R

Sebastian Landwehr

Overview

1. Introduction to R
2. Statistical Computing
3. Graphics
4. Strengths
5. Weaknesses

Introduction to R

What is R?

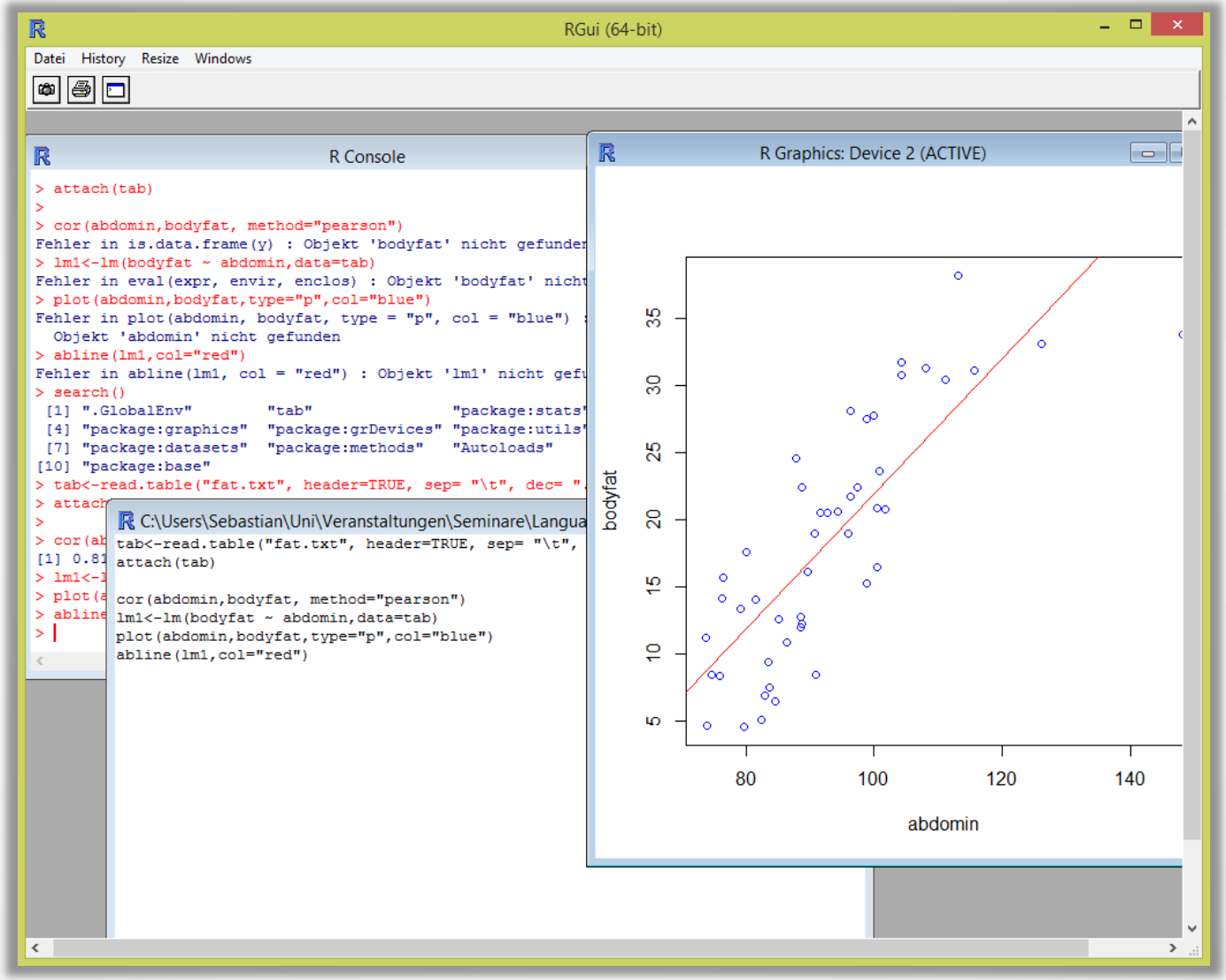
- Language and software suite for
  - Statistical Computing
  - Data Analysis in General
  - Graphics
- GNU project
- Free software (GPL)
- Cross-platform (Unix, Windows, MacOS)
- Even web-based via RStudio!
- Similar to the S language
- Extensible via packages

