COMPSTAT 2010

IASC-ERS, CNAM, Paris, 26. August 2010

» Interactive Graphics Interfacing Statistical Modelling and Data Exploration«

Adalbert Wilhelm

JACOBS
UNIVERSITY

Content

Goal: Improving the links between visual data exploration and statistical modelling

Focus: Regression models

Interfacing visualization and statistical modelling

Interface 1: Visualizing raw data

- Exploratory analysis for model building
  - data patterns
  - subgroups/clusters
  - outliers

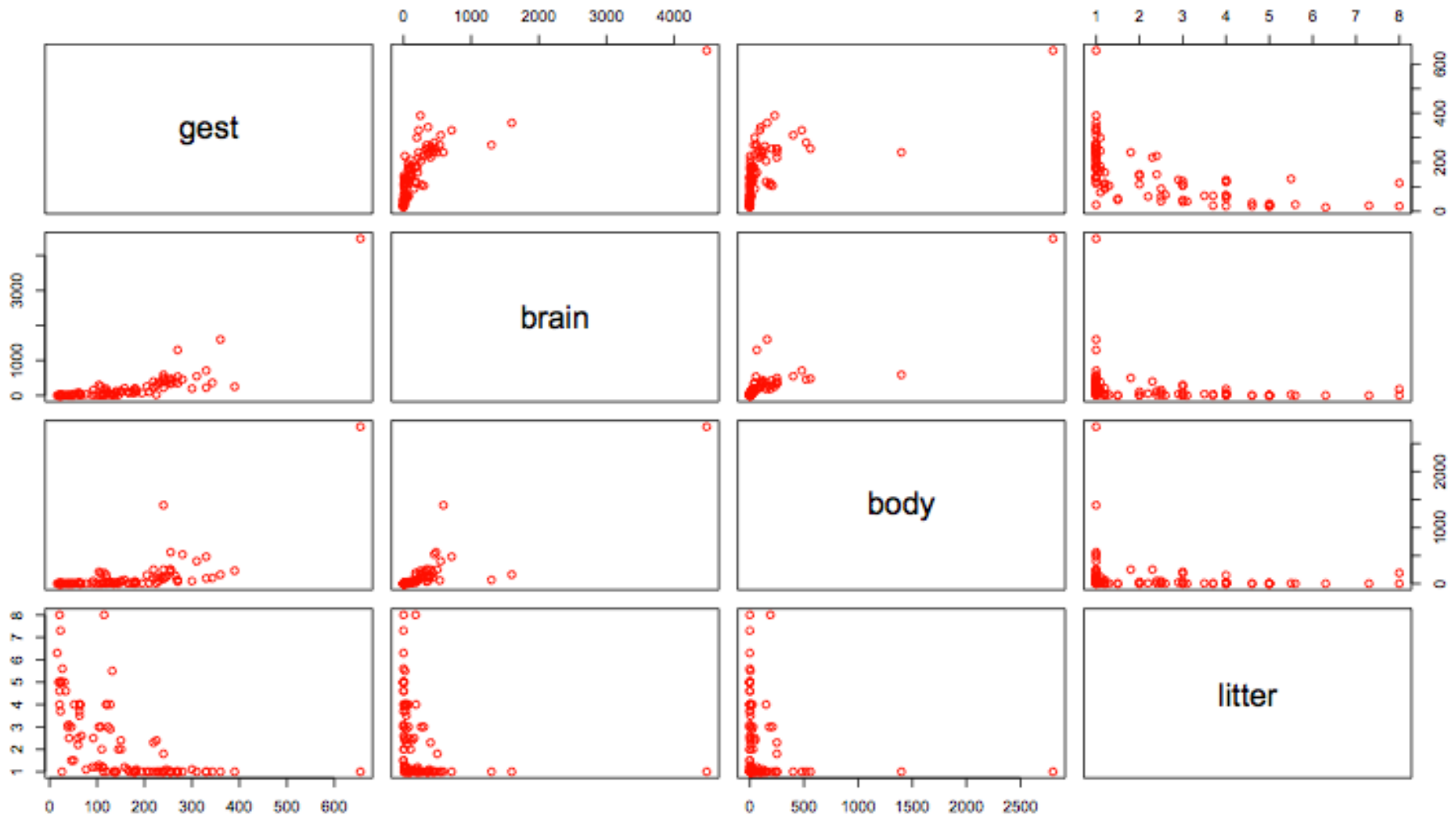
Interface 2: Visualizing raw data and models

- visual goodness of fit, residual diagnostics

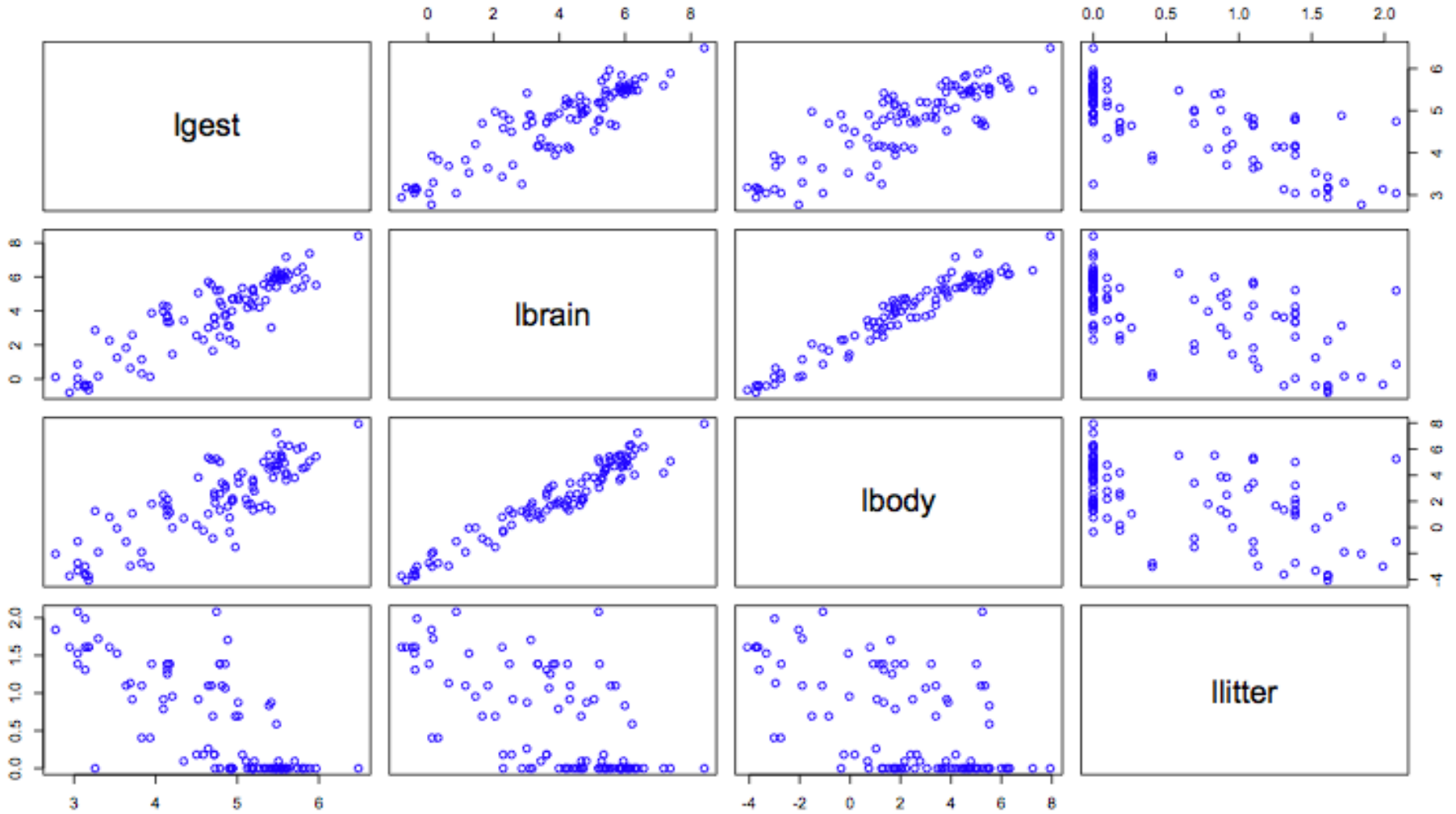
Interface 3: Visualizing model parameters

