WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

LORET: Logistic Regression Trees
Mob for Discrete Category Models

Contents

- Discrete Category Models
- Logistic Models
- Segmentation with Trees
- Mob Algorithm
- Technical Aspects of Logistic Regression
- LORET
- Example

Discrete Category Models - I

- Models to describe, explain and predict the categories of a discrete variable from a finite set of two or more alternatives.
- Discrete categories models statistically relate the category to attributes of the object that exhibits the category, to attributes of the category and to attributes of the alternatives available.
- The probability of a category is modelled rather than the category itself.

Discrete Category Models - II

Examples:

- Binary: $y_i \in \{0, 1\}$ with observed categories 0 or 1. Information to utilize is the category.
- Multinomial: $y_i \in \{0, 1, 2, \ldots, K\}$ where ordering of the k categories is irrelevant. Observed ategories are 0 through K. Information to utilize is the category. There can be correlation between the categories or not.
- Ordinal: $y_i \in \{0, 1, 2, \ldots, K\}$ where ordering of the k choices carries information. Observed categories are 0 through K. Information to utilize is the category and the ordering of the categories. There can be correlation between the categories or not.

Discrete Category Models - III

A discrete category model is made up of:

- Category set with finite cardinality and exhaustive and mutually exclusive elements
- Category probability $P(Y_i = k) = p_{ik}$ when the category Y of object i is alternative k
- Probability model as a function of alternative- (v) and object (x)-specific variables and unknown parameter vector β, $p_{ik} = g(x_i, v_{ik}, v_{il}; \beta)$ with $l \neq k$ and $g(\cdot)$ a (useful but arbitrary) function.
- Discrete observation follows from a data generating process governed by the probabilities based on the concept of a latent underlying continuous variable.

Discrete Category Models - IV

Latent variable view (often also called utility theory view):

- Define a latent variable U_{ik} that stands for the meaning a category k has for object i
- We observe $y_{ik} = 1$ iff $U_{ik} > U_{il}, l \neq k$ else $y_{ik} = 0$
- The latent value U_{ik} is an additive function of a combination of $x's$ and $v's$ with unknown but given β and a latent error term ϵ_{ik}, so $U_{ik} = f(x_i, v_{ik}, v_{il}; \beta) + \epsilon_{ik}$.
- Then for $k \neq l, l \neq m$:

$$\begin{aligned} p_{ik} &= P(U_{ik} > U_{il}) \\ &= P(f(x_i, v_{ik}, v_{il}; \beta) + \epsilon_{ik} > f(x_i, v_{il}, v_{im}; \beta) + \epsilon_{il}) \\ &= P(\epsilon_{il} - \epsilon_{ik} < f(x_i, v_{ik}, v_{il}; \beta) - f(x_i, v_{il}, v_{im}; \beta)) \end{aligned}$$

Discrete Category Models - V

Notable properties of these models

- We look at the probability that the difference in latent errors is smaller than differences in the structural part of the latent variable.
- The U are on a latent difference scale so there is no absolute point.
- The scale if not properly defined in units, so it usually gets normalized to the variance of the ϵ which need not be the same for different data sets.

Logistic Models - I

How to estimate a discrete category model?

- Use a linear function for the $f(\cdot)$
- Use the logistic function for the $g(\cdot)$
- Assume that $\epsilon_{ik} - \epsilon_{il}$ follow a logistic distribution with mean 0 (or equivalently that the individual errors are Gumbel or Type I Generalized Extreme Value distributions).

Logistic Models - II

Then the model can be set up as

$$p_{ik} = \frac{\exp\left(x_{ik}^{\top}\beta\right)}{\sum_{l=1}^{L}\exp\left(x_{il}^{\top}\beta\right)}$$

where

- $x_{ik} = h(x_i, v_{ik}, v_{il})$ is a design vector for person i and alternative k expressing all the effects of interest.
- $f(x_{ik};\beta)$ is linear in β (β collects all the parameters of interest); we call the $x_{ik}^{\top}\beta$ the linear predictor.

These models can be seen as belonging to the wider class of Generalized Linear Models. The vector β is usually estimated by maximum likelihood ($\hat{\beta}_{ML}$).

