Leveraging Open Source R for Predictive Analytics

Jim Porzak

Predictive Analytics Summit
San Diego
February, 2011

ancestry.com

Outline

- What is R?
- Doing Predictive Analytics with R.
- R Challenges.
- Getting Started in R.
- Questions/Discussion

What is R?

"R is a free software environment for statistical computing and graphics."

- www.r-project.org

R *is* Free!

As is beer

As is freedom

As is lunch

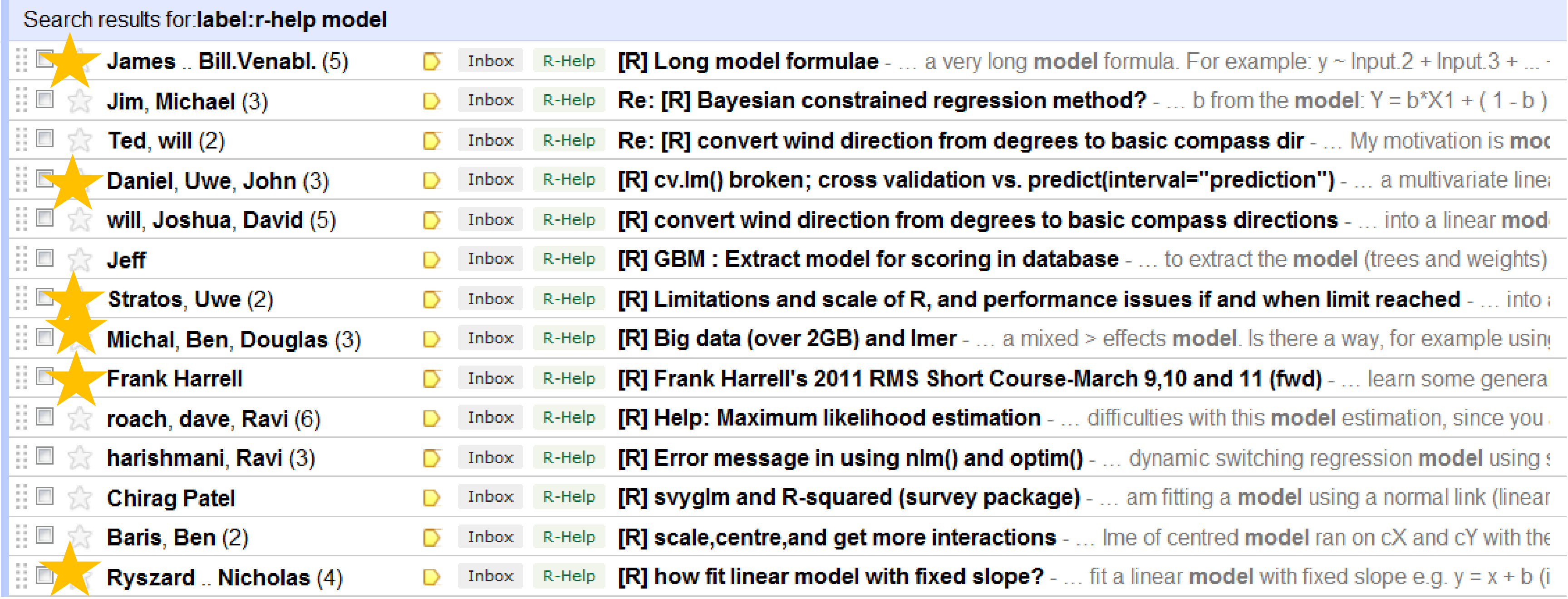
R *is* Community!

As in Grass Roots

- 20 core developers, ~ 2000 package developers
- 20 benefactors, 35 supporting institutions, 50 donors, ~ 200 members
- Annual international conferences since 2004

As in Help

r-help mail list questions answered by one of the staRs!



Search results for:**label:r-help model**

☐ ★	**James** .. **Bill.Venabl.** (5)	Inbox	R-Help	**[R] Long model formulae** - ... a very long **model** formula. For example: y ~ Input.2 + Input.3 + ... -
☐ ☆	**Jim**, **Michael** (3)	Inbox	R-Help	**Re: [R] Bayesian constrained regression method?** - ... b from the **model**: Y = b*X1 + (1 - b)
☐ ☆	**Ted**, **will** (2)	Inbox	R-Help	**Re: [R] convert wind direction from degrees to basic compass dir** - ... My motivation is **moc**
☐ ★	**Daniel**, **Uwe**, **John** (3)	Inbox	R-Help	**[R] cv.lm() broken; cross validation vs. predict(interval="prediction")** - ... a multivariate line
☐ ☆	**will**, **Joshua**, **David** (5)	Inbox	R-Help	**[R] convert wind direction from degrees to basic compass directions** - ... into a linear **mod**
☐ ☆	**Jeff**	Inbox	R-Help	**[R] GBM : Extract model for scoring in database** - ... to extract the **model** (trees and weights)
☐ ★	**Stratos**, **Uwe** (2)	Inbox	R-Help	**[R] Limitations and scale of R, and performance issues if and when limit reached** - ... into a
☐ ★	**Michal**, **Ben**, **Douglas** (3)	Inbox	R-Help	**[R] Big data (over 2GB) and lmer** - ... a mixed > effects **model**. Is there a way, for example usin
☐ ★	**Frank Harrell**	Inbox	R-Help	**[R] Frank Harrell's 2011 RMS Short Course-March 9,10 and 11 (fwd)** - ... learn some genera
☐ ☆	**roach**, **dave**, **Ravi** (6)	Inbox	R-Help	**[R] Help: Maximum likelihood estimation** - ... difficulties with this **model** estimation, since you
☐ ☆	**harishmani**, **Ravi** (3)	Inbox	R-Help	**[R] Error message in using nlm() and optim()** - ... dynamic switching regression **model** using s
☐ ☆	**Chirag Patel**	Inbox	R-Help	**[R] svyglm and R-squared (survey package)** - ... am fitting a **model** using a normal link (linear
☐ ☆	**Baris**, **Ben** (2)	Inbox	R-Help	**[R] scale,centre,and get more interactions** - ... lme of centred **model** ran on cX and cY with the
☐ ★	**Ryszard** .. **Nicholas** (4)	Inbox	R-Help	**[R] how fit linear model with fixed slope?** - ... fit a linear **model** with fixed slope e.g. y = x + b (i

R *is* Hot!