- model comparisons and model interpretation

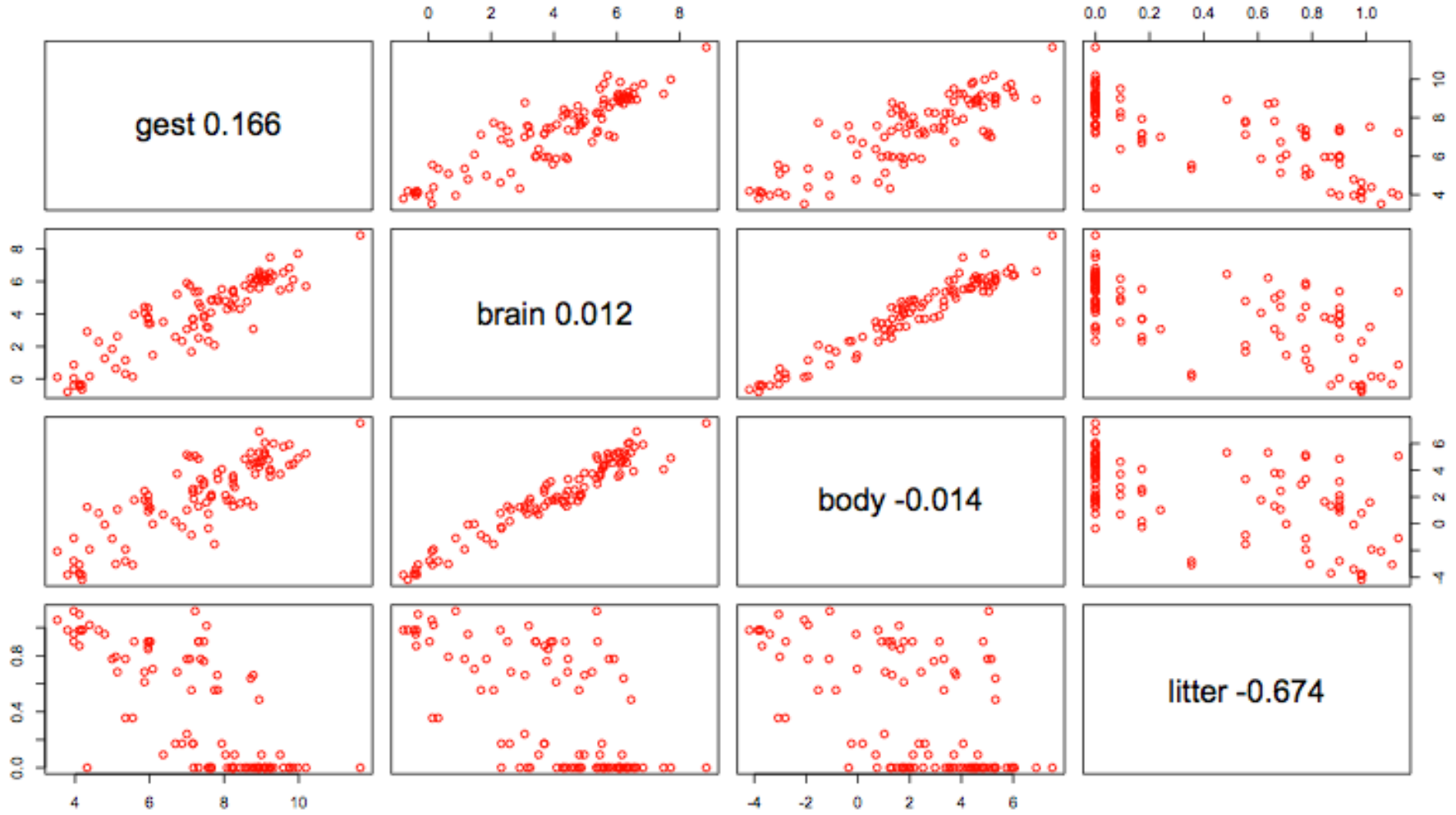
Interface 1: Visualizing raw data



Looking for structure, outliers and transformations

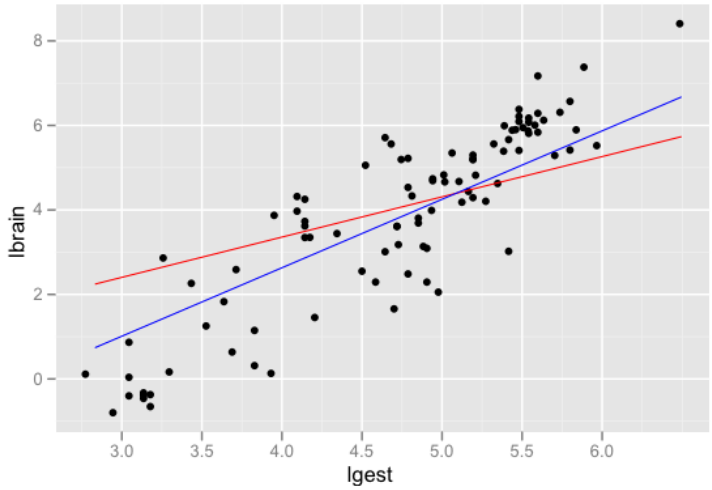


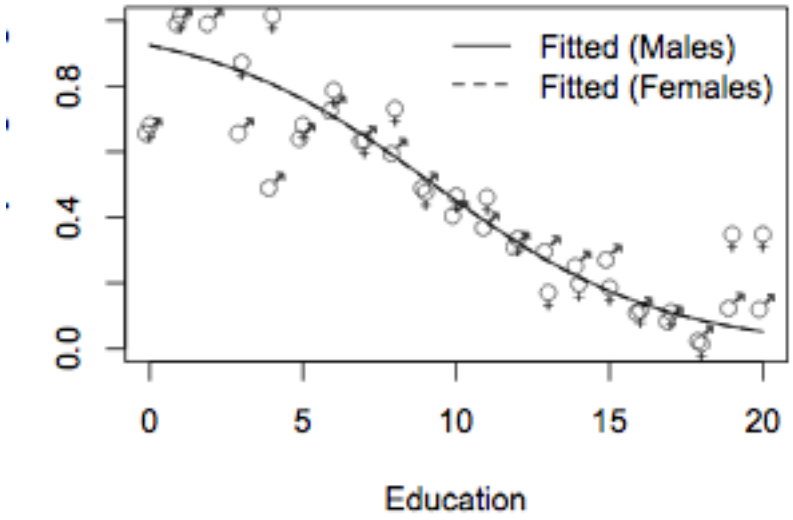
Looking for structure, outliers and transformations

