

# Integrable e-lements for Statistics Education

Wolfgang Härdle

Sigbert Klinke

Uwe Ziegenhagen

Institute für Statistics and Econometrics

Humboldt-Universität zu Berlin

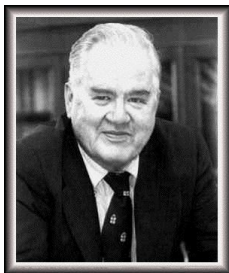
<http://ise.wiwi.hu-berlin.de>

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



*"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-elements 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-elements in statistics education?



# Outline

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





## 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

## Undergraduate/Bachelor

Statistics I & II  
(STAT)

6 hours/week, 400 students

## Graduate/Master

Multivariate  
Statistics  
I & II (MVA)

6 hours/week, 80 students

Non- and  
Semiparametric  
Models I & II (SPM)

4 hours/week, 30 students

Statistics of  
Financial Markets  
I & II (SFM)

6 hours/week, 20 students

Computerbased  
Statistics I & II  
(CBS)

4 hours/week, 25 students

XploRe Introductory  
Course (XIC)

2 hours/week 20 students

Numerics  
Introductory Course  
(NIC)

2 hours/week 20 students

## PhD

Applied Quantitative  
Methods (AQM)

2 hours/week, 20 students

Quantitative Finance  
Seminar (QFS)

2 hours/week, 20 students

Weierstrass Seminar  
(WEI)

2 hours/week, 20 students

Statistical Tools for  
Finance and  
Insurance (STF)

2 hours/week 20 students

Advanced Methods  
in Finance (AMF)

2 hours/week 20 students



## 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

Regressionsanalyse

8-25

## Residual Sum of Squares (RSS)

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

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

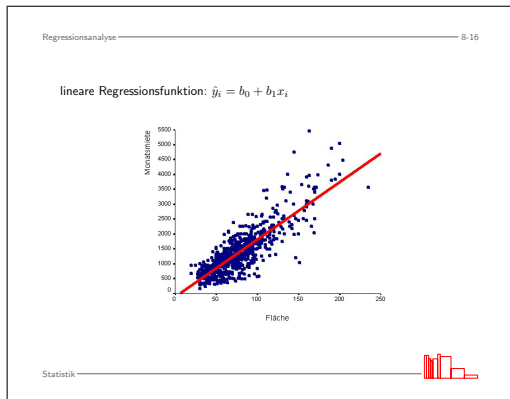
$$\frac{\partial S(b_0, b_1)}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) \stackrel{!}{=} 0$$

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

Statistik



# PDF Slides for Undergraduate Studies



- 815 Berlin flats
- X: m<sup>2</sup>
- Y: 1m rent





## 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



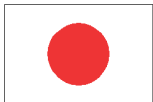
Czech



Polish



Indonesian



Japanese



Chinese



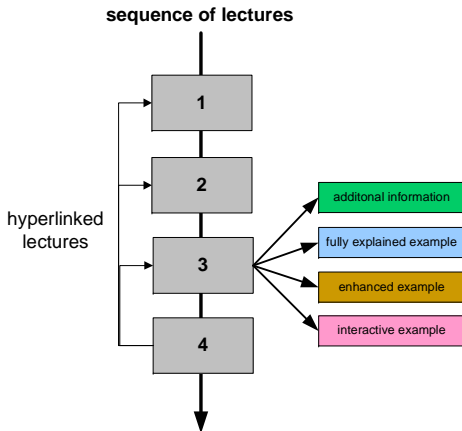
Portuguese



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

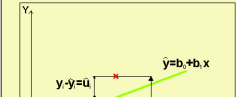
### One-dimensional linear regression function

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):

