Visualizing Non-Parametric and Parametric Model Trees

Achim Zeileis Torsten Hothorn Kurt Hornik

http://statmath.wu-wien.ac.at/~zeileis/

Overview

- Motivation
- Two algorithms for growing trees
  - Conditional inference trees
  - Model-based recursive partitioning
- Visualization
  - Univariate displays
  - Bivariate displays
- Examples
- Implementation in R
- Summary

Motivation

Starting point: Popularity of classification and regression trees stems mainly from two features:

1. *interpretability*, enhanced by visualizations of the fitted decision trees,
2. *predictive power* in non-linear regression relationships.

Idea: Enhance decision trees (especially for exploration) by adding statistical graphics for the models fitted in the leafs to the tree displays. This facilitates to communicate complex regression problems, particularly to non-statisticians.

Growing trees

The suggested visualization techniques are applicable much more generally, but are readily implemented in R for two algorithms:

- *conditional inference trees* (CTree), learns non-parametric tree models where response variable and partitioning variables can be measured at arbitrary scales. Visualizations are aimed particularly at univariate responses at different scales.
- *model-based recursive partitioning* (MOB), learns parametric tree models based on M-type estimators (e.g., ML or OLS). Visualizations are aimed particularly at trees based on regression models.

Conditional inference trees

1. Test the global null hypothesis of independence between any of the partitioning variables and the response. If there is some overall dependence, select the variable with strongest association to the response.

2. Compute the split point(s) that locally optimize the association measure.

3. Split this node into daughter nodes and repeat the procedure.

Model-based recursive partitioning

1. Fit the model once to all observations in the current node by optimization of an objective function (e.g., log-likelihood, sum of squares).
2. Assess whether the parameter estimates are stable with respect to every partitioning variable. If there is some overall instability, select the variable associated with the highest parameter instability, otherwise stop.
3. Compute the split point(s) that locally optimize the objective function.
4. Split this node into daughter nodes and repeat the procedure.

Univariate displays

Decision trees are easily visualized by their associated tree graph.

Inner nodes are typically just summarized by their label (and p value).

Edge labels describe the splits.

Univariate graphical displays can be used instead of textual summaries in terminal nodes:

- *numeric:* boxplot, histogram, kernel density.
- *categorical:* barplots (besides or stacked).
- *censored:* Kaplan-Meier curves.

Bivariate displays

For model-based trees, the nodes can include partial plots of the response against (each of) the regressor(s) along with (projected) model fits:

- *numeric* $\sim$ *numeric:* scatterplot.
- *numeric* $\sim$ *categorical:* parallel boxplots.
- *categorical* $\sim$ *categorical:* mosaic plots, spineplots.
- *categorical* $\sim$ *numeric:* spinograms, CD plots.

