors of the model that you used in question 1. Are the underlying regression assumptions satisfied?

De Adjusted R-squared is gelijk aan 0,6412. Dit wil zeggen dat we 64% van de wijzigingen van het aantal verkeersslachtoffers kunnen verklaren. Het resterende gedeelte (36%) kunnen we niet verklaren aan de hand van ons model maar zijn bijvoorbeeld te wijten aan uitzonderlijke weersomstandigheden.

Het volgende waar we heen zullen kijken is de Interpolation Plot. De stippellijn op deze grafiek geeft het werkelijke aantal slachtoffers weer. De volle lijn geeft het aantal verkeersslachtoffers ~~we~~ die voorspeld zijn door het model.

Actuals and Interpolation

<http://www.freestatsitics.org/blog/index.php?v=date/2007/Nov/14/t1195074007ni07puuvacjlu0w.htm>

We zien ook hier een dalende trend op lange termijn. Op het einde zien we een structurele breuk, dit is het effect van de seatbelt law (-395). Er treedt ook een

Snapshot of “Blogged” Computation

The screenshot shows a web browser window with the following elements:

- Browser Title Bar:** Blog & Share Statistical Computations at FreeStatistics.org
- Address Bar:** <http://www.freestatics.org/blog/index.php?v=date/2007/Nov/14/t1195074007ni07puuvacjlu0w.htm>
- Navigation Buttons:** Home » date » 2007 » Nov »
- Action Buttons:** Print, PDF, TeX, Statistics, Search, Edit, Post Comment, **Reproduce**, Reuse
- Section Header:** WS 8 - Q1 (3)
- Metadata:**
 - R Software Module: `rwasp_multipleregression_wasp` (opens new window with default values)
 - Title produced by software: Multiple Regression
 - Date of computation: Wed, 14 Nov 2007 14:04:40 -0700
- Citation:**

Cite this page as follows:
Statistical Computations at FreeStatistics.org, Office for Research Development and Education, URL <http://www.freestatics.org/blog/date/2007/Nov/14/t1195074007ni07puuvacjlu0w.htm>, Retrieved Sun, 03 Aug 2008 09:58:02 +0000
- Keywords:** IsPrivate?, User-defined keywords:
- Data:**

Dataseries X:
» [Textbox](#) « » [Textfile](#) « » [CSV](#) «

1687 0 1508 0 1507 0 1385 0 1632 0 1511 0 1559 0 1630 0 1579 0 1653 0 2152 0 2148 0 1752 0 1765 0 1717 0 1558 0 1575 0 1520 0 1805 0 1800 0 1719 0 2008 0 2242 0 2478 0 2030 0 1655 0 1693 0 1623 0 1805 0 1746 0 1795 0 1926 0 1619 0 1992 0 2233 0 2192 0 2080 0 1768 0 1835 0 1569 0 1976 0 1853 0 1965 0 1689 0 1778 0 1976 0 2397 0 2654 0 2097 0 1963 0 1677 0 1941 0 2003 0 1813 0 2012 0 1912 0 2084 0 2080 0 2118 0 2150 0 1608 0 1503 0 1548 0 1382 0 1731 0 1798 0 1779 0 1887 0 2004 0 2077 0 2092 0 2051 0 1577 0 1356 0 1652 0 1382 0 1519 0 1421 0 1442 0 1543 0 1656 0 1561 0 1905 0 2199 0 1473 0 1655 0 1407 0 1395 0 1530 0 1309 0 1526 0 1327 0 1627 0 1748 0 1958 0 2274 0 1648 0 1401 0 1411 0 1403 0 1394 0 1520 0 1528 0 1643 0 1515 0 1685 0 2000 0 2215 0 1956 0 1462 0 1563 0 1459 0 1446 0 1622 0 1657 0 1638 0 1643 0 1683 0 2050 0 2262 0 1813 0 1445 0 1762 0 1461 0 1556 0 1431 0 1427 0 1554 0 1645 0 1653 0 2016 0 2207 0 1665 0 1361 0 1506 0 1360 0 1453 0 1522 0 1460 0 1552 0 1548 0 1827 0 1737 0 1941 0 1474 0 1458 0 1542 0 1404 0 1522 0 1385 0 1641 0 1510 0 1681 0 1938 0 1868 0 1726 0 1456 0 1445 0 1456 0 1365 0 1487 0 1558 0 1488 0 1684 0 1594 0 1850 0 1998 0 2079 0 1494 0 1057 1 1218 1 1168 1 1236 1 1076 1 1174 1 1139 1 1427 1 1487 1 1483 1 1513 1 1357 1 1165 1 1282 1 1110 1 1297 1 1185 1 1222 1 1284 1 1444 1 1575 1 1737 1 1763 1
- Footer:** Text written by user: Done

