Introduction to R
A Statistical Computing Environment

J.C. Wang

Department of Statistics
Western Michigan University

September 19, 2008 / Statistics Seminar
Outline

1. Introduction
   - What is R
   - R Environment

2. Examples
   - Statistics
   - Graphics
   - Modeling
Outline

1 Introduction
   - What is R
   - R Environment

2 Examples
   - Statistics
   - Graphics
   - Modeling
What is R

- a language and environment for statistical computing and graphics
- a dialect of S language (John Chambers, 1998 ACM Software System Award winner)
- initially written by Robert Gentleman and Ross Ihaka
- actively developed/improved by R core team and many contributors
What is R

- a language and environment for statistical computing and graphics
- a dialect of S language (John Chambers, 1998 ACM Software System Award winner)
- initially written by Robert Gentleman and Ross Ihaka
- actively developed/improved by R core team and many contributors
What is R

- a language and environment for statistical computing and graphics
- a dialect of S language (John Chambers, 1998 ACM Software System Award winner)
- initially written by Robert Gentleman and Ross Ihaka
- actively developed/improved by R core team and many contributors
What is R

- a language and environment for statistical computing and graphics
- a dialect of S language (John Chambers, 1998 ACM Software System Award winner)
- initially written by Robert Gentleman and Ross Ihaka
- actively developed/improved by R core team and many contributors
A Programming Environment

- a true computer language
- fully planned and coherent
- allows users to add additional functionality by defining new functions
- easy to extend by packages
A Programming Environment

- a true computer language
- fully planned and coherent
- allows users to add additional functionality by defining new functions
- easy to extend by packages
A Programming Environment

- a true computer language
- fully planned and coherent
- allows users to add additional functionality by defining new functions
- easy to extend by packages
A Programming Environment

- a true computer language
- fully planned and coherent
- allows users to add additional functionality by defining new functions
- easy to extend by packages
Integrated Suite of Software Facilities

including

- an effective data handling and storage facility
- a suite of operators for calculations on arrays/matrices
- a large collection of intermediate tools for data analysis
- graphical facilities for data analysis and display
- a well-developed, simple and effective programming language
Integrated Suite of Software Facilities

including

- an effective data handling and storage facility
- a suite of operators for calculations on arrays/matrices
- a large collection of intermediate tools for data analysis
- graphical facilities for data analysis and display
- a well-developed, simple and effective programming language
Integrated Suite of Software Facilities

including

- an effective data handling and storage facility
- a suite of operators for calculations on arrays/matrices
- a large collection of intermediate tools for data analysis
- graphical facilities for data analysis and display
- a well-developed, simple and effective programming language
Integrated Suite of Software Facilities

including

- an effective data handling and storage facility
- a suite of operators for calculations on arrays/matrices
- a large collection of intermediate tools for data analysis
- graphical facilities for data analysis and display
- a well-developed, simple and effective programming language
Integrated Suite of Software Facilities

including

- an effective data handling and storage facility
- a suite of operators for calculations on arrays/matrices
- a large collection of intermediate tools for data analysis
- graphical facilities for data analysis and display
- a well-developed, simple and effective programming language
Outline

1. Introduction
   - What is R
   - R Environment

2. Examples
   - Statistics
   - Graphics
   - Modeling
Descriptive Statistics

- mean, median, variance and standard deviation
- summary (5-number summary and mean)
- quantile
- table and flat-table of categorical variables
- column sum/column means, row sum/row means
Descriptive Statistics

- mean, median, variance and standard deviation
- summary (5-number summary and mean)
- quantile
- table and flat-table of categorical variables
- column sum/column means, row sum/row means
Descriptive Statistics

- mean, median, variance and standard deviation
- summary (5-number summary and mean)
- quantile
- table and flat-table of categorical variables
- column sum/column means, row sum/row means
Descriptive Statistics

- mean, median, variance and standard deviation
- summary (5-number summary and mean)
- quantile
- table and flat-table of categorical variables
- column sum/column means, row sum/row means
Descriptive Statistics

- mean, median, variance and standard deviation
- summary (5-number summary and mean)
- quantile
- table and flat-table of categorical variables
- column sum/column means, row sum/row means
Distributions and Random Variates

- density: \( d \text{dist}r \)
- probability (cdf): \( p \text{dist}r \)
- quantile (inverse cdf): \( q \text{dist}r \)
- random variates: \( r \text{dist}r \)
- \( \text{dist}r \) include: \( \text{normal}, t, f, \text{chisquare}, \text{binomial}, \text{poisson}, \ldots \)
Distributions and Random Variates

- density: \( ddistr \)
- probability (cdf): \( pdistr \)
- quantile (inverse cdf): \( qdistr \)
- random variates: \( rdistr \)
- \( distr \) include: normal, t, f, chisquare, binomial, poisson, ...
Distributions and Random Variates