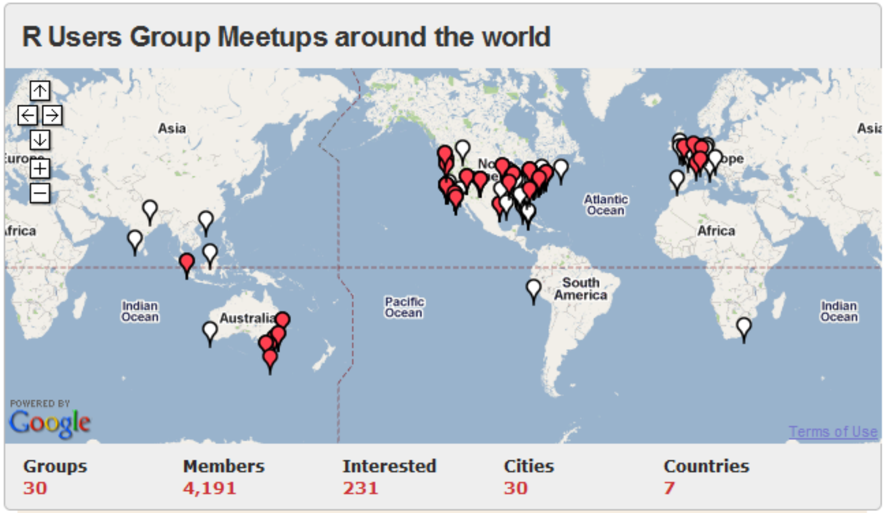
Hype:

Tipping point: Jan 6, 2009 New York Times *Inside Technology* feature story on R.

Bloggs:

>140 summarized at r-bloggers.com

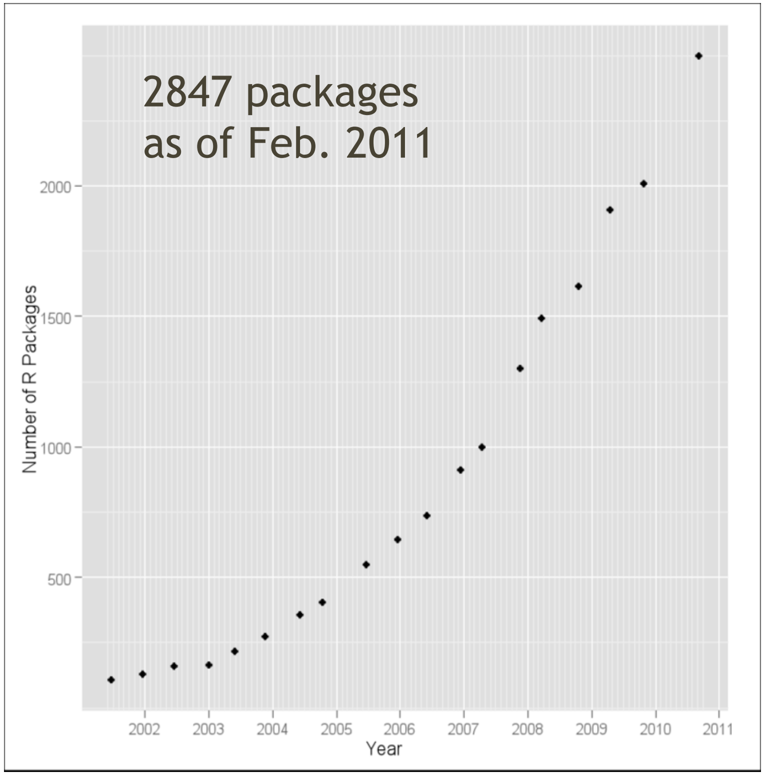
User Groups



R Users Group Meetups around the world

Groups	Members	Interested	Cities	Countries
30	4,191	231	30	7

Rate of Development



2847 packages
as of Feb. 2011

# R Packages released on CRAN (Comprehensive R Archive Network) by year – does not include related R packages released elsewhere, e.g. BioConductor.

Predictive Analytics in R

Predictive analytics is the art of finding actionable statistical models to identify individual risks and opportunities.

Building Models & Making Predictions

Some R (pseudo) code:

```
# Build the model
 MyModel <- SomeMethod (ModelFormula, Data, Parameters,…)
# Evaluate the model
 MyScores <- predict(MyModel, NewData)
```

Or, for classification:

```
 MyClasses <- predict(MyModel, NewData)
```

Some model formulas:

```
 y ~ x                    ## straight line
 y ~ 0 + x                ## straight line through origin
 log(y) ~ x1 + x2 + …     ## transformed response
 y ~ A*B*C - A:B:C        ## 3-factor, main & 2-way inter.
```

A Sampling of Modeling Methods in R

Classification:

Linear Discriminant Analysis: lda, Linda

Quadratic Discriminant Analysis: qda, QdaCov

Stabalized Linear Discriminant Analysis: alda

Heteroscedasic Discriminant Analysis: hda

Shrinkage Linear Discriminant Analysis: sda

Sparse Linear Discriminant Analysis: sparseLDA

Stepwise Discriminant: stepLDA

Stepwise Diagonal Discriminant Analysis: addaLDA, sddaQDA

Regularized Discriminant Analysis: rda

Mixture Discriminant Analysis: mda

Sparse Mixture Discriminant Analysis: smda

Penalized Discriminant Analysis: pda, pda2

High Dimensional Discriminant Analysis:hdda

Flexible Discriminant Analysis (MARS basis): fda

Bagged FDA: bagFDA

Logistic/Multinomial Regression: multinom, plr

LogitBoost: logitBoost

Logistic Model Trees: LMT

Rule-Based Models: J48, OneR, PART, JRip

Logic Forests: logforest

Bayesian Multinomial Probit Model: vbmpRadial

Nearest Shruken Centroids: pam, scrda

Naive Bayes: nb

Generalized Partial Least Squares: gpls

Learned Vector Quantization: lvq

ROC Curves: rocc

Methods supported by Max Kuhn’s caret package (classification & regression training). There are more on CRAN!

Regression:

Linear Least Squares: lm, lmStepAIC

Principal Component Regression: pcr

Independent Component Regression: icr

Robust Linear Regression: rlm

Neural Networks: neuralnet

Quantile Regression Forests: qrf

Rule Based Models: M5Rules

Projection Pursuit Regression: ppr

Penalized Linear Models: penalized, lars, lars2, enet,lasso, foba

Relevance Vector Machines: rvmLinear, rvmRadial, rvmPoly

Supervised Principal Components: superpc

Dual Use:

Generalized linear model: glm,glmStepAIC

Generalized additive model: gam, gamLoess, gamSpline

Recursive Partitioning: rpart, ctree, ctree2, gbm

Boosted Trees: gbm, blackboost, ada

Other Boosted Models: glmboost, gamboost,

Random Forests: rf, parRF, cforest

Bagging: treebag, bag, logicBag

Other Trees: nodeHarvest, partDSA

Multivariate Adaptive Regression Splines (MARS): earth, mars

Bagged MARS: bagEarth

Logic Regression: logreg

Elastic Net (glm): glmnet

Neural Networks: nnet, pcaNNet

Partial Least Squares: pls

Sparse Partial Least Squares: spls

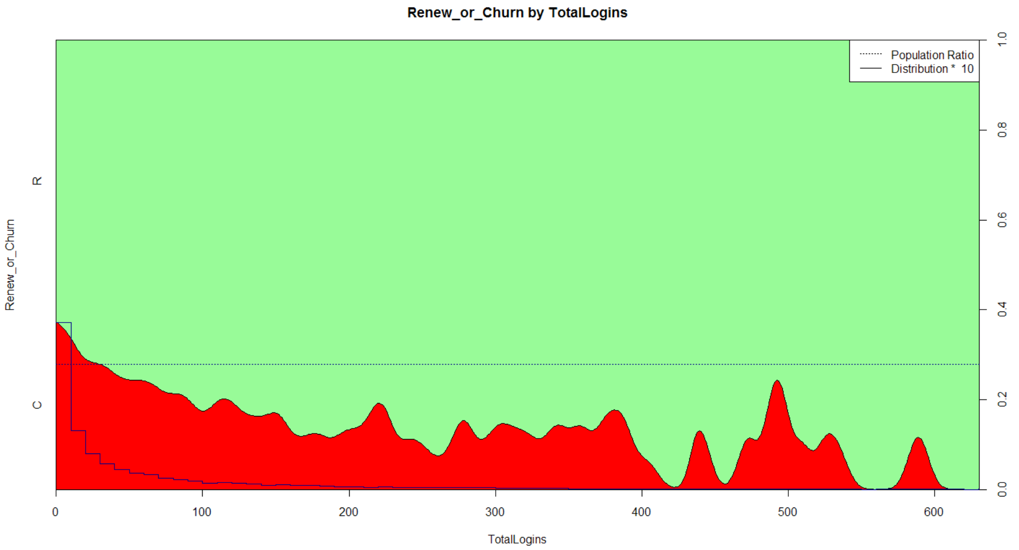
Support Vector Machines: svmLinear, svmRadial, svmPoly