The R Language

- Full-featured programming language
- Conditionals
- Loops
- User-defined functions
- IO
- …
- C, C++ and Fortran can be called at runtime

➢ System is extended by packages written in R

User Interface



Statistical Computing

Statistical Computing

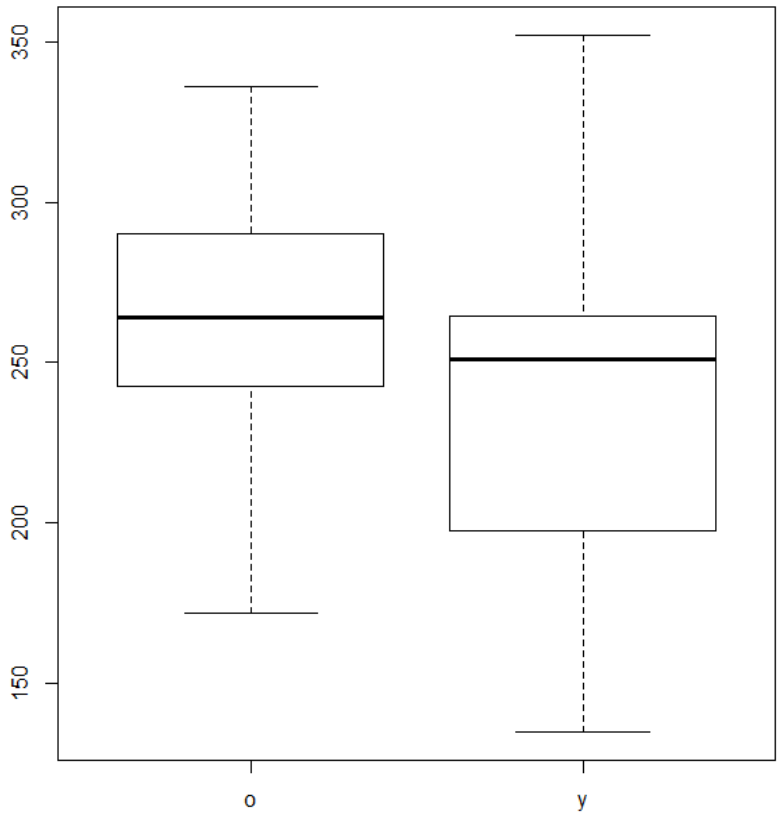
- Descriptive statistics
- All kinds of methods to analyze tabular data
- Calculate statistical parameters
  - Mean, deviation, quantiles, …
- Regression
- Contingency tables
- Plots
- Loads of packages created by
  - A large user base
  - Experts in statistics

1. Calculating Parameters

- Create a table of data
  - ```
    tab <- data.frame(
      Age=c("o","y","y","o","y","y","y","o","o","y","o",
            "y","y","o","y","y","o","o","o","y","o","o"),
      Cholesterol=c(294,222,251,254,269,235,286,246,239,173,277,
            135,260,286,252,352,336,208,311,156,172,264))
    ```
- Calculate mean and standard deviation
  - ```
    > mean(tab$Cholesterol)
    [1] 249
    > sd(tab$Cholesterol)
    [1] 55.38523
    ```
- Calculate mean for each age group
  - ```
    > tapply(tab$Cholesterol, tab$Age, mean)
          o        y
    262.4545 235.5455
    ```
- Calculate quantile
  - ```
    > tapply(tab$Cholesterol, tab$Age, quantile, p=0.25)
        o     y
    242.5 197.5
    ```

Plotting the Data

```
> boxplot(tab$Cholesterol ~ tab$Age);
```



2. Statistical Tests

- Are computer scientists taller than 1,80m?
  - `> x=c(188, 172, 194, 178, 170, 184, 198, 189, 175, 184)`
- Apply a one-sample t-Test
  - `> t.test(x, mu=180)`

```
        One Sample t-test

data:  x
t = 1.0817, df = 9, p-value = 0.3075
alternative hypothesis: true mean is not equal to 180
95 percent confidence interval:
 176.508 189.892
sample estimates:
mean of x
    183.2
```

- Cannot say that true mean differs from 1,80m!

