WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Peeking Into The Black Box

Gaining Insight With Recursive Partitioning of GLMs

Outline

1. Introduction
   - Motivation
   - Examples
2. Model-Based Recursive Partitioning
   - Basic Idea
   - Technical Details
   - Implementation
3. Gaining insight with Mobs
   - Voting behaviour
   - Debt amortization
   - Beautiful professors
4. Summary

Parametric models

- We often fit parametric models to our data that impose strong restrictions on the relationship between inputs and outputs
- Mainly linear relationships of input and output or functions of the outputs are used
  - LS models
  - Maximum Likelihood models (GLM)
- These approaches have the advantage to be easily interpretable
- They sometimes lack flexibility and predictive power

Highly predictive models

- A number of nonparametric and/or nonlinear approaches have been introduced to model data more flexibly
  - Neural Networks
  - Ensemble Methods
  - Kernel Methods
- These approaches usually exhibit high predictive power
- Are often difficult to interpret (“black box” methods)

Trees

- Trees are somewhere in between these two groups
  - Nonlinear and often nonparametric nature
  - More flexible and higher predictive power than classic parametric models
  - Easier to interpret and visualise than modern prediction models
- Classic idea of trees
  - (Hard) partition the input space $\mathscr{Z}$ into a set of disjoint rectangles
  - Fit a constant in every partition

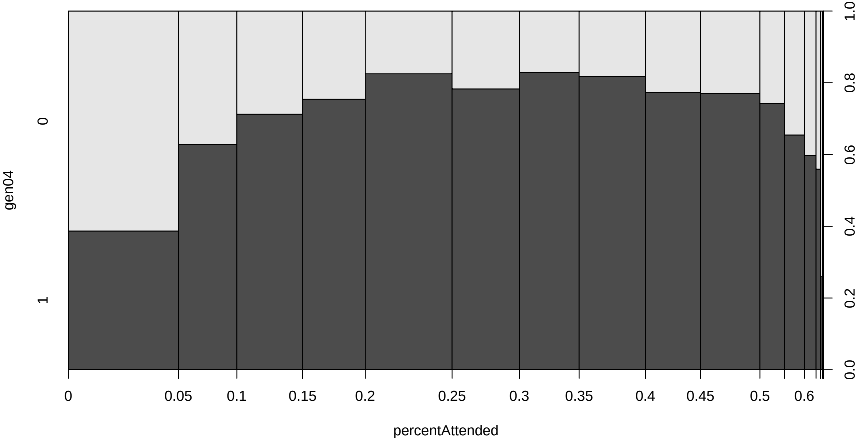
Trees with parametric models

- We want to present an approach that allows to fit a parametric model of interest in the leaves of a tree
- Advantages
  - A specific relationship between dependent and independent variables can be captured by a parametric model
  - A certain model can be assumed and diffential model functioning can be detected
  - Allows graphical representation
  - General framework for model fitting, splitting, pruning

Example 1: Voting behaviour - I

- We looked at data from 2004's general election in Ohio (Bush vs. Kerry)
  - Sample consists of 19634 people
  - Aggregate voting records
  - Demographic, behavioural and institutional covariates
  - Target: Voted in 2004 (yes/no)
- It is "known" that the more often a person went voting in the past, the more likely she will do so in the future
- This is a **logistic regression** problem

