Integrable e-lements for Statistics Education*

Wolfgang Härdle, Sigbert Klinke and Uwe Ziegenhagen

Humboldt-Universität zu Berlin
Center for Applied Statistics and Economics

Abstract

Without doubt modern education in statistics must involve practical, computer-based data analysis but the question arises whether and how computational elements should be integrated into the canon of methodological education. Should the student see and study high-level programming code right at the beginning of his or her studies? Which technology can be presented during class and which computational elements can re-occur (at increasing level of complexity) during the different courses?

In this paper we address these questions and discuss where e-techniques have their limits in statistics education.

Keywords electronic books, hypertext, e-supported teaching, statistical software

JEL codes I21, C19,

*This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 ”Economic Risk”.

1 Introduction

Since the entering of modern computing equipment into schools and universities there have been increased efforts to use computers not only for research and numerical computations but also for the education of students. While traditional textbooks on statistics are usually restricted to small examples, computers offer great opportunities to enrich the teaching of statistics by the means of explaining animations or on-the-fly computations of large real-world datasets.

But each new technology does not only hold opportunities and advantages there may also be hazards or risks. And a sentence stated by John Tukey in 1965 should be taken into consideration: "Each new generation of computers offers us new possiblities, at a time when we are far from using most of the possibilities offered by those already obsolete."

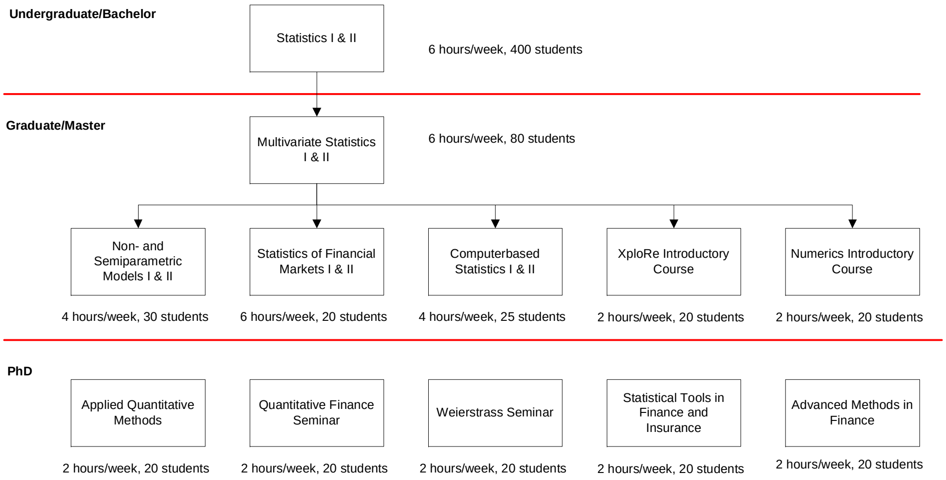
We will outline our thoughts about the integration of electronic and computational elements into statistics along the courses tought at the School for Business and Economics of Humboldt-Universität zu Berlin.

Students following different course programs are taught at the Institute for Statistics and Econometrics at Humboldt-Universität zu Berlin, German Business Administration and Economics students as well as students from international bachelor or master classes or the math department.

After finishing the bachelor level, where statistics is taught in a two-term lecture three hours a week, the students who choose statistics as one main subject are required to take a course on multivariate data analysis, that enables them to understand the basics of probability theory and to analyze high-dimensional data by the means of cluster-, principal component- and factor analysis.

On the basis of this lecture the student has further options: While *Non- and Semiparametric Modelling* focuses on nonparametric density estimation and regression the student learns in the computational statistics courses the applied data analysis with SPSS and real world data. Supplementary to the *Computerbased Statistics* courses are the *XploRe Introductory Course* and the *Numerics Introductory Course* focussing on the practical work with XploRe respectively the numerical details of selected algorithms.

Statistics of financial markets is the third specialisation taught at Humboldt-Universität, covered are theoretical and practical aspects of option pricing, risk-management and time series modelling.