2. Statistical Tests

- Testing the average size of computer scientists
  - `> x=c(188, 172, 194, 178, 170, 184, 198, 189, 175, 184)`
- Apply a one-sample t-Test
  - `> t.test(x, mu=170)`

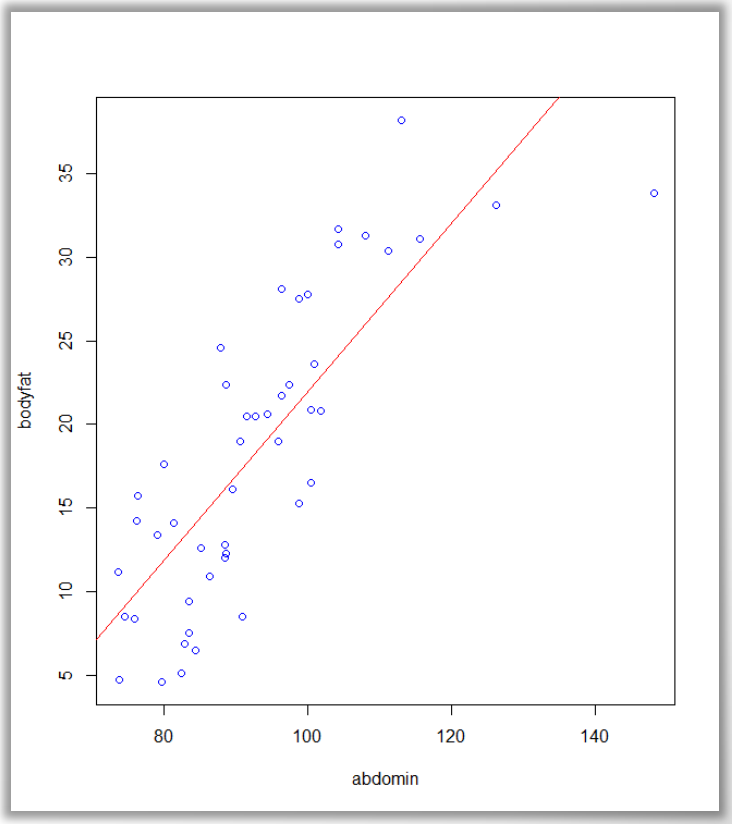
```
        One Sample t-test

data:  x
t = 4.4621, df = 9, p-value = 0.001572
alternative hypothesis: true mean is not equal to 170
95 percent confidence interval:
 176.508 189.892
sample estimates:
mean of x
    183.2
```

- True mean differs from 1,70m!

3. Regression

- Is there a correlation between body fat and the abdominal girth (Bauchumfang)?
- Calculate the correlation coefficient by Pearson
  - `> cor(abdomin, bodyfat, method="pearson")`
    `[1] 0.810928`
- Perform a **linear** regression of `bodyfat` against `abdomin`
  - `> lm1 <- lm(bodyfat ~ abdomin, data=tab)`
- Plot the data in a diagram
  - `> plot(abdomin, bodyfat, type="p", col="blue")`
- Add the regression function
  - `> abline(lm1,col="red")`



