

Ecological Data Analysis with R

Glen Sargeant

Northern Prairie Wildlife Research Center

Jamestown, North Dakota

glen_sargeant@usgs.gov

Hosted by the

USGS Western Fisheries Research Center

Seattle, Washington

24-25 March, 2009

What is R?

- R is not a statistics "program" in the traditional sense.
- R is an **open-source** implementation of the **S programming language** and an **environment** for statistical computing.

- S was developed at AT&T Bell Laboratories by John Chambers and colleagues.
 - Currently a commercial product (S-Plus) developed and supported by TIBCO Corporation of Palo Alto, California
- R originated as freeware written by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand.
 - Currently a GNU project maintained, developed, and documented by the R Development Core Team and an extensive international user-community



What makes R(S) unique?

- Most statistical software is based on a batch-processing model.
 - Users select from statistical methods and options provided by programmers.
 - Standardized output is typically the end product of an analysis.
 - Code is proprietary, hidden from users, and not easily modified.

Advantage: Relative ease of "typical" use

R(S) is based on a different model

- Programmers still try to anticipate needs and default output often is useful, but...
 - Standard output often is *not* an end product, but fodder for additional analysis
 - Code is freely available, written in the R language, and can easily be modified

Emphasis: Flexibility and extensibility; What John Chambers called "Programming with Data"

Strengths of R

- 1 Transparency, flexibility, and extensibility have helped make R the dominant medium for statistical research.
- 2 Use by academics for research and cost (free) have helped make R a popular teaching package.
- 3 Use for teaching and cost have encouraged professional use and development of a large user community.
- 4 Widespread use, extensibility, and a network for distributing improvements have fostered development and documentation.
- 5 R is becoming the *de facto* standard for ecological data analysis.

Objectives

- Identify the **minimum** set of concepts that must be mastered to use the R language effectively
- Pursue a **coordinated** understanding of objects, functions, and the structure of the R language
- Facilitate communication **with** and **about** R
- **Briefly** introduce **many** useful functions and practice key skills



- The R language
- Creating and saving objects
- Importing data
- Manipulating objects

Part II: Examples

- Summarizing univariate data
- Summarizing multivariate data
- Linear models
- Statistical graphics
- Spatial data and S4 classes
- Programming with functions
- Dates and times

Part III: Extending capabilities of R

- Package development
- Programming environments
- Report preparation with \LaTeX
- Presentations with BEAMER

- No prior knowledge of R
 - Coverage will be extensive but self-contained
- General knowledge of introductory statistics
 - Focus will be on implementation of statistical methods with R, *not* on statistics

For those with prior experience

- Coordination of concepts
- Improved workflow
- Useful functions and programming tricks

Acknowledgements

These notes reflect my understanding of R, which derives, in turn, from the following sources:

- Documentation distributed with R by the R Development Core Team.
- Package documentation and inspection of contributed code from the **Comprehensive R Archive Network (CRAN)**.
- Books on R and articles published in *R News*.
- Programming hints gleaned from internet forums.
- Insights of co-instructors.
- Feedback from workshop participants.

The correct citation for R is as follows:

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

The use of trade, product, or firm names does not imply endorsement by the U.S. Government.