Undergraduate/Bachelor
Statistics I & II
6 hours/week, 400 students
Graduate/Master
Multivariate Statistics I & II
6 hours/week, 80 students
Non- and Semiparametric Models I & II
Statistics of Financial Markets I & II
Computerbased Statistics I & II
XploRe Introductory Course
Numerics Introductory Course
4 hours/week, 30 students
6 hours/week, 20 students
4 hours/week, 25 students
2 hours/week, 20 students
2 hours/week, 20 students
PhD
Applied Quantitative Methods
Quantitative Finance Seminar
Weierstrass Seminar
Statistical Tools in Finance and Insurance
Advanced Methods in Finance
2 hours/week, 20 students
2 hours/week, 20 students
2 hours/week, 20 students
2 hours/week, 20 students
2 hours/week, 20 students

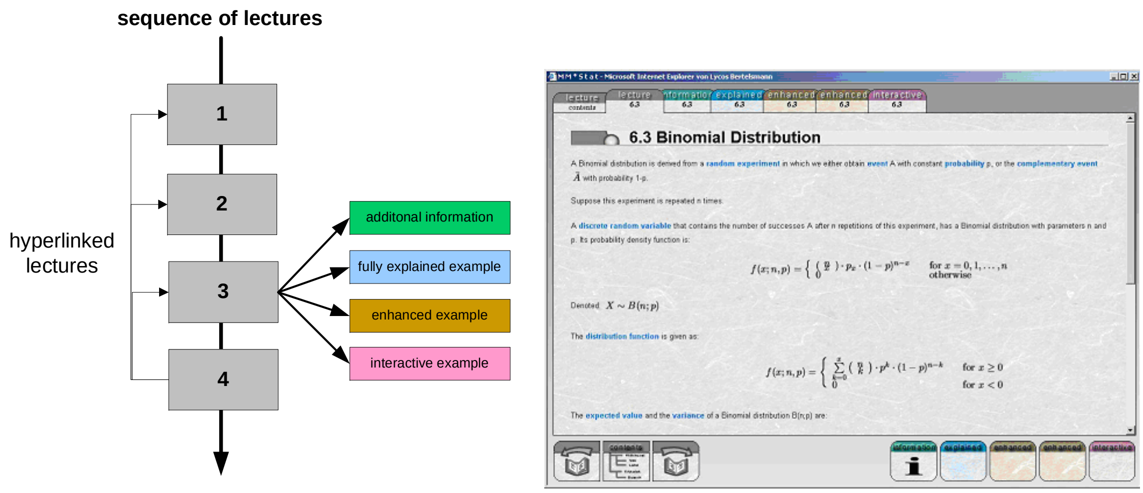
Figure 1: Typical one year cycle of statistics courses in Berlin

At the master level there are currently five courses also aimed at PhD students: In *Applied Quantitative Methods* and the *Weierstrass Seminar* recent developments of mathematical statistics and data analysis are discussed, *Advanced Methods in Finance*, the *Quantitative Finance Seminar* and *Statistical Tools in Finance and Insurance* deal with the field of financial mathematics.

2 Traditional and Modern Teaching Material

The main components of MM*Stat (Müller, M., Rönz, B., Ziegenhagen, U., 2000), that was developed to support the undergraduate teaching of statistics, are *lecture units*, *additional information* and *(non)interactive examples*, presented in a style similar to filing cards. This structure, called *MD*Booklet*, has also been used for *Numerical Methods in Statistics* and *Finance Introductory Course* (http://www.quantlet.com/mdstat/products.html). The lecture units contain common and well-known topics as basic concepts of statistics, basics of probability calculus and sampling theory.

Each "lecture" filing card provides the basic concepts of the methods considered together with definitions, important formulas and graphics. "Fully explained", "Enhanced" and "Interactive" examples show how these formulas can be applied. The first explain standard classbook examples in detail while the latter discuss important aspects of more advanced examples. The interactive examples are one of the main features of MM*Stat. By using a combination of embedded Java-based *XploRe Quantlet Client* (Borak, S.,



6.3 Binomial Distribution

A Binomial distribution is derived from a random experiment in which we either obtain event A with constant probability p, or the complementary event $\bar{A}$ with probability 1-p.

Suppose this experiment is repeated n times.