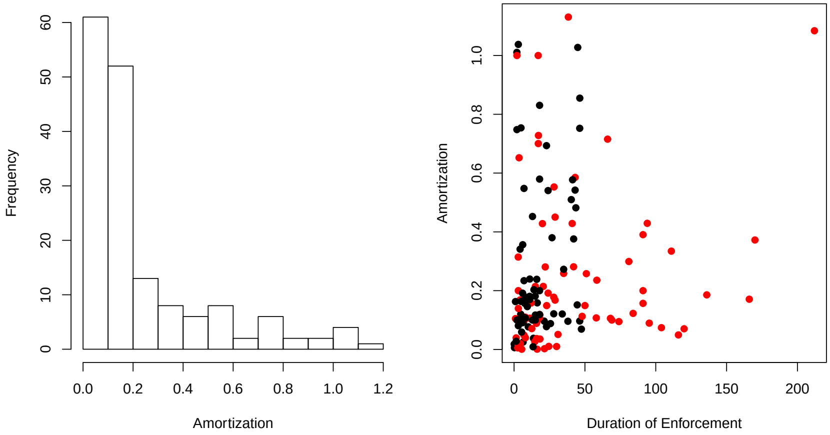
Example 1: Voting behaviour - II



Example 2: Debt amortization - I

- Data from an NPO that advises people who are in debt on and helps with legal, economic, psychosocial issues
  - Sample consists of 165 people
  - Two Austrian provinces
  - Target: Amortization rate relative to the original claim
- We were interested in the relationship between NPO's advice and amortization rates
- Additionally we look at the duration of the enforcement
- We settled for a **parametric "time-to-event" analysis**
  - Amortization rate was "time"
  - Failure to pay more/insolvency/bankruptcy/amortization was "event"

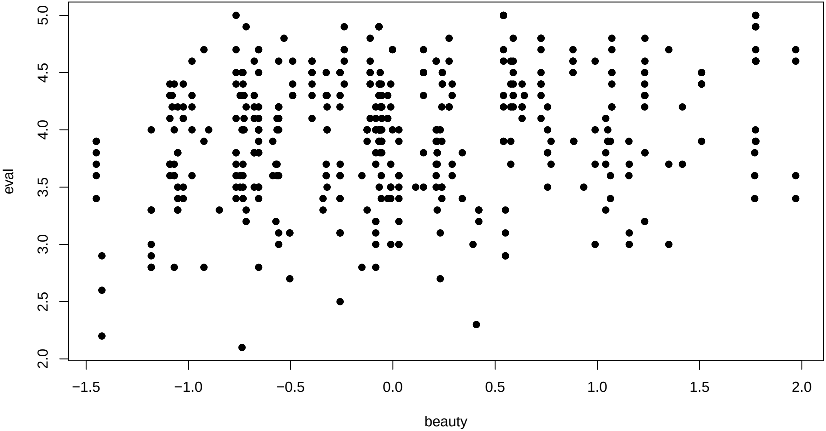
Example 2: Debt amortization - II



Example 3: Teacher performance - I

- Data are from Hamermesh & Parker (2005)
- We are interest in if there is a relationship between professors' good looks and their teaching evaluation
  - Sample consists of 463 courses at University of Texas, Austin
  - Standardized measure of beauty is included
  - Target: Average teaching evaluation per course on scale 1-5
- We used a **weighted linear regression** model with the number of students that evaluated in each course as weights

Example 3: Teacher performance - II



Basic idea

Model-based Recursive Partitioning (Zeileis, et al. 2008) works like this:

- Fit a model for the dependent random variable vector $Y_i, i = 1, \ldots, n$
- Assess the stability of the parameter estimates over partitioning variable vectors $Z_j, j = 1, \ldots, l$
- Split data set according to the variable and its value with the highest parameter instability
- Repeat recursively until either no significant instability can be found or another criterion is met

Fitting the model - Estimation I

- A parametric (possibly multivariate) model $\mathcal{M}(Y, \theta),\ Y \in \mathcal{Y}$ with k-dimensional parameter vector $\theta \in \Theta$ is to be fitted to all observations $Y_i (i = 1, \dots, n)$.
- To get the parameter estimates $\hat{\theta}$ we minimize an objective function $\Psi(Y, \theta)$ (or set its first partial derivatives $\psi(Y, \theta)$ to zero)

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \Psi(Y_i, \theta).$$

Fitting the model - Estimation II

- We assume the existence of a single and stable true θ_0
- Under weak regularity conditions

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d.} \mathcal{N}(0, V(\theta_0))$$

Here, $V(\theta_0) = \{A(\theta_0)\}^{-1}B(\theta_0)\{A(\theta_0)\}^{-1}$, A is the expectation of the derivative of ψ and B the variance of ψ.

- This is the standard M-estimation approach (Huber, 1981) to model fitting approach and includes procedure like
  - Maximum Likelihood (ML)
  - Least Squares (OLS, WLS or GLS)
  - Quasi-ML
  - General M-estimation