Gaussian Processes: gaussprLinear, gaussprRadial, gaussprPoly

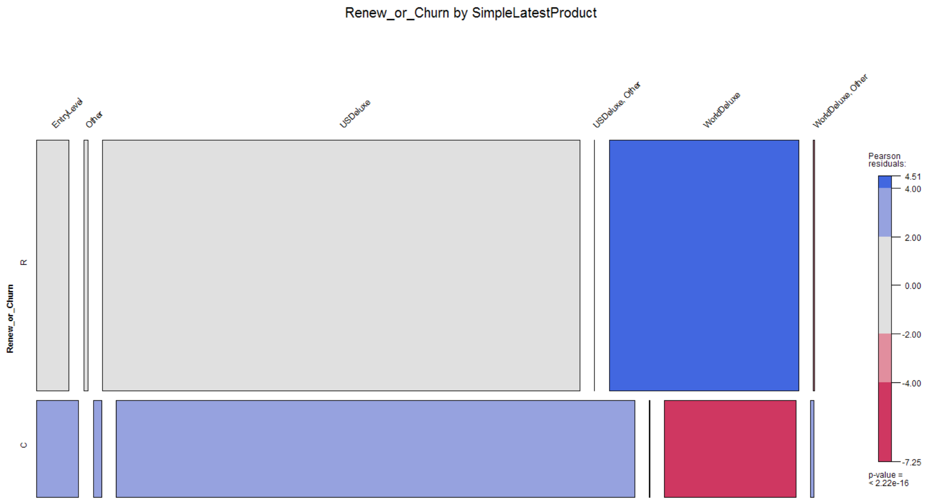
k Nearest Neighbors: knn

Visualizing Candidate Predictors

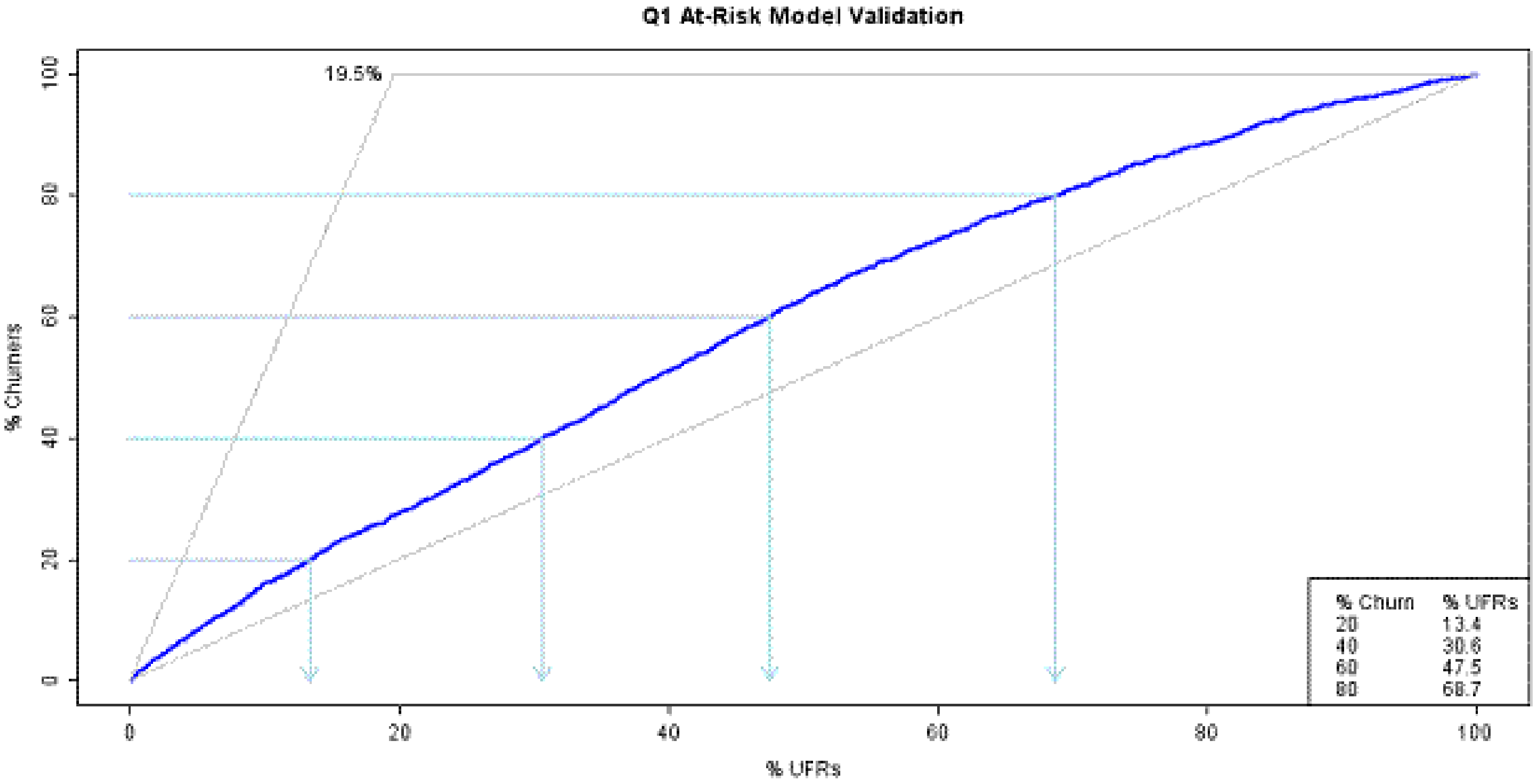
CDP for continuous predictors.



Mosaic plot for categorical predictors.