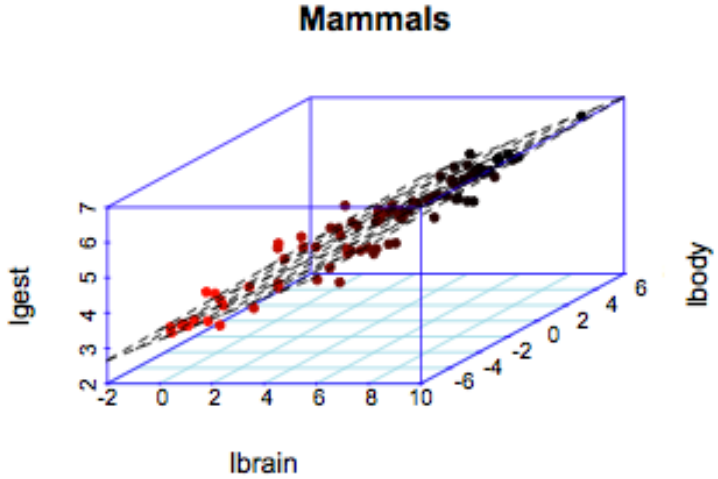


Interface 2: Visualizing raw data and models

Graphical representation of linear and generalized linear models

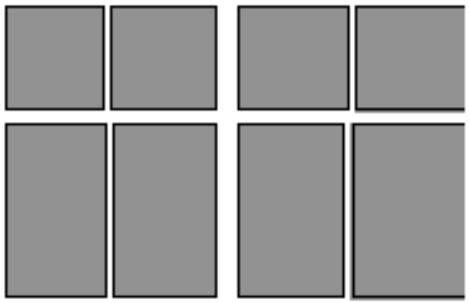




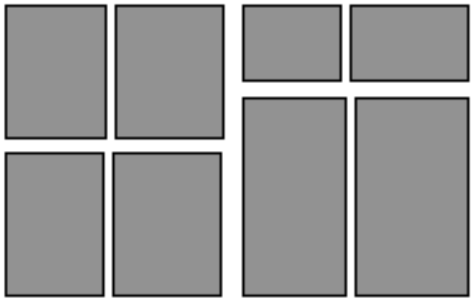


Interface 2: Visualizing raw data and models

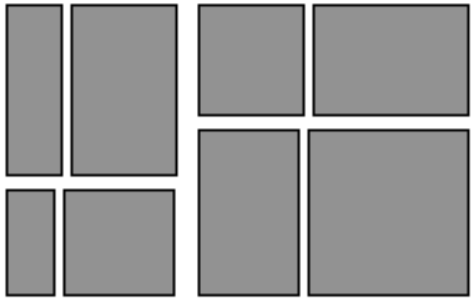
3-dim.
Independence



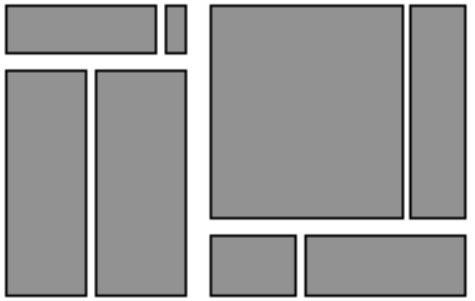
Partial Independence



Conditional Independence

