



Leveraging Open Source R for Predictive Analytics

Jim Porzak

Predictive Analytics Summit
San Francisco
November, 2010

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

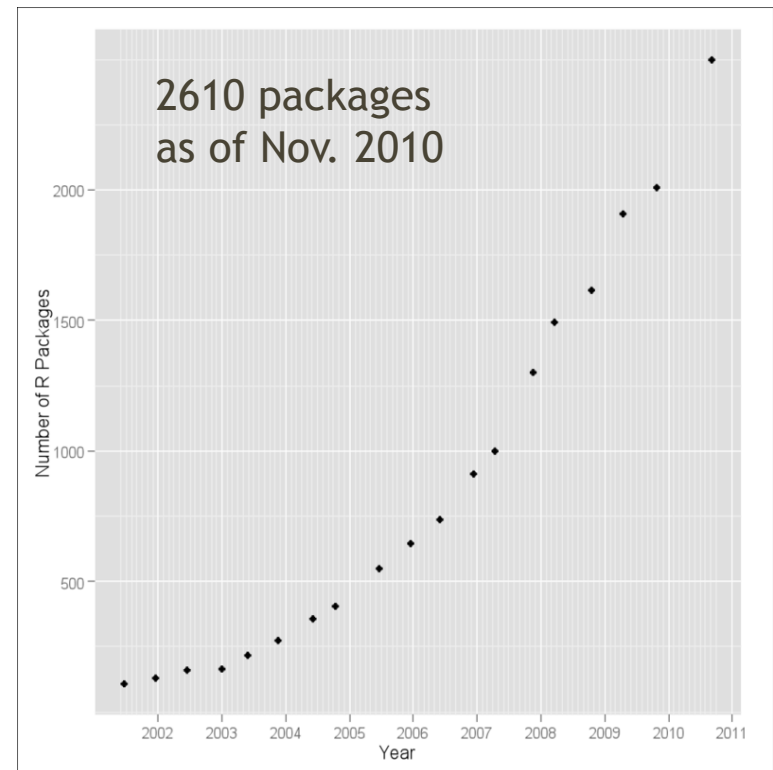
~130 summarized at r-bloggers.com

User Groups



R Users Group Meetup Group Interested in a new R Users Group Meetup Group

Rate of Development



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:

```
MyModel <- SomeMethod (ModelFormula, Data,  
  Parameters,...)
```

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

Or, for classification:

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

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:

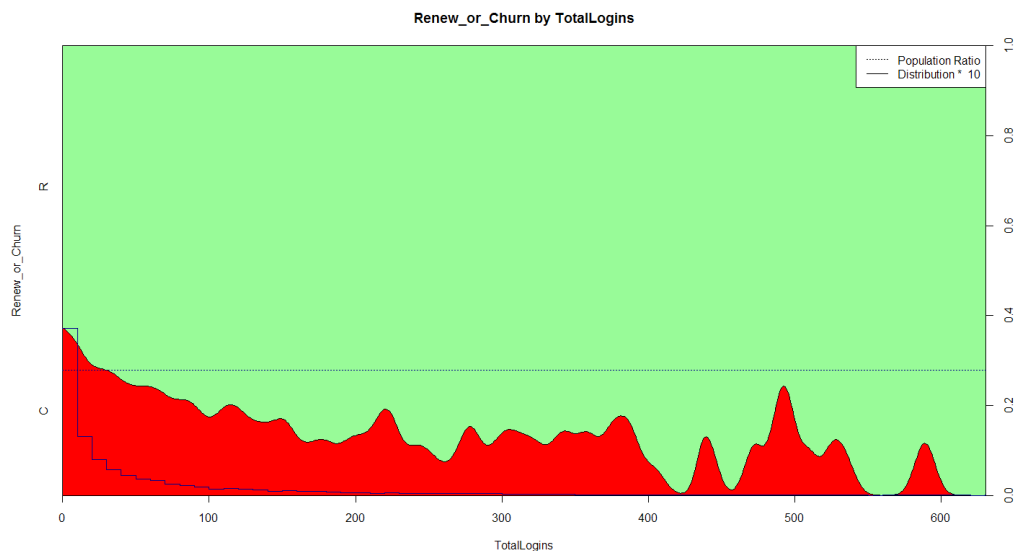
Linear Least Squares: lm, lmStepAIC
Principal Component Regression: pcr
Independent Component Regression: icr
Robust Linear Regression: rlm
Neural Networks: neuralnet
Quantile Regression Forests: qrf
Rule Based Models: M5Rules
Projection Pursuit Regression: ppr
Penalized Linear Models: penalized, lars, 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