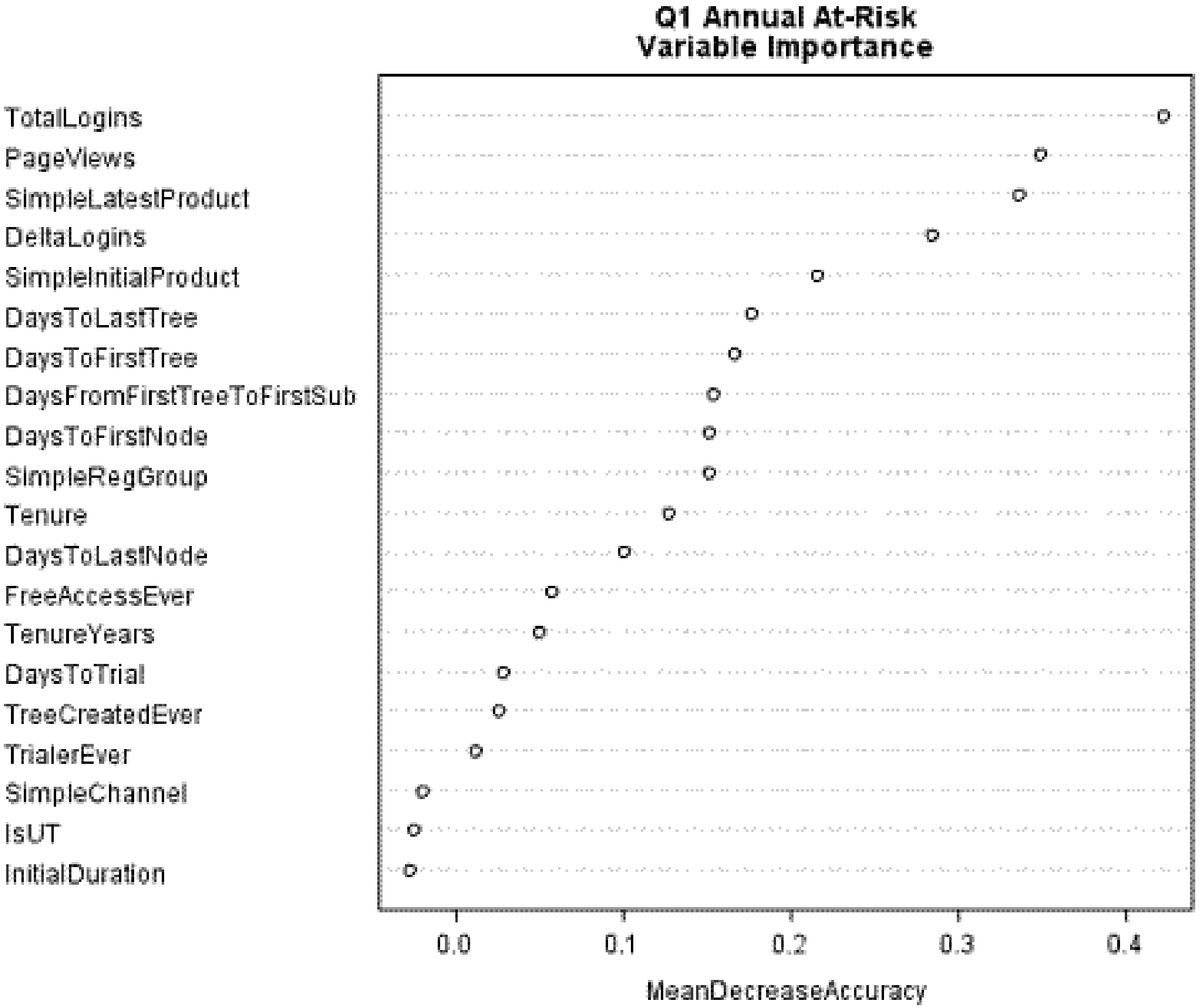


Visualizing Results I – Lift Plot



% Churn	% UFRs
20	13.4
40	30.6
60	47.5
80	68.7

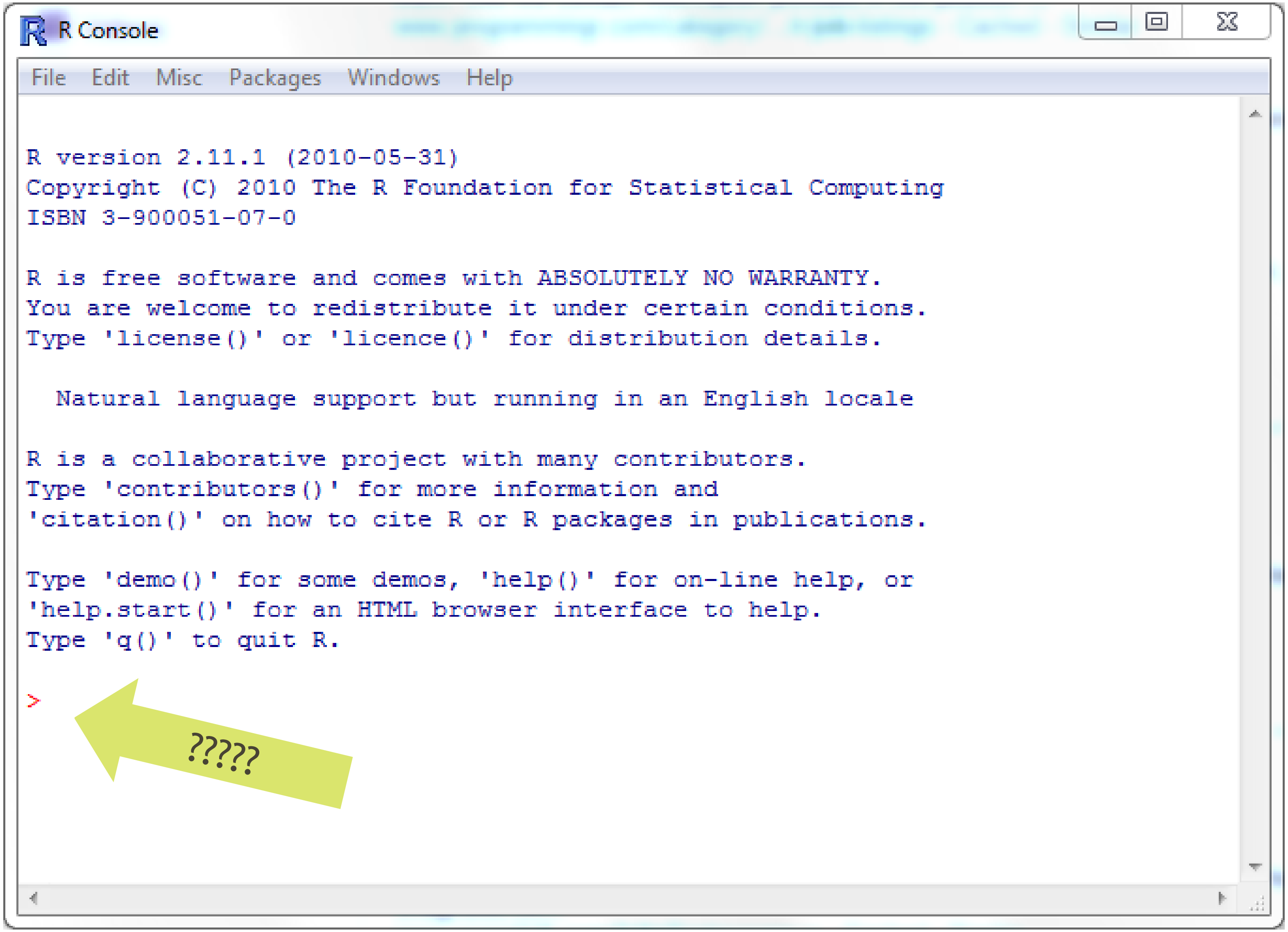
Visualizing Results II – Variable Importance



R Challenges

The “no free lunch” part.