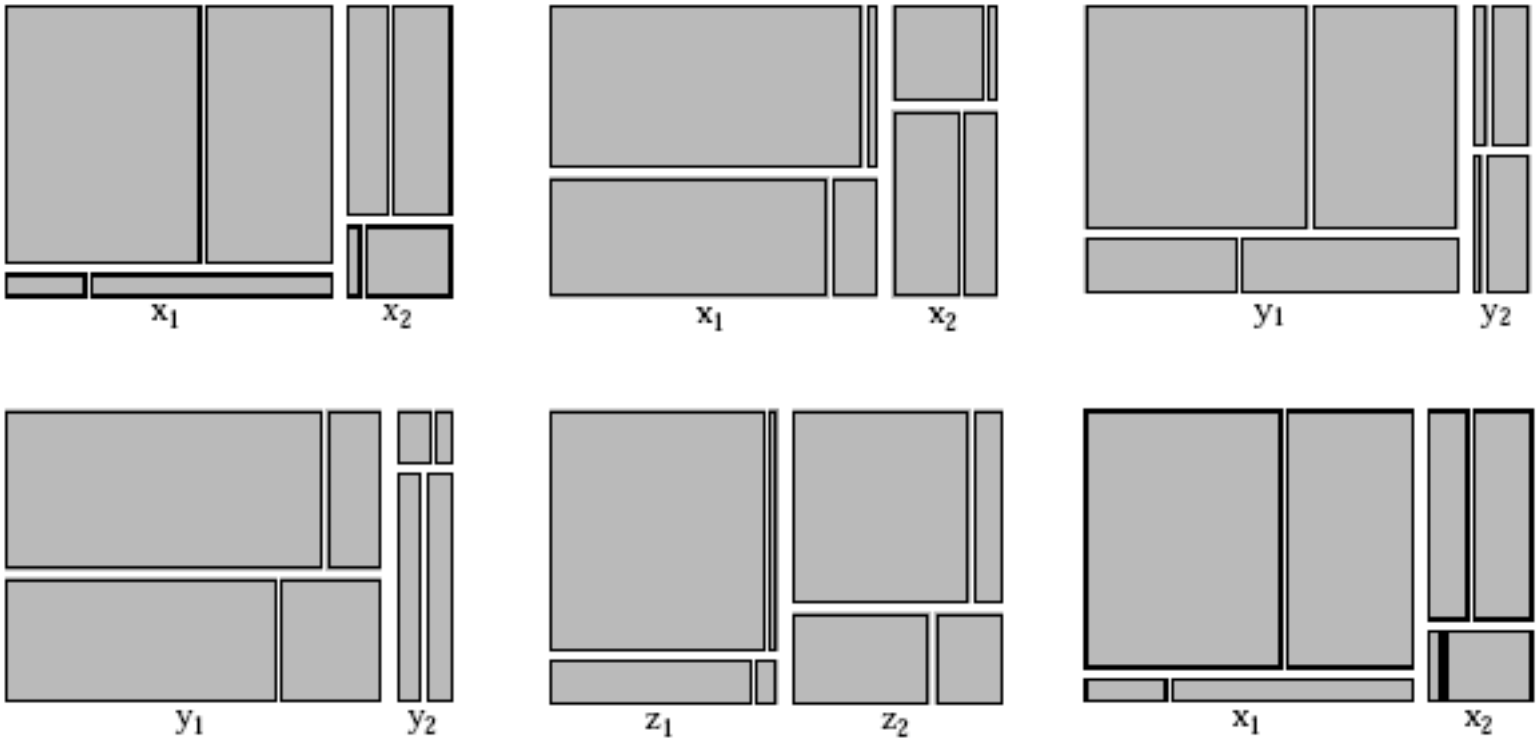


no 3way Interaction



Theus & W. 1996
Theus & Lauer 1998

Interface 2: Visualizing raw data and models

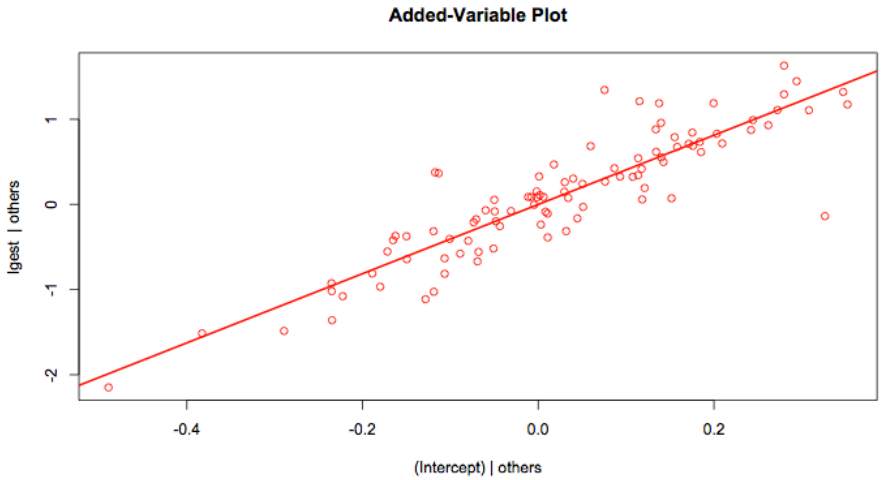


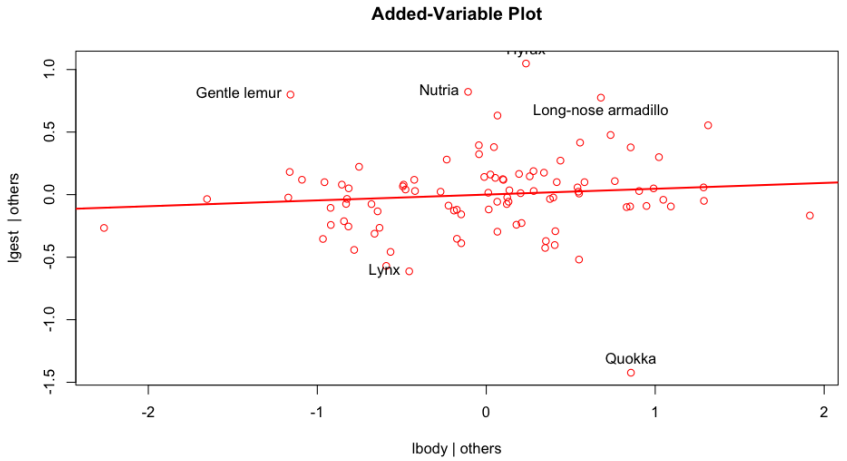
$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

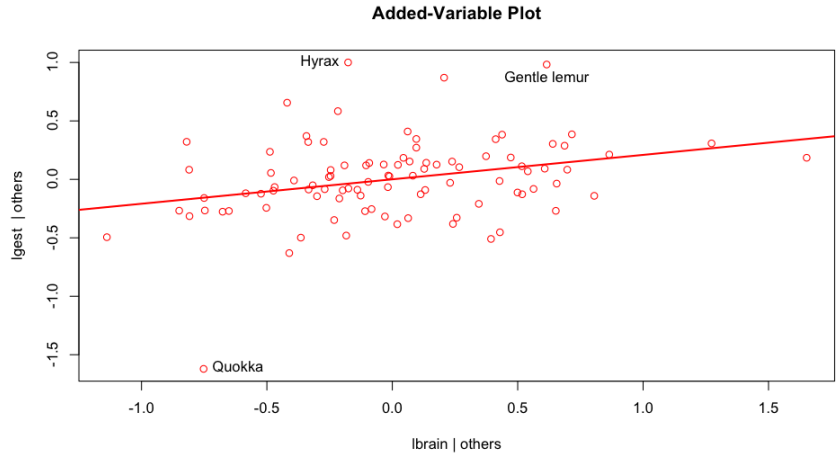
No-three-way interaction

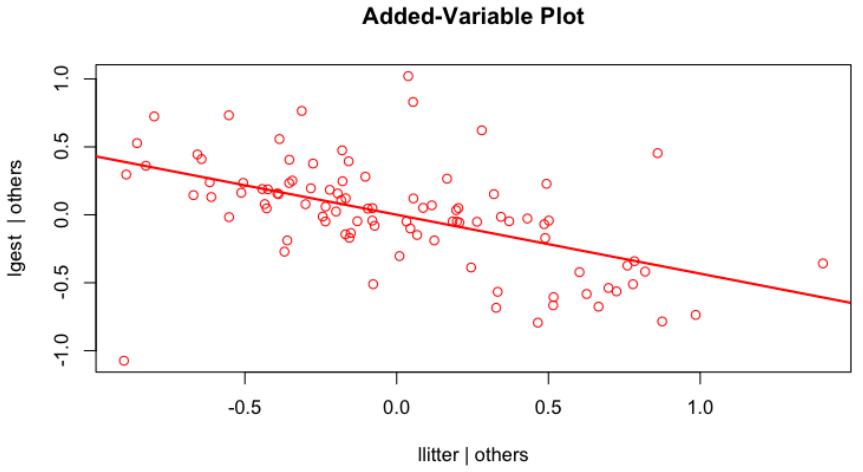
Interface 2: Visualizing raw data and models