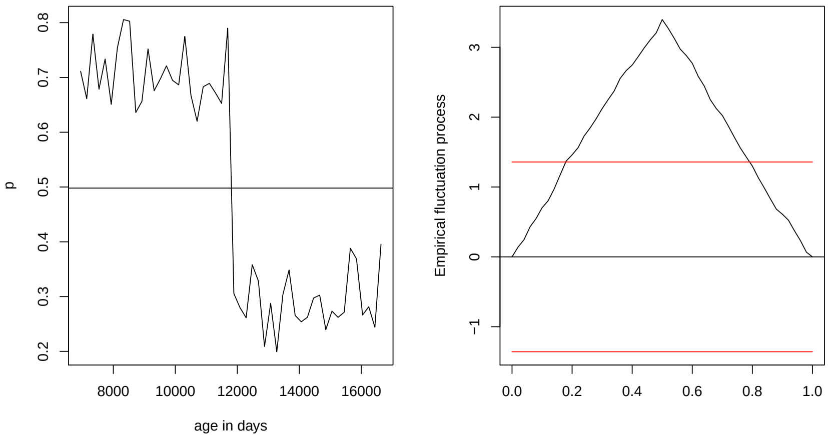
Assessing parameter instability - I

- The parametric model $\mathcal{M}(Y, \theta)$ may have a poor fit for the whole data set but separate models for segments defined by $Z_j, (j = 1, \ldots, l)$ may fit better
- We assess parameter stability w.r.t. Z_j by means of **Generalized M-Fluctuation tests** for an ordering of $Z_j, \sigma(Z_{ij})$
- The empirical fluctuation process of cumulative deviations of the score function $\psi(Y, \hat{\theta})$ with respect to the ordering permutation $\sigma(Z_{ij})$ is

$$W_j(t, \hat{\theta}) = \hat{B}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \psi(Y_{\sigma(Z_{ij})}, \hat{\theta}) \quad (0 \leq t \leq 1).$$

- Zeileis et al. (2007) showed that under parameter stability, $W_j(.) \xrightarrow{d.} W^0(.)$ holds. W^0 is a k-dimensional Brownian bridge

Assessing parameter instability - II



Assessing parameter instability - III

- Test statistics are of the form: $\lambda(W_j)$
- The Null distribution is therefore the asymptotic distribution of $\lambda(W_0)$
  - **Metric variables**: SupLM statistic (Andrews, 1993)

$$\lambda_{supLM}(W_j) = \max_{i=\underline{\imath},\dots,\overline{\imath}} \left(\frac{i}{n} \cdot \frac{n-i}{n} \right)^{-1} \left\| W_j \left(\frac{i}{n} \right) \right\|_2^2 ,$$

    with the interval $[\underline{\imath}, \overline{\imath}]$ over which the potential instability point is shifted (typically defined by requiring some minimal segment size $\underline{\imath}$ and $\overline{\imath} = N - \underline{\imath}$)
  - Maximization of single-shift LM statistics for all possible breakpoints in $[\underline{\imath}, \overline{\imath}]$
  - Limiting distribution is a squared, k-dimensional tied-down Bessel process

Assessing parameter instability - IV

- Test statistics are of the form: $\lambda(W_j)$
- The Null distribution is therefore the asymptotic distribution of $\lambda(W_0)$
  - **Categorical variables:** χ^2 statistic (Hjort & Koning, 2002):

$$\lambda_{\chi^2}(W_j) = \sum_{c=1}^{C} \frac{|I_c|^{-1}}{n} \left\| \Delta_{I_c} W_j \left(\frac{i}{n} \right) \right\|_2^2,$$

    where $\Delta_{I_c} W_j$ is the increment of the empirical fluctuation process over the observations in category $c = 1, \ldots, C$
  - Invariant to reordering of and within categories
  - Captures instability for splitting data according to C categories
  - Limiting distribution is χ^2 with $df = k(C - 1)$

Splitting

- After instability has been detected, the data set is split into 2 subsets for the variable Z_j that exhibited highest parameter instability
- The split point is found such that

$$\sum_{i \in I_1} \Psi(Y_i, \theta_1) + \sum_{i \in I_2} \Psi(Y_i, \theta_2)$$

  is optimized for two rival segmentations by an exhaustive search over all pairwise comparisons of possible splits (runtime $O(n)$).
- More than binary splits are possible
- The usage of p-values (Bonferroni-corrected because of multiple testing) serves as some kind of pre-pruning of the tree

Implementation I

Model-based Recursive Partitioning can be carried out with the function `mob()` in the R package `party`. Currently only Generalised Linear Models are implemented.