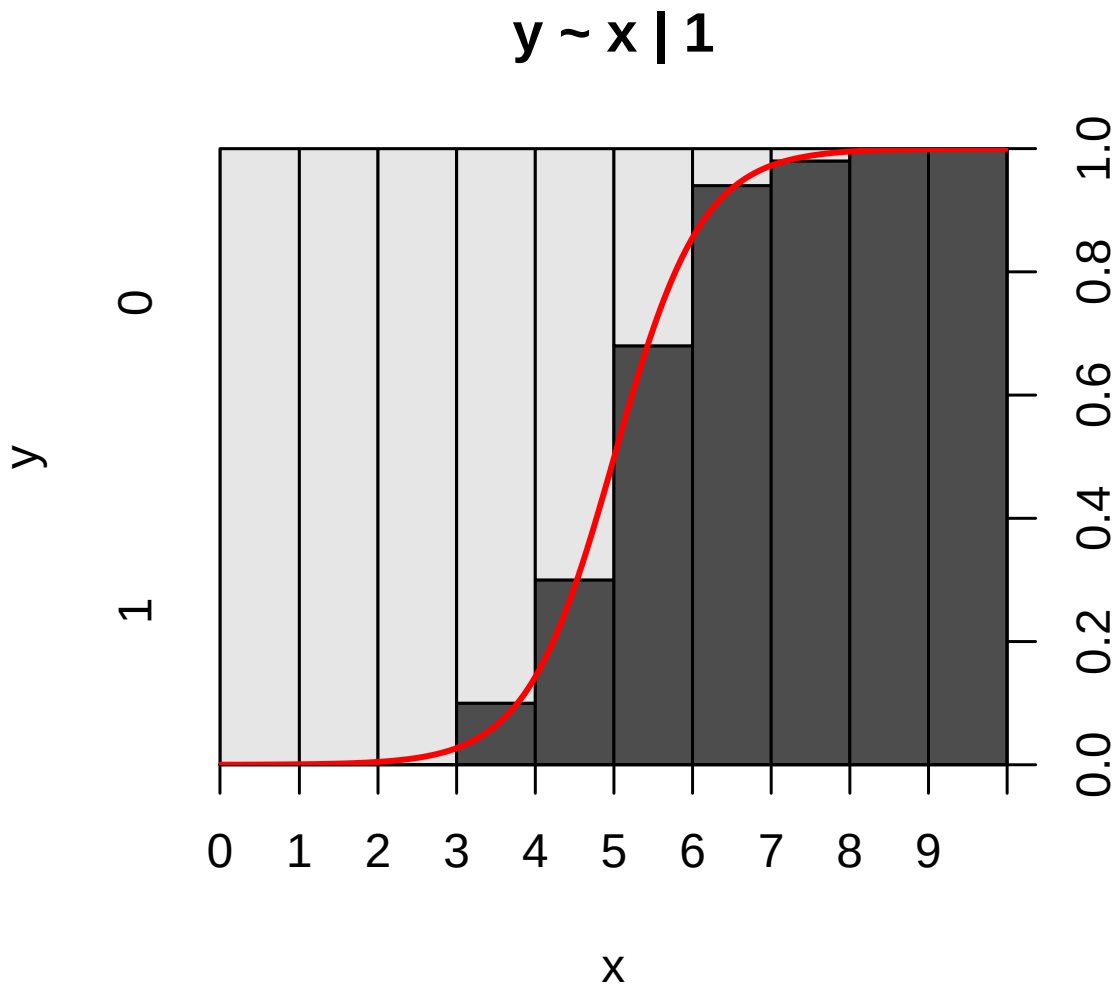
Example: Voting behaviour

We look at data from 2004's general election in Ohio (Bush vs. Kerry - Rusch et al., 2013).

- Sample consists of 19600 people
- Aggregate voting records
- Target: Voted in 2004 (yes/no)

The more often a person went voting in the past, the more likely she will do so in the future.

Example: Logistic Regression



Logistic Models - III

The family of logistic model includes the binary logistic model, the multinomial logistic model, the proportional odds model (or cumulative logit link model), the conditional logit model, the nested logit models and the mixed logit model.