A discrete random variable that contains the number of successes A after n repetitions of this experiment, has a Binomial distribution with parameters n and p. Its probability density function is:

$$f(x;n,p) = \begin{cases} \binom{n}{x} \cdot p_x \cdot (1-p)^{n-x} & \text{for } x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

Denoted $X \sim B(n;p)$

The distribution function is given as:

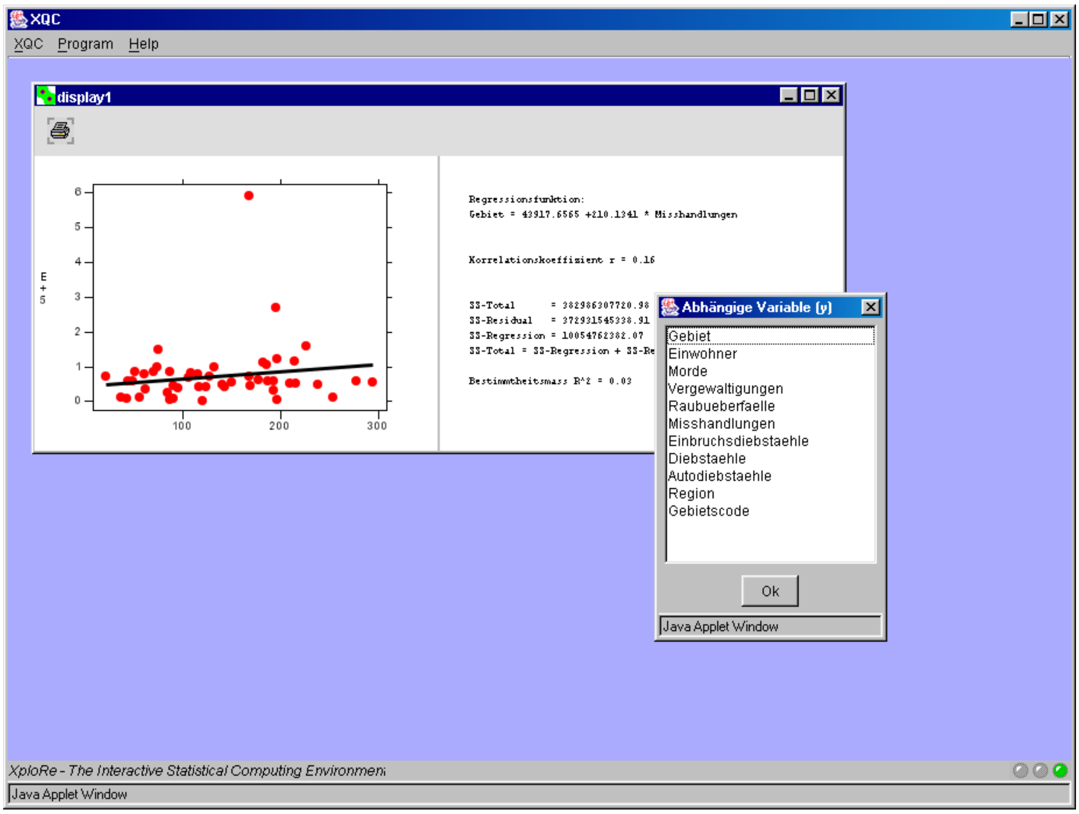
$$f(x;n,p) = \begin{cases} \sum_{k=0}^{x} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

The expected value and the variance of a Binomial distribution B(n;p) are:

Figure 2: MM*Stat: Layer Architecture and Screenshot

Härdle, W., Lehmann, H., 2005) and a remote or locally installed *XploRe Quantlet Server* the student can compute distribution functions, histograms and test results. Compared with classical textbooks where effects of parameter changes can only be presented as sequence of graphics or tables, the user can experience the results of different settings in realtime.

The repetition of concepts and methods introduced earlier has been implemented in two different ways: To see or recall the definition of a used term, a glossary has been written and important words within the different lectures inside MM*Stat are linked to this glossary. Furthermore the student can complete multiple choice questions by selecting radio buttons, the results are given instantly.



```
Regressionsfunktion:
Gebiet = 43917.6565 +210.1341 * Misshandlungen


Korrelationskoeffizient r = 0.16


SS-Total       = 382986307720.98
SS-Residual    = 372931545338.91
SS-Regression  = 10054762382.07
SS-Total = SS-Regression + SS-Re

Bestimmtheitsmass R^2 = 0.03
```