Two callout boxes are present:

- A box pointing to the **Reproduce** button with the text: "Reproduce at wessa.net"
- A box pointing to the citation text with the text: "Cite the computation as follows"

Feedback (Peer Review)

ABS/SHW: Case: the Seatbelt Law - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://193.190.124.21/moodle/mod/workshop/viewassessment.php?aid=8325

Parallels Business Autom... Server Status Page - Fre... Gmail - Inbox (23) - patric... Free Statistics ABS/SHW: Case: the Sea...

Element 1: Evaluate Q1.	Weight: 1.00
Grade: Excellent <input type="radio"/> <input type="radio"/> <input checked="" type="radio"/> <input type="radio"/> <input type="radio"/> Very Poor	
Feedback: Soms wat onduidelijke uitleg over de p-waarde. Ik mis ook wat concrete uitleg over de verschillende onderdelen van het model zelf.	
Element 2: Evaluate Q2.	
Grade: Excellent <input type="radio"/> <input type="radio"/> <input checked="" type="radio"/> <input type="radio"/> <input type="radio"/> Very Poor	
Feedback: Hier staat alleen een link naar een berekening. Je had de r-squared waarden moeten bespreken. Ook had je kunnen nagaan of de voorspellingsfouten zich aan bepaalde assumpties voldaan hadden. Zo zijn er twee voorwaarden voor een goed model: het gemiddelde van de voorspellingsfouten moet 0 zijn en het gemiddelde moet constant zijn. Als we de voorspellingsfouten invullen in de central tendency calculator kunnen we zien dat aan deze assumpties niet voldaan is. Dus er is nog verbetering mogelijk aan ons model.	

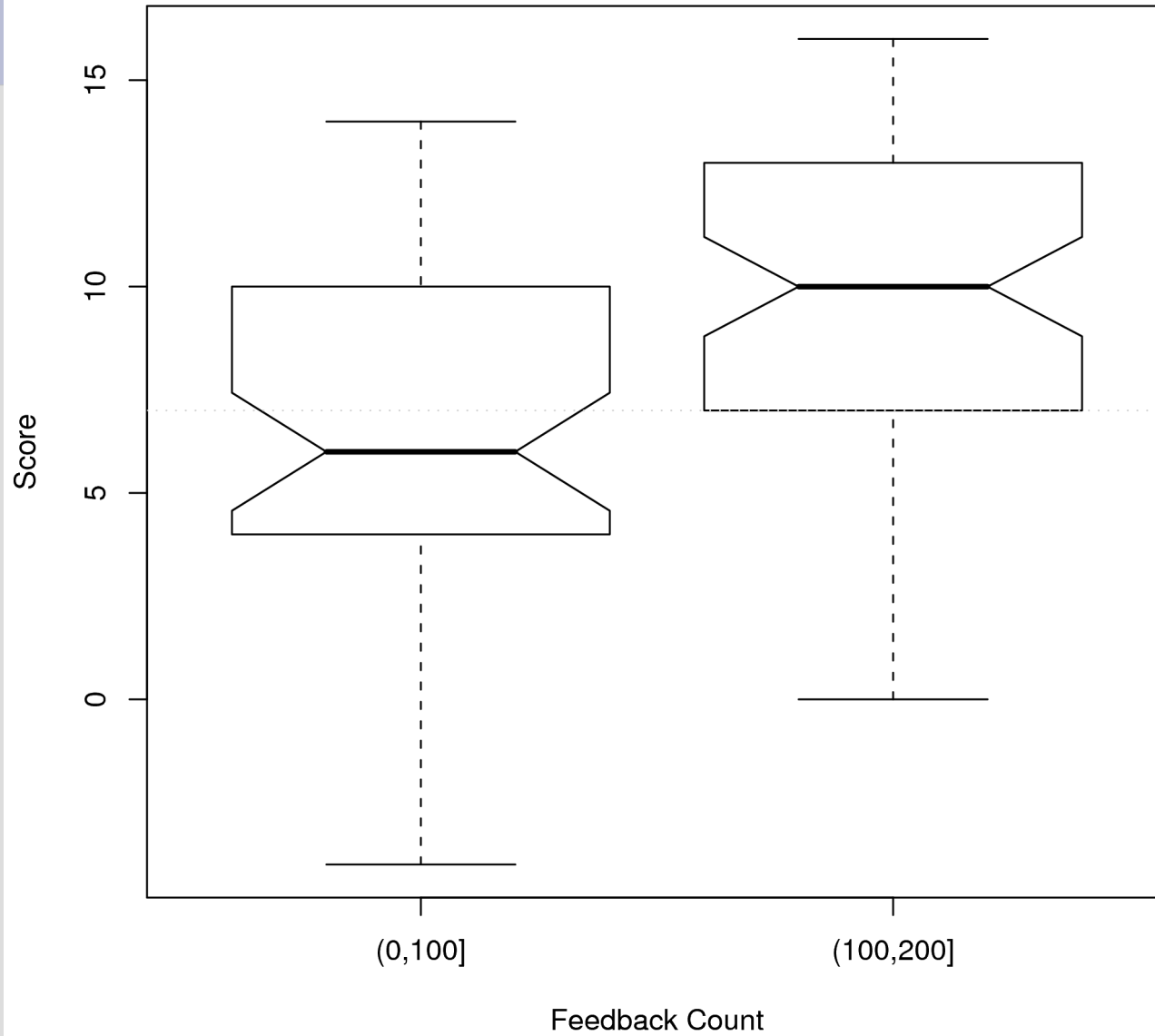
"Geen titel"