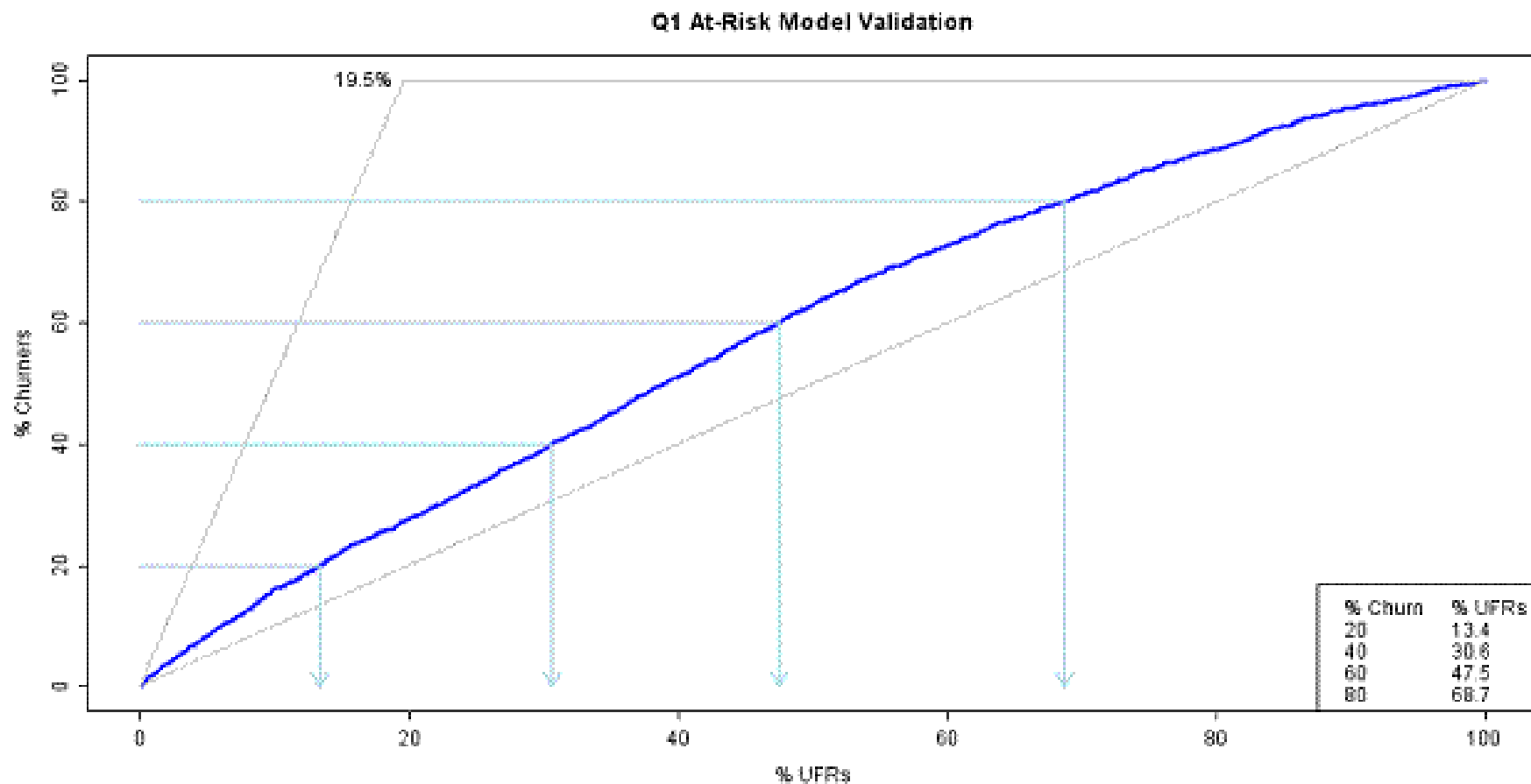
CDP for continuous predictors.