- density: $ddistr$
- probability (cdf): $pdistr$
- quantile (inverse cdf): $qdist$
- random variates: $rdist$
- $distr$ include: normal, t, f, chisquare, binomial, poisson,...
Distributions and Random Variates

- density: \texttt{ddistr}
- probability (cdf): \texttt{pdistr}
- quantile (inverse cdf): \texttt{qdist}
- random variates: \texttt{rdistr}
- \texttt{distr} include: \texttt{normal}, \texttt{t}, \texttt{f}, \texttt{chisquare}, \texttt{binomial}, \texttt{poisson}, ...
Distributions and Random Variates

- density: \texttt{ddistr}
- probability (cdf): \texttt{pdistr}
- quantile (inverse cdf): \texttt{qdistr}
- random variates: \texttt{rdistr}

\texttt{distr} include: \texttt{normal, t, f, chisquare, binomial, poisson, ...}
Simple Plots

examples

- scatterplots (plot)
- boxplots (boxplot)
- histograms (hist)
Simple Plots

examples

- scatterplots (plot)
- boxplots (boxplot)
- histograms (hist)
Simple Plots

examples

- scatterplots (plot)
- boxplots (boxplot)
- histograms (hist)
Exploratory Plots

examples

- scatterplot matrix (pairs)
- condition plot (coplot)
Exploratory Plots

examples

- scatterplot matrix (pairs)
- condition plot (coplot)
Publication-quality Plots
lattice package

- **univariate**: barchart, bwplot, stripplot, dotplot, histogram, qqmath, densityplot
- **bivariate**: xyplot, qq
- **trivariate**: levelplot, contourplot, cloud, wireframe
- **hypervariate**: splom, parallel
- **others**: rfs, tmd
Publication-quality Plots
lattice package

- **univariate**: barchart, bwplot, stripplot, dotplot, histogram, qqmath, densityplot
- **bivariate**: xyplot, qq
- **trivariate**: levelplot, contourplot, cloud, wireframe
- **hypervariate**: splom, parallel
- **others**: rfs, tmd
Publication-quality Plots
lattice package

- **univariate**: `barchart`, `bwplot`, `stripplot`, `dotplot`, `histogram`, `qqmath`, `densityplot`
- **bivariate**: `xyplot`, `qq`
- **trivariate**: `levelplot`, `contourplot`, `cloud`, `wireframe`
- **hypervariate**: `splom`, `parallel`
- **others**: `rfs`, `tmd`
Publication-quality Plots
lattice package

- **univariate**: barchart, bwplot, stripplot, dotplot, histogram, qqmath, densityplot
- **bivariate**: xyplot, qq
- **trivariate**: levelplot, contourplot, cloud, wireframe
- **hypervariate**: splom, parallel
- **others**: rfs, tmd
Publication-quality Plots
lattice package

- **univariate**: barchart, bwplot, stripplot, dotplot, histogram, qqmath, densityplot
- **bivariate**: xyplot, qq
- **trivariate**: levelplot, contourplot, cloud, wireframe
- **hypervariate**: splom, parallel
- **others**: rfs, tmd
Modeling Tools

- ANOVA model (`aov`, `anova`)
- linear model (`lm`)
- nonlinear model (`nls`)
- time-series model
- factor model (`factanal`)
- linear/nonlinear mixed-effect model (`nlme` in package `nlme`)
Modeling Tools

- ANOVA model (`aov`, `anova`)
- linear model (`lm`)
- nonlinear model (`nls`)
- time-series model
- factor model (`factanal`)
- linear/nonlinear mixed-effect model (`nlme` in package `nlme`)
Modeling Tools

- ANOVA model (*aov, anova*)
- linear model (*lm*)
- **nonlinear model** (*nls*)
- time-series model
- factor model (*factanal*)
- linear/nonlinear mixed-effect model (*nlme in package nlme*)
Modeling Tools

- ANOVA model (`aov`, `anova`)
- linear model (`lm`)
- nonlinear model (`nls`)
- time-series model
- factor model (`factanal`)
- linear/nonlinear mixed-effect model (`nlme` in package `nlme`)
Modeling Tools

- ANOVA model (aov, anova)
- linear model (lm)
- nonlinear model (nls)
- time-series model
- factor model (factanal)
- linear/nonlinear mixed-effect model (nlme in package nlme)
Modeling Tools

- ANOVA model (aov, anova)
- linear model (lm)
- nonlinear model (nls)
- time-series model
- factor model (factanal)
- linear/nonlinear mixed-effect model (nlme in package nlme)
Projects

- Data Expo 2009: Airline on time data
  Aim: to provide a graphical summary of important features of the data set
- John Chambers Award
Projects

- Data Expo 2009: Airline on time data
  Aim: to provide a graphical summary of important features of the data set
- John Chambers Award
R Links

- R Homepage
- Comprehensive R Archive Network
- One R Tip a Day
- R Graph Gallery
- R Wiki
- R_note: The Exploration of Statistical Software R
- Technical Notes on the R Programming Language
- John Chambers Award
- Data Expo 2009