- They are popular and often work quite well
- They have the advantage to be easily interpretable and they do not tend to overfit

But

- They sometimes lack flexibility and predictive power when functional form is too rigid
- They can fail in reproducing the underlying structures (e.g., if data show additional complexity or heterogeneity)

Segmentation - I

One way to allow for extra flexibility by assuming that there is a number of segments of observations and that within each segment a logistic model applies.
These segments are found with only little assumptions so flexibility is introduced thus. The difficulty is to find those segments.
Three main approaches:

- Finite Mixture Models
- Clustering and ex post model fitting
- Top down partitioning

We will consider the last approach.

Segmentation - II

Top down partitioning:

- Divide the data set into subsets that are maximally homogenuous within and maximally heterogenuous between the subsets (for a suitably defined criterion of optimality)
- Repeat this for the subsets

This is the idea of a tree (or recursive partitioning) algorithm, which (hard) partitions the input space $\mathcal{Z}$ into a set of disjoint, bordering polytopes.

Trees are

- Non-linear, non-parametric models
- Allow for automatic interaction detection
- The segmentation is learned rather than specified
- Often more flexible and show higher predictive power than simple parametric models

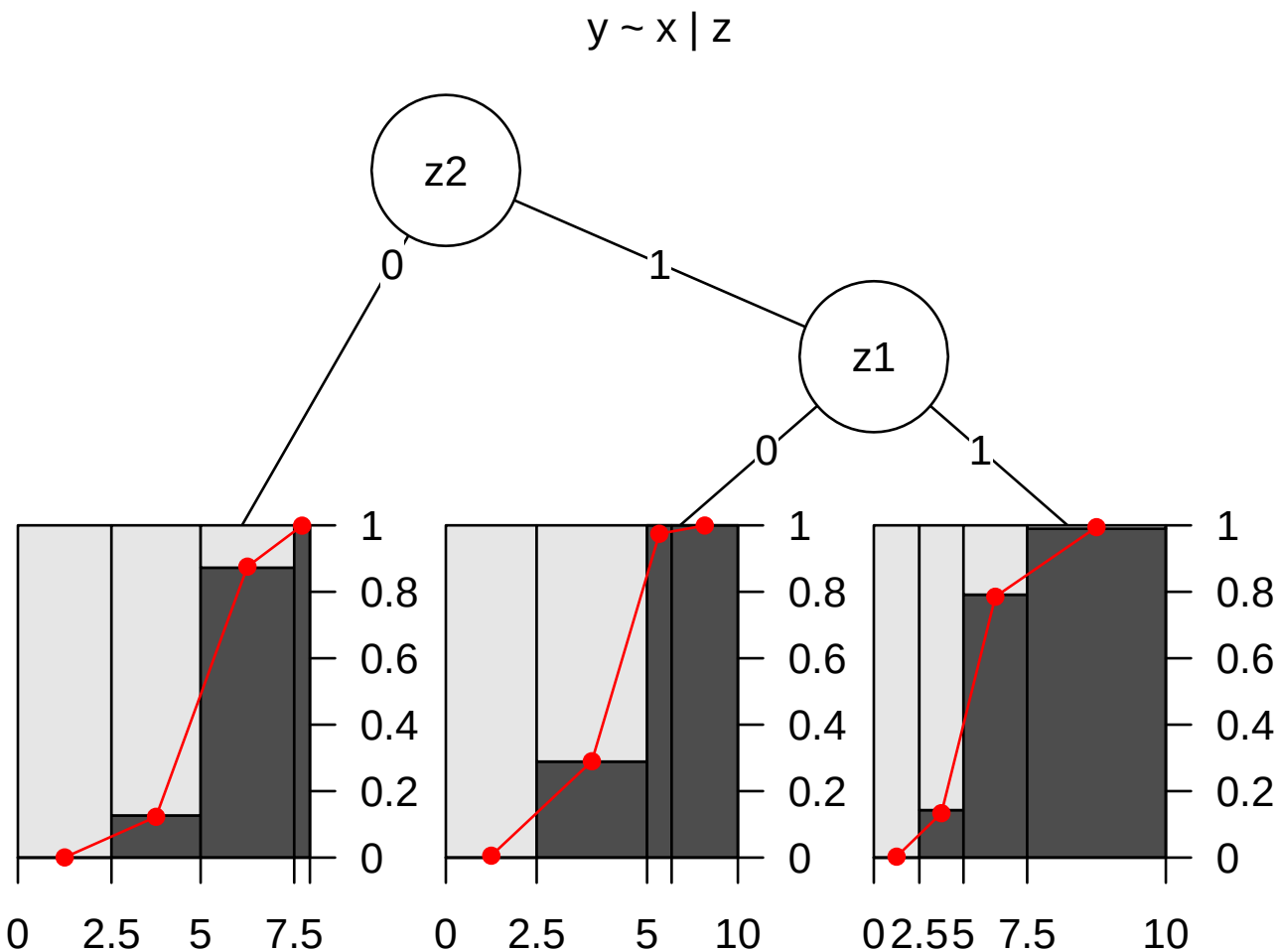
Trees

- Idea of classic regression or classification trees
  - Partition the predictor space into a set of disjoint polytopes by splitting along variables
  - Fit constant in every partition (e.g., the majority class)
- To find split variables and the split points, a loss function is minimized (e.g. a node purity criterion or a negative log-likelihood)
- Tree algorithms yield a piece-wise constant function that approximates the real functional form
- Algorithms usually work in a greedy, hierarchical fashion to find the partitions
- Pruning: Resulting trees are eventually stopped from growing further (pre-pruning) or are reduced to a subtree (post-pruning)

Model Trees

- The idea of trees can be used to allow more more flexibility for logistic models
  - Employ a logistic regression model and
  - Find subsets in the data for which it fits well
- Basically fitting the discrete category model of interest into the leaves of a tree ("model trees")
- Properties:
  - A specific relationship can be captured by the discrete category model
  - Segments based on covariates can capture interactions / additional heterogeneity
  - Number of segments does not need to be known or assumed
  - Segmentation is hard and exclusive (i.e., observation belongs to segment or not)
