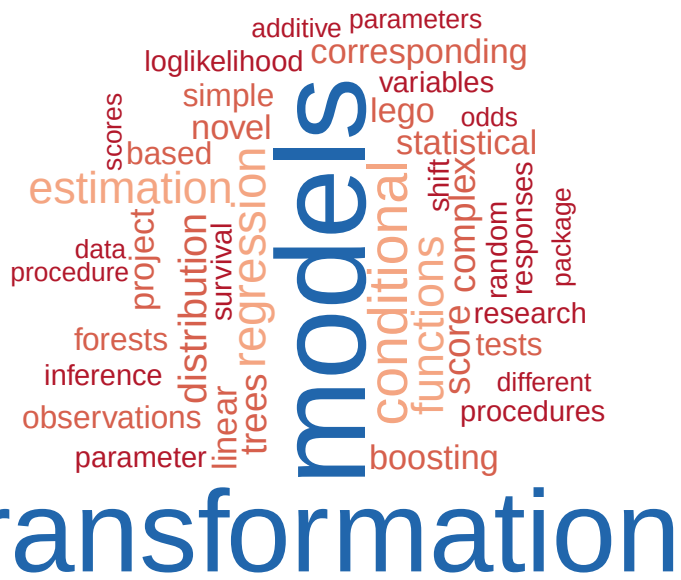
University of Zurich[UZH]

EBPI Epidemiology, Biostatistics and Prevention Institute

Transformation Models: Pushing the Boundaries

Torsten Hothorn



Università degli Studi di Brescia, 2022-09-27

On the Menu Today

- What are transformation models?
- What are they capable of?
- What is their connection to trees and forests?

Transformation Models

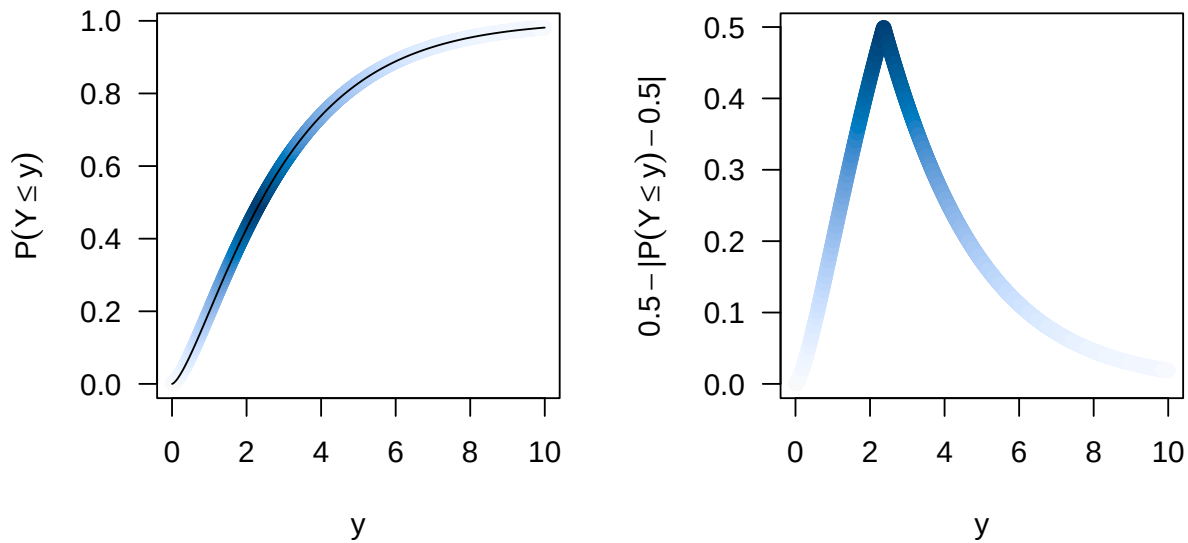
Models for Distributions, not Means

Regression:

$$\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x})$$

not only

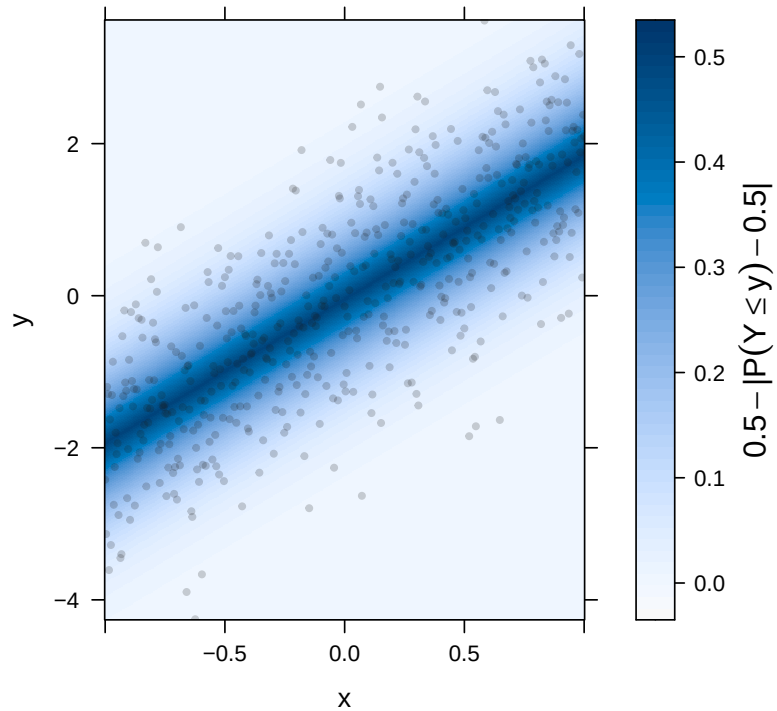
$$\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$$



The Normal Linear Regression Model

$$Y = \alpha + \boldsymbol{x}^\top \gamma + \sigma Z, \quad Z \sim \mathrm{N}(0, 1)$$

- everything but “normal”
- most special case
- $Y \mid \boldsymbol{x} \sim \mathrm{N}(\alpha + \boldsymbol{x}^\top \gamma, \sigma^2)$
- no way escaping normal land



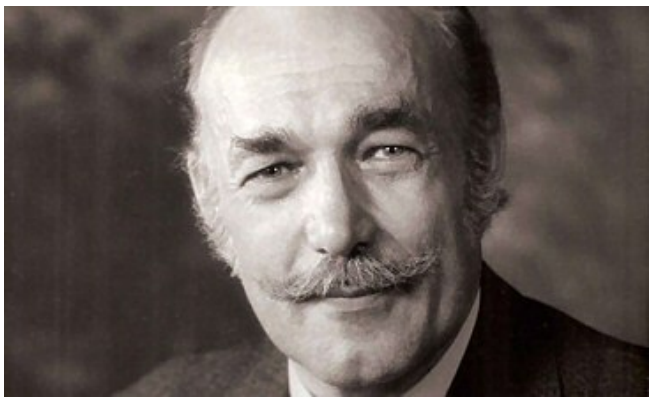
Generalisation I

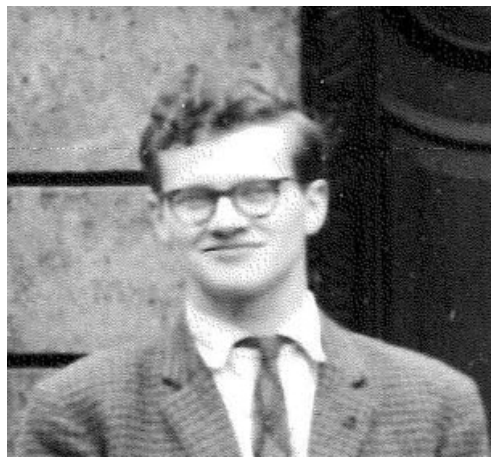
$$
\begin{aligned}
Y &= \alpha + \boldsymbol{x}^\top \gamma + \sigma Z, \quad && Z \sim \mathsf{N}(0, 1) \\
\iff \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) &= \alpha + \boldsymbol{x}^\top \gamma \quad , && Y \mid \boldsymbol{X} = \boldsymbol{x} \sim \mathsf{N}(,) \\
\hookrightarrow g(\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})) &= \alpha + \boldsymbol{x}^\top \gamma \quad , && Y \mid \boldsymbol{X} = \boldsymbol{x} \sim \mathsf{ExpFam}(,)
\end{aligned}
$$

