

Integrable e-lements for Statistics Education

Wolfgang Härdle

Sigbert Klinke

Uwe Ziegenhagen

Institute für Statistics and Econometrics

Humboldt-Universität zu Berlin

<http://ise.wiwi.hu-berlin.de>

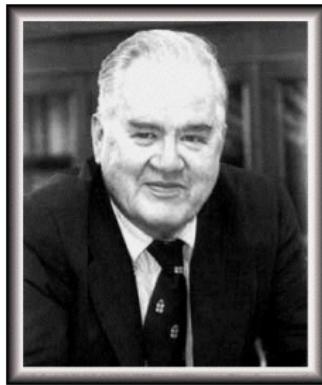
<http://www.case.hu-berlin.de>



CENTER FOR APPLIED STATISTICS AND ECONOMICS

"Each new generation of computers offers us new possibilities, at a time when we are far from using most of the possibilities offered by those already obsolete."

John W. Tukey (1965)



e-lements in Statistics Education

Modern education in statistics must involve practical computer-based data analysis.

Questions

- Which elements do re-occur during different courses?
- Which technology can be presented during class, which not?
- High-level code at the beginning of the studies?
- Where are the limits of e-lements in statistics education?



Outline

- Introduction ✓
- Statistics courses
- MM*Stat and e-stat
- Electronic books
- XploRe and Yxilon
- Limits of e-lements



CASE Courses

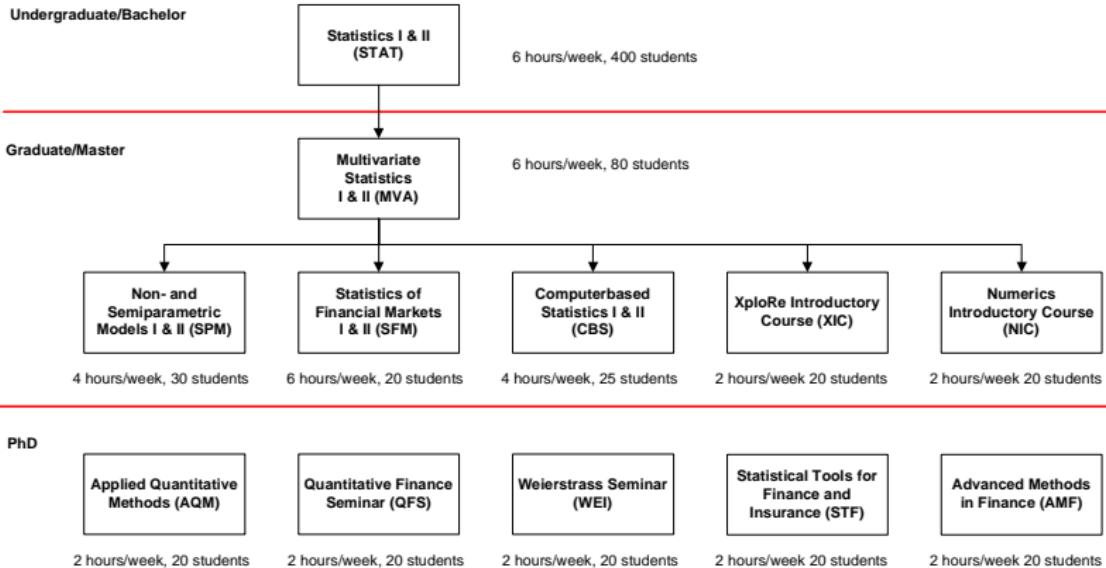


Students with different backgrounds are taught at ISE:

- German Business Administration and Economics
- BA/MA courses in Economics and Statistics
- Students from math and other science departments



Layout of Studies



Statistics I & II

- Basic probability theory
- Random variables
- Discrete and continuous distributions
- Point and interval estimation
- OLS-Regression
- Analysis of timeseries



Multivariate Statistics I & II

- Histograms and kernel density estimation
- Matrix algebra and multivariate distribution
- Principal component and discriminant analysis
- Cluster analysis and multidimensional scaling
- Factor analysis and projection pursuit