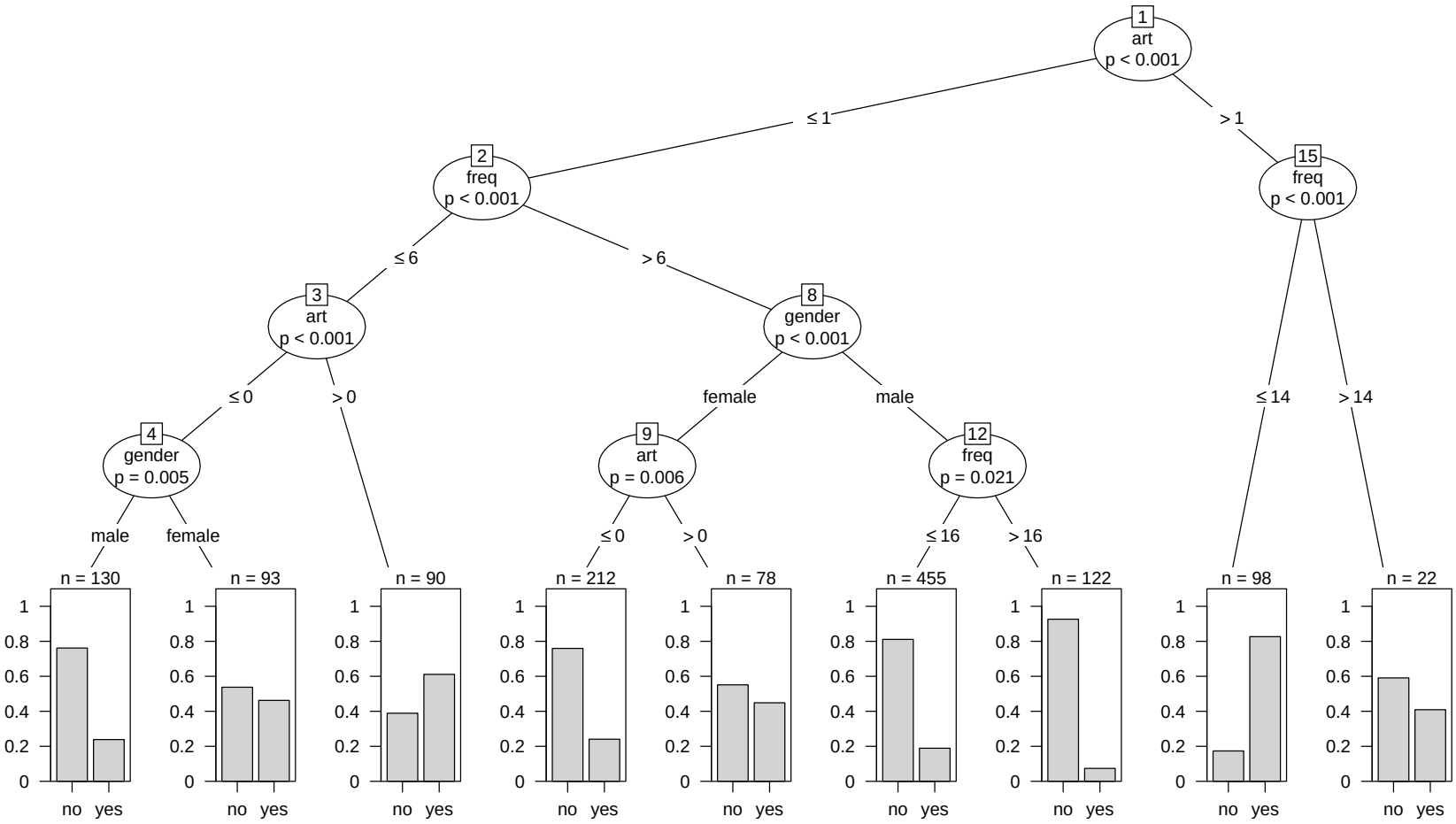
Examples: CTree

- *Classification tree:* Customer choice for an art book advertised by the Bookbinder's Book Club depending on socio-economic covariates and customer history.

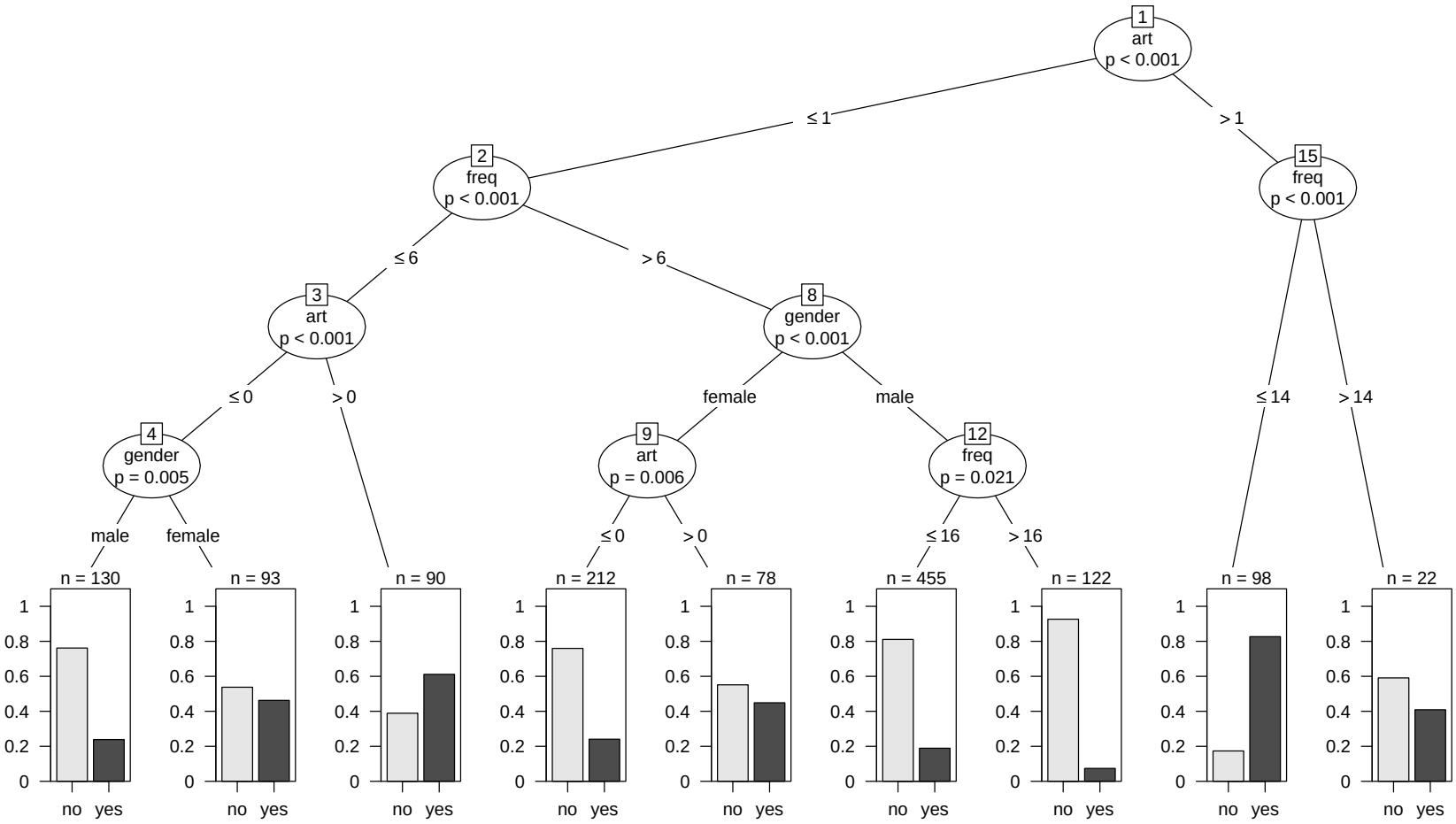
- *Regression tree:* Abundance of tree pipits depending on environmental factors.