- Fitting a Mob

```
> model <- mob(Y ~ X1 + X2 + ... + XK | Z1 + Z2 + ... + ZL, model = glinearModel,
+     family = family(), control = mob_control())
```

- Control the Recursive Partitioning Algorithm: `mob_control()`
- Extract a leaf: `nodes(model,leafNumber)`
- Available standard S3 functions: `summary(),plot(),predict(),print(),coef()`

Implementation - II

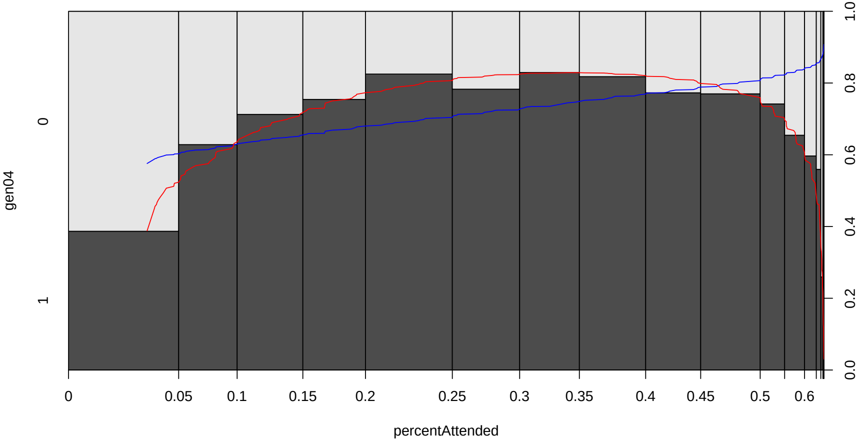
```
> fmPID <- mob(diabetes ~ glucose | pregnant + mass + age, model = glinearModel,
+     family = binomial())
> print(nodes(fmPID, c(2, 4, 5)))

[[1]]
2)*  weights = 167
Terminal node model
Binomial GLM with coefficients:
(Intercept)      glucose
   -9.95151      0.05871


[[2]]
4)*  weights = 304
Terminal node model
Binomial GLM with coefficients:
(Intercept)      glucose
   -6.70559      0.04684


[[3]]
5)*  weights = 297
Terminal node model
Binomial GLM with coefficients:
(Intercept)      glucose
   -2.77095      0.02354
```