Done

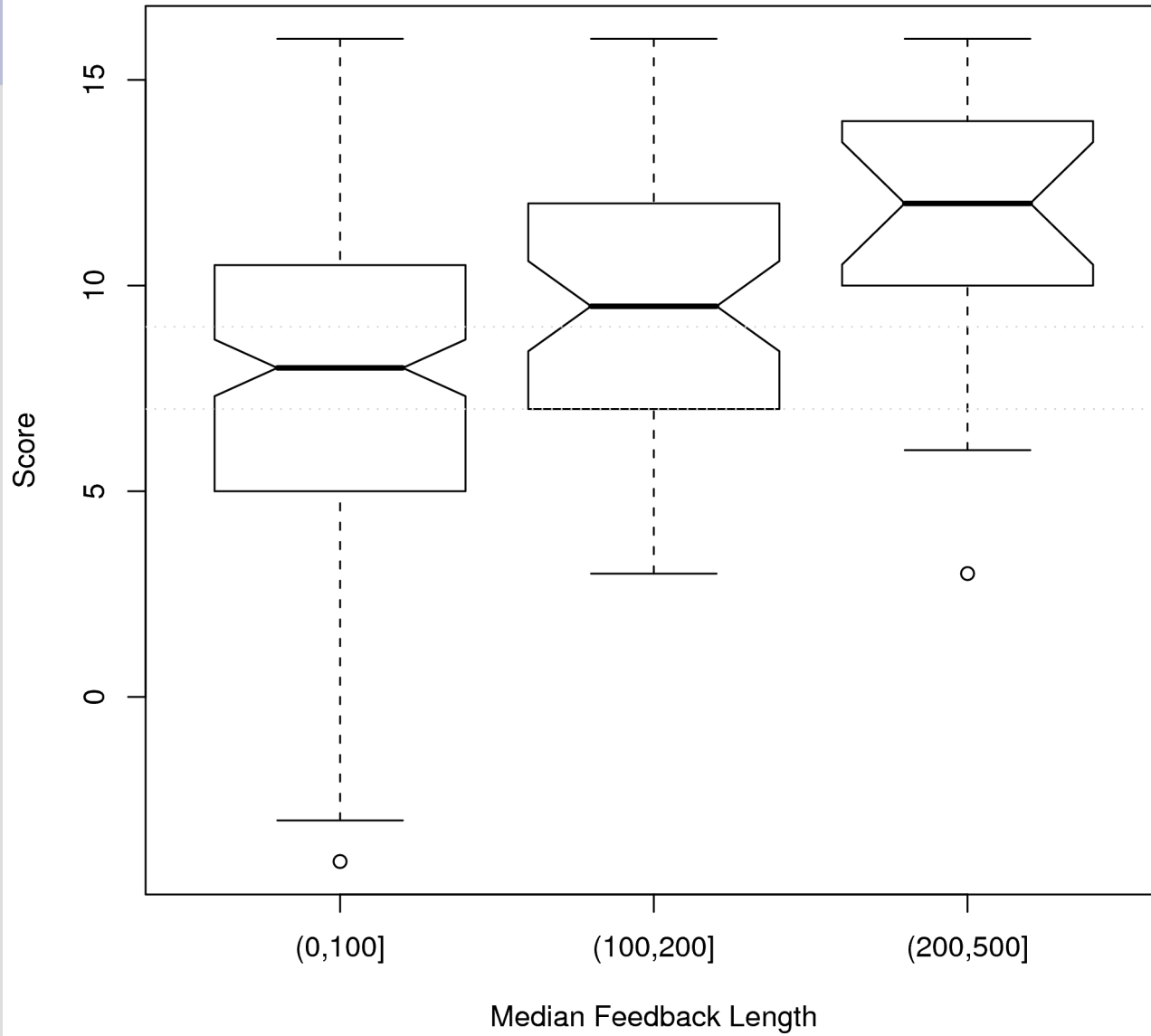
Submitting Peer Review (feedback) is a good learning activity – not a good grading procedure