- *Survival tree:* Breast cancer survival depending on prognostic factors.

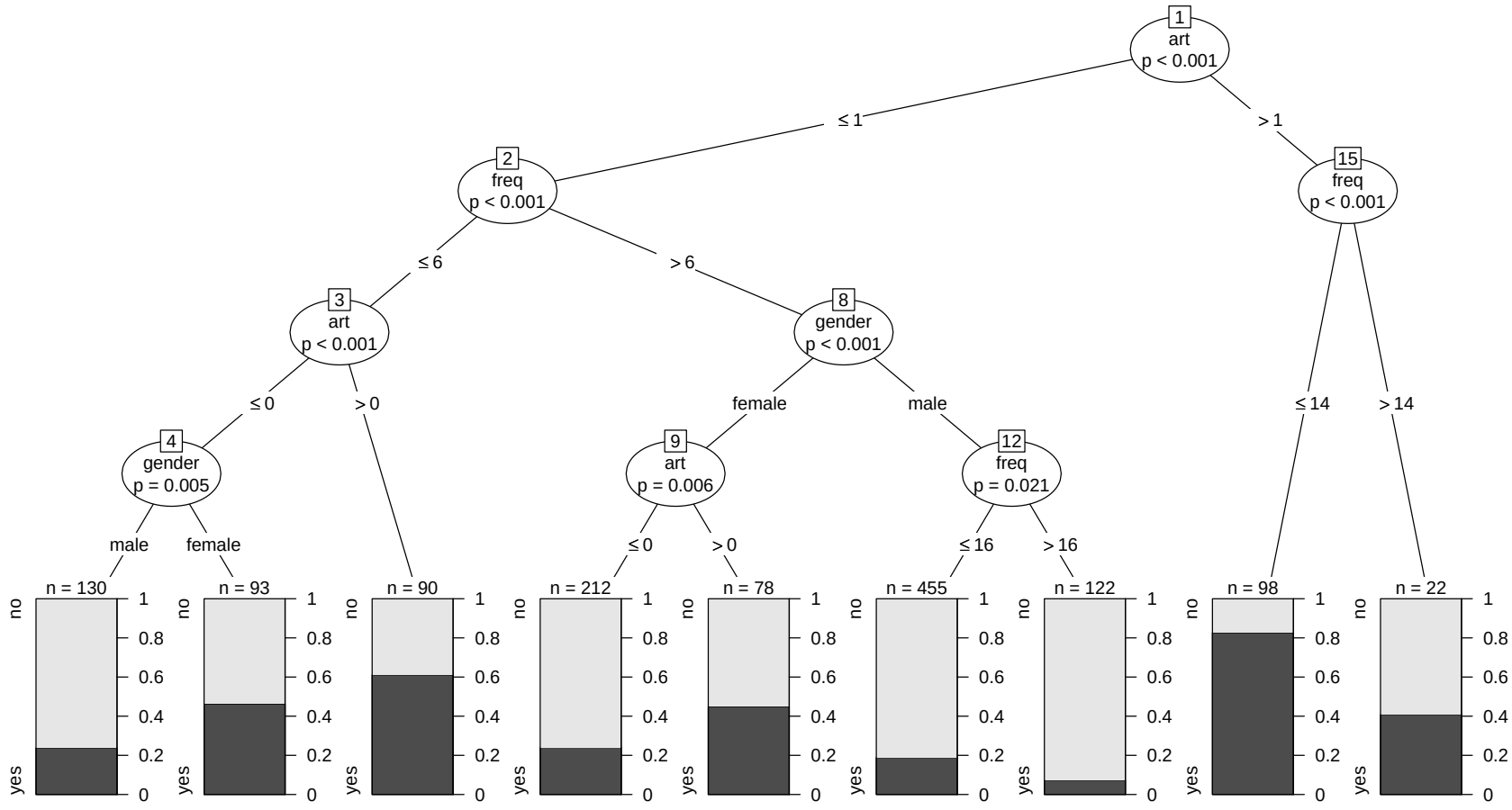
Examples: Customer choice



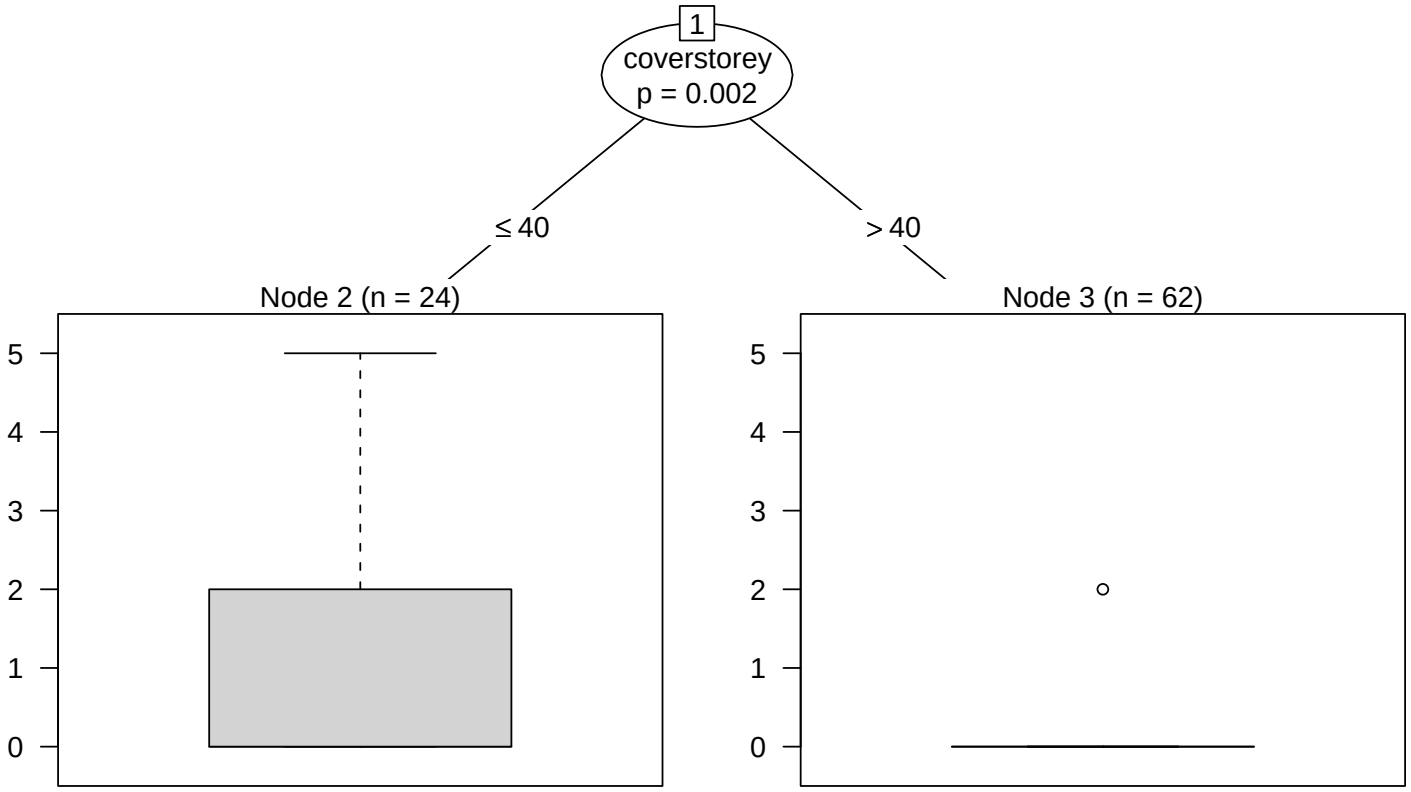