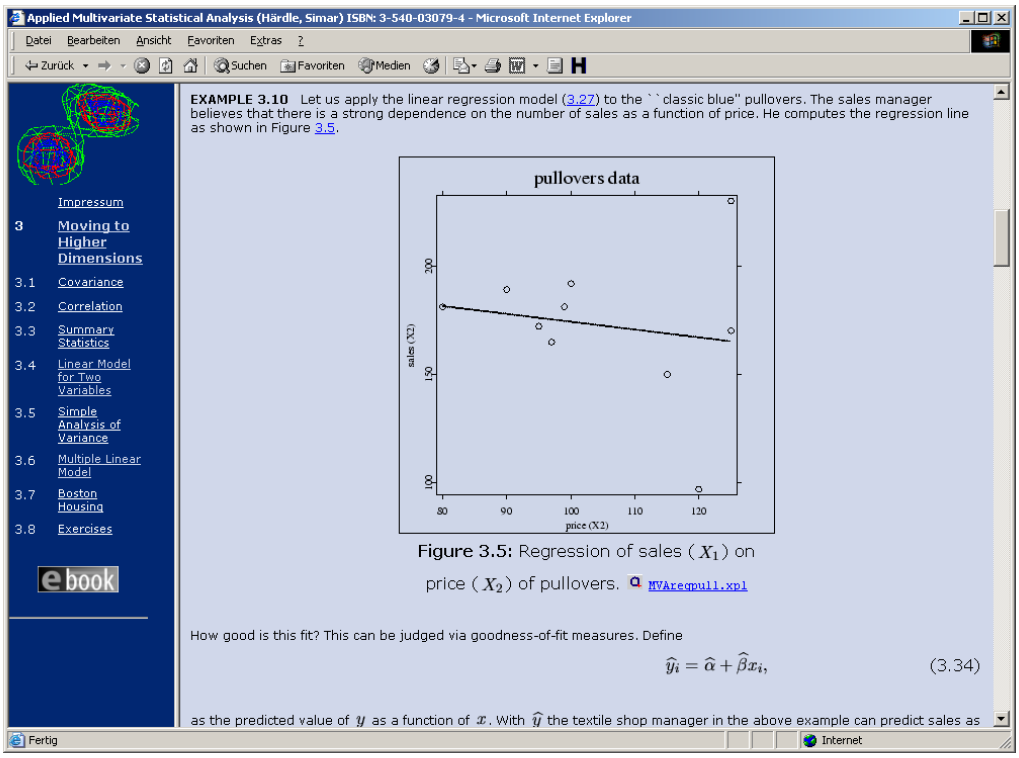
Figure 3: Example for linear regression in MM*Stat

The target group of MM*Stat, the undergraduate students, have different levels of literacy concerning computers, while some are firm in common office software tools others have little experience. So it was decided to hide the source code of the interactive examples from the user, to lower the inhibition threshold of using the software. Figure 3 shows a typical example for linear regression, which allows the user to choose the dependent and independent variable and returns the regression plot and the parameters of the model.

The combination of XploRe Quantlet Server and Client has not only been used in MM*Stat, furthermore it is used in the DoSS@d system (Mori, Y., Yamamoto, Y. and Yadohisa, H., 2003) and the books published by the members of the authors' institute (http://www.xplore-stat.de/ebooks/ebooks.html) also use this framework, although in a different way.

As mentioned the source code for the interactive examples was hidden from the user in MM*Stat, from the target group of the advanced books, graduate and PhD students on the one, researchers and professional on the other hand, we assume a certain level of computational literacy. In each book respectively the corresponding slides selected pages contain links to HTML-pages that contain the source code of the example as well as links to two different implementations of this example.

These two implementations differ in that sense that the 'edit' page calls the XploRe Quantlet Client in editor mode, which means that the user can manipulate the source



Applied Multivariate Statistical Analysis (Härdle, Simar) ISBN: 3-540-03079-4 - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück
Suchen
Favoriten
Medien
Impressum
3 Moving to Higher Dimensions
3.1 Covariance
3.2 Correlation
3.3 Summary Statistics
3.4 Linear Model for Two Variables
3.5 Simple Analysis of Variance
3.6 Multiple Linear Model
3.7 Boston Housing
3.8 Exercises
e book
pullovers data
sales (X2)
100
150
200
80
90
100
110
120
price (X2)
Fertig
Internet

EXAMPLE 3.10 Let us apply the linear regression model (3.27) to the ``classic blue'' pullovers. The sales manager believes that there is a strong dependence on the number of sales as a function of price. He computes the regression line as shown in Figure 3.5.