  - Can aleviate the "high bias" problem of GLM-type models and the "high variance" problem of trees

Example: Logistic Regression Tree



Model Trees - Algorithms

- Many algorithms for model trees have been published over the last 30+ years in machine learning and statistics, including
  - GUIDE, LOTUS (e.g. Loh, 2002; Chan and Loh, 2005)
  - LMT (Landwehr et al., 2005)
  - Functional Trees (Gama, 2004)
  - MOB (Zeileis et al., 2008)
- They differ mainly in
  - Different properties of tree induction
  - Which node models are possible and how they are fitted

Model Trees - MOB

The MOB algorithm of Zeileis et al. (2008) offers a general, coherent framework for model fitting, splitting and pruning.
For discrete category models it works like this

1. Fit a model to all observations Y_i in the current node $R(i \in R)$
2. Assess parameter stability with respect to every possible ordering of the partitioning variable vectors $Z_{ik}, k = 1, \ldots, K; i \in R$
3. Select the covariate associated with the highest significant instability
4. Compute the binary split that optimizes the sum of the objective functions for the daughter nodes
5. Repeat recursively until no split variables are found or any other stopping criterion is fulfilled

MOB - Details I

A parametric model $\mathcal{M}(Y, \theta),\ Y \in \mathcal{Y}$ with l-dimensional parameter vector $\theta \in \Theta$ is to be fitted to all observations $Y_i (i = 1, \ldots, n)$.
To get the parameter estimates $\hat{\theta}$ we minimize an objective function $\rho(Y, \theta)$, i.e.

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{N} \rho(Y_i, \theta).$$

If there is a true θ_0, asymptotically $\sqrt{n}(\hat{\theta} - \theta_0) \sim \mathcal{N}(0, V(\theta_0))$ under weak regularity conditions.
This is Huber's M-estimation approach and includes

- Maximum Likelihood (ML)
- Least Squares (OLS or WLS)
- Quasi-ML
- General M-Estimation

MOB - Details II

- $\mathcal{M}(Y, \theta)$ for the whole data set may be less good the separate models for subsets defined by $\mathcal{Z}_|$.
- One can assess parameter stability w.r.t. to Z_j by means of Generalized M-Fluctuation tests for any ordering ofZ_j
- This is done by applying a scalar function $\lambda(.)$ to the empirical fluctuation process of cumulative deviations of the score function with respect to the ordering permutation, $W_j(t, \hat{\theta})$.
- Zeileis (2006) showed that under parameter stability, $W_j(.) \sim W^0(.)$ (a k-dimensional Brownian bridge).
- One can use $\lambda(W_j)$ as a test statistic. The Null distribution is asymptotic distribution of $\lambda(W_0)$ (the supLM statistics of Andrews (1993) is used for metric covariates, χ^2-statistic for categorical and ordinal covariates (Hjort & Koning (2002), Merkle & Zeileis (2014)).