Examples: Customer choice



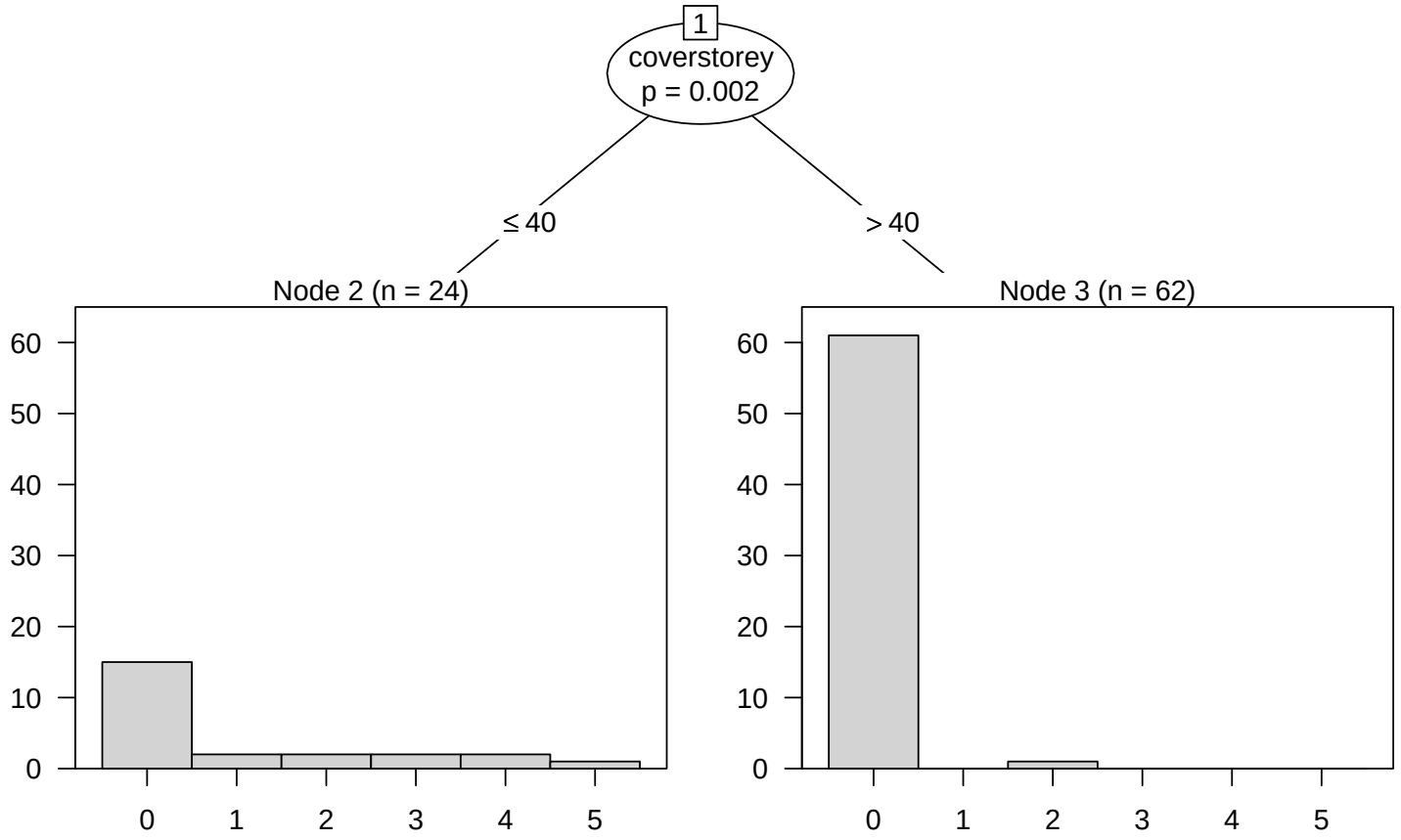
Examples: Customer choice



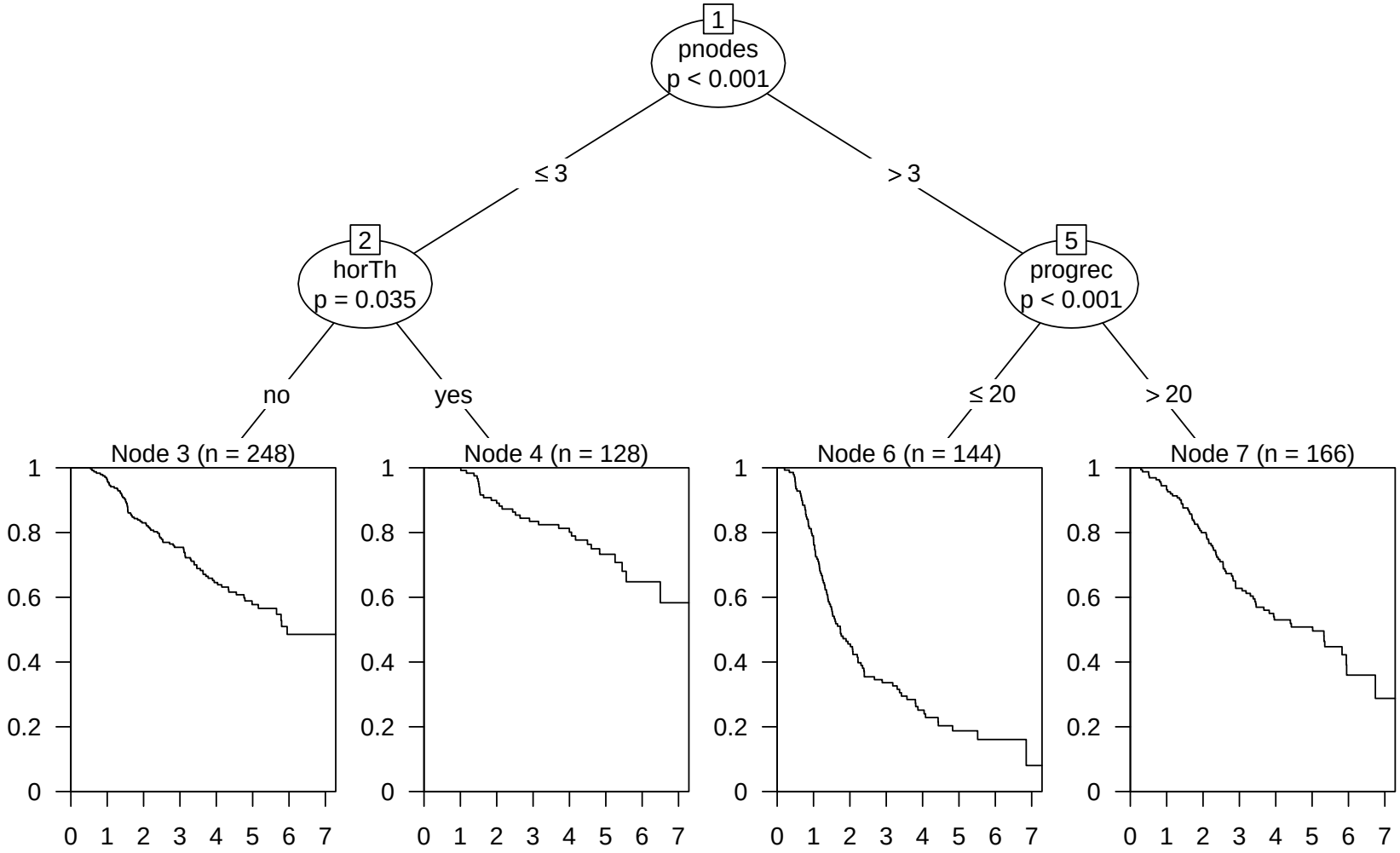
Examples: Tree pipits