Figure 3.5: Regression of sales (X_1) on price (X_2) of pullovers. MVAregpull.xpl

How good is this fit? This can be judged via goodness-of-fit measures. Define

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i, \qquad (3.34)$$

as the predicted value of y as a function of x. With $\widehat{y}$ the textile shop manager in the above example can predict sales as

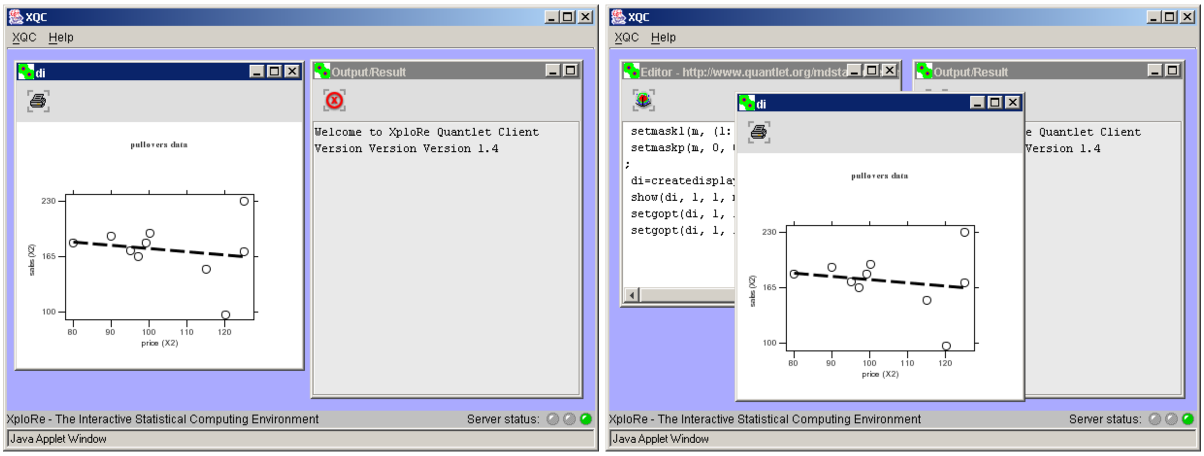
Figure 4: Applied Multivariate Statistics: HTML page with link to an example

code of the example. From MM*Stat they vary from the textual component. While at undergraduate level regression is taught using a notation with sums, the more abstract matrix-based approach is used for graduate students.

Figure 5 depicts two screenshort from the book *Applied Multivariate Analysis* by W. Härdle and L. Simar, on the left hand side the 'run' version of the quantlet, on the right the editable 'edit' version.

3 XploRe, Xplore Quantlet Server and Yxilon

Besides SPSS which is used for the computational statistics courses the software package XploRe is used for educating students at Humboldt-Universität. Developed in cooperation with MD*Tech (http://www.mdtech.de), XploRe is a full-featured statistical programming language. Using a matrix-oriented approach in combination with a C-style syntax a great variety of diverse statistical problems can be handled conveniently.



```
Welcome to XploRe Quantlet Client
Version Version Version 1.4
```

```
 setmaskl(m, (1:
 setmaskp(m, 0, 
;
 di=createdispla
 show(di, 1, 1, 
 setgopt(di, 1, 
 setgopt(di, 1, 
```

```
e Quantlet Client
Version 1.4
```

Figure 5: Applied Multivariate Statistics: *execute* and *edit* versions of an example