How Reproducible Research Leads to Non-Rote Learning Within a Socially Constructivist E-Learning Environment, Proceedings of the 7th European Conference on e-Learning (ECEL'08), Cyprus

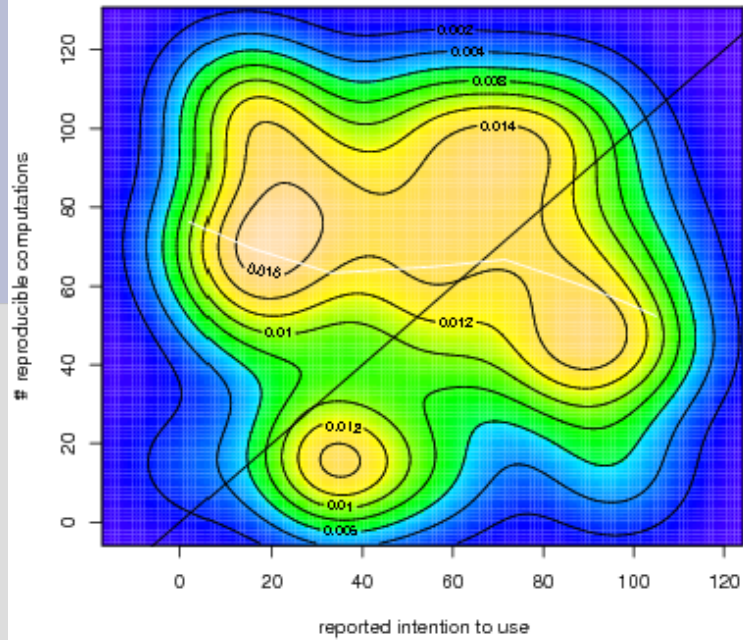
Score by Feedback Count



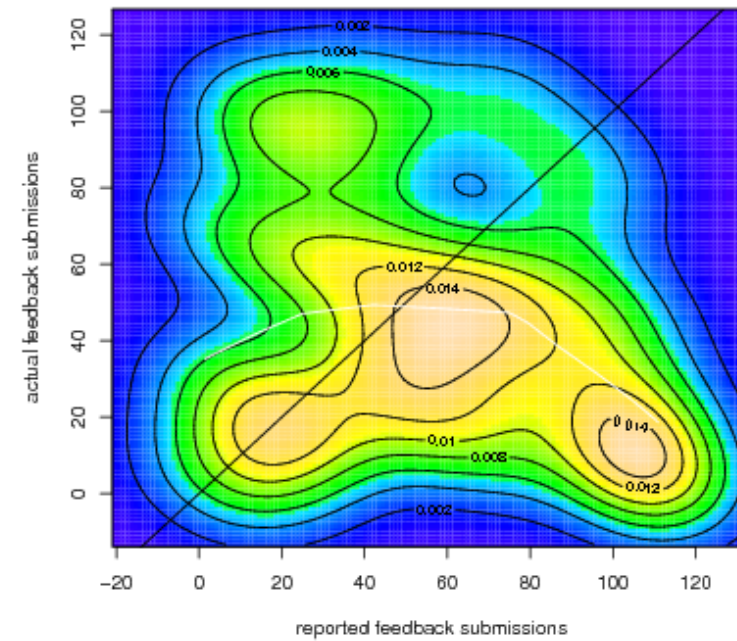
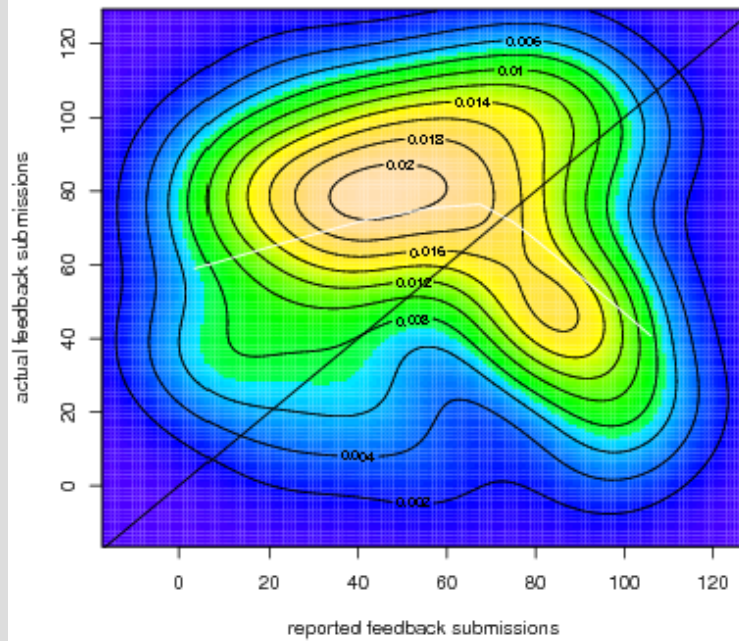