Statistics of Financial Markets I & II

- Options and derivatives
- Black-Scholes model
- Exotic options
- Financial time series
- Value at Risk and copulae



Applied Quantitative Methods

- Analysis and interpretation of multivariate data
- Generalized linear models
- Statistical process control and trend detection
- Computer intrusion detection by statistical means
- Architecture of internet search engines



PDF Slides for Undergraduate Studies

Residual Sum of Squares (RSS)

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min \quad | \quad \hat{y}_i = b_0 + b_1 x_i$$

$$S(b_0, b_1) = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 \rightarrow \min_{b_0, b_1}$$

$$\frac{\partial S(b_0, b_1)}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) \doteq 0$$

$$\frac{\partial S(b_0, b_1)}{\partial b_1} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i \doteq 0$$

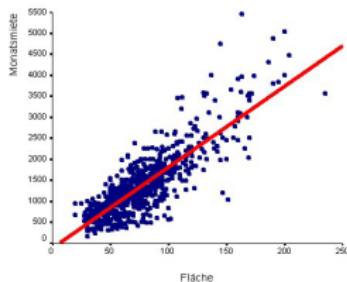


PDF Slides for Undergraduate Studies

Regressionsanalyse

8-16

lineare Regressionsfunktion: $\hat{y}_i = b_0 + b_1 x_i$



- 815 Berlin flats
- X: m^2
- Y: 1m rent

Statistik



MM*Stat

- support studies at undergraduate level
- HTML-based 'filing cards'
- embedded JavaScript and Java applets
- published by Springer and MHSG (<http://www.mhsg.de>)



MM*Stat Translations



German



English



French



Spanish



Italian



Czech



Polish



Indonesian



Japanese



Chinese



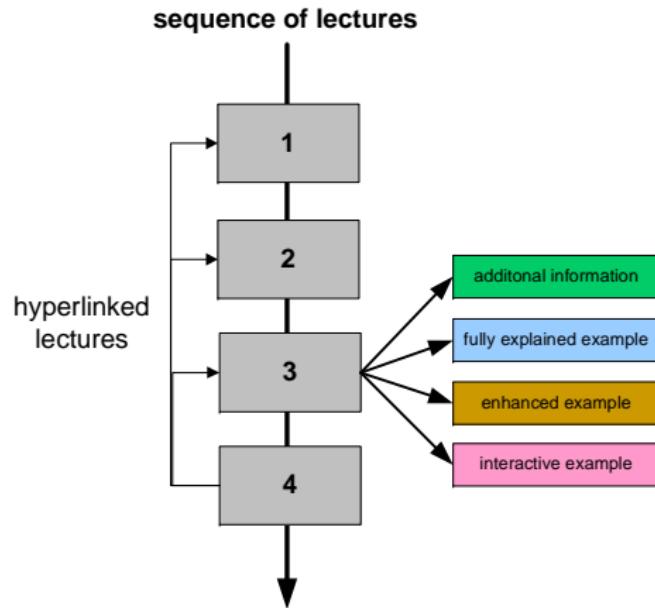
Portuguese



Dutch



Different Layers of Learning



MM*Stat Lecture

Statistics - Scientific data analysis made easy - Microsoft Internet Explorer

Lecture contents Lecture 11.2

11.2 One-Dimensional Regression Analysis

One-dimensional linear regression function

A simple linear **regression function** has the following form:

$$E(y_i|x_i) = b_0 + b_1 x_i \quad i = 1, \dots, n$$

In this equation, x_i represents the observed values of a random variable X (fixed) and b_0 and b_1 are unknown regression parameters.