MOB - Details IV

- 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 the sum of the objective function for two separate fits of the model in the two segments is optimized, i.e.

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

  for two rival segmentations by an exhaustive search over all pairwise comparisons of possible splits.
- The usage of p-values (Bonferroni-corrected) serves as pre-pruning/regularisation of the tree.

Logistic Models - Details I

Logistic models as we use them are defined by the following combination

- A distribution assumption and a given expected value from an exponential family: $p_{ik} = E(Y_i = k) = \mu_{ik}, Y_i \sim \text{Multinomial}(p_{i1}, \ldots, p_{iK}, n), p_{ik} \geq 0, \sum_k p_{ik} = 1.$
- A link function, that utilizes the logit (which is the canonical link for this exponential family)
- Linear predictor for the linked expected value $\text{logit}(\mu_{ik}) = x_{ik}^\top \beta_k$

Logistic Models - Details II

The logit links used can be different:

- Baseline logit: Category K is the baseline (used in binary and multinomial logits models).

$$\log\left(\frac{\mu_{ik}}{\mu_{iK}}\right) = \alpha_k + x_i^\top \beta_k, \quad k = 1, \ldots, K-1$$

- Cumulative Logit: Category k or less is contrasted with all categories larger k (used in ordered logit models)

$$\log\left(\frac{\sum_{k=1}^{l} \mu_{ik}}{1 - \sum_{k=1}^{l} \mu_{ik}}\right) = \alpha_k + x_i^\top \beta, \quad k = 1, \ldots, K-1$$

Logistic Models - Details III

Logistic models can be estimated by maximum likelihood from an exponential family probability mass function $\rho(y_i, \theta) = f(y_i; \theta)$ by doing

$$\arg\max_{\theta} \sum_{i=1}^{n} \log f(y_i; \theta)$$

The log-likelihood of the above has the form (in the canonical representation)

$$l(\theta) = \langle s, \theta \rangle - c(\theta)$$

with s being a r-dimensional vector statistic (the vector of natural or sufficient statistics), θ a r-dimensional vector parameter (natural or canonical parameter) and

$$\langle s, \theta \rangle = \sum_{r=1}^{R} s_r \theta_r.$$

Logistic Models - Details IV

We are interested in some linear submodel that is defined by a linear predictor for the cannonical parameter and is of the form $\theta = X\beta$ which has the log likelihood

$$\begin{aligned} l_{sub}(\beta) &= \langle s, X\beta \rangle - c(X\beta) \\ &= \langle X^\top s, \beta \rangle - c_{sub}(\beta) \end{aligned}$$

Here, X is the (fixed) model or design matrix and β are the parameters of interest. The linear submodel is itself an exponential family with the natural statistic being $X^\top s$ and canonical parameter β.

In general the log-likelihood of the exponential family is strictly concave in this natural characterisation and thus the MLE will be unique if it exists and it will lie in the interior of the parameter space.

Logistic Models - Separation I

- There are conditions under which the MLE in a discrete exponential family is not finite (i.e. does not exist)
- In logit models the most prevalent condition is referred to as (quasi-)complete separation
- Separation occurs formally when there is a vector β such that for all i that belong to class r and for $r, t = 1, \ldots, K, r \neq t$

$$(\beta_r - \beta_t)^\top x_i \geq 0$$

  where strict inequality is called complete separation and the equality (for at least one triplet i, r, t) quasi-complete separation.
- It means for a (linear combination of some subset of the) predictor a vector β either seperates points so that the points are correctly allocated or fall onto the separating hyperplane.

Logistic Models - Separation II

What to do in general?