Added Variable Plots

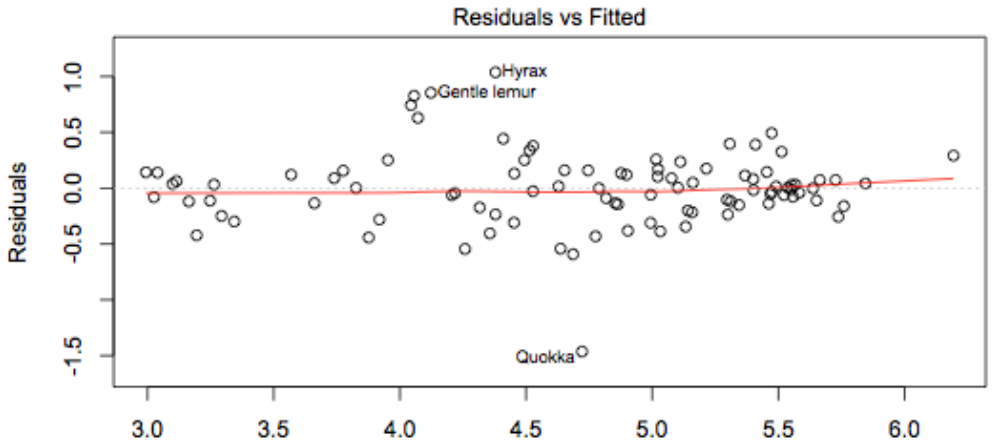


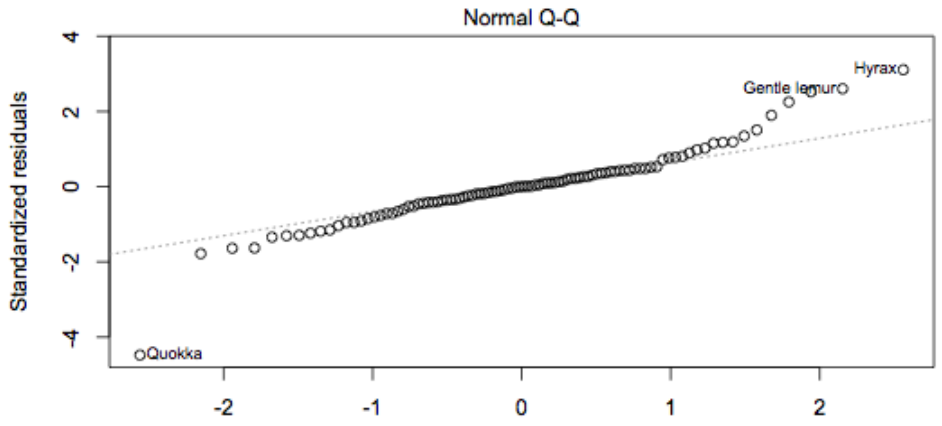


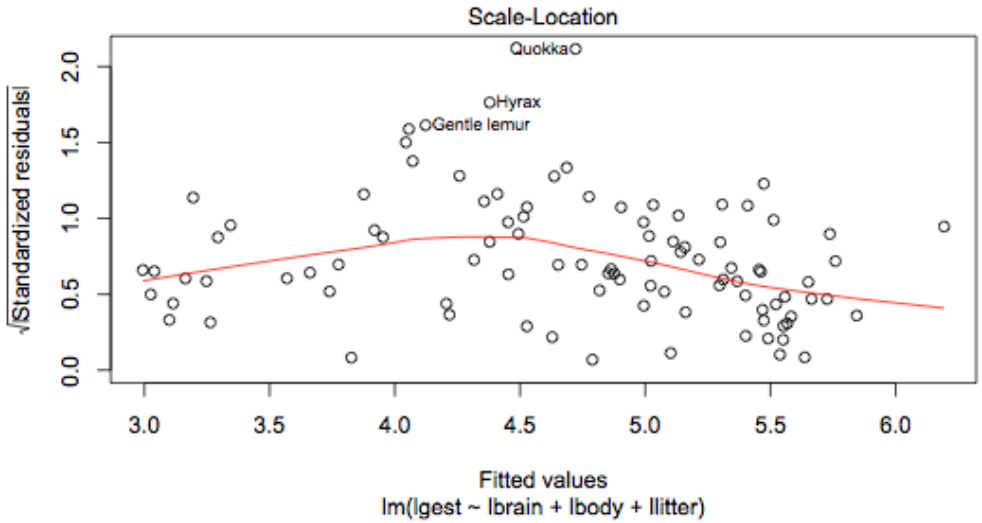


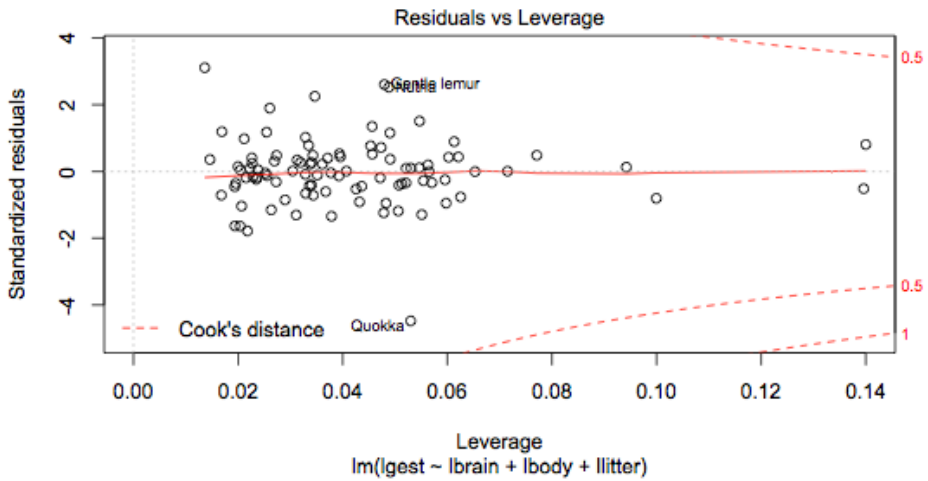


Interface 2: Visualizing raw data and models





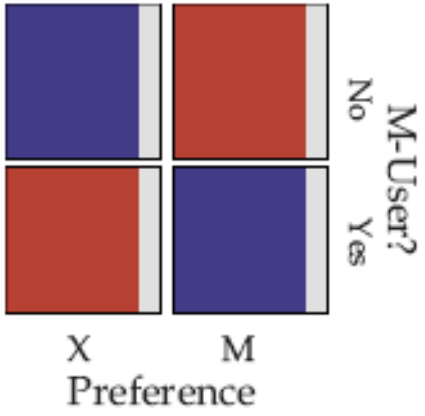




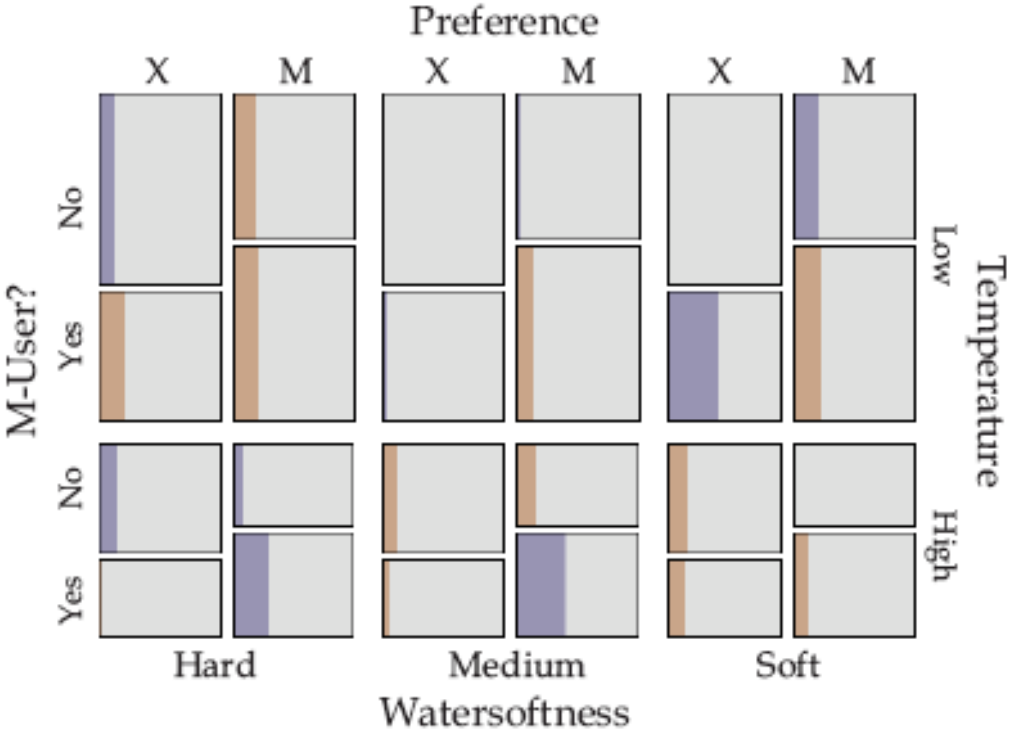
Interface 2: Visualizing raw data and models

Superimposing residual information:

Theus & Lauer (1998), Theus & W. (1996)



Add Interaction:
Preference : M-User?
$G^2 = 20.5$
New Model:
Preference : M-User?
$G^2 = 22.4$; df = 17; p = 0.17



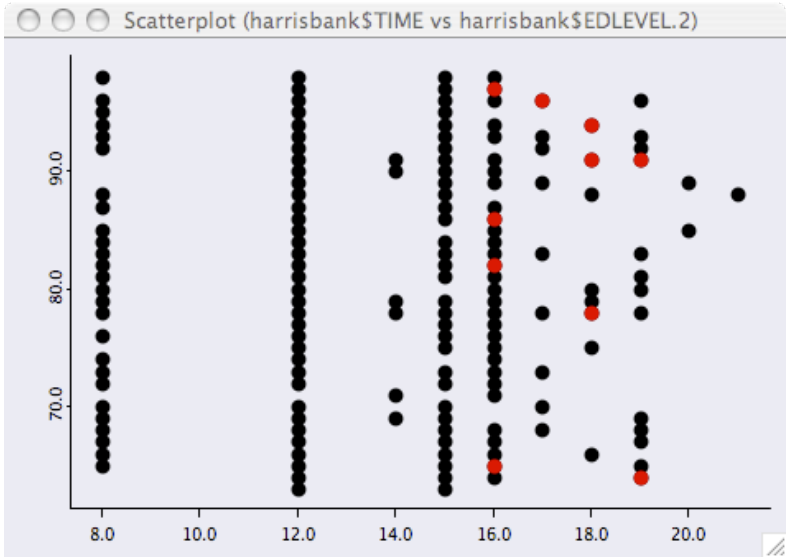
Interface 2: Visualizing raw data and models