The actual observed values y_i ($i = 1, \dots, n$) can be obtained by summing residual u_i and $E(y_i|x_i)$ (as you can see on the picture):

$$y_i = E(y_i|x_i) + u_i = b_0 + b_1 x_i + u_i \quad i = 1, \dots, n$$

Y

$y - \hat{y} = \hat{u}$

$\hat{y} = b_0 + b_1 x$

contents explained enhanced enhanced interactive



MM*Stat Explained

Statistics - Scientific data analysis made easy - Microsoft Internet Explorer

Lecture contents Lecture 11.2 explained 11.2

Example for one-dimensional linear regression

Now, we examine the monthly net income and monthly expenditures on living of 10 two-person households.

Household	1	2	3	4	5	6	7	8	9	10
Net income in DM x_1	3,500	5,000	4,300	6,100	1,000	4,800	2,900	2,400	5,600	4,100
Expenditures in DM y_1	2,000	3,500	3,100	3,900	900	3,000	2,100	1,900	2,900	2,100

These observations are drawn in the following scatterplot. You can see that the net income of a household has a positive influence of the household's expenditures and that this dependence can be estimated by means of a linear regression function.

The scatterplot shows the relationship between net income (x_1) and expenditures (y_1). The x-axis is labeled "net income" and the y-axis is labeled "expenditures". The data points show a clear positive correlation, and a straight line is drawn through them, representing the linear regression model.

Lecture contents Lecture 11.2 explained 11.2



MM*Stat Enhanced

Statistics - Scientific data analysis made easy - Microsoft Internet Explorer

Lecture contents Lecture 11.2 explained enhanced

The following measures were collected for 74 different types of cars:

X_1 - price
 X_2 - mpg (miles per gallon)
 X_3 - headroom (in inches)
 X_4 - rear seat clearance (distance from front seat back to the rear seat,in inches)
 X_5 - trunk space (in cubic feet)
 X_6 - weight (in pound)
 X_7 - length (in inches)
 X_8 - turning diameter (clearance required to make a U-turn, in feet)
 X_9 - displacement (in cubic inches)

The dependence of **turning diameter** (X_8) on the **length** (X_7) of a car can be depicted in a **scatterplot**. Every car is represented in the diagram by a single point (X_7, X_8). Moreover, an estimated **regression line** is added in the picture (it is drawn in black).

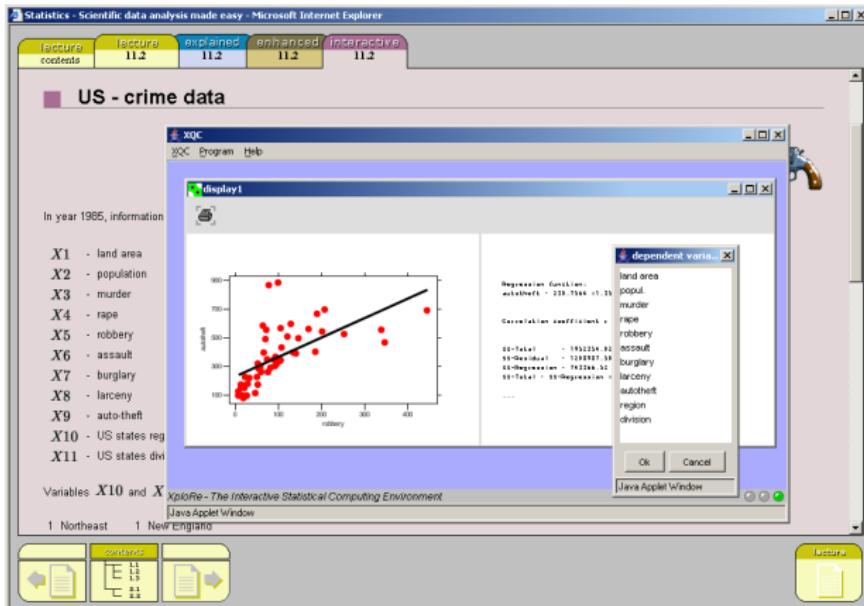
Scatterplot showing the relationship between turning diameter (X8) and length (X7). The x-axis ranges from 100 to 200 inches, and the y-axis ranges from 10 to 80 feet. A positive linear regression line is drawn through the data points. The plot includes the following statistics:

- regression function: turn-diam = 7.1739 + 0.1735 * lenght
- correlation coefficient: r = 0.90
- SS-Total = 1361.96
- SS-Residual = 259.06

Lecture Students E-mail Help



MM*Stat Interactive



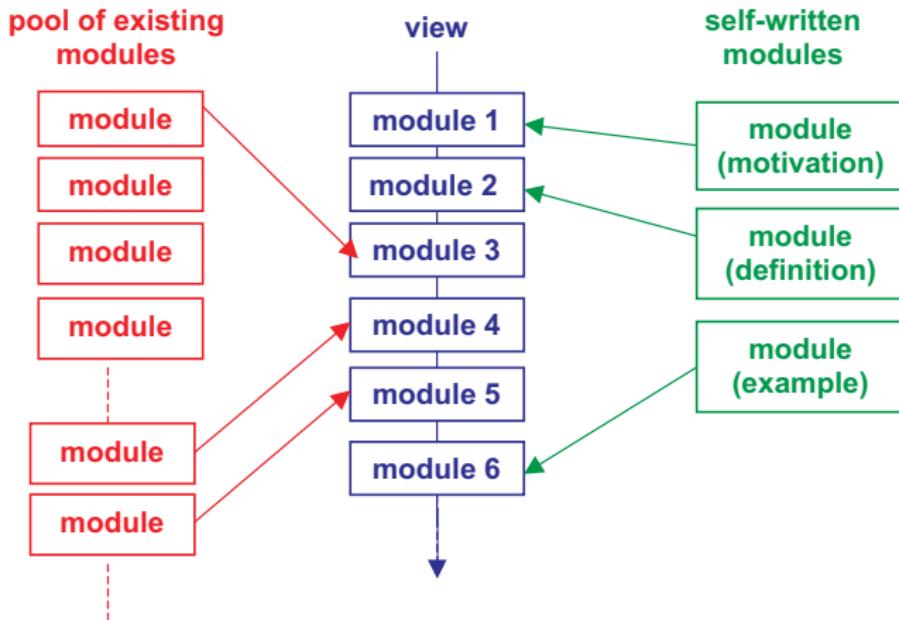
e-stat

- Developed by team of seven German universities
- Funded by BMBF
- XML-based
- Statistical content broken into small modules
- Example: regression analysis
 1. actual motivation
 2. explanation of general purpose
 3. specification of regression model
 4. listing of properties
 5. estimation techniques

e.stat



e-stat structure



e-stat example

EMILeA-stat - Microsoft Internet Explorer

Stöbern | Bearbeiten | Ansicht | Elemente | Extras | ? Hilfe

Stöbern in EMILeA-stat

Inhaltsverzeichnis

- EMILeA-stat Modulwelt
- Amtliche Statistik
- Assoziation
- Beschreibende Statistik
- Entropie
- Explorative Datenanalyse
- Finanzmathematik
- Lineare Strukturgleichungen
- Machine Learning
- Mathematische Grundlagen
- Methodenkritische Begleitung zu PISA 2000
- Numerische Methoden
- Qualitätsoptimierung
- Robuste Statistik
- Schließende Statistik
- Sequenzielle Methoden
- Statistik der Finanzmärkte
- Stochastik in der Schule
- Stochastische Prozesse
- Veralgemeinerte lineare Modelle
- Versicherungsmathematik
- Wahrscheinlichkeitsrechnung
- Wirtschafts- und Bevölkerungsstatistik
- Zeitreihenanalyse

Beschreibende Statistik > Regressionsanalyse > Regression > lineare Regression

Motivation | Bezeichnung | Bemerkung (Anwendung bei nichtlinearen Zusammenhängen) | Level A | Level B | Level C

Motivation zur linearen Regression

In der Praxis treten häufig Fragestellungen auf, bei denen die Abhängigkeitsstruktur zweier metrischer Merkmale X und Y untersucht werden soll. Meistens kann bereits aufgrund der jeweiligen Situation davon ausgegangen werden, dass das Merkmal X in einer bestimmten Weise auf das Merkmal Y einwirkt.