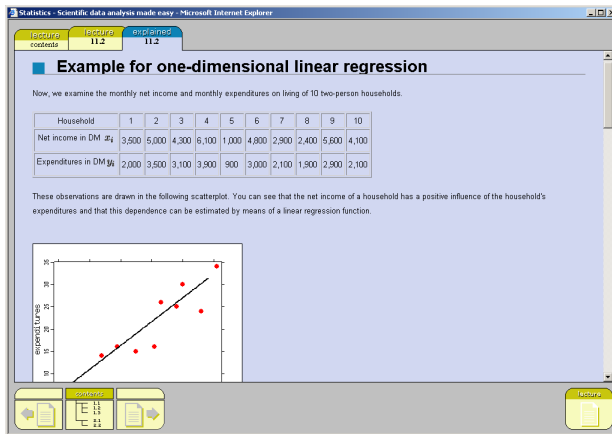
$$y_i = E(y_i|x_i) + u_i = b_0 + b_1 x_i + u_i \quad i = 1, \dots, n$$


Navigation icons: back, forward, search, and other controls.

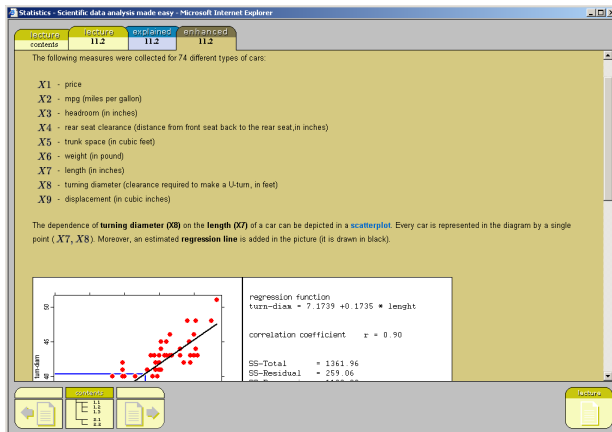
Buttons: explained, enhanced, enhanced, interactive.



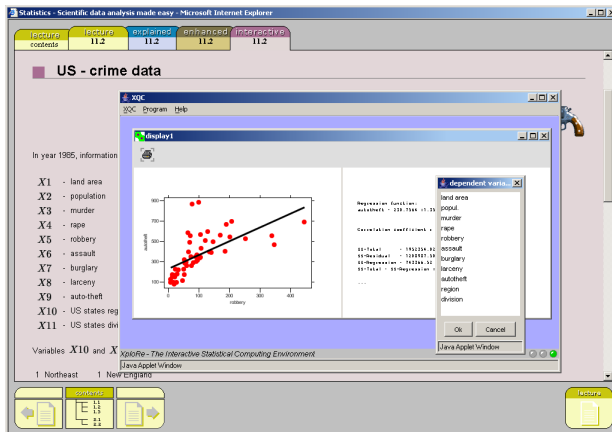
# MM\*Stat Explained



# MM\*Stat Enhanced



# MM\*Stat Interactive



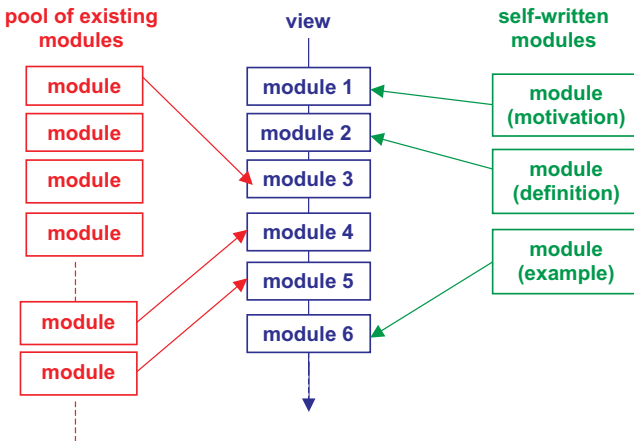
## e-stat

- ▣ 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 structure



# e-stat example

EMILeA-stat - Microsoft Internet Explorer

Stöbern in EMILeA-stat

Inhaltsverzeichnis

- EMILeA-stat Modulkwelt
  - ☐ Amtliche Statistik
  - ☐ 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
  - ☐ Verallgemeinerte lineare Modelle
  - ☐ Versicherungsmathematik
  - ☐ Wahrscheinlichkeitsrechnung
  - ☐ Wirtschafts- und Bevölkerungsstatistik
  - ☐ Zeitreihenanalyse