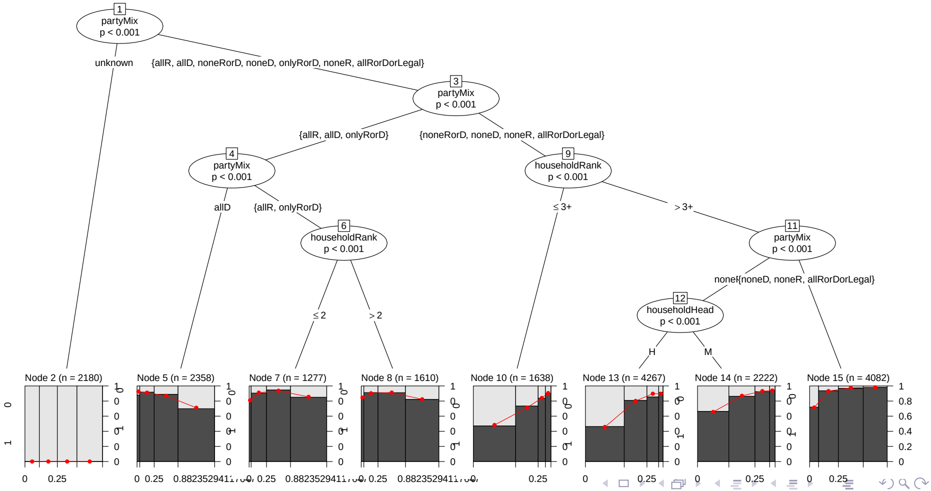
Example 1: Voting behaviour - I



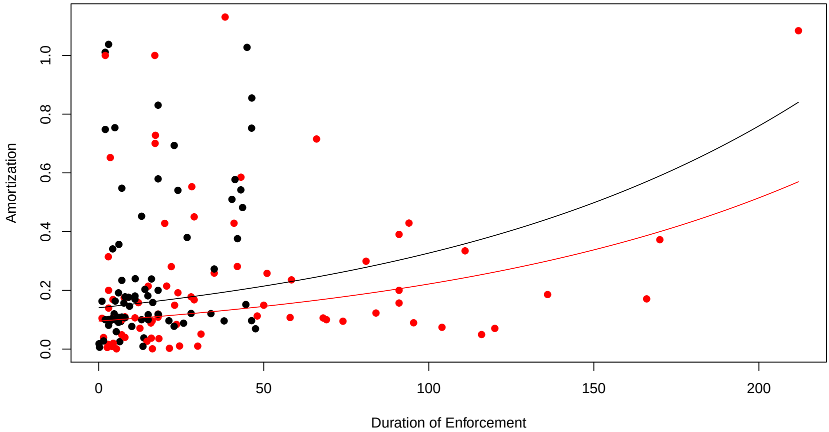
Example 1: Voting behaviour - II

- **Logistic Regression Model**
- Voting probability explained by the relative frequency of attended elections so far
- We had a number of covariates whose influence were not clear
  - Age in days, gender
  - Party affiliation, party makeup of household, rank and position in household
  - Income, education
  - Donation to various causes (health, environment etc.)
  - Federal contribution in certain years
  - Computer owner, home owner

Example 1: Voting behaviour - III



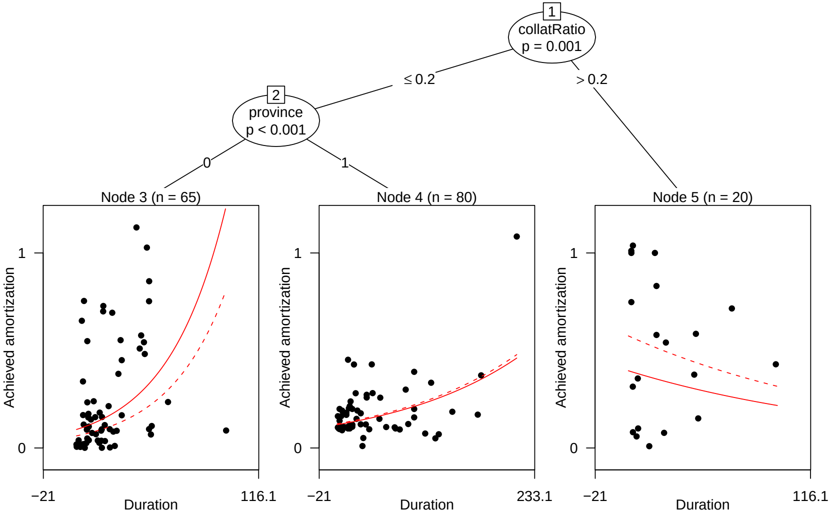
Example 2: Debt amortization - I



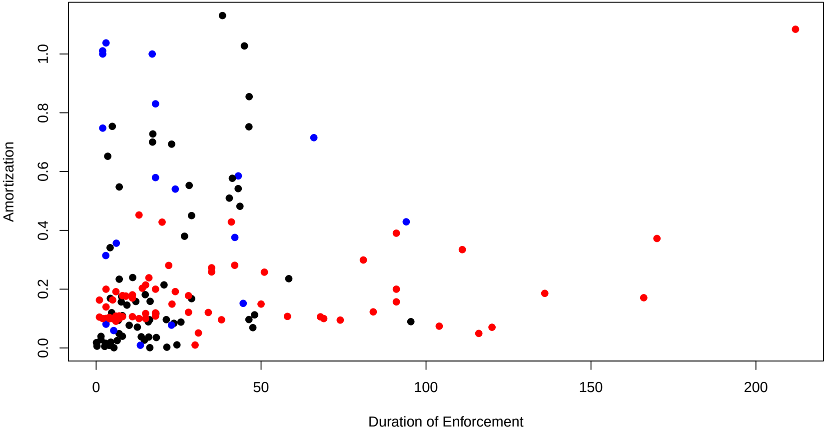
Example 2: Debt amortization - II

- **Weibull Regression Model**
- Debt amortization rate until “failure” explained by the duration of the enforcement and if the person was advised
- Additional covariates
  - Gender, province
  - Contact frequency
  - Liability at begin of the enforcement, current liability
  - Securities, collateralization ratio