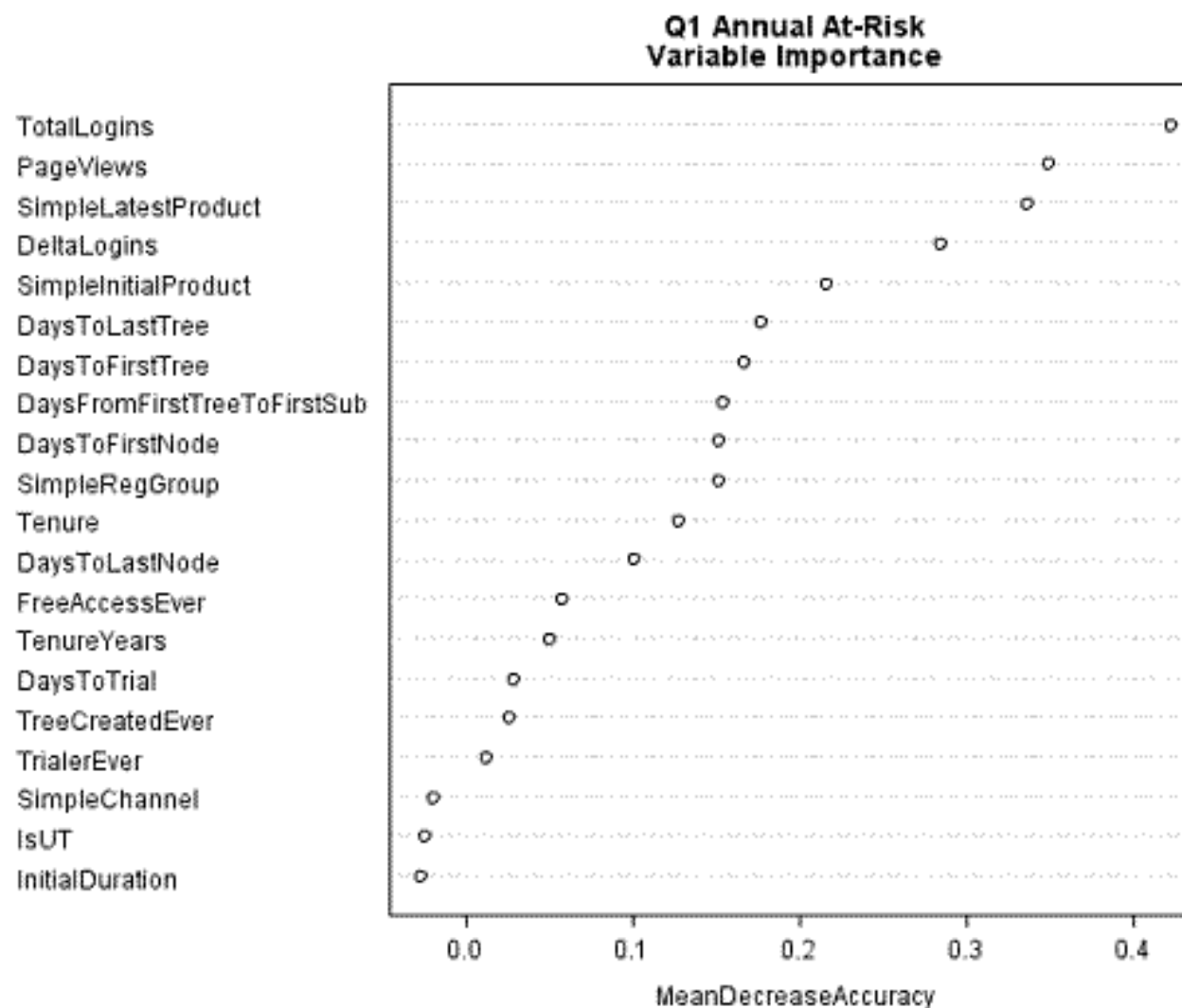
Mosaic plot for categorical predictors.