Generalized Linear Models

By J. A. NELDER and R. W. M. WEDDERBURN

Rothamsted Experimental Station, Harpenden, Herts





Generalisation II

$$
\begin{aligned}
Y &= \alpha + \boldsymbol{x}^\top \gamma + \sigma Z, \quad Z \sim \mathrm{N}(0,1) \\
\iff \frac{Y - \alpha}{\sigma} &= \boldsymbol{x}^\top \frac{\gamma}{\sigma} + Z, \quad Z \sim \mathrm{N}(0,1) \\
\iff \frac{Y - \alpha}{\sigma} &= \boldsymbol{x}^\top \beta + Z, \quad Z \sim \mathrm{N}(0,1) \\
\hookrightarrow h(Y) &= \boldsymbol{x}^\top \beta + Z, \quad Z \sim \mathrm{N}(0,1) \\
\hookrightarrow h(Y) &= \boldsymbol{x}^\top \beta + Z, \quad Z \sim
\end{aligned}
$$

Transformation models, $Z \in \mathbb{R}$ with absolute continuous log-concave density f_Z, $h : \mathbb{R} \to \mathbb{R}$ nondecreasing

An Analysis of Transformations (1964)

An Analysis of Transformations

By G. E. P. Box and D. R. Cox

University of Wisconsin *Birkbeck College, University of London*

[Read at a Research Methods Meeting of the Society, April 8th, 1964, Professor D. V. Lindley in the Chair]

Conceptually more powerful but, at the time, hard to compute and thus restricted to

$$h(y \mid \lambda) = \begin{cases} \frac{y^\lambda + 1}{\lambda} & \lambda > 0 \\ \log(y) & \lambda = 0 \end{cases} \quad \text{with } Z \sim \mathsf{N}(0, \sigma^2)$$

"Box-Cox" power transformation

Conditional Distribution Functions

$$\begin{aligned} h(Y) &= \boldsymbol{x}^\top \beta + Z \\ \hookrightarrow \mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) &= F_Z(h(y) - \boldsymbol{x}^\top \beta) \end{aligned}$$

also allows discrete models via step-function h

Linear transformation models: Proportional hazards, proportional odds, ...

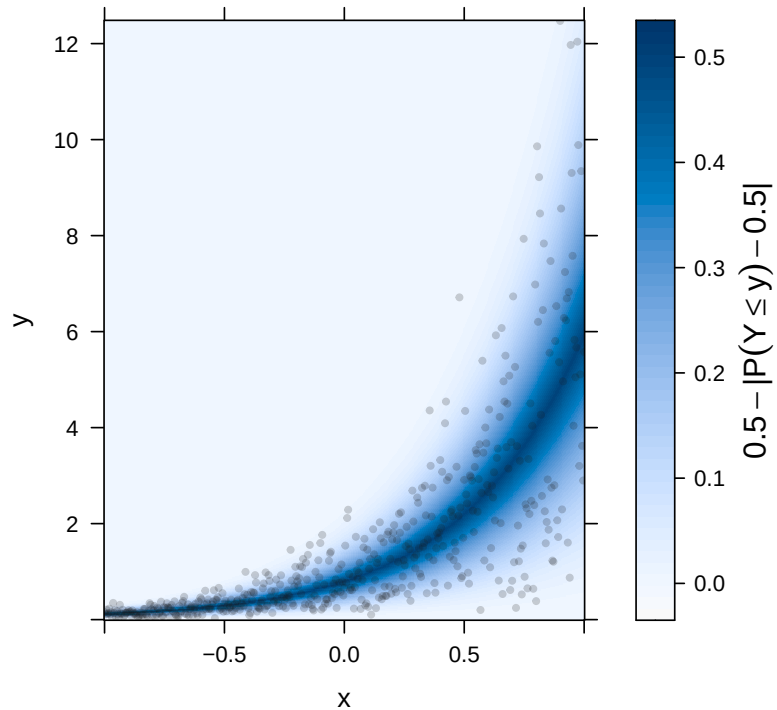
aka Probabilistic index models:

$$\mathbb{P}(Y_1 < Y_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = m_Z((\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top \beta)$$

Cox Models

$$\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) \quad = \quad 1 - \exp(-\exp(h(y) + \boldsymbol{x}^\top \beta))$$