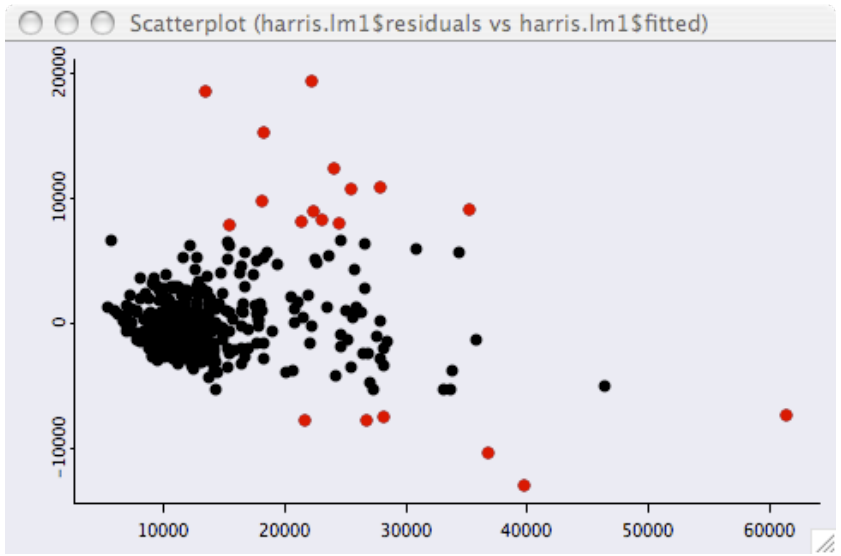
- Model assumptions and model improvement
  - Heteroscedasticity
  - Moderator effects

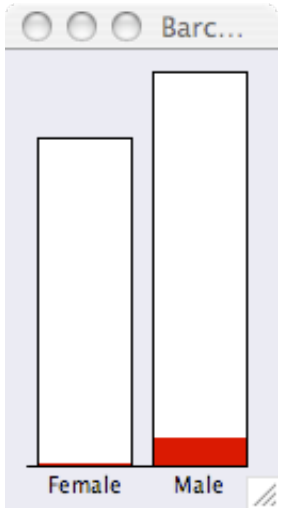
Explaining heteroscedasticity

SALNOW ~SALBEG + WORK

SALNOW = 1540 + 1.92 SALBEG -107.74 WORK

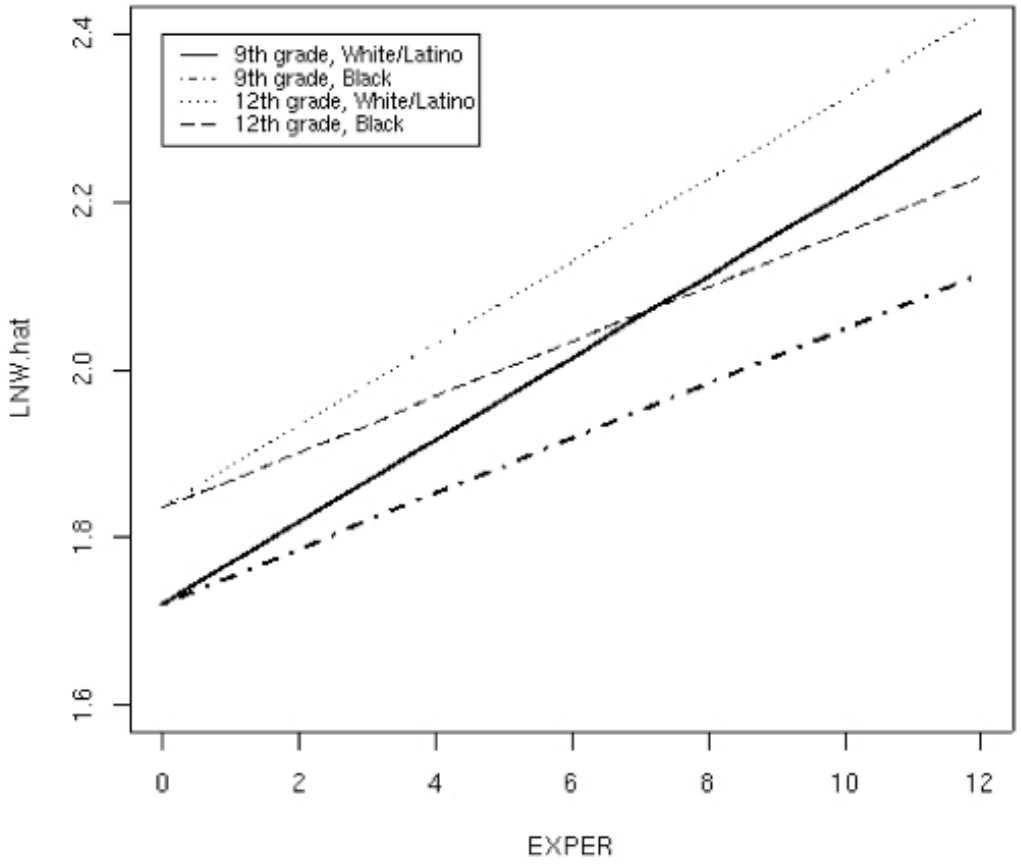






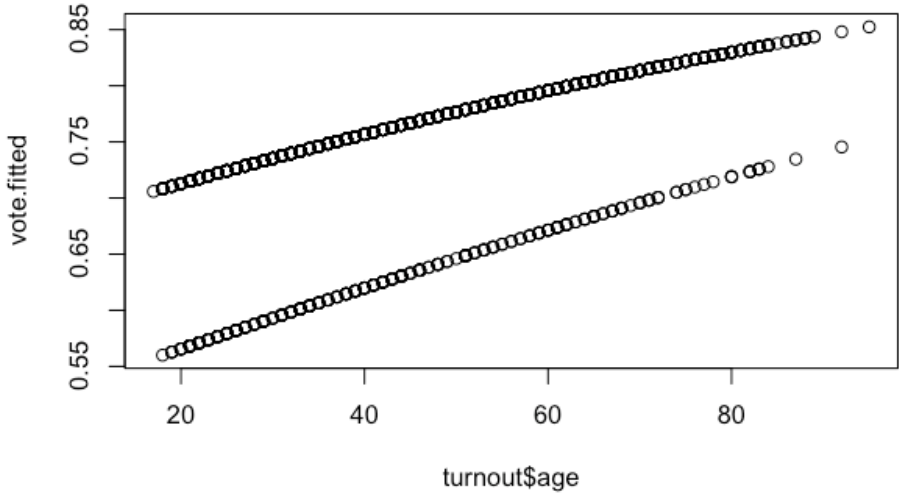
Moderation effects: Modelling of longitudinal data: Wages-Data (Singer & Wiblett, 2003)

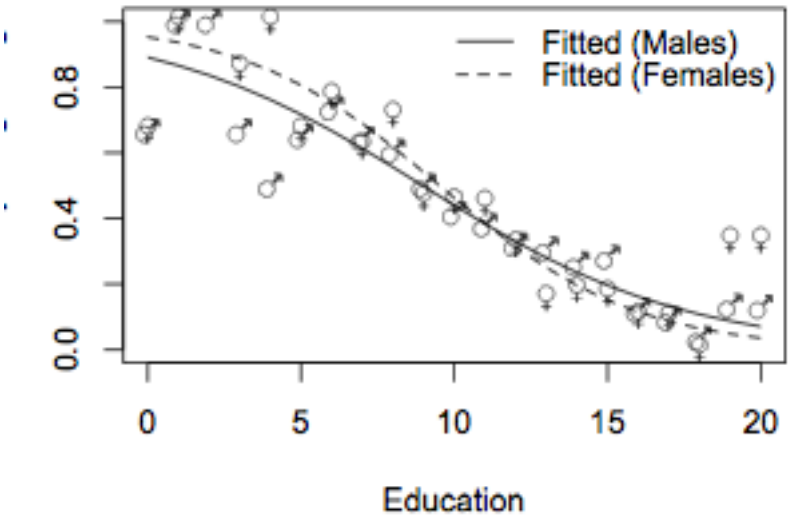
- OLS
- Wage increases with experience
- Slope depending on education level AND ethnicity