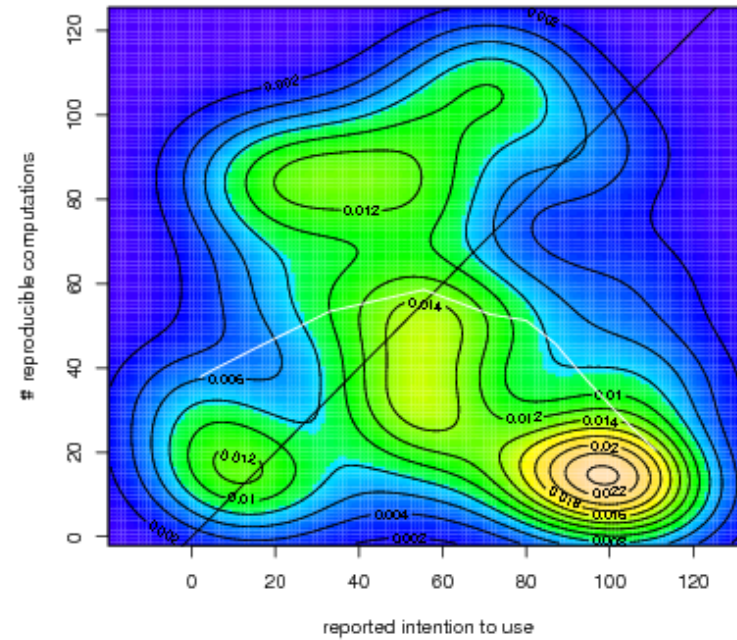
Score by Median Feedback Length



Female Bachelor Students



Male Bachelor Students



Conclusions & Future work

- Reproducible Computing can be made easy (for students)
- RC improves statistics learning
- RC allows us to research learning activities (based on actual – not reported – data)

