Commonly Used Statistical Packages in R

A package in R is a related set of capabilities, functions, help pages, and sometimes data that are bundled together. R comes with a basic set of packages with the initial download of the software, but multiple packages exist, and much of R's flexibility derives from the extensive set of user-developed packages.

This document will list and describe several commonly used packages for statistical analysis, data manipulation, and graphing. Throughout this document, text written `in this font` will indicate text that you should type verbatim into the R console. Text *in italics* will indicate text that you should substitute with the correct filename, pathname, command name, or other information specific to your needs.

For a complete list of packages supported by CRAN (the Comprehensive R Archive Network) see:

http://cran.r-project.org/web/packages/

Note: At the time this document was written, 2448 packages are available.

For a list of the packages that are included in the R download, see:

http://www.r-project.org/

You can also use the following command in the R console, to see what packages you have installed by default:

```
> getOption("defaultPackages")
```

Table of Contents

Installing and Loading Packages

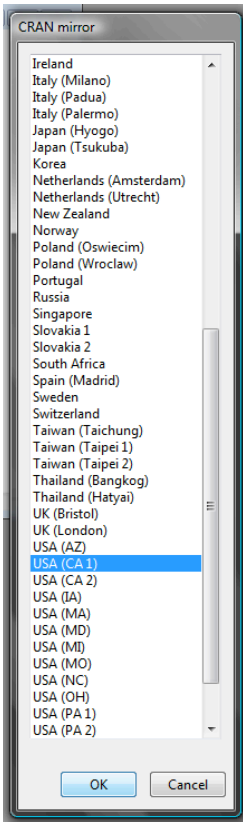
There are two main options for installing packages in R. First, you can download and install a package using the install.packages command:

```
> install.packages("package name")
```

Make sure to include the quotation marks around the package name (either single or double quotes will work).

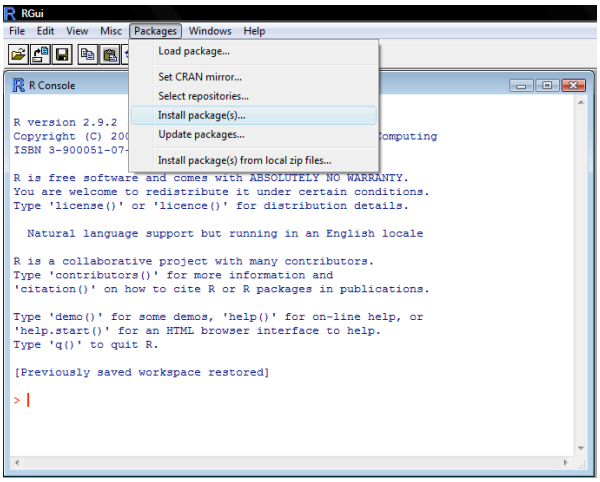
The first time you install a package in a session, a dialog box will open, asking you to choose a CRAN mirror for this session. Mirrors are sites around the world that store the packages, and from which you

can download them. Choose a mirror that is close to you geographically to minimize download time, (if you are at or near Stanford, USA (CA1), which is UC Berkeley, is a good choice) and click OK.



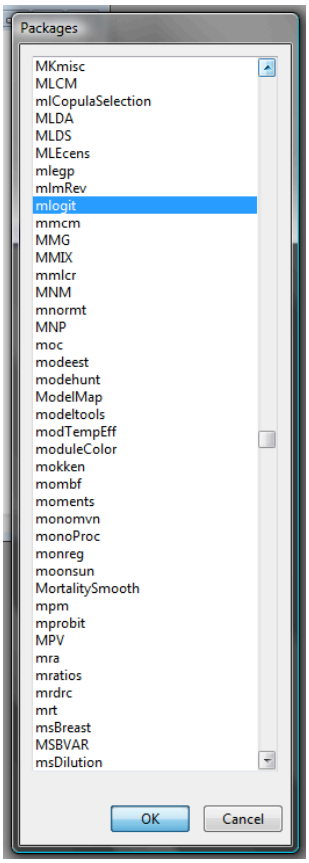
R will now automatically install the package.

Alternatively, in the R window, you can choose the "Packages" drop-down menu, and the "Install Package(s)…" option.



As shown above, this will open a dialog box asking you to choose a CRAN mirror for this session. Choose a mirror that is close to you geographically to minimize download time, and click OK.

A second dialog box will open, asking you which package you wish to install. Select the desired package name and click OK.



Once a package has been installed, you do not need to reinstall it. However, you will need to load it into a library in each session when you wish to use it. You can load a package using:

```
> library(package name)
```

Note: whereas `install.packages()` requires quotation marks around the package name, `library()` takes no quotation marks.

Commonly Used Packages

The packages you will find helpful will vary depending on the tasks you are trying to accomplish. The following list is not intended to be exhaustive, but rather is meant as a starting place, pointing you toward some packages we have found useful. (Notes on each package derive from package descriptions available at http://cran.r-project.org/web/packages/)

AER: Applied Econometrics with R. Functions, data sets, examples, demos, and vignettes for the book Christian Kleiber and Achim Zeileis (2008), *Applied Econometrics with R*. The book includes examples for a wide range of econometric models, from classical linear regression models for cross-section, time series or panel data and the common non-linear models of microeconometrics, such as logit, probit, tobit models as well as regression models for count data, to recent semiparametric extensions.

car: This package accompanies J. Fox and S. Weisberg, *An R Companion to Applied Regression*, Second Edition. It includes functions for: ANOVA analysis, matrix and vector transformations, printing readable tables of coefficients from several regression models, creating residual plots, tests for autocorrelation of error terms, and many other general interest statistical and graphing functions.