Beschreibende Statistik > Regressionsanalyse > Regression >

1 lineare Regression

Motivation Bezeichnung Bemerkung (Anwendung bei nichtlinearen Zusammenhängen)

### 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.

XploRe® SPSS Fernanfertigung als PDF



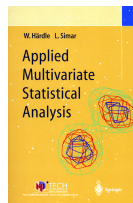
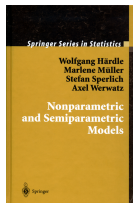
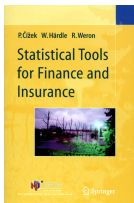
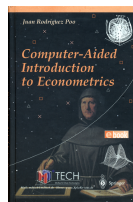
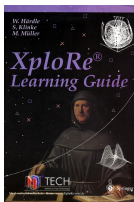
## MD\*Book Architecture

Aim:

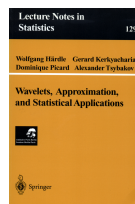
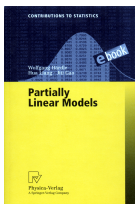
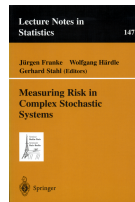
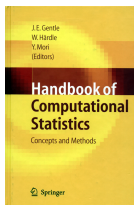
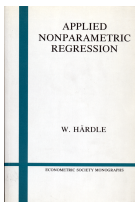
- Generate different formats from one  $\text{\LaTeX}$  source
- PDF, PS, HTML, MD\*booklet
- Add interactive examples, visualizing the theory
- for printed books download versions available  
(incl. XploRe Client/Server)