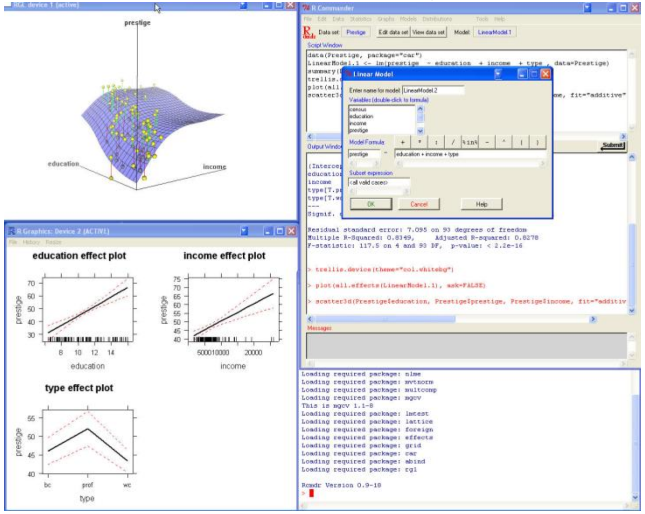
The Dreaded Blank Screen!



```
R version 2.11.1 (2010-05-31)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Various Menu Driven Front Ends Help!

The R Commander by John Fox



```
data(Prestige, package="car")
LinearModel.1 <- lm(prestige ~ education + income + type , data=Prestige)
```

Linear Model: Enter name for model: LinearModel.2

Variables (double-click to formula): census, education, income, prestige

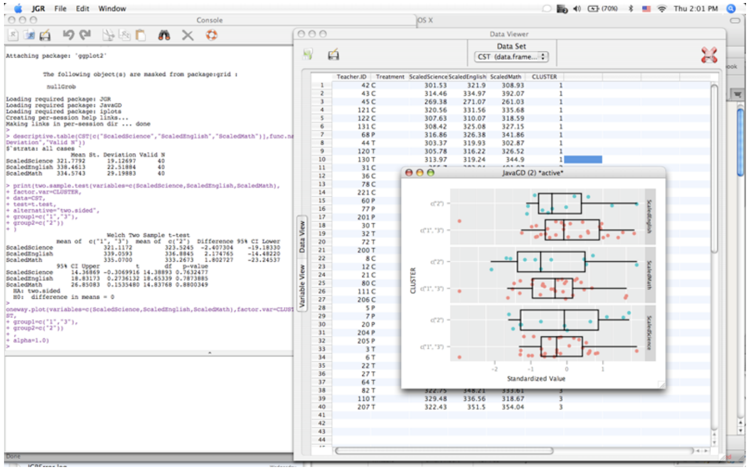
Model Formula: prestige ~ education + income + type

Subset expression: <all valid cases>

```
Residual standard error: 7.095 on 93 degrees of freedom
Multiple R-Squared: 0.8349,     Adjusted R-squared: 0.8278
F-statistic: 117.5 on 4 and 93 DF,  p-value: < 2.2e-16

> trellis.device(theme="col.whitebg")

> plot(all.effects(LinearModel.1), ask=FALSE)

> scatter3d(Prestige$education, Prestige$prestige, Prestige$income, fit="additiv
```

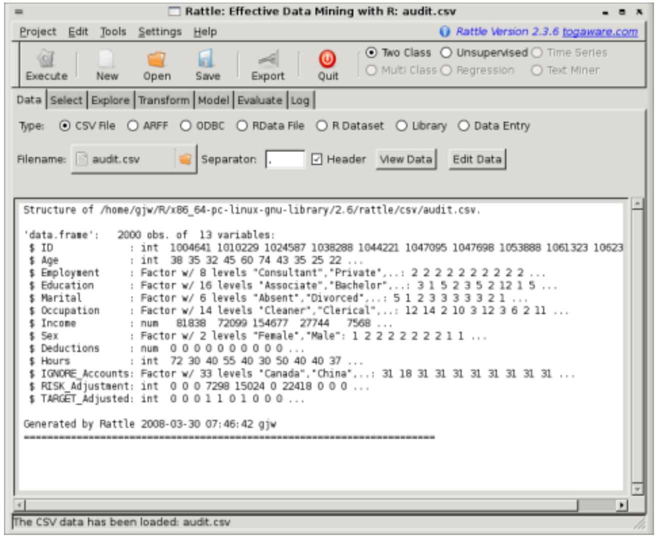
```
Loading required package: nlme
Loading required package: mvtnorm
Loading required package: multcomp
Loading required package: mgcv
This is mgcv 1.1-8
Loading required package: lmtest
Loading required package: lattice
Loading required package: foreign
Loading required package: effects
Loading required package: grid
Loading required package: car
Loading required package: abind
Loading required package: rgl

Rcmdr Version 0.9-18
>
```

Deducer by Ian Fellows



```
Attaching package: 'ggplot2'


        The following object(s) are masked from package:grid :

         nullGrob

Loading required package: JGR
Loading required package: JavaGD
Loading required package: iplots
Creating per-session help links...
Making links in per-session dir ... done
>
> descriptive.table(CST[c("ScaledScience","ScaledEnglish","ScaledMath")],func.na
Deviation","Valid N"))
$`strata: all cases `
                  Mean St. Deviation Valid N
ScaledScience 321.7792      19.12697      40
ScaledEnglish 338.4613      22.51884      40
ScaledMath    334.5743      29.19883      40