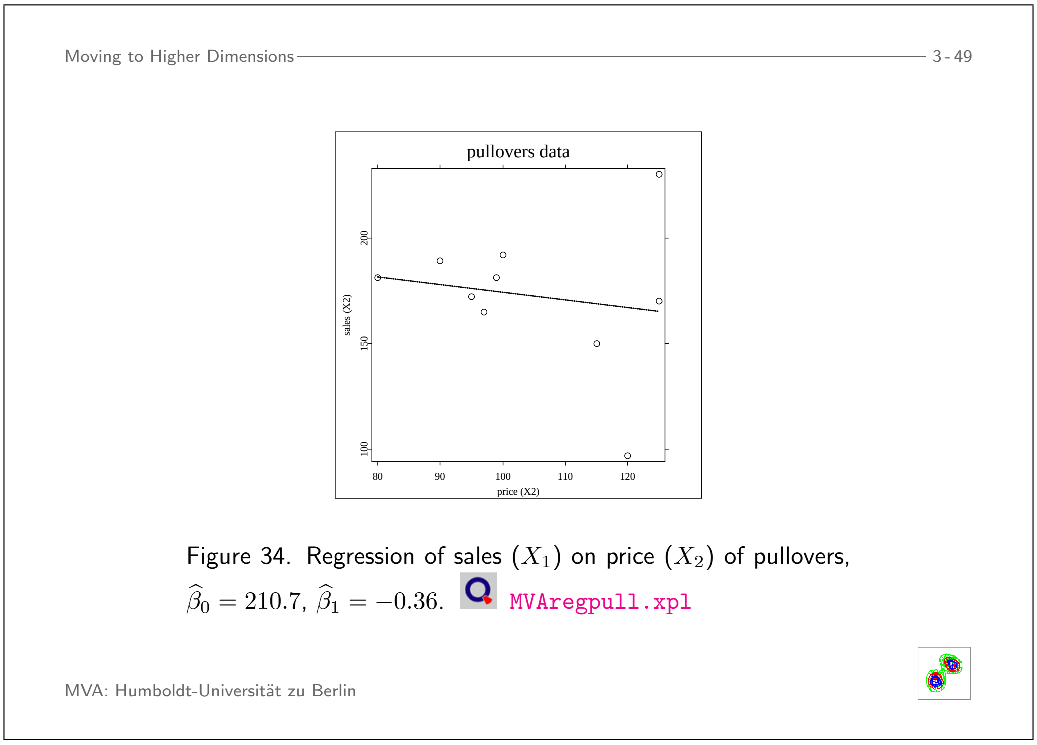
The implementation of XploRe incorporates ideas from the requirements of statistical software mentioned by Chambers and Lang (1999): usage from multiple front-ends, extensibility on language and native core level, interactive graphics and the inclusion of existing code (`C`, `Fortran`).

With Yxilon (Guril, Y., Klinke, S., Ziegenhagen, U., 2005) MD*Tech and Humboldt-Universität zu Berlin agreed on a new research project with the aim to supply a freely available statistical environment and to allow the implementation of recent developments in software technology while remaining fully compatible with XploRe.

The most dramatic change compared with XploRe or other packages as *R* or *Jasp* (Fujiwara, T., Ikunori K. et al. 2000) is the way the code is executed: Besides the interpretation by a so called runtime environment the idea is to compile directly to C++ and Java. Further changes are:

- published under Free-BSD license
- fully independent from machine architecture by using platform-independent protocols and software frameworks
- improved modularity to allow convenient exchange of single components
- integration of dynamic data sources as databases and webservices
- increased computing performance simplifying the language definitions
- improved integration into standard office and business software

The software architecture of Yxilon is shown in Figure 9. Via a graphical or non-graphical user interface the user accesses the system. All objects and information are held in the object database, that communicates with the other parts as the parser or runtime



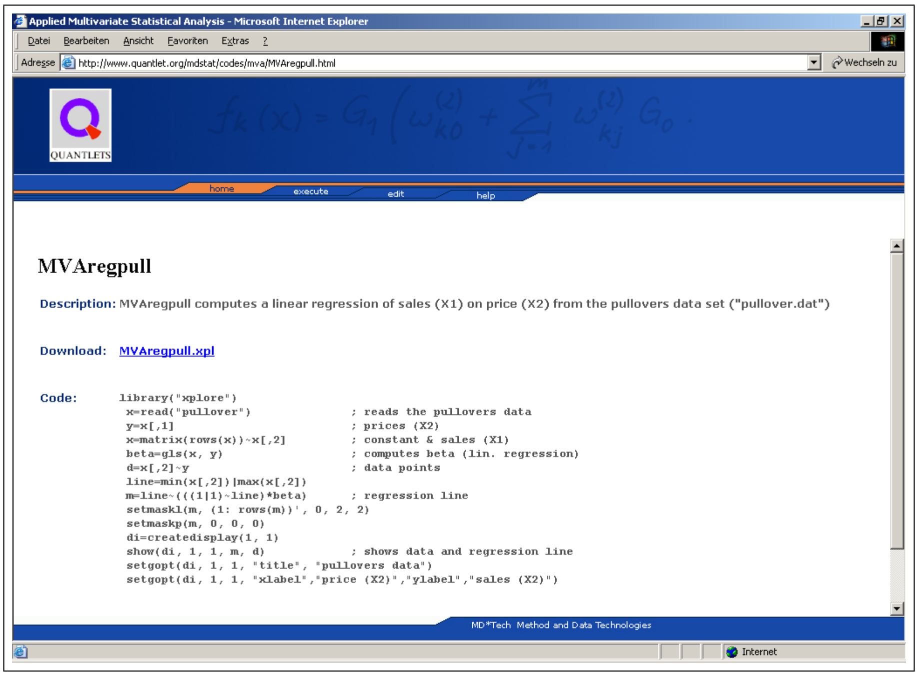
Moving to Higher Dimensions 3-49

Figure 34. Regression of sales (X_1) on price (X_2) of pullovers, $\widehat{\beta}_0 = 210.7$, $\widehat{\beta}_1 = -0.36$. MVAregpull.xpl

