Statistics and Mathematics
Wirtschaftsuniversität Wien
Wirtschaftsuniversität Wien, Augasse 2-6, A-1090 Wien, Austria
WIRTSCHAFTS UNIVERSITÄT
WIEN

APPLIED STATISTICS WITH R

May/June 2006

John Fox
McMaster University
Hamilton, Ontario, Canada

jfox@mcmaster.ca
http://socserv.socsci.mcmaster.ca/jfox/

What is S?

- A statistical programming language and computing environment.
- Two major implementations:
  1. S-PLUS: commercial, for Windows and Unix/Linux.
  2. R: free, open-source, for Windows, Macintoshes, Unix/Linux.

How does a statistical programming environment differ from a statistical package (such as SPSS)?

- A package is oriented toward combining instructions and rectangular datasets to produce (voluminous) printouts and graphs. Routine, standard data analysis is easy; innovation or nonstandard analysis is hard or impossible.
- A programming environment is oriented toward transforming one data structure into another. Programming environments such as S are *extensible*. Standard data analysis is easy, but so are innovation and nonstandard analysis.

Why use S?

- S has become the standard statistical software among statisticians. Consequently, new statistical methods are often first available in S.
- There is a great deal of built-in statistical functionality and many add-on libraries (“packages” in R) available that extend the basic functionality.
- S creates fine statistical graphs with relatively little effort.
- S is very well designed.
- S software is of very high quality.
- S is easy to use.

Which implementation of S should I use?

- The similarities between S-PLUS and R are more striking than the differences.
- S-PLUS has a graphical user interface for statistical functions (for those who like this kind of thing); the **Rcmdr** package in R provides a basic-statistics GUI.
- S-PLUS keeps working data in disk files (rather than in memory), making it possibly more suitable for the analysis of very large datasets. But R is able to interface with data-base management systems.
- Some advanced features of R (such as its so-called *scoping rules*) are more convenient.
- The current development of R is more dynamic.
- There is a Macintosh version of R.
- R is *free*.

This Course:

The purpose of this course is to show how to accomplish a variety of statistical tasks in R.

- Depending upon the topic, the statistical content is assumed known or an overview will be provided.
- With the exception of structural-equation modeling (which will use the **sem** package in R), most of the material for the course is drawn from J. Fox, *An R and S-PLUS Companion to Applied Regression*, Sage, 2002.
- More advanced students may prefer W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S, Fourth Edition*. New York: Springer, 2002
- Additional materials and links are available on the web site for the book: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion/index.html.
- The book is associated with an R package (called **car**) that implements a variety of methods helpful for analyzing data with linear and generalized linear models.
- Other references and materials are given on the course web site: http://statmath.wu-wien.ac.at/courses/StatsWithR/index.html.