> print(two.sample.test(variables=c(ScaledScience,ScaledEnglish,ScaledMath),
+ factor.var=CLUSTER,
+ data=CST,
+ test=t.test,
+ alternative="two.sided",
+ group1=c("1","3"),
+ group2=c("2"))
+ )
                        Welch Two Sample t-test
              mean of  c("1", "3")  mean of  c("2")  Difference 95% CI Lower
ScaledScience             321.1172          323.5245  -2.407304    -19.18330
ScaledEnglish             339.0593          336.8845   2.174765    -14.48220
ScaledMath                335.0700          333.2673   1.802727    -23.24537
              95% CI Upper          t        df   p-value
ScaledScience     14.36869 -0.3069916 14.38893 0.7632477
ScaledEnglish     18.83173  0.2736132 18.65339 0.7873885
ScaledMath        26.85083  0.1535480 14.83768 0.8800349
  HA: two.sided
  H0:  difference in means = 0
>
oneway.plot(variables=c(ScaledScience,ScaledEnglish,ScaledMath),factor.var=CLUST
ST,
+ group1=c("1","3"),
+ group2=c("2"))
+ ,
+ alpha=1.0)
>
```

Rattle by Graham Williams


Rattle: Effective Data Mining with R: audit.csv
Project Edit Tools Settings Help
Rattle Version 2.3.6 togaware.com
Execute
New
Open
Save
Export
Quit
Data Select Explore Transform Model Evaluate Log
View Data
Edit Data

● Two Class ○ Unsupervised ○ Time Series ○ Multi Class ○ Regression ○ Text Miner

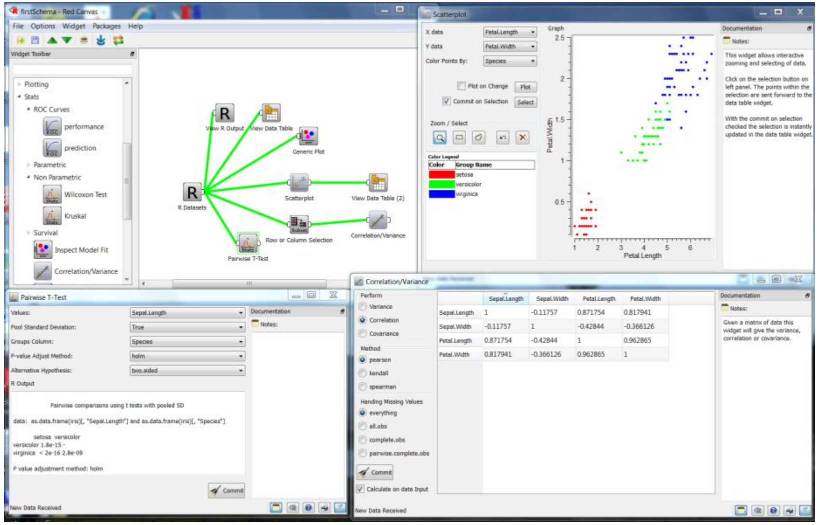
Type: ● CSV File ○ ARFF ○ ODBC ○ RData File ○ R Dataset ○ Library ○ Data Entry

Filename: audit.csv Separator: , ☑ Header

```
Structure of /home/gjw/R/x86_64-pc-linux-gnu-library/2.6/rattle/csv/audit.csv.

'data.frame':   2000 obs. of  13 variables:
 $ ID             : int  1004641 1010229 1024587 1038288 1044221 1047095 1047698 1053888 1061323 10623
 $ Age            : int  38 35 32 45 60 74 43 35 25 22 ...
 $ Employment     : Factor w/ 8 levels "Consultant","Private",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Education      : Factor w/ 16 levels "Associate","Bachelor",..: 3 1 5 2 3 5 2 12 1 5 ...
 $ Marital        : Factor w/ 6 levels "Absent","Divorced",..: 5 1 2 3 3 3 3 3 2 1 ...
 $ Occupation     : Factor w/ 14 levels "Cleaner","Clerical",..: 12 14 2 10 3 12 3 6 2 11 ...
 $ Income         : num   81838  72099 154677  27744   7568 ...
 $ Sex            : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 2 2 2 1 1 ...
 $ Deductions     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Hours          : int  72 30 40 55 40 30 50 40 40 37 ...
 $ IGNORE_Accounts: Factor w/ 33 levels "Canada","China",..: 31 18 31 31 31 31 31 31 31 31 ...
 $ RISK_Adjustment: int  0 0 0 7298 15024 0 22418 0 0 0 ...
 $ TARGET_Adjusted: int  0 0 0 1 1 0 1 0 0 0 ...

Generated by Rattle 2008-03-30 07:46:42 gjw
======================================================================
```

The CSV data has been loaded: audit.csv

Red-R by Parikh & Covington



Scatterplot: X data: Petal.Length; Y data: Petal.Width; Color Points By: Species

☐ Plot on Change ☑ Commit on Selection

This widget allows interactive zooming and selecting of data.

Click on the selection button on left panel. The points within the selection are sent forward to the data table widget.

With the commit on selection checked the selection is instantly updated in the data table widget.

Pairwise T-Test

Values: Sepal.Length

Pool Standard Deviation: True

Groups Column: Species

P-value Adjust Method: holm

Alternative Hypothesis: two.sided

R Output