MVA: Humboldt-Universität zu Berlin

Figure 6: Applied Multivariate Statistics: slide with link to an example

environment. The parser takes the sourcecode provided by the user and converts them to either Java/C++ code or a binary format directly readable by the runtime engine. The advantage of source conversion is that the high-level, interpreted XploRe code, which is slower than binary code, can be compiled by the Java or C++ compiler to fast running machine code. Especially for computing-intensive applications as bootstrapping and simulation we expect significant improvements concerning execution time.



Applied Multivariate Statistical Analysis - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Adresse http://www.quantlet.org/mdstat/codes/mva/MVAregpull.html
Wechseln zu
QUANTLETS
home
execute
edit
help

MVAregpull

Description: MVAregpull computes a linear regression of sales (X1) on price (X2) from the pullovers data set ("pullover.dat")

Download: MVAregpull.xpl

Code:

```
library("xplore")
 x=read("pullover")               ; reads the pullovers data
 y=x[,1]                          ; prices (X2)
 x=matrix(rows(x))~x[,2]          ; constant & sales (X1)
 beta=gls(x, y)                   ; computes beta (lin. regression)
 d=x[,2]~y                        ; data points
 line=min(x[,2])|max(x[,2])
 m=line~(((1|1)~line)*beta)       ; regression line
 setmaskl(m, (1: rows(m))', 0, 2, 2)
 setmaskp(m, 0, 0, 0)
 di=createdisplay(1, 1)
 show(di, 1, 1, m, d)             ; shows data and regression line
 setgopt(di, 1, 1, "title", "pullovers data")
 setgopt(di, 1, 1, "xlabel","price (X2)","ylabel","sales (X2)")
```

Figure 7: Applied Multivariate Statistics: homepage of the linked example

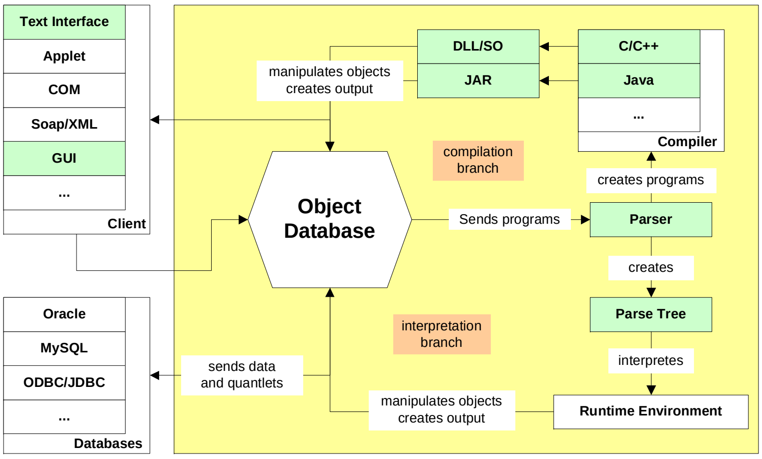
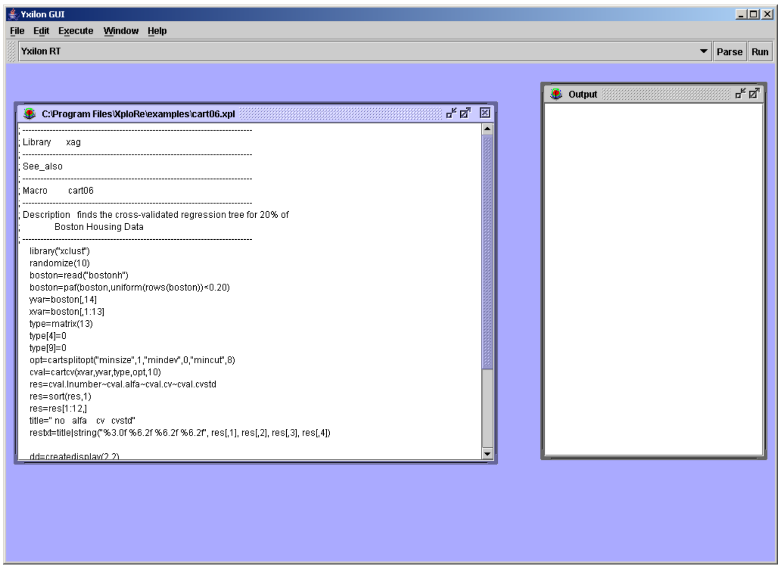