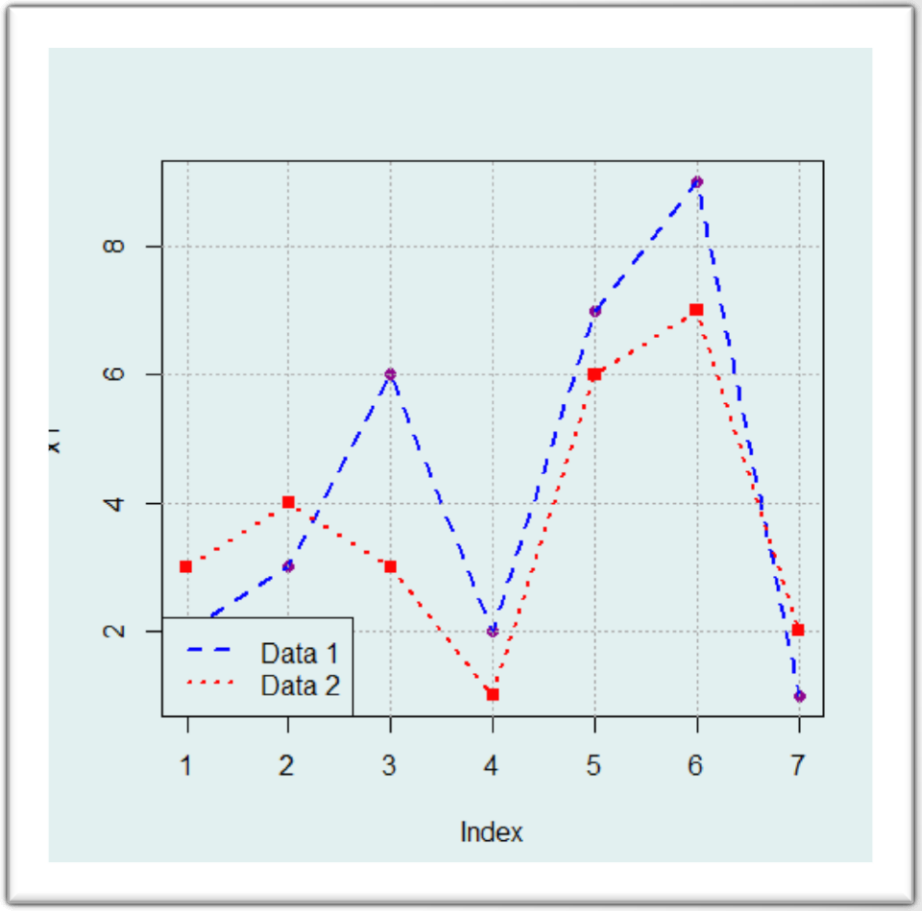
Graphics

The Base Graphics System

- Plotting points, lines, functions etc.
- Statistical diagrams
  - Boxplots
  - Bar charts
  - Pie charts
- 3D graphics
- System is based on an old S system
- Very flexible but still restricted in many aspects
  - Many external packages available!
- Graph is constructed by consecutively adding elements

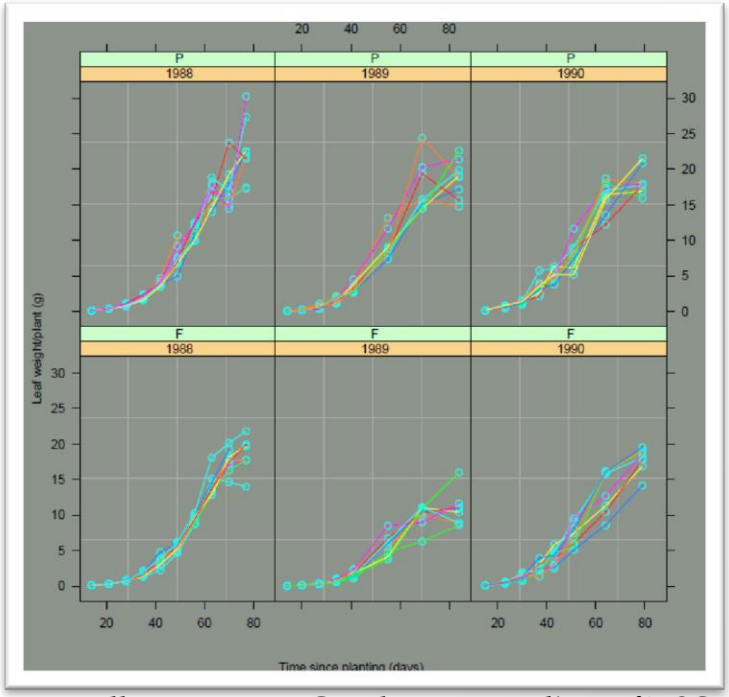
Example

```
> par(bg="azure2")
> x1 <- c(2,3,6,2,7,9,1)
> plot(x1, col="darkmagenta", pch=16)
> lines(x1, col="blue", lty=2, lwd=2)
> grid()
> x2 <- c(3,4,3,1,6,7,2)
> points(x2, pch=15, col="red")
> lines(x2, col="red", lty=3, lwd=2)
> legend("bottomleft",
      legend=c("Data 1","Data 2"),
      col=c("blue","red"),
      lty=2:3, lwd=2:2)
```