Design: Regression modeling, testing, estimation, validation, graphics, prediction, and typesetting by storing enhanced model design attributes in the fit. Design is a collection of about 180 functions that assist and streamline modeling, especially for biostatistical and epidemiologic applications. It also contains new functions for binary and ordinal logistic regression models and the Buckley-James multiple regression model for right-censored responses, and implements penalized maximum likelihood estimation for logistic and ordinary linear models. Design works with almost any regression model, but it was especially written to work with logistic regression, Cox regression, accelerated failure time models, ordinary linear models, the Buckley-James model, and generalized least squares for serially or spatially correlated observations.

foreign: Functions for reading and writing data stored by statistical packages such as Minitab, S, SAS, SPSS, Stata, Systat, and others and for reading and writing dBase files.

ggplot2: ggplot2 is a plotting system for R, based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad parts. It takes care of many of the fiddly details that make plotting a hassle (like drawing legends) as well as providing a powerful model of graphics that makes it easy to produce complex multi-layered graphics. Extensive documentation and examples here http://had.co.nz/ggplot/.

igraph: A package for creating and manipulating undirected and directed graphs. It includes implementations for classic graph theory problems like minimum spanning trees and network flow, and also implements algorithms for some recent network analysis methods, like community structure search. Excellent capabilities for visualizing network data, samples available here. See the igraph website and the Stanford network analysis labs for examples.

lme4: Linear mixed-effects models using S4 classes. Fit linear and generalized linear mixed-effects models (also referred to as multi-level models and hierarchical linear models).

lattice: A widely used data visualization framework based on principles developed by Cleveland, Tufte, and Tukey. See package author Deepayan Sarkar's introduction to the package. Perhaps most widely used is the implementation of Cleveland's dotplots; see Jacoby's webpage for an excellent overview on how to use and produce dotplots.

MASS: Functions and datasets to support Venables and Ripley, *Modern Applied Statistics with S* (4th edition). Includes many useful functions and data examples, including functions for estimating linear models through Generalized Least Squares, fitting negative binomial linear models, robust fitting of linear models, and Kruskal's non-metric multidimensional scaling.

mice: Multiple Imputation by Chained Equations. Multiple imputation is a technique for analyzing incomplete datasets where missing data are a concern.

mlogit: Estimation by maximum likelihood of the multinomial logit model, with alternative-specific and/or individual specific variables. Includes the commands mlogit.data, which reshapes a data.frame in a suitable form for the mlogit function.

moments: Functions to calculate: moments, Pearson's kurtosis, Geary's kurtosis and skewness; tests related to them.

MNP: MNP fits the Bayesian multinomial probit model via Markov chain Monte Carlo. The multinomial probit model is often used to analyze the discrete choices made by individuals recorded in survey data. The MNP software can also fit the model with different choice sets for each individual, and complete or partial individual choice orderings of the available alternatives from the choice set.

muhaz: Hazard Function Estimation in Survival Analysis. This package is for producing a smooth estimate of the hazard function for censored data.

Rcmdr: R Commander. A version of R with drop-down menus for performing basic statistical analyses. Although this package is limited in what it can do, it is good for people who are new to R. It produces R syntax for each task run from the drop-down menus so that the user can learn how to program in R. Please see the Guide to R Commander on the SSDS website: http://www.stanford.edu/group/ssds/cgi-bin/drupal/content/software-services-getting-started-guides-documents

reshape: A well-documented framework for reformatting your data. It provides a framework for converting data from "long" to "wide" format and vice-versa, and also helps one prepare data for visualization with lattice or ggplot2. See http://had.co.nz/reshape/.

sampleSelection: Provides methods for estimating sample selection models, also known as generalized tobit or Heckman selection models. Includes Heckman two-step, maximum likelihood estimation, and calculation of Inverse Mills Ratios.

sandwich: Model-robust standard error estimators for cross-sectional, time series and longitudinal data.

sem: Structural Equation Models. This package contains functions for fitting general linear structural equation models (with observed and unobserved variables) by the method of maximum likelihood, and for fitting structural equations in observed-variable models by two-stage least squares.

statnet: provides routines for the statistical modeling of network data, including exponential random graph models (ERGMs), latent position and cluster models, and permutation models. It also provides a framework for managing network data and producing visualizations, though most people prefer **igraph** for that purpose. See the statnet website and the Stanford network analysis labs for examples.

survival: Survival Analysis: descriptive statistics, two-sample tests, parametric accelerated failure models, Cox model. Delayed entry (truncation) allowed for all models; interval censoring for parametric models. Case-cohort designs.

tm: provides a comprehensive text mining framework for R. The Journal of Statistical Software article Text Mining Infrastructure in R gives a detailed overview and presents techniques for count-based analysis methods, text clustering, text classification and string kernels.

topicmodels: extends **tm** allowing for the probabilistic modeling of term frequency occurrences in documents. The package implements latent Dirichlet allocation (LDA) and the correlated topic model (CTM), and provides a framework for implementing other models or estimation strategies. The package vignette provides many useful examples.

truncreg: Truncated Regression Models: Estimation of models for truncated variables by maximum likelihood.

vcd: Visualizing Categorical Data.Visualization techniques, data sets, summary and inference procedures aimed particularly at categorical data. Special emphasis is given to highly extensible grid graphics. The package was inspired by the book "Visualizing Categorical Data" by Michael Friendly.

For More Information and Assistance

R Documentation and Books

Please see the document “Resources for Learning R” on SSDS website.

SSDS Software Services at Stanford

Software Services provides technical support for statistical software users at Stanford. Users can ask questions or make appointments with the consultants via our website. For more information or to contact us, see the web at:

http://ssds.stanford.edu/

Note: this document is based on R 2.9.2 for Windows, and R 2.10.0 for Macintosh.

Social Science Data and Software
Document revised: 9/21/2010