- New features (social interaction, collaboration)
- RC for scientists
- RC for scientific publishing

Some References

- **J. Buckheit and D. L. Donoho.** Wavelab and reproducible research. In A. Antoniadis, editor, Wavelets and Statistics. Springer-Verlag, 1995.
- **Peter J. Green.** Diversities of gifts, but the same spirit. The Statistician, pages 423–438, 2003.
- **T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M.L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander.** Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286:531–537, 1999.
- **David L. Donoho, Xiaoming Huo,** BeamLab and Reproducible Research, International Journal of Wavelets, Multiresolution and Information Processing, 2004
- **Roger D. Peng, Francesca Dominici, and Scott L. Zeger,** Reproducible Epidemiologic Research, American Journal of Epidemiology, 2006
- **R. Gentleman,** Reproducible Research: A Bioinformatics Case Study, Bioconductor
- **R. Gentleman,** Applying Reproducible Research in Scientific Discovery, BioSilico, 2005
- **Jan de Leeuw,** Reproducible Research: the Bottom Line, 2001, online

Some References

- **Roger Koenker, Achim Zeileis**, Reproducible Econometric Research (A Critical Review of the State of the Art), Department of Statistics and Mathematics Wirtschaftsuniversität Wien, Research Report Series, Report 60, November 2007
- **Robert Gentleman, Duncan Temple Lang**, Statistical Analyses and Reproducible Research, <http://www.bepress.com/bioconductor/paper2>
- **Schwab, M., Karrenbach, N. and Claerbout, J.** Making scientific computations reproducible, Computing in Science & Engineering, 2 (6), pp. 61-67, 2000.
- **Robert Gentleman**, Some Perspectives on Statistical Computing, online
- **Leisch, F.**, “Sweave and beyond: Computations on text documents”, Proceedings of the 3rd International Workshop on Distributed Statistical Computing, 2003, Vienna, Austria, ISSN 1609-395

Some more references

- Wessa P., E. van Stee (2008), The Xycoon Stock Market Game: Virtual Learning Environment of Real-Life Laboratory, Proceedings of the International Conference of Education, Research and Innovation (ICERI 2008), *submitted*
- Poelmans S., P. Wessa, K. Milis, E. Bloemen, C. Doom (2008), Usability and Acceptance of E-Learning in Statistics Education, based on the Compendium Platform, Proceedings of the International Conference of Education, Research and Innovation (ICERI 2008), *submitted*
- Poelmans S., P. Wessa, K. Milis, E. Bloemen, C. Doom (2008), The Impact of Gender on the Acceptance of Virtual Learning Environments, Proceedings of the International Conference of Education, Research and Innovation (ICERI 2008), *submitted*
- Wessa P. (2008), Let us free statistics of irreproducible research, Statistics Seminar at Simon Fraser University, Vancouver, Canada
- Wessa P. (2008), How to Research the Effectiveness of Constructivist Statistics Education? An Approach based on Reproducible Computing, Applied Statistics 2008, to be submitted to Advances in Methodology and Statistics
- Wessa P. (2008), A framework for statistical software development, maintenance, and publishing within an open-access business model, Computational Statistics
- Wessa P. (2008), Learning Statistics based on the Compendium and Reproducible Computing, Proceedings of the International Conference on Education and Information Technology (ICEIT'08), Berkeley, San Francisco, USA
- Wessa P. (2008), How Reproducible Research Leads to Non-Rote Learning Within a Socially Constructivist E-Learning Environment, Proceedings of the 7th European Conference on e-Learning (ECEL'08), Cyprus
- Wessa P. (2008), Measurement and Control of Statistics Learning Processes based on Constructivist Feedback and Reproducible Computing, Proceedings of the 3rd International Conference on Virtual Learning (ICVL '08), Romania
- Wessa P. (2008), A Compendium of Reproducible Research about Descriptive Statistics and Linear Regression, URL <http://www.wessa.net/download/tutorial.pdf>
- Wessa P. (2008), A Compendium of Reproducible Research about Time Series Analysis, URL <http://www.wessa.net/download/tutorial1.pdf>

All documents will be available at <http://www.freestatistics.org/index.php?action=10> in the near future.