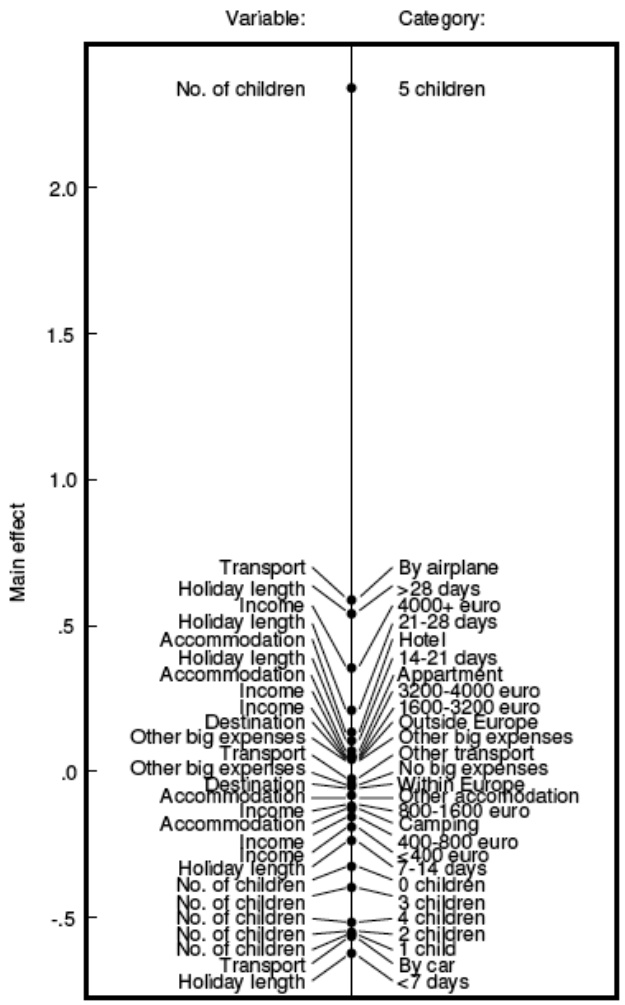


Moderation effects:

- Logistic Regression
- Fixed effects of nominal variables







Interface 3: Visualizing model parameters

Groenen & Koning (2004)

ANOVA

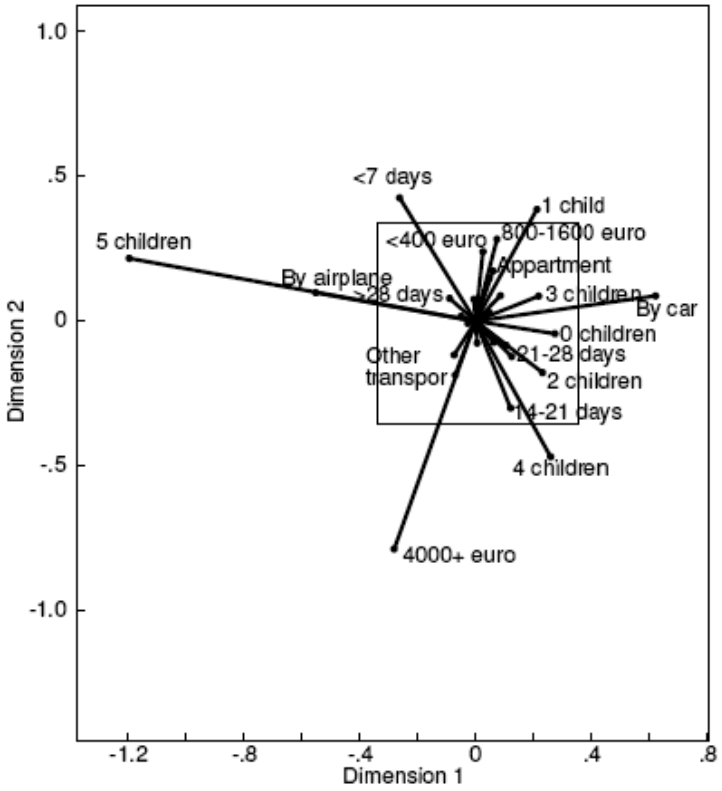
Ranking importance of main effects

Interface 3: Visualizing model parameters

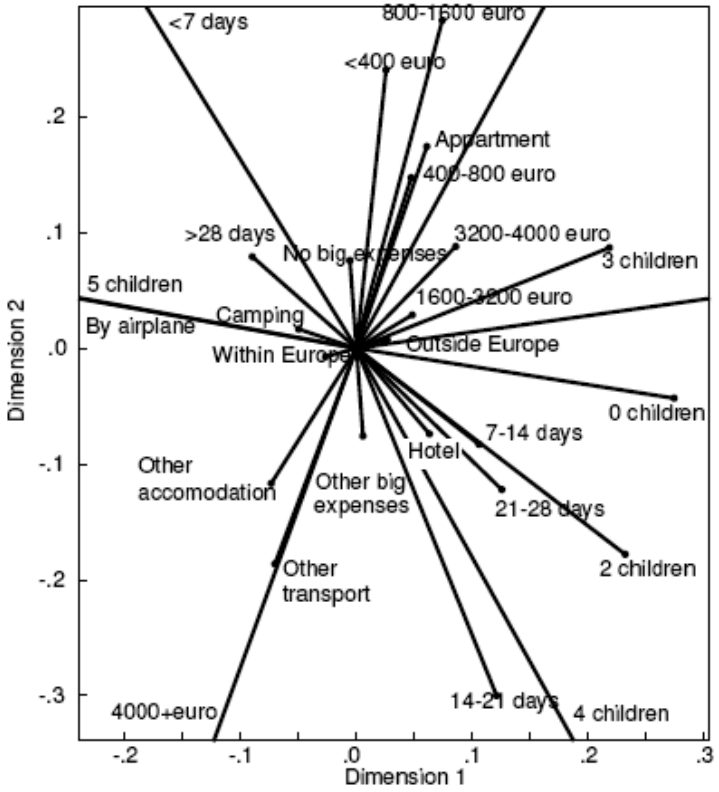
Groenen & Koning (2004)

ANOVA

Graphical representation of interactions



a.Interaction plot.