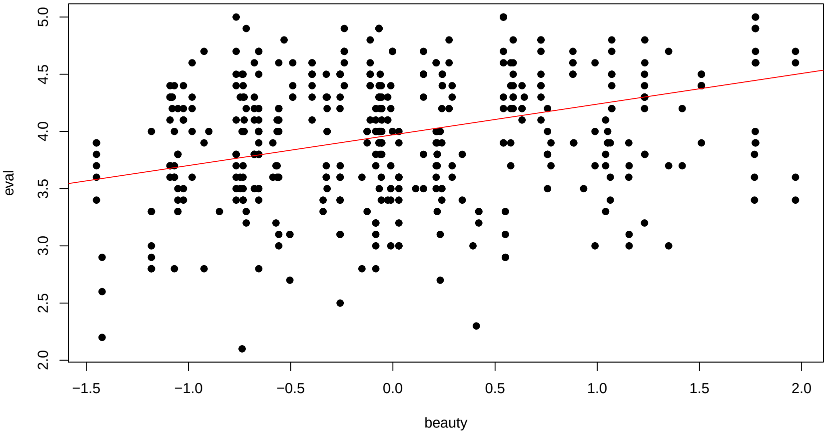
Example 2: Debt amortization - III



Debt amortization - IV



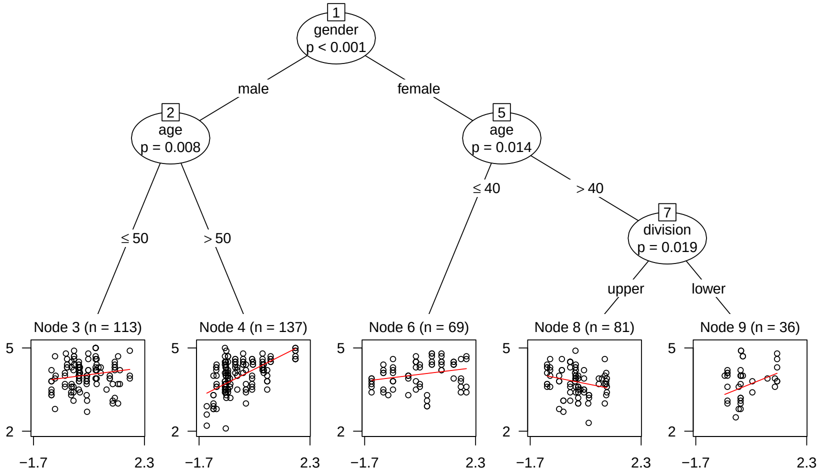
Example 3: Beautiful professors - I



Example 3: Beautiful professors - II

- **Weighted Least Squares Regression**
- Performance evaluation explained by beauty score
- Additionally, we can use these covariates
  - Gender, Age
  - Minority
  - On tenure track
  - Native speaker
  - Lower division course

Example 3: Beautiful professors - III



Further examples

Model-based Recursive Partitioning has already been successfully applied to

- Bradley-Terry models (Strobl et al., 2010)
- Item Response models (Strobl et al., 2010)
- Functional differential equation models (Jank et al., 2008)

And there is more to come...

Summary

- Model-based Recursive Partitioning allows growing trees with a parametric model in leaves
- Works for all M-estimators
- Automatically detects interactions and differential model functioning
- Enables visualisation of complex segmented models
- Especially useful if there is an a-priori model to be partitioned

References

- Andrews.(1993). Econometrica, 61, 821-856.
- Hamermesh & Parker.(2005). Economics of Education Review, 24, 369-376.
- Hjort & Koning.(2002). Nonparametric Statistics, 14, 113-132.
- Huber.(1981). Robust Statistics.
- Jank, Shmueli, Wang.(2008). Statistical methods in E-Commerce Research.
- Strobl, Kopf, Zeileis.(2010). In preparation.
- Strobl, Wickelmaier, Zeileis.(2010). Journal of Educational and Behavioral Statistics.
- Zeileis, Hothorn & Hornik.(2008). Journal of Computational and Graphical Statistics, 17, 492-514.
- Zeileis & Hornik.(2007). Statistica Neerlandica, 61, 488-508.

Thank you for your attention!

Thomas Rusch
Department of Finance, Accounting and Statistics
Institute for Statistics and Mathematics
email: `thomas.rusch@wu.ac.at`
URL: `http://statmath.wu.ac.at/~tr`

WU Wirtschaftsuniversität Wien
Augasse 2–6, A-1090 Wien

Achim Zeileis
Department of Statistics
email: `Achim.Zeileis@R-project.org`
URL: `http://statmath.wu.ac.at/~zeileis`

Universität Innsbruck
Universitätsstrasse 15, A - 6020 Innsbruck