Wenn Überlegungen einen linearen Zusammenhang zwischen beiden Merkmalen nahe legen (d. h. es wird angenommen, dass $a, b \in \mathbb{R}$ mit $Y = a + bX$ existieren), können auf der Basis eines Datensatzes im Rahmen eines **linearen Regressionsmodells**

$$Y = f(X) + \varepsilon = a + bX + \varepsilon$$

plausible Schätzwerte für die beiden Parameter $a, b \in \mathbb{R}$ mittels der **Methode der kleinsten Quadrate** ermittelt werden. Der auf diese Weise geschätzte Zusammenhang kann dann z. B. für Prognosezwecke verwendet werden.



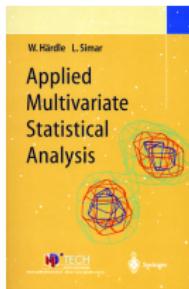
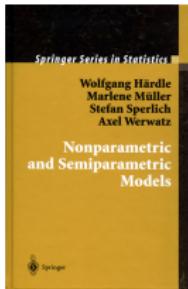
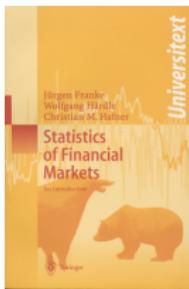
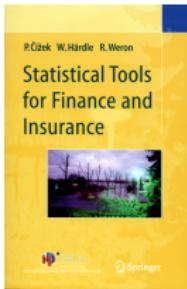
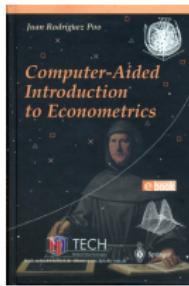
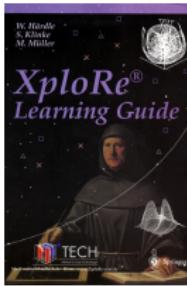

MD*Book Architecture

Aim:

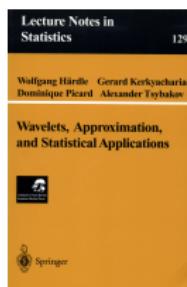
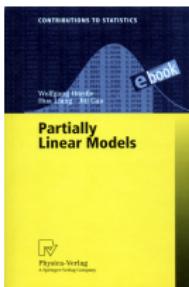
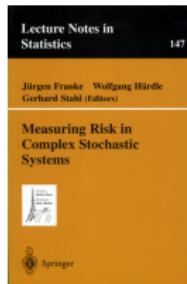
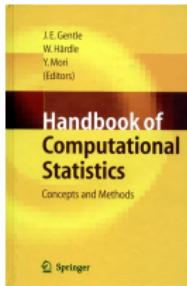
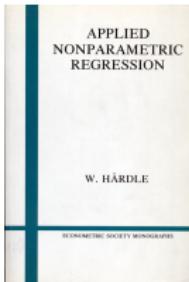
- Generate different formats from one L^AT_EX source
- PDF, PS, HTML, MD*booklet
- Add interactive examples, visualizing the theory
- for printed books download versions available
(incl. XploRe Client/Server)