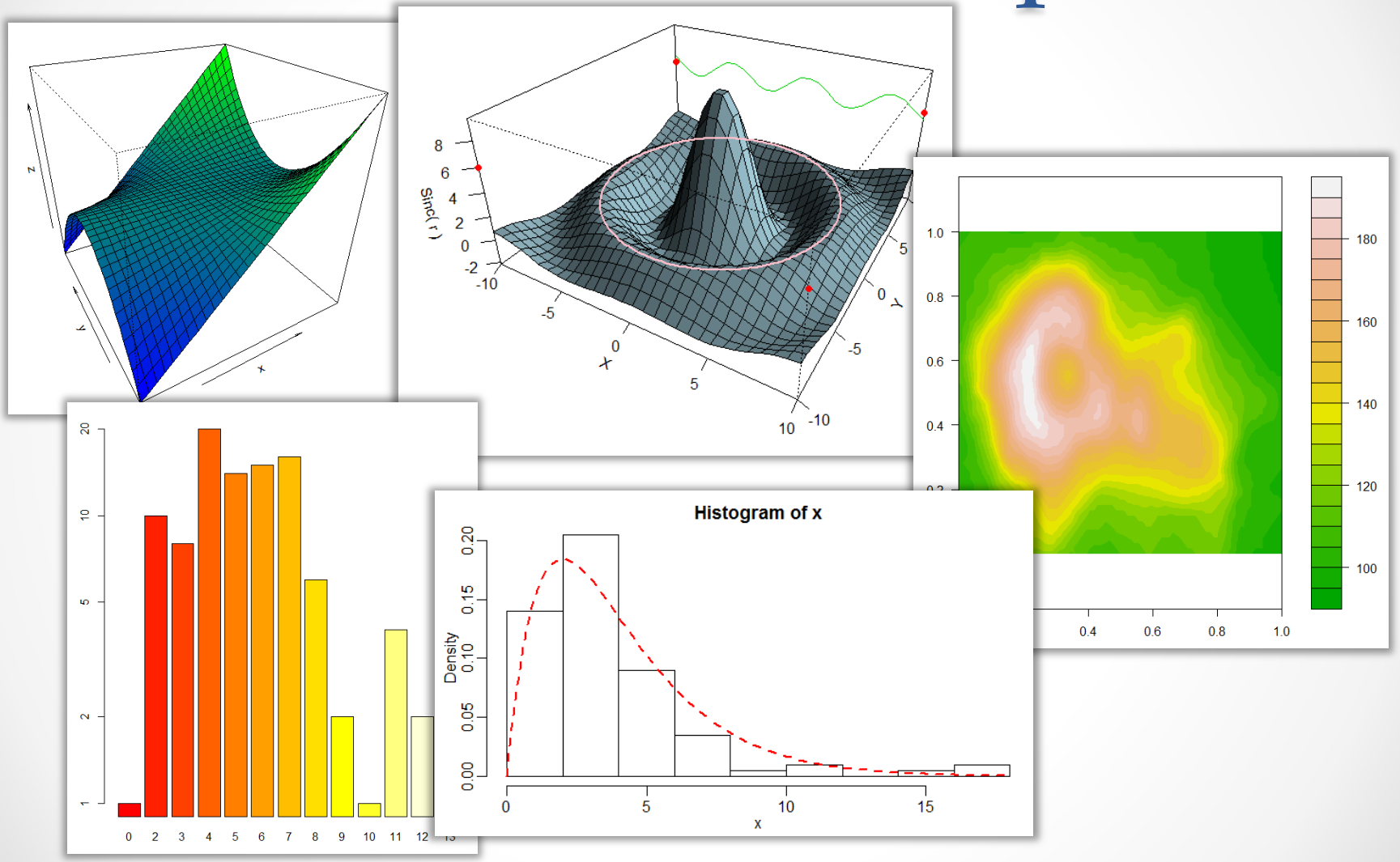
Lattice

- Package as an alternative to the base R graphics system
- Goals:
  - Produce graphics similar to the Trellis Graphics package in S-Plus
  - Improve aspects of base R
- Flexible viewports
- Coordinate systems and units
- Interaction and Customization
- Extensibility



Murrell, P. R Lattice Graphics. *Proceedings of DSC*, (2001), 2.

Advanced Examples



Strengths and Weaknesses

Strengths

- Free software
- Cross-platform and web-based
- Publication-quality graphics
- Up-to-date
  - Many new analysis methods first appear in R
- Highly readable language
- CRAN Repository
- Communication with other tools
  - Load and save to various file formats
  - Can be called within Python
- Popular
  - Large user base in industry and academic fields

Weaknesses

- No built-in parallelization
  - But packages available
  - `foreach`, `Rmpi`, `snow`, `snowfall`
- Swiss Cheese Phenomenon
  - Cascading dependencies between packages
- Stores data in RAM memory only
  - But packages available
  - `bigmemory`: Store big matrices as pointers to C++ or in a file
  - `ff`: Store data in a file
- Performance
- Steep learning curve

Thank You ;-)