# Printed and Electronic Books



# Printed and Electronic Books



# Applied Multivariate Analysis - Slide

Moving to Higher Dimensions

3-49

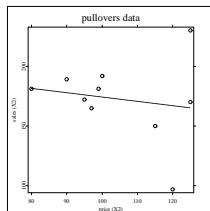


Figure 34. Regression of sales ( $X_1$ ) on price ( $X_2$ ) of pullovers,

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

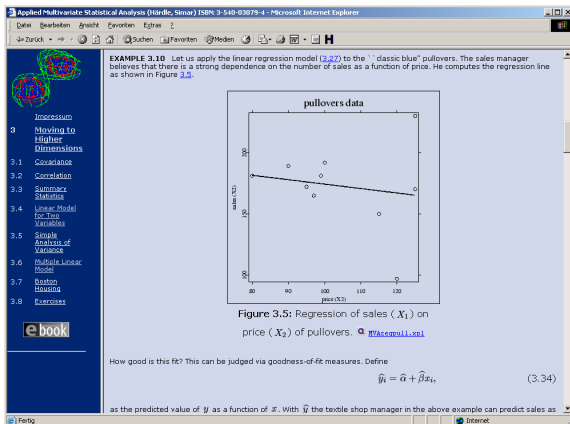


MVAregpull.xpl

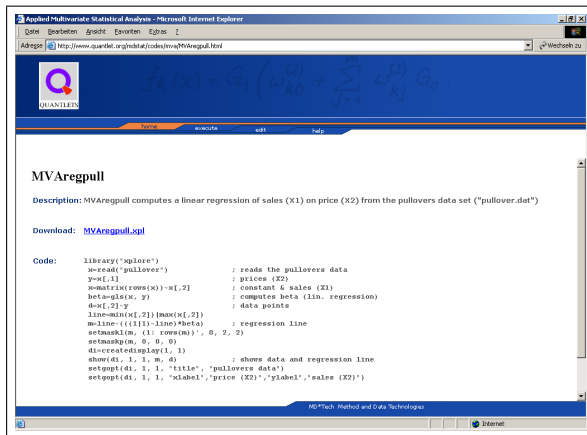
MVA: Humboldt-Universität zu Berlin



# Applied Multivariate Analysis - Online



# HTML-page for Electronic Examples



**Quantiles**

$$f_k(x) = C_k \left( \omega_{k0} + \sum_{j=1}^m \omega_{kj} G_j \right)$$

**MVAregrpull**

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

Download: [MVAregrpull.xpl](#)

Code:

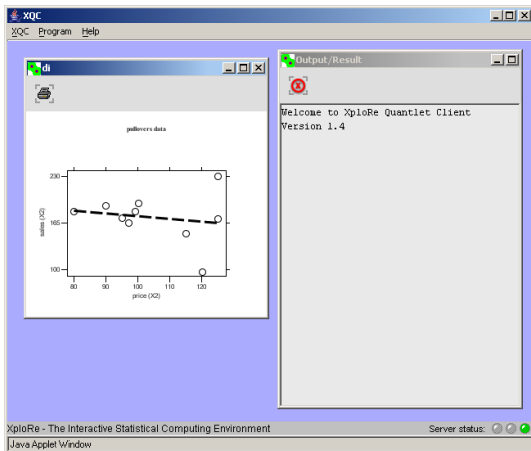
```
library("xplore")
x=read("pullover")      ; reads the pullovers data
y=x[,1]                 ; prices (X2)
n=matrix(rows(x))-x[,2] ; constant & sales (X1)
beta=glm(x, y)           ; computes beta (lin. regression)
d=x[,2]-y                ; data points
line=min(x[,2]) : max(x[,2])
m=line-(((1|1)-line)*beta) ; regression line
setmask(m, (1: rows(m))', 0, 2, 2)
setmask(m, 0, 0, 0)
dis=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 (X1)")
```

HO-Tech Method and Data Technologies

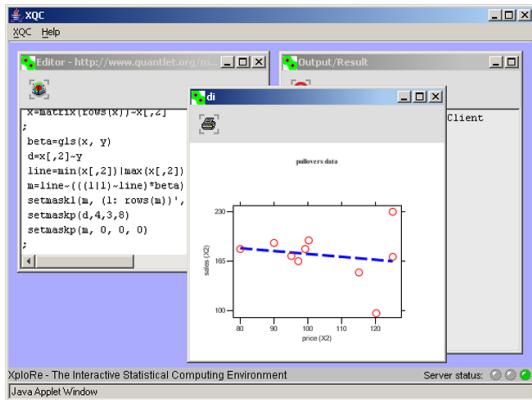




## Interactive Example for MVA - Run



## 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

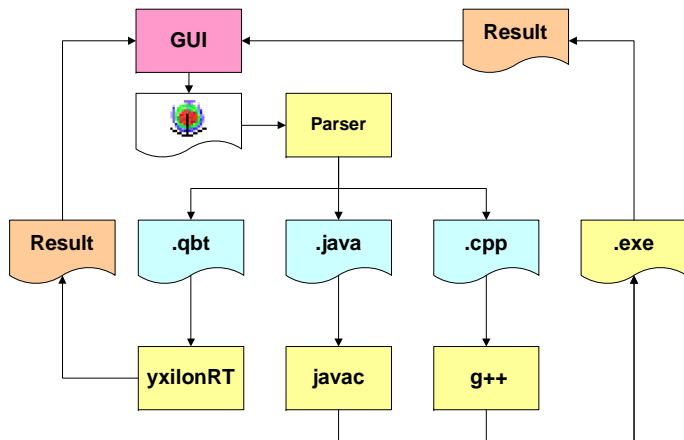


## Yxilon

- ▣ 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



e-lements



## 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.