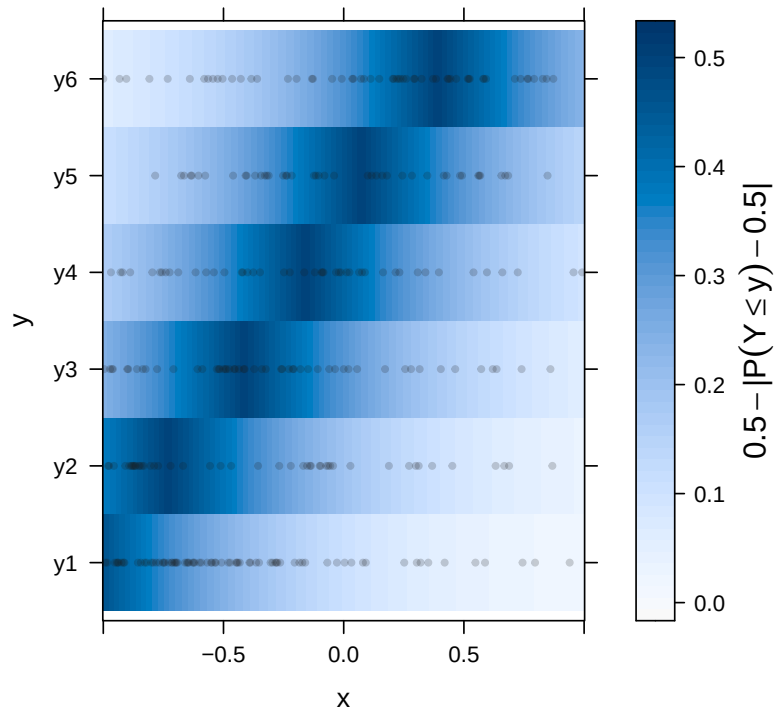
- the most prominent transformation model
- h is baseline log-cumulative hazard
- $\boldsymbol{x}^\top \beta$ is log-hazard ratio
- partial likelihood profiles out h
- np score test: log-rank
- no censoring in *model*, only in *likelihood*



Proportional-odds Models

Ordinal outcome at categories $y_1 < y_2 < \cdots < y_K$
$h(y_k) = \vartheta_k, k = 1, \ldots, K-1$ with $F_Z = \text{expit}$

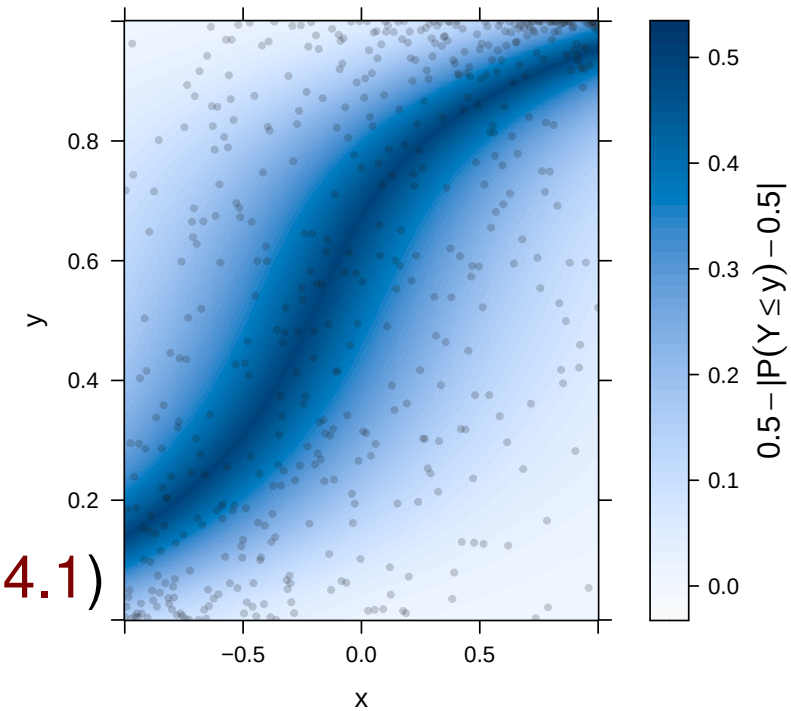
- Proportional-odds model
- Popular for ordinal data analysis
- $\boldsymbol{x}^\top \beta$ is log-odds ratio
- *Simultaneous* estimation of h (via ϑ_k) and β



Conditional Outcome Logistic Regression / Ordinal Regression Model

$$\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) \quad = \quad \text{expit}(h(y) - \boldsymbol{x}^\top \beta)$$

