Multiple linear regression background

(STATS544.2: Applied multiple linear regression)

Gunnar Stefansson

August 13, 2013

The basics

Want to describe or predict a dependent y variable from several independent/exploratory x variables.
Main package for this course: R
Examples will be from engineering, biology, economics, education etc.

For info on fields which use R, see http://cran.r-project.org/web/views/
For more economic data sets, see e.g. http://www.micecon.org/ or http://eu.wiley.com/legacy/wileychi/verbeek2ed/datasets.html
For economic applications with R see e.g. http://cran.r-project.org/web/views/Econometrics.html

Figure: Icecream data.

Plotting

The first step in analysing data should always consist of plots

Some R plot functions:

```
plot              - basic scatter plot
lines, points     - add lines or points to plot
pairs             - plot for all pairs in a data frame
histogram, truehist - various histograms
boxplot           - box and whiskers plot
```

More on R: See `http://r-project.org` and STATS240 on the tutor-web.

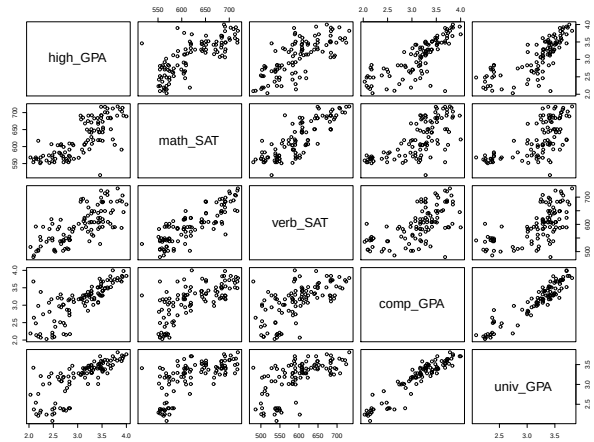

Figure: **Example (education):** Typical first plot using the pairs command. Comparing scores from several exams.

Data sets

Data are stored in files, which are then read into data frames. Data files (or data sets) should be structured as simple rectangular tables before they are read into R

The data set is commonly just a table of numbers, possibly with a single header line.

We think of the data frame in the same way - usually as a table of numbers.

Many data sets are available as built-in data sets in R

Data can also be read directly from a URL

R is commonly used in medicine.

A typical data set is the ais data set in the DAAG library.

```
library(DAAG)
data(ais)
pairs(ais[,c("rcc","wcc","bmi","ht","wt","pcBfat","lbm")])
```

Figure: Data set ais from the DAAG library (see http://cran.r-project.org/web/packages/DAAG/DAA

MULREG case studies

In a MULREG course you will be asked to do analyze certain data sets (case studies).
Depending on the course, these may be assigned by the instructor or you may need to find your own.

Example data sets:
`http://www.stats4stem.org/data-sets.html`
`http://tgax14.rhi.hi.is/html/data/biol/\protect\begingroup\immediate\write\@unused`
UCI Machine Learning Repository `http://archive.ics.uci.edu/ml/`
Competitions: `https://www.kaggle.com/c/informs2010/`
Info on fields which use R, see http://cran.r-project.org/web/views/
For more economic data sets, see e.g. `http://www.micecon.org/` or `http://eu.wiley.com/legacy/wileychi/verbeek2ed/datasets.html`
Economic applications `http://cran.r-project.org/web/views/Econometrics.html`

association vs causation

Evidence of a relationship does not show causation

Example (biology): A data set of several quantities from Icelandic waters can be found at `http://tgax14.rhi.hi.is/html/data/biol/borecol.txt`.
This can be read into R using

```
b<-read.table("http://tgax14.rhi.hi.is/html/data/biol/borecol.txt",header=T)
```

Just typing "b" shows the content..