1. Leave it
2. Use a bias corrected version
3. Set a prior on the estimate

The problem separation poses for us is

- Numerical Issue 1: Point estimate and errors will be finite but large and inaccurate
- Numerical Issue 2: Score calculated from the solution may fluctuate artificially
- Inferential issues: Frequentist inference can have pretty bad properties

We need a way to

- Diagnose the types of separation
- (Eventually) Fix inference issues

Generic Direction of Recession I

How to diagnose? Solve a linear program, but

- Checking for quasi-complete separation directly is numerically hard
- We want some theory to exploit for the checks that helps also fix inferential issues

Geyer (2009) defines the generic direction of recession (GDOR), a vector δ such that

$$t \mapsto l(\beta + t\delta)$$

is a strictly increasing function for each fixed β and there exists a $\hat{\beta}$ such that

$$\lim_{t \to \infty} l(\hat{\beta} + t\delta) = \sup_{\beta \in R^R} l(\beta)$$

In this case the MLE can be understood as a $\hat{\beta}$ sent to infinity in the direction δ.

Generic Direction of Recession II

- MLE does not exist in the conventional sense if and only if a GDOR exists (Geyer, 2009).
- This implies that the MLE for β exists and is unique if and only if the GDOR is the null vector.
- This implies further that (forthcoming)
  - If an element in δ is zero than the corresponding element in β exists and is unique
  - If an element in δ is non-zero than this is a GDOR and the corresponding element in β is infinite into the direction set up by δ
- So we can say that if we have separation we must find a GDOR somewhere.

Generic Direction of Recession III

- For detecting separation by the GDOR we exploit that the parameter space spanned by a polyhedron and the problem of separation leads to an MLE at the boundary of the polyhedron.
- An exact linear program that looks for the GDOR can check for whether this occurs for a given model and data set by considering the problem in a half-space representation.
- Half-spaces are parts into which a hyperplane divides an affine space specified by a linear inequality derived from the equation of the hyper plane (can be open and closed).
- In that sense a GDOR exists if the vectors either are in the closed half-space (quasi-complete separation) or in the open half-space (complete separation).
- We can look for half-spaces and thus diagnose the cases with infinite precision rational arithmetic.

LORET: Logistic Regression Trees

All that we said so far culminates in LORET which is

- MOB for models of the discrete category fashion
- With a built-in separation check and handling based on the GDOR
- (Eventually) With GDOR based valid frequentist inference

LORET: Implementation

The implementation is based on `mob()` from partykit and uses a number of workhorses for the logistic model fitting

- `clm()` from ordinal for ordinal models
- `multinom()` from nnet for multinomial models with subject specific covariates
- `mlogit()` from mlogit for general multinomial models

For separation checks we have the object-oriented wrapper function `separation_check()` that relies on `linearity()` and `lpcdd()` from the rcdd package.
For the GDOR based inference we plan to develop our own package (which is doing this for any type of discrete exponential family and corresponding R base functions).

Voter Targeting: Political Campaigns

- Political campaigning is a multi-million dollar business and increasingly so.
- A large part of that money is spent on mobilizing voters to turnout
- Rusch et al. (2013) proposed a framework of logistic regression trees for prediction and identification of likely voters and efficient campaign ressource allocation
- Data and target were get-out-the-vote in a presidential election
- Smaller campaigns or elections often have a harder time mobilizing supporters

Voter Targeting: Data

- Data are from Ohio for the 2004 US primary elections
- Proprietary data set that was retrieved from Aristotle, Inc.
  - Sample consists of 19640 people
  - Voting records from 1990 to 2004
  - Demographic, behavioural and institutional covariates
    - Age in days, gender
    - Party affiliation, party makeup of household, rank in household,
    - Income, education
    - Donation to various causes (health, environment etc.)
    - Federal contribution in certain years
    - Computer owner, home owner
  - Target: Vote in Primary 2004 (D/R/I/N)
- Standard targeting variables for the logistic model are primarily
  - Individual historic voting records
  - Age

Targeting Voters: Data

- Standard model and a number of variables for which we do not really know their effect
- Perfect for model trees
- But there is a chance that partitioning will lead to nodes with separation

For instance, look at the outcome and the party mix of one's household

	PARTY_MIX							
PPP04	unknown	allR	allD	allOther	noneD	allRorD	noneR	AllRorDorL
N	87	22	10	332	74	2	49	4
D	14	1	95	2	0	15	57	6
R	9	93	0	0	65	11	0	7
Y	8	0	1	26	7	0	3	0