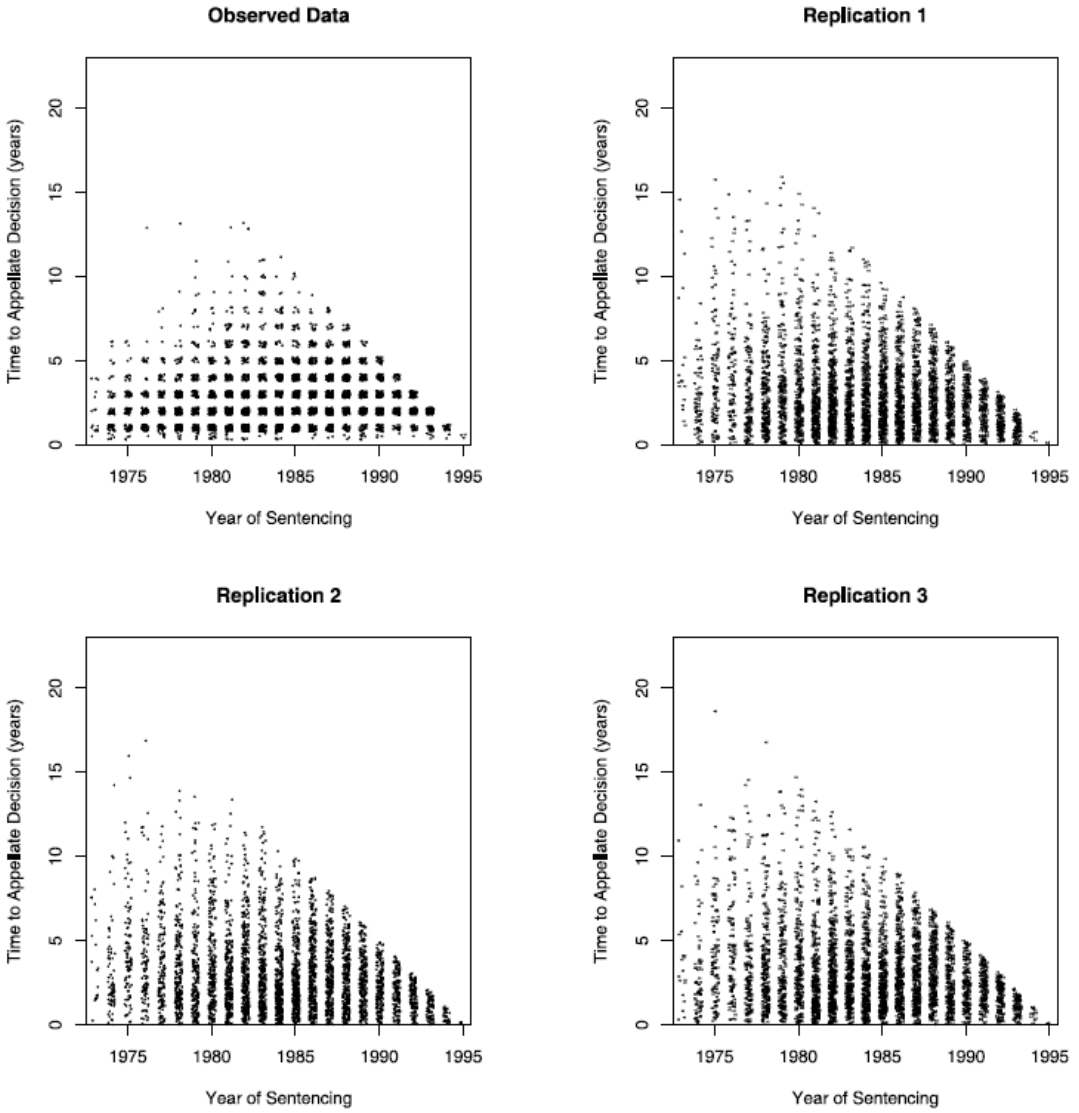
b. Zoomed part of the interaction plot.

Interface 3: Visualizing model parameters
Comparison: Model - Data
(Gelman, 2004)

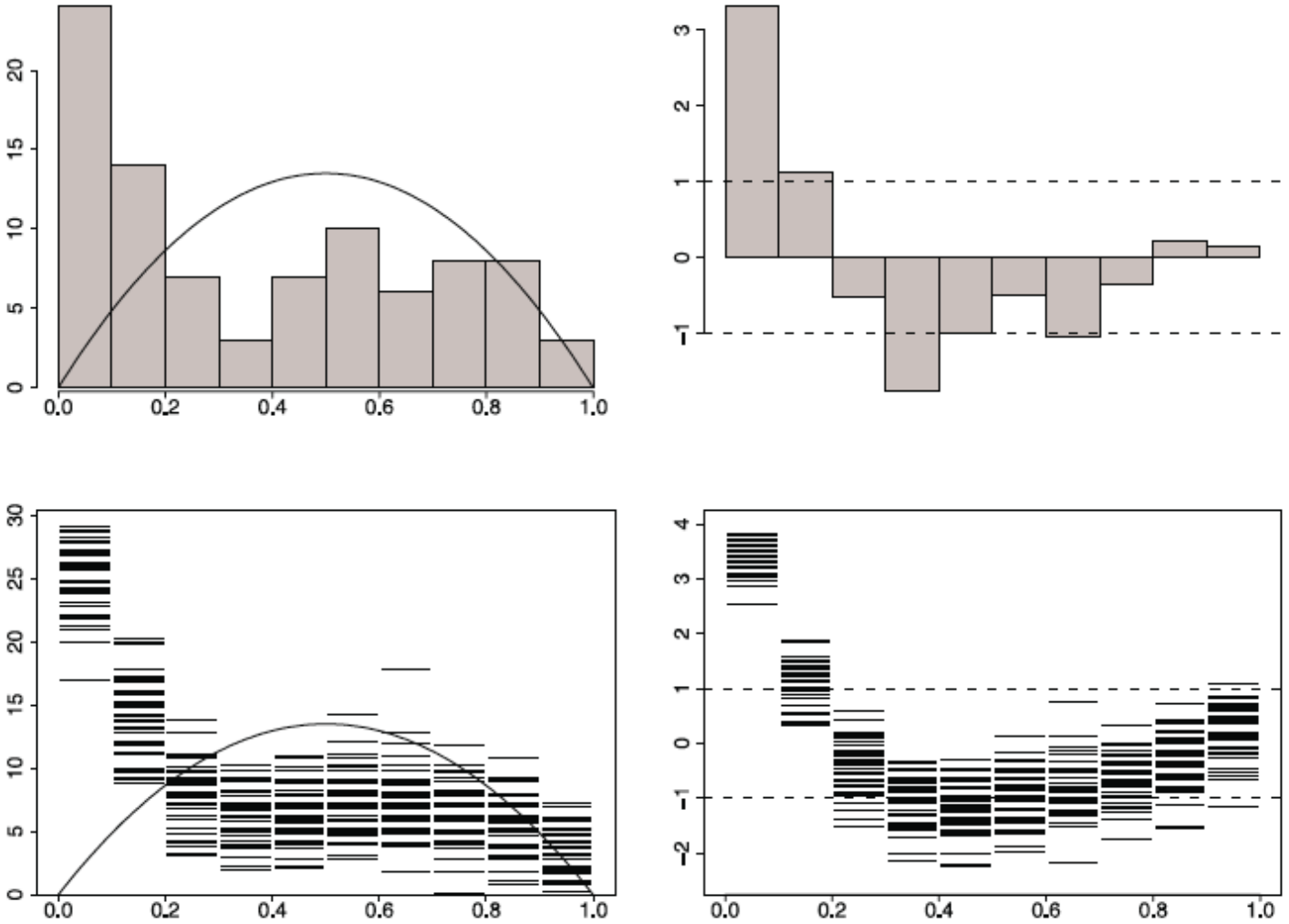
Graphics as a model check

a posteriori predictive distribution

resampling



Interface 3: Visualizing model parameters (Gelman, 2004)



Summary

Various interfaces between visualization and statistical modelling

Graphics as means

- to detect model improvements/deficits
- to understand model structure
- to interpret parameters

Different graphical techniques

- rich selection of basic graphics
- interactivity
- trellis display (grid layouts)

Integrated working environment (iplots, iplots Xtreme, jjplot ?)

Complementarity of EDA and statistical modeling

References:

- [1] A. Gelman, Exploratory Data Analysis for Complex Models (with Discussion by Andreas Buja and Rejoinder). Journal of Computational and Graphical Statistics 13, 755-787, (2004).
- [2] J.W. Tukey, Analyzing data: Sanctification or detective work? American Psychologist, 24, 83-91 (1969).
- [3] J. Bowers & K. W. Drake, EDA for HLM: Visualization when Probabilistic Inference Fails. Political Analysis, 13, 301-326, (2005).
- [4] P. Groenen, & A.J. Koning, A new model for visualizing interactions in analysis of variance. Econometric Institute Report, No EI 2004-06 Revision Date: 2009-07-29, Erasmus University Rotterdam, Econometric Institute, http://econpapers.repec.org/RePEc:dgr:eureir:1765001189, (2004).