Examples: Tree pipits



Examples: Breast cancer survival

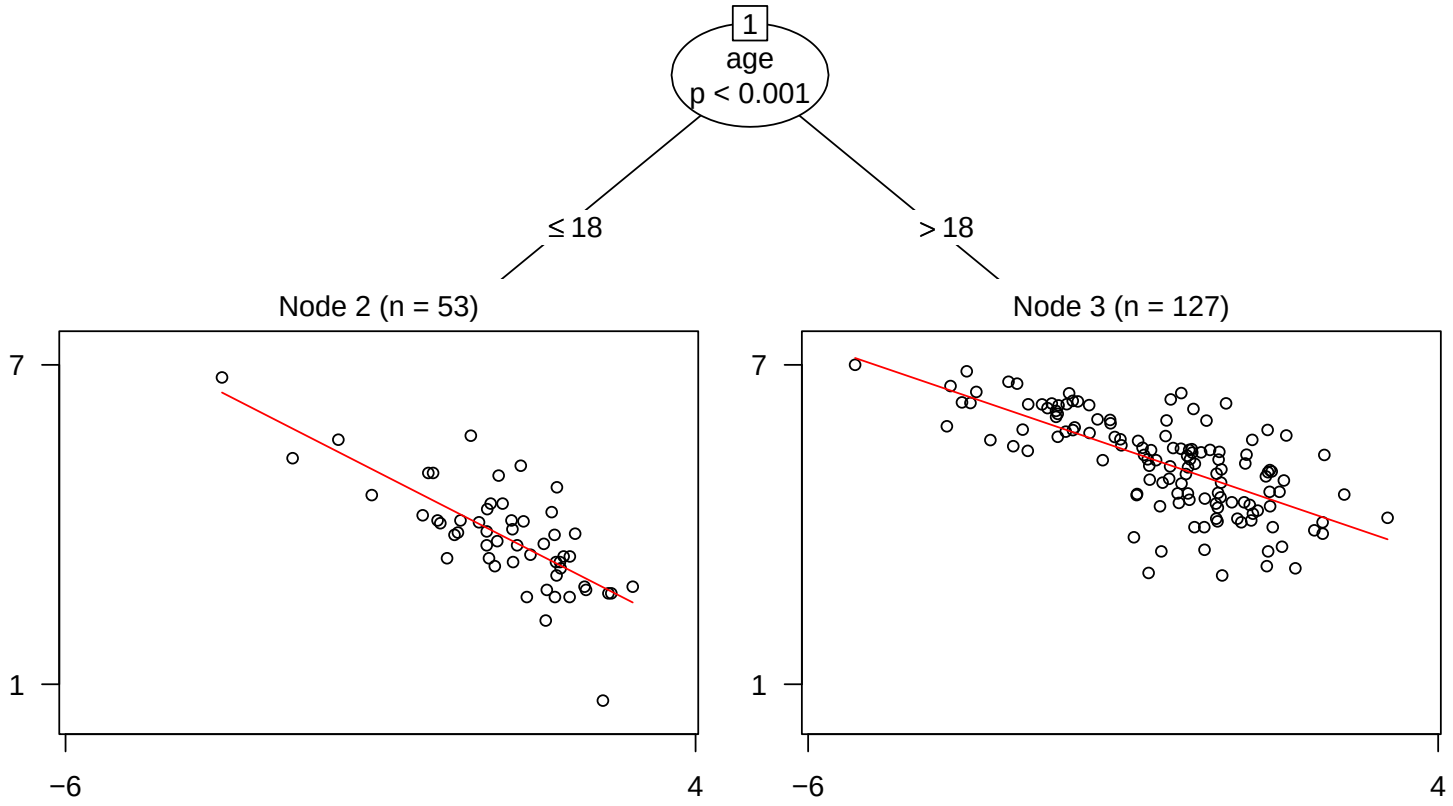


Examples: MOB

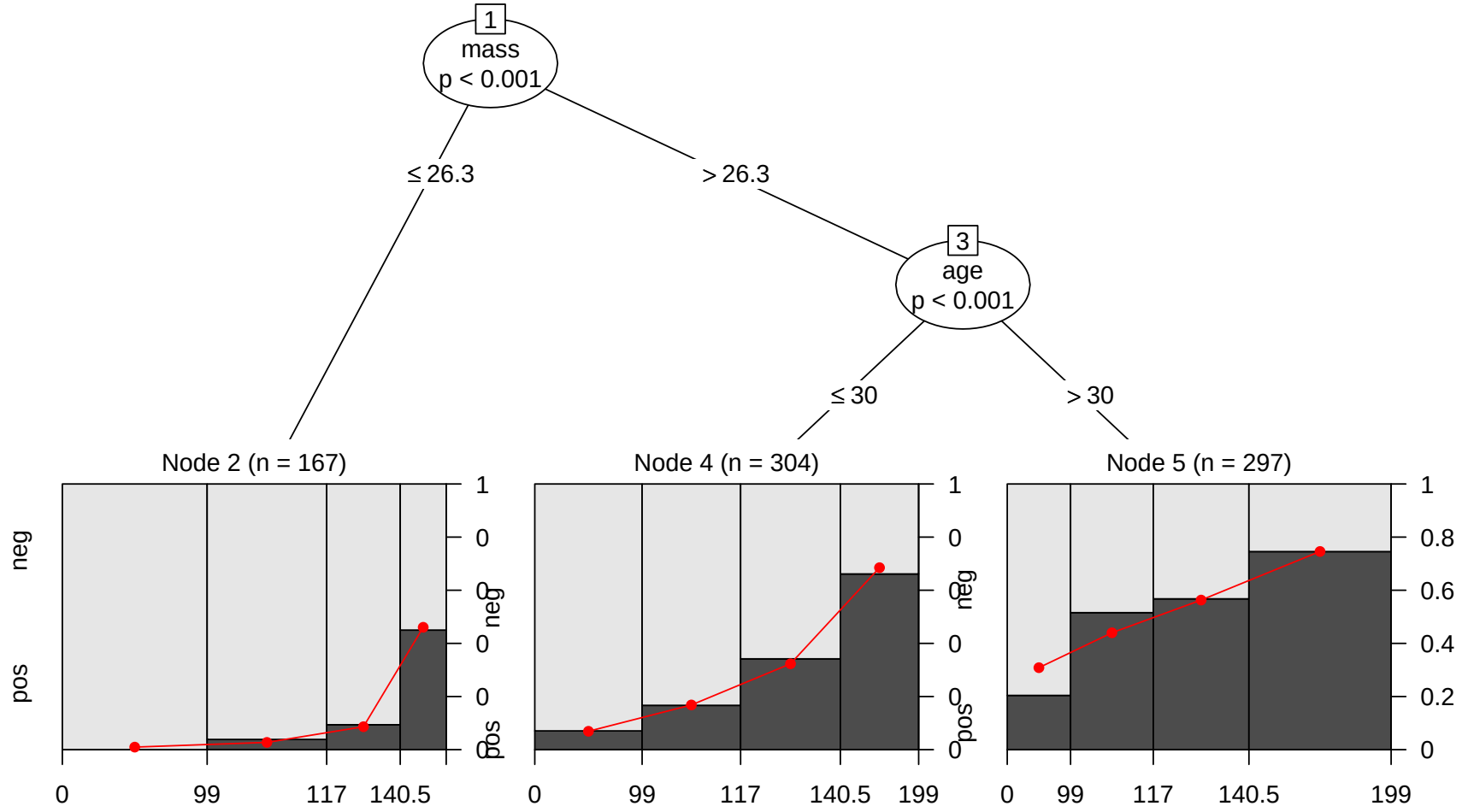
Each of the following models is segmented by further covariates:

- *Linear regression:* Demand for economic journals (library subscriptions) by price per citation.
- *Logisitc regression:* Outcome of diabetes test by plasma glucose concentration.
- *Weibull survival regression:* Breast cancer survival by node positive breast nodes and hormonal therapy.