Printed and Electronic Books



Printed and Electronic Books



Applied Multivariate Analysis - Slide

Moving to Higher Dimensions

3 - 49

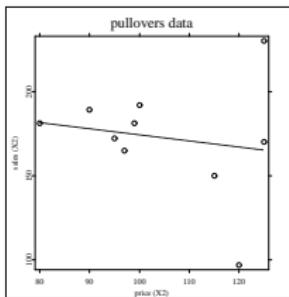


Figure 34. Regression of sales (X_1) on price (X_2) of pullovers,

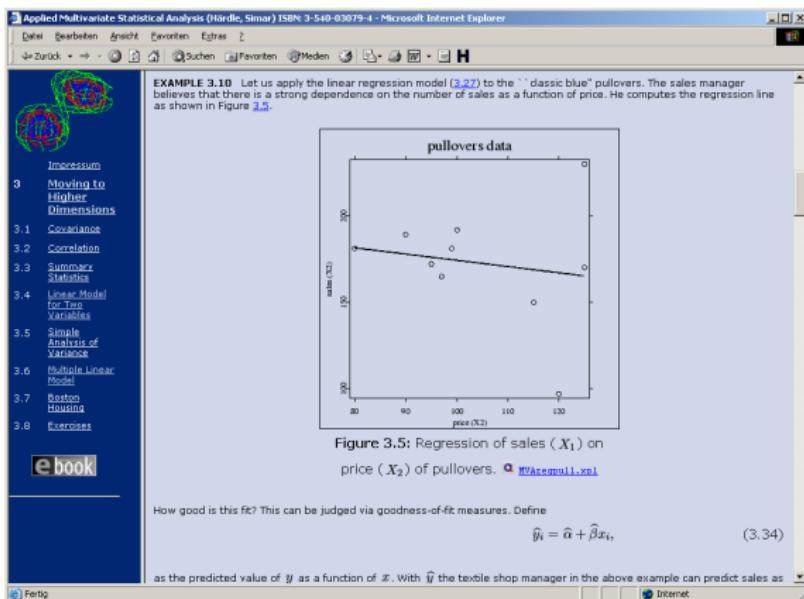
$$\hat{\beta}_0 = 210.7, \hat{\beta}_1 = -0.36.$$

 [MVAregpull.xpl](#)

MVA: Humboldt-Universität zu Berlin



Applied Multivariate Analysis - Online



HTML-page for Electronic Examples

Screenshot of a Microsoft Internet Explorer window displaying an HTML page for electronic examples. The page title is "Applied Multivariate Statistical Analysis - Microsoft Internet Explorer". The URL in the address bar is "http://www.quantlet.org/nldstat/codes/mvaregpull.html". The page content includes:

QUANTILET

$$f_k(x) = G_1 \left(\omega_{k0}^{(1)} + \sum_{j=1}^J \omega_{kj}^{(1)} G_0 \right)$$

MVAreppull

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

Download: [MVAreppull.xls](#)

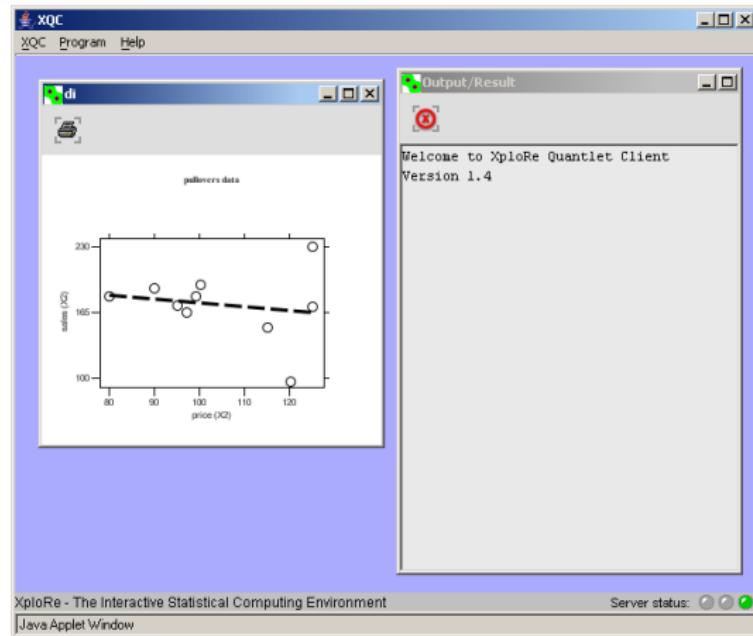
Code:

```
library("xplor4")
x<-read("pullovers")
y<-x[,1]                                : prices (X2)
x<-matrix(rbind(x)-x[,2])                : constant & sales (X1)
beta<-ols(x, y)                          : computes beta (lin. regression)
d<-x[,2]-1                                : data points
l1<-x[,1]+(x[,2]-1)*max(x[,2])
m<-l1-(1/(1+1)-l1)*beta                 : regression line
setmask1(m, 1, rows(m), 0, 2, 2)
setmask2(m, 0, 0, 0)
di<-createDisplay(1, 1)
show(di, 1, 1, m, d)                      : shows data and regression line
setopt(di, 1, 1, "title", "pullovers data")
setopt(di, 1, 1, " xlabel", "price (X2)", " ylabel", "sales (X2)")
```