Figure 8: Architecture of the Yxilon Framework (green components in development)


Yxilon GUI
File Edit Execute Window Help
Yxilon RT
Parse
Run
C:\Program Files\XploRe\examples\cart06.xpl
Output

```
; -----------------------------------------------------------------------
; Library      xag
; -----------------------------------------------------------------------
; See_also
; -----------------------------------------------------------------------
; Macro        cart06
; -----------------------------------------------------------------------
; Description   finds the cross-validated regression tree for 20% of
;               Boston Housing Data
; -----------------------------------------------------------------------
   library("xclust")
   randomize(10)
   boston=read("bostonh")
   boston=paf(boston,uniform(rows(boston))<0.20)
   yvar=boston[,14]
   xvar=boston[,1:13]
   type=matrix(13)
   type[4]=0
   type[9]=0
   opt=cartsplitopt("minsize",1,"mindev",0,"mincut",8)
   cval=cartcv(xvar,yvar,type,opt,10)
   res=cval.lnumber~cval.alfa~cval.cv~cval.cvstd
   res=sort(res,1)
   res=res[1:12,]
   title=" no   alfa    cv   cvstd"
   restxt=title|string("%3.0f %6.2f %6.2f %6.2f", res[,1], res[,2], res[,3], res[,4])

   dd=createdisplay(2,2)
```

Figure 9: Screenshot of the Yxilon graphical user interface

4 Limits of e-lements in statistics education

Electronic and computational elements in education can be limited from a variety of different reasons as the technical implementation or the educational approach.

MM*Stat relies on a complex framework of HTML, Cascading Style Sheets, Java and JavaScript. While HTML and Java were already well-established standards, JavaScript and Cascading Style Sheets were implemented in different ways for each browser, so extensive work for each major webbrowser had to be invested.

From the educational point of view each use of electronic media should be questioned carefully. An analysis by Brandes, U. (2004) showed no significant improvements in the results of those students who used MM*Stat in comparison to those who used classical textbooks. The educational limits also lie in the way the information is presented. The classical way of using a textbook or blackboards also offers the opportunity to include on-the-fly notes from either the teacher or the student. Electronic teaching solutions either do not offer this capability at all or require significant technical ressources.

References

Borak, S., Härdle, W., Lehmann, H. (2005), Working with the XQC, IN: Statistical Tools for Finance and Insurance, editors: Cizek, P., Härdle, W., Weron, R., Springer Verlag

Brandes, U. (2004). Statistische Bewertung und Analyse der Klausurergebnisse Statistik

(Statistical analysis of exam results in statistics undergraduate courses), diploma thesis, Humboldt-Universität zu Berlin

Chambers, J. and Lang, D. T. (1999). Omegahat – A Component-based Statistical Computing Environment, Proceedings of the 52nd Session of the ISI, Helsinki, Finland.

Fujiwara, T., Ikunori K., Nakano, J., Yoshikazu, Y. (2000). A Statistical Package Based on Pnuts, In: COMPSTAT. Proceedings in Computational Statistics, Physica Verlag

Guril, Y., Klinke, S., Ziegenhagen, U. (2005). Yxilon – a Modular Open-Source Statistical Programming Language, In: Proceedings of the 55th Session of the International Statistical Institute (ISI), Sydney, Australia

Mori, Y., Yamamoto, Y. and Yadohisa, H. (2003). Data-oriented Learning System of Statistics based on Analysis Scenario/Story (DoLStat). Bulletin of the International Statistical Institute (ISI).

Müller, M., Rönz, B., Ziegenhagen, U.(2000). The Multimedia Project MM*Stat for Teaching Statistics, In: COMPSTAT. Proceedings in Computational Statistics, Physica Verlag