Visualizing Results I – Lift Plot



Visualizing Results II – Variable Importance

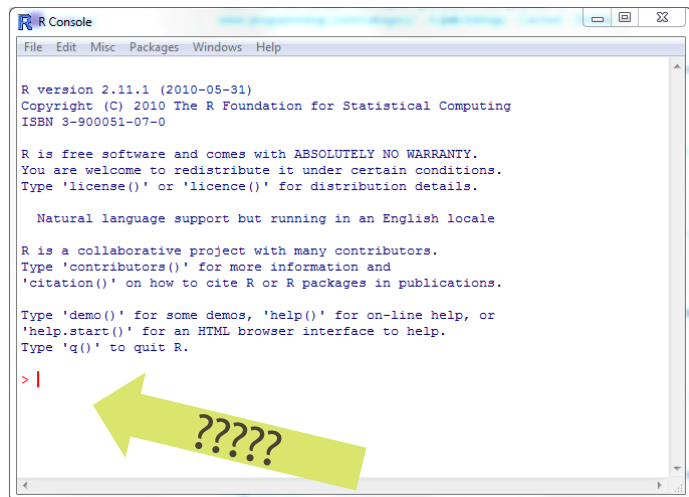




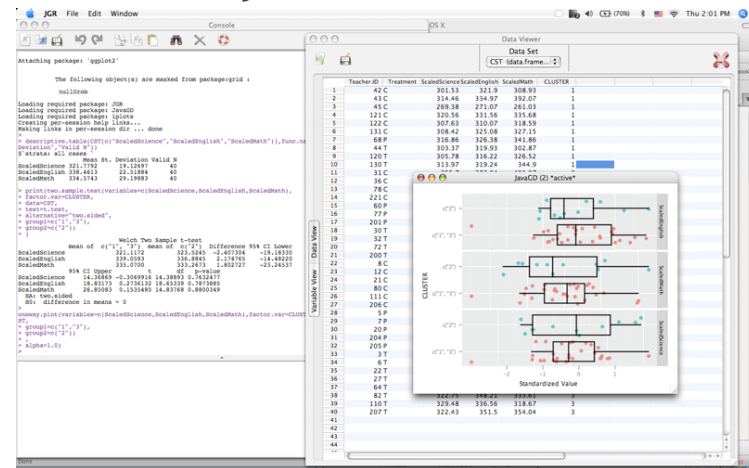
R Challenges

The “no free lunch” part.

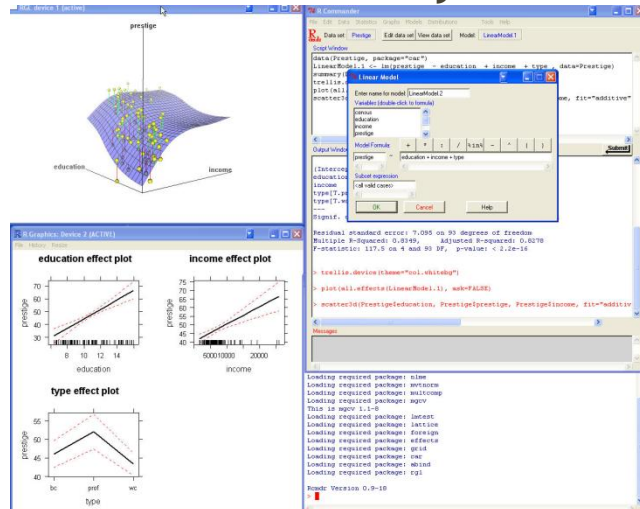
The Dreaded Blank Screen!