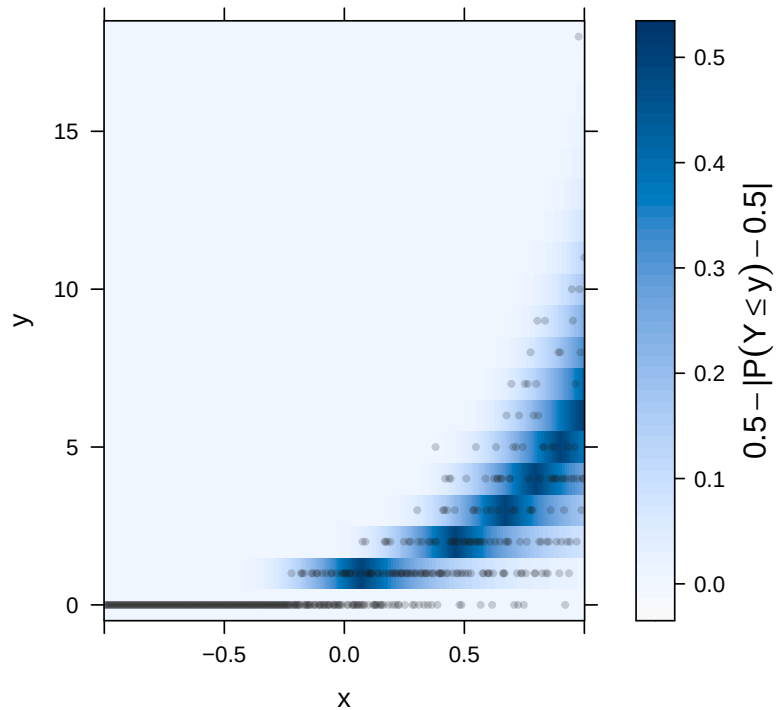
- Continuous proportional-odds model
- h is baseline log-odds function
- $\boldsymbol{x}^\top \beta$ is log-odds ratio
- np score test: Wilcoxon
- parametric: Colr (10.12688/f1000research.12934.1)
- nonparametric: orm (10.1002/sim.7433)



Count Transformation Models

Count outcome $Y \in \{0, 1, 2, \dots\}$
$h(k) := h(\lfloor y \rfloor) \quad \forall k \leq y < k + 1$ with $h : \mathbb{R} \to \mathbb{R}$

- Smooth $h(\cdot \mid \vartheta)$ evaluated discretely
- More flexible than Poisson/NB
- Discrete count likelihood
- 10.1111/2041-210X.13383



Distribution Regression

Quantile regression:

$$Q(\tau \mid \boldsymbol{X} = \boldsymbol{x}) \quad = \quad \alpha(\tau) + \boldsymbol{x}^\top \boldsymbol{\delta}(\tau)$$

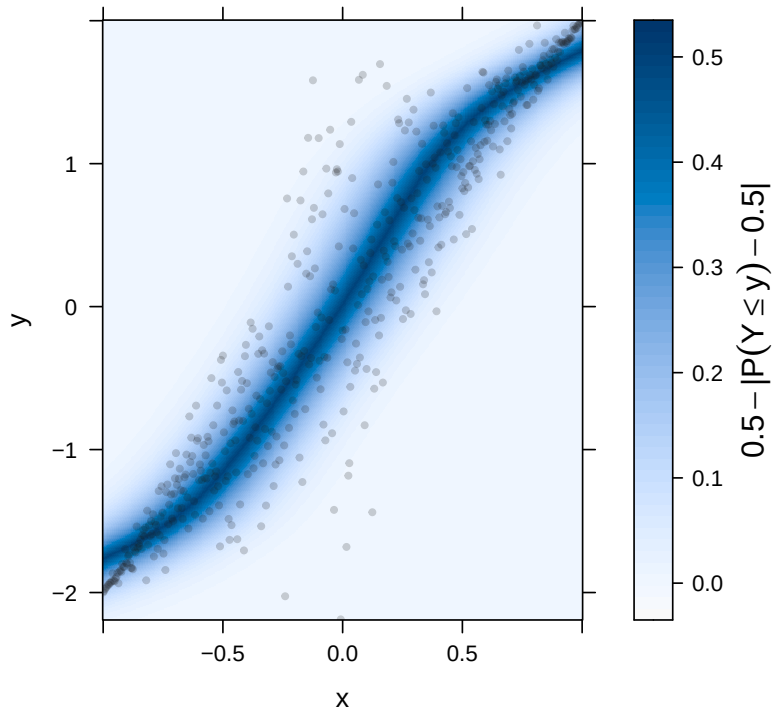
Distribution regression:

$$F_Z^{-1}(\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x})) \quad = \quad h(y) - \boldsymbol{x}^\top \boldsymbol{\beta}(y)$$

i.e. quantile regression on the log-cumulative hazard / odds or probit scale

Distribution Regression