Targeting Voters: Separation Check

We can check if that is really separation with the function `separation_check()`. For

```
R> model <- multinom(PPP04~PARTY_MIX,data=voterss)
R> separation_check(data,model)
```

```
Separation for Category:  D R Y

Overlap for Category:
```

Conversely for

```
R> model <- multinom(PPP04~ageY,data=voterss)
```

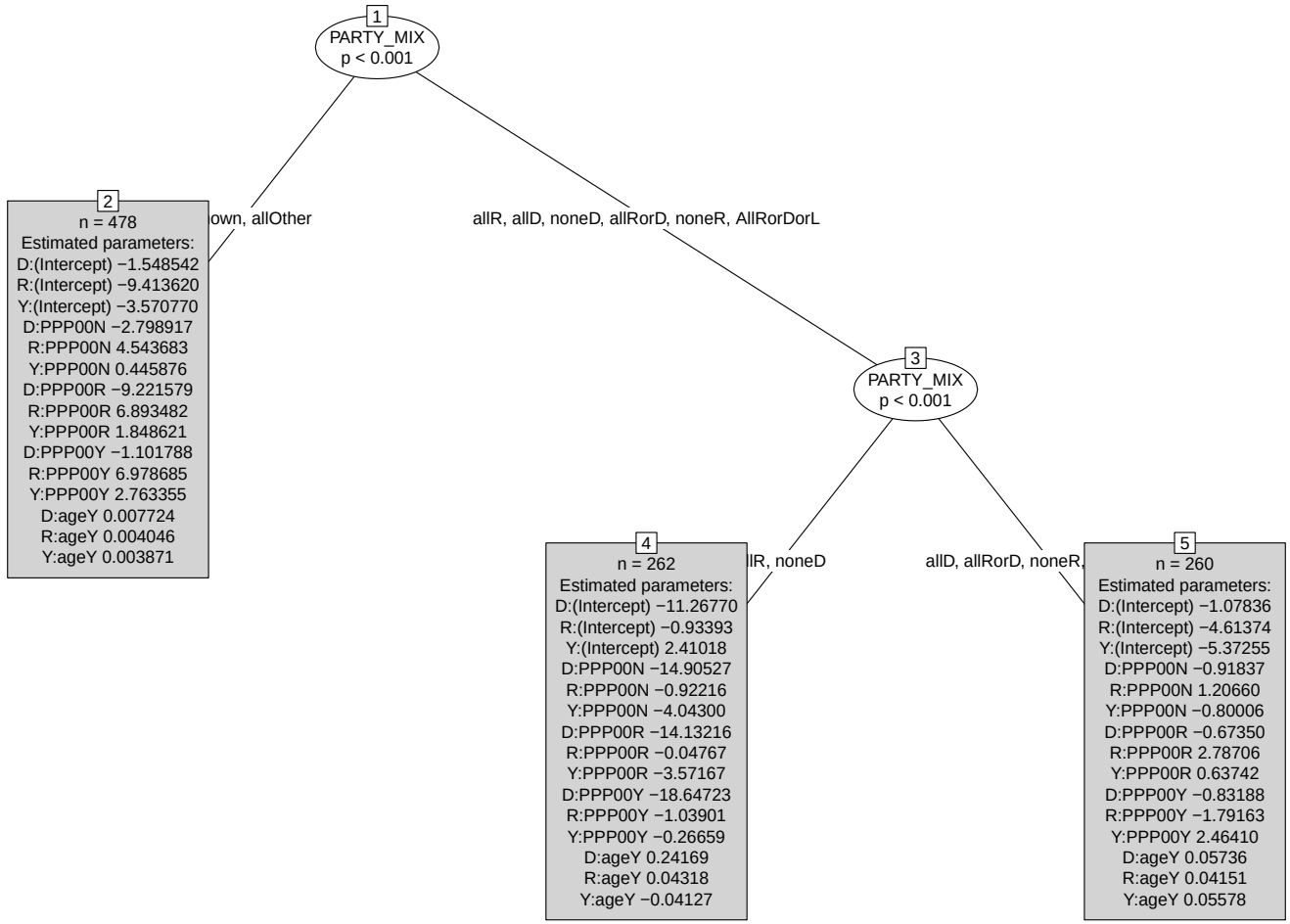
```
Separation for Category:

Overlap for Category:  D R Y
```