```
             Pairwise comparisons using t tests with pooled SD

data:  as.data.frame(iris)[, "Sepal.Length"] and as.data.frame(iris)[, "Species"]

           setosa  versicolor
versicolor 1.8e-15 -
virginica  < 2e-16 2.8e-09

P value adjustment method: holm
```

New Data Received

Correlation/Variance

Perform: ○ Variance ● Correlation ○ Covariance

Method: ● pearson ○ kendall ○ spearman

Handing Missing Values: ● everything ○ all.obs ○ complete.obs ○ pairwise.complete.obs

☑ Calculate on data Input

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1	-0.11757	0.871754	0.817941
Sepal.Width	-0.11757	1	-0.42844	-0.366126
Petal.Length	0.871754	-0.42844	1	0.962865
Petal.Width	0.817941	-0.366126	0.962865	1

Given a matrix of data this widget will give the variance, correlation or covariance.

New Data Received

The Big Problem Challenge

"R can only solve problems that fit in RAM."

1) Fix the problem:
- Make friends with a DBMS (ODBC, JDBC, mySQL, SQLite)
- Sample down

2) Fix R:
- Clever coding – biglm, bigmemory, biganalytics
- Parallel & High Performance Computing – see Dirk Eddelbuettel's task view (MapReduce, GPU's, cloud, & more)
- Revolution Analytics RevoScaleR

Organizational Issues

- Policy & Regularity Constraints
- Introducing R Into Your Organization
- Integrating with Existing Infrastructure

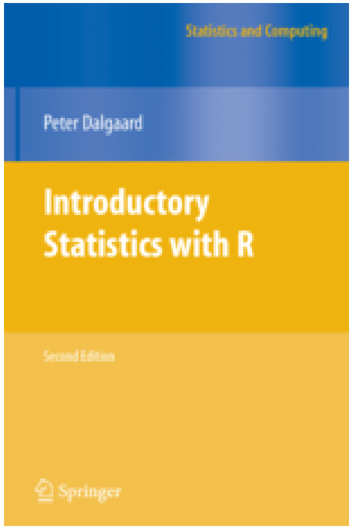
Getting Started in R

R Links

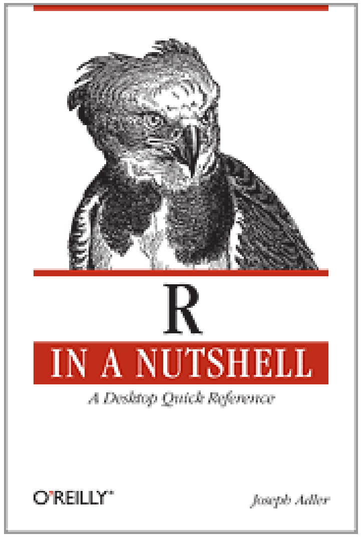
- R Homepage: www.r-project.org
  - The official site of R
- R Foundation: www.r-project.org/foundation
  - Central reference point for R development community
  - Holds copyright of R software and documentation
- To Download, find your local CRAN mirror (The Comprehensive R Archive Network)
  - We use: cran.cnr.berkeley.edu
  - Find yours at: cran.r-project.org/mirrors.html
  - Current Binaries
  - Current Manuals & FAQs
  - The R Journal
  - CRAN Task Views – especially finance, machine learning, cluster, Econometrics, & Robust
  - Links to related projects and sites
- R Bloggers
  - John Quick's R Tutorial Series
  - David Smith's Revolution Analytics blog
  - Tal Galili's R bloggers aggregation of other R bloggers.
- The R Wiki
- UCLA Resources to help you learn R

Jim's Favorite R Books

Introductory Statistics with R
Series: Statistics and Computing
Dalgaard, Peter
2nd ed., 2008, XVI, 364 p., Softcover
ISBN: 978-0-387-79053-4

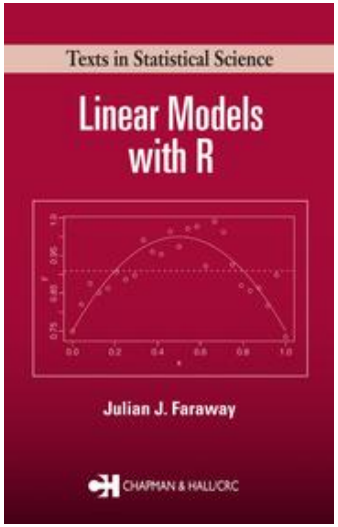


R in a Nutshell
A Desktop Quick Reference
By Joseph Adler
Publisher: O'Reilly Media
Released: December 2009
Pages: 640

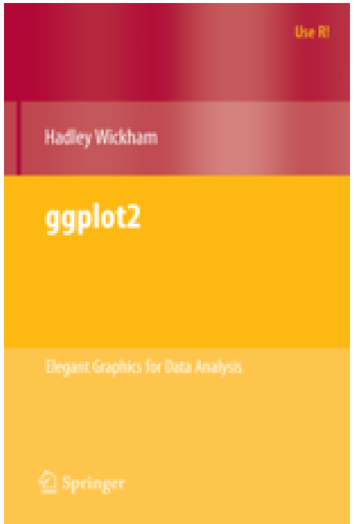


Linear Models with R
Julian J. Faraway,
University of Bath, United Kingdom
Cat. #: C4258
ISBN: 9781584884255
Publication Date: August 12, 2004

Also his:
Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models



ggplot2
Elegant Graphics for Data Analysis
Series: Use R
Wickham, Hadley
2nd Printing., 2009, VIII, 216 p., Softcover
ISBN: 978-0-387-98140-6



For a 2nd opinion, see Books for learning the R language on stackoverflow.

Getting Help

- For Free
  - On-Line
    - The R-Help mail list
    - R on stackoverflow.com
    - Many of bloggers & user groups have help forms
  - Your local R User Group (aka RUG)
    - Meetups
    - List on R Wiki
- For $’s
  - Revolution Analytics
  - Mango Solutions
  - Post need on R Jobs
  - Announce need at your local RUG

Questions? Comments?

Now would be the time!

jporzak@ancestry.com