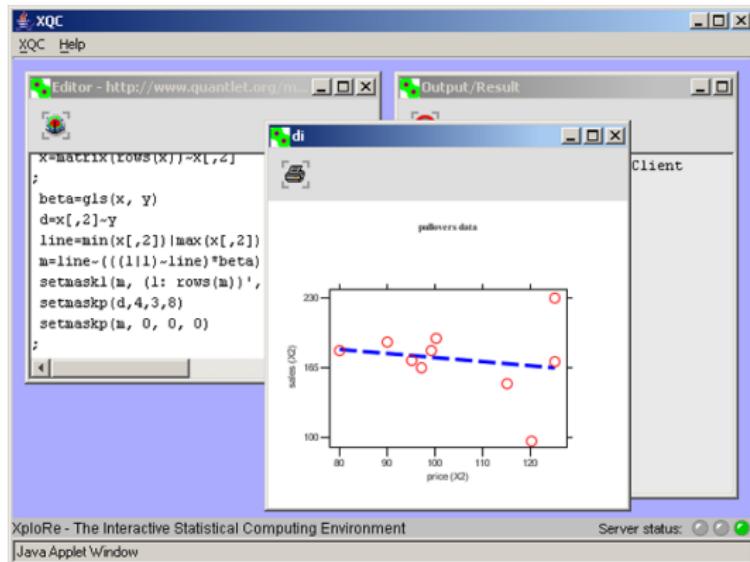
At the bottom of the page, there is a footer with the text "HD*Tech Method and Data Technologies" and a small logo of a blue robot holding a book.



Interactive Example for MVA - Run



Interactive Example for MVA - Edit



XploRe

- developed by Humboldt-Universität zu Berlin and MD*Tech
- C-style syntax, procedural approach
- available as batch, standalone and Client/Server on Win32 and Unix/Linux
- strong focus on non-parametric and quantitative finance methods