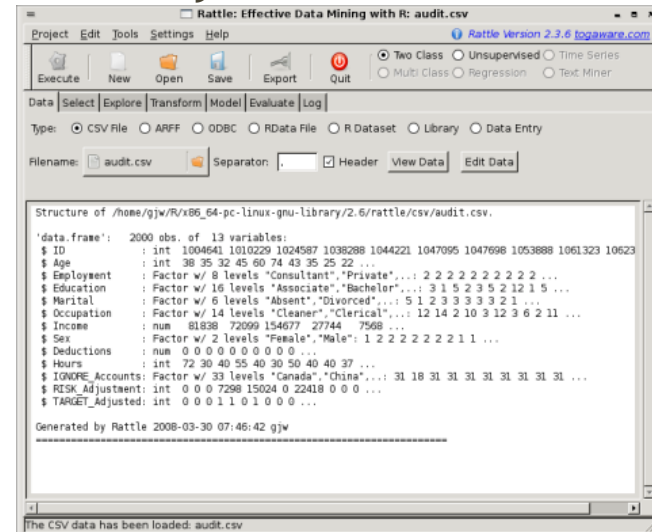
Deducer by Ian Fellows



The R Commander by John Fox



Rattle by Graham Williams



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

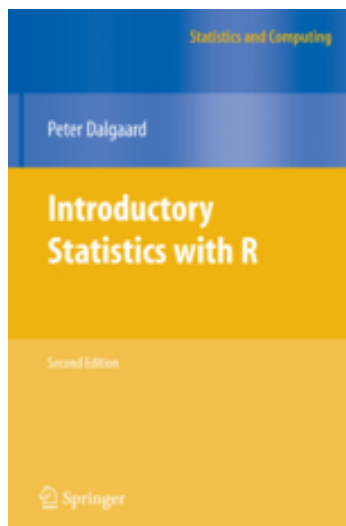
Introductory Statistics with R

Series: [Statistics and Computing](#)

Dalgaard, Peter

2nd ed., 2008, XVI, 364 p., Softcover

ISBN: 978-0-387-79053-4



R in a Nutshell

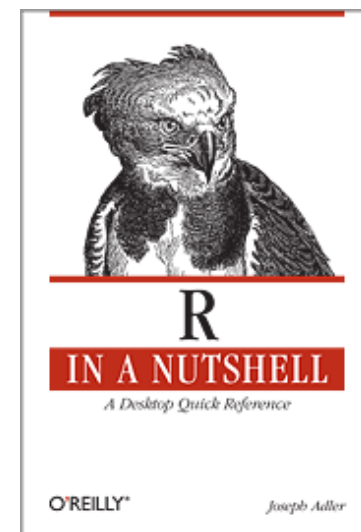
A Desktop Quick Reference

By [Joseph Adler](#)

Publisher: O'Reilly Media

Released: December 2009

Pages: 640



Linear Models with R

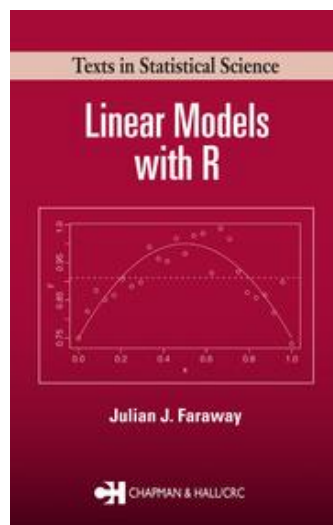
Julian J. Faraway,

University of Bath, United Kingdom

Cat. #: C4258

ISBN: 9781584884255

Publication Date: August 12, 2004



ggplot2

Elegant Graphics for Data Analysis

Series: [Use R](#)

Wickham, Hadley

2nd Printing., 2009, VIII, 216 p.,

Softcover

ISBN: 978-0-387-98140-6



Also his:

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

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