Targeting Voters: Multinomial LORET

We use `separation_check()` internally in the mob wrapper to decide whether we should split further or not. If separation is detected, no further splitting occurs.

```
R> modt <- multinomtree(PPP04~PPP00+ageY|PARTY_MIX+SEX+HOH_FLAG, data=vot
```

Targeting Voters: Multinomial LORET



Targeting Voters: Notice Anything Odd?

Node 5:

```
Call:
multinom(formula = PPP04 ~ PPP00 + ageY)

Coefficients:
  (Intercept)  PPP00N  PPP00R  PPP00Y    ageY
D      -1.078 -0.9184 -0.6735 -0.8319 0.05736
R      -4.614  1.2066  2.7871 -1.7916 0.04151
Y      -5.373 -0.8001  0.6374  2.4641 0.05578

Std. Errors:
  (Intercept) PPP00N PPP00R PPP00Y    ageY
D      0.6583 0.3792 0.5242  1.374 0.01234
R      1.5926 1.1299 1.1610  9.070 0.02325
Y      2.3486 1.4851 1.5142  1.879 0.03838

Residual Deviance: 386.6
AIC: 416.6
```

There is separation for “R”/”Y”. I missed it but the check caught it.

```
          votss$PPP00
votss$PPP04  D  N  R  Y
          N 13 43  8  1
          D 80 69 21  3
          R  1  8  9  0
          Y  1  1  1  1
```

Conclusion

- We presented LORET (LOgistic REgression Trees), a combination of discrete category models with mob.
- The distinguishing characteristics are
  - Allows to fit many different logistic models (i.e., all from glm, mlogit, multinom and ordinal)
  - Mob algorithm delivers a coherent recursive partitioning framework for fitting, splitting and pruning
  - Existence of the MLE and separation is checked for and handled in a general fashion to avoid artefacts
  - Object-oriented, flexible, fast implementation for R
- TO DO:
  - A single wrapper for it all (`loret()`)
  - Faster check for separation
  - Inference based on the GDOR (and possibly corrected for the tree estimation)

References

- Andrews, D. (1993) Tests for Parameter Instability and Structural Change with Unknown Change Point, *Econometrica*, 61:821-856.
- Chan, K., and Loh, W. (2004) LOTUS. An Algorithm for Building Accurate and Comprehensible Logistic Regression Trees, *Journal of Computational and Graphical Statistics*, 13:826-852.
- Chaudhuri, P., Lo, W., Loh, W., and Yang, C. (1995) Generalized Regression Trees. *Statistica Sinica*, 5:641-666.
- Gama, J. (2004) Functional trees. *Machine Learning*, 55:219-250.
- Geyer C. (2009). Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, 3:259-289.
- Hjort, N. and Koning, A. (2002) Tests for Constancy of Model Parameters Over Time, *Nonparametric Statistics*, 14:113-132
- Merkle, E., Fan, J. and Zeileis, A. (2014). Testing for Measurement Invariance with Respect to an Ordinal Variable. *Psychometrika*, 79:569-584

References

- Landwehr, N., Hall, M., and Eibe, F. (2005) Logistic Model Trees. *Machine Learning*, 59:161-205.
- Loh, W. (2002) Regression Trees with Unbiased Variable Selection and Interaction Detection. *Statistica Sinica*, 12:361-386.
- McCullough, P. & Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, 2nd Ed.
- Rusch, T., Lee, I., Hornik, K., Jank, W. and Zeileis, A. (2013). Influencing elections with statistics: Targeting voters with logistic regression trees. *Annals of Applied Statistics*, 7:161-1639.
- Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17:492-514.

Thank you for inviting a Vorarlberger!

Thomas Rusch
Competence Center for Empirical Research Methods
WU Wirtschaftsuniversität Wien
Welthandelsplatz 1, A-1020 Wien

email: `thomas.rusch@wu.ac.at`

URL: `http://wu.ac.at/methods/team/dr-thomas-rusch`