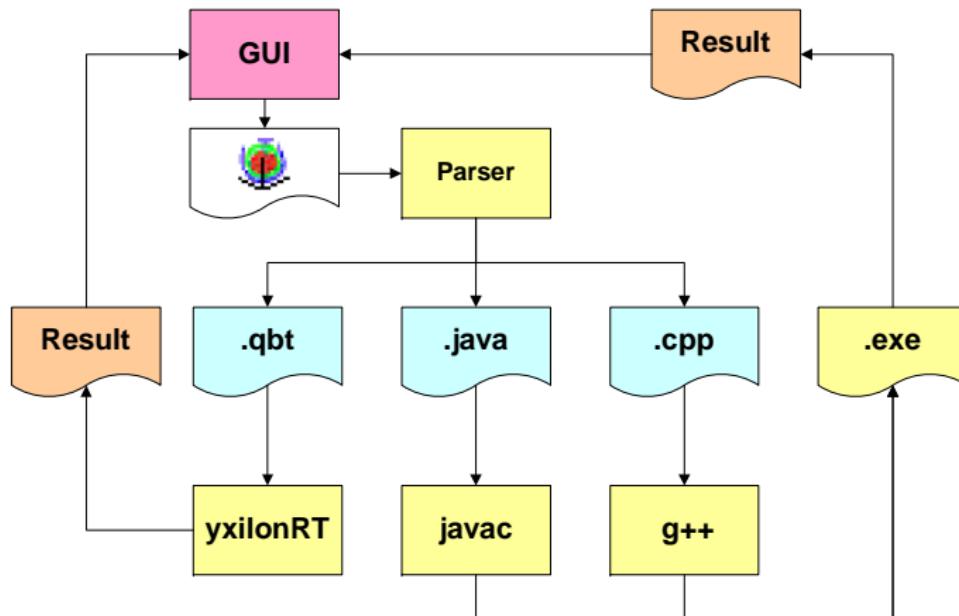


Yxilon

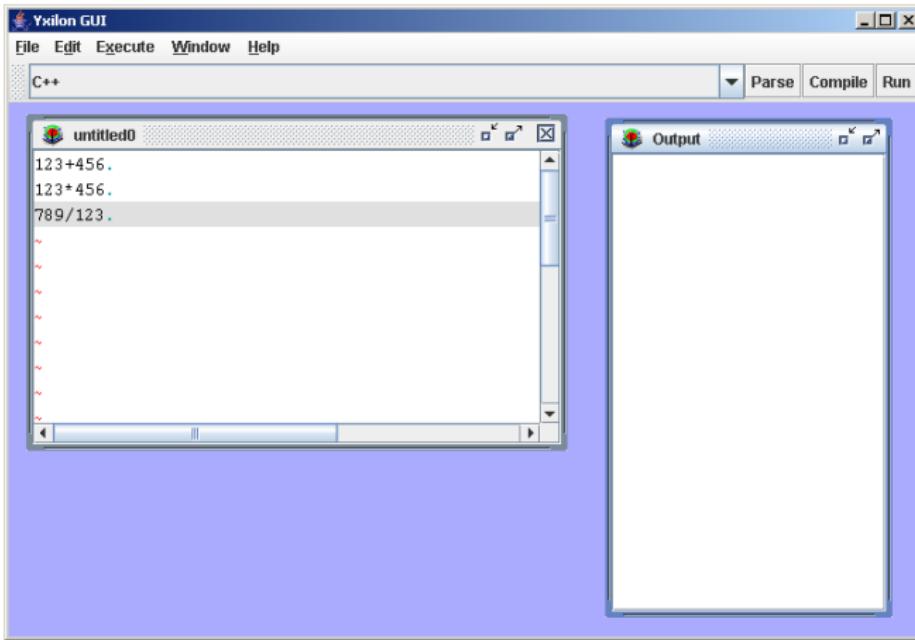
- Multiple front-ends: commandline, GUI, embedded into Excel
- Extensibility on language and native core level
- Read/write data and to calculate across networks
- Support for databases and interactive graphics
- Inclusion of existing code (C, Fortran, XploRe)



Compilation



Yxilon Screenshot



Technical Limitations

- e-tools need software architecture
 - ▶ Easy to handle and less powerful?
 - ▶ Powerful and complex
 - ▶ strategic decision, trade-off must be found
- MM*Stat: HTML, CSS and JavaScript are browser-dependent
- JAVA platform-independent, only intersection of functionality



Psychological Barriers

- computer knowledge among students still diverse
- psychological barriers to use e-lements for learning
- 90s: hype to teach online failed
- Do students want to learn online?



Educational Limitations

What cannot be taught via e-lements?

- complex data analysis has several steps
 - ▶ explorative and descriptive analysis
 - ▶ ANOVA or PCA
 - ▶ regression
 - ▶ statistical tests
- single steps are teachable well
- 'big picture' may get lost
- Statistical thinking cannot be taught



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
- Chambers, J. and Lang, D. T. (1999)
 $\hat{\Omega}$ – A Component-based Statistical Computing Environment
Proceedings of the 52nd Session of the ISI, Helsinki
- 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 ISI, Sydney



Mori, Y., Yamamoto, Y. and Yadohisa, H. (2003)

Data-oriented Learning System of Statistics based on Analysis Scenario/Story

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



Tukey, T. (1965)

The Technical Tools of Statistics

American Statistician 19, 23-28.