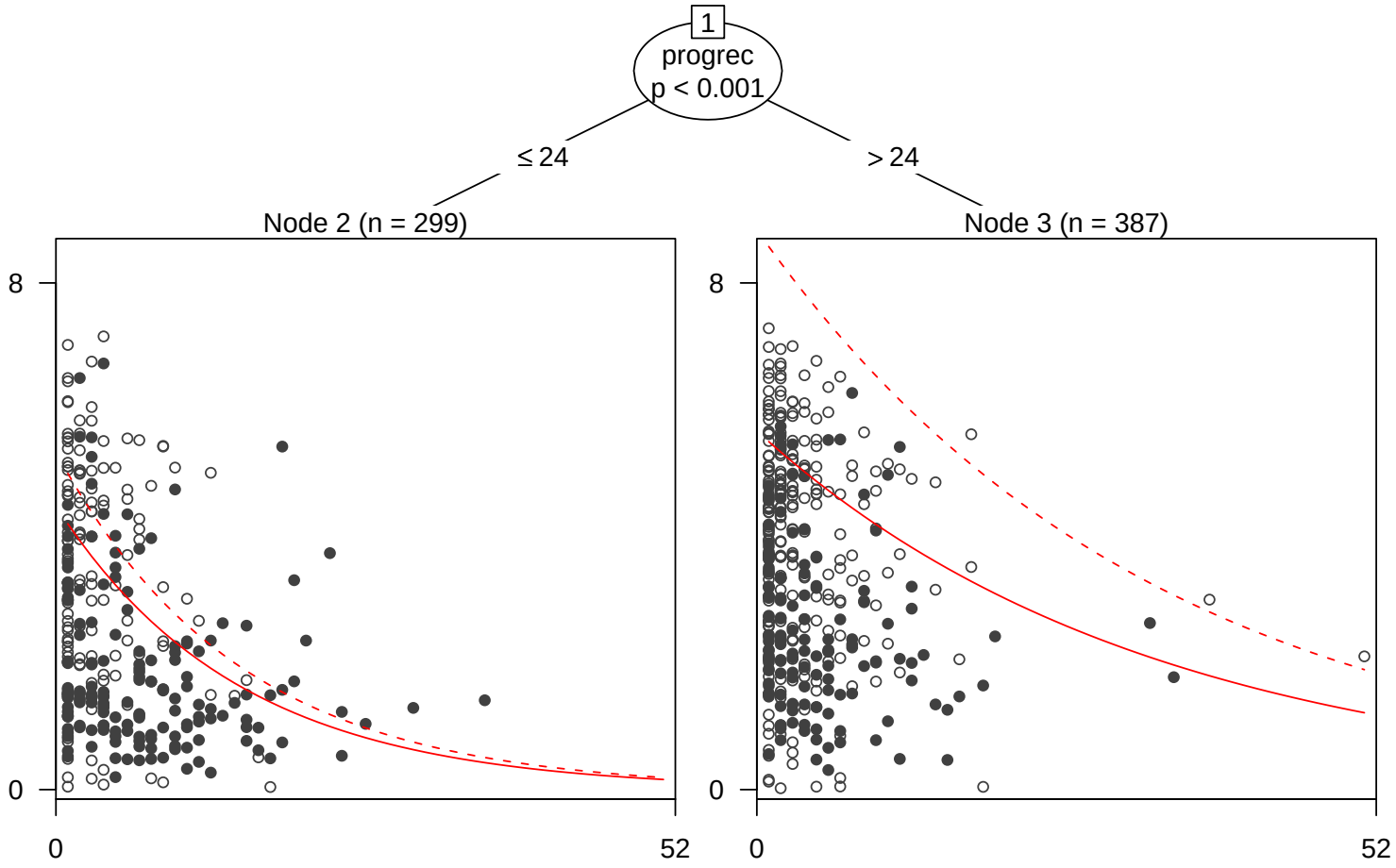
Examples: Demand for econ. journals



Examples: Pima Indians diabetes



Examples: Breast cancer survival 2



Implementation in R

All trees and associated visualizations can be easily generated with the R package **party**, a recursive *party*tioning lab, available from

http://CRAN.R-project.org/

Visualizations are made possible by the **grid** graphics system: a viewport is created for each node of the tree and painted by a panel function. New panel functions can be plugged in by the user, flexible defaults are chosen automatically.

Implementation in R

CTree is provided by function `ctree()`:

```
fmBBBC <- ctree(choice ~ ., data = BBBClub)
plot(fmBBBC)
```

MOB is provided by function `mob()`:

```
fmPID <- mob(diabetes ~ glucose | pregnant + pressure + triceps +
  insulin + mass + pedigree + age, data = PimaIndiansDiabetes,
  model = glinearModel, family = binomial())
plot(fmPID)
```

Summary

Tree models are popular due to their interpretability and predictive power.

Prediction: In purely predictive settings, single trees are often outperformed by ensemble methods, boosting, random forests or support vector machines.

Exploration: However, trees are still an excellent measure to communicate complex regression problems to non-statisticians. Graphical representations of tree graphs coupled with standard statistical graphics enhance the interpretability and make trees even more intellegible.

References

Hothorn T, Hornik K, Zeileis A (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, **15**(3), Forthcoming.

Zeileis A, Hothorn T, Hornik K (2005). "Model-based Recursive Partitioning." *Report 19*, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series. URL http://epub.wu-wien.ac.at/.

Zeileis A, Hothorn T, Hornik K (2006). "Evaluating Model-based Trees in Practice." *Report 32*, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series. URL http://epub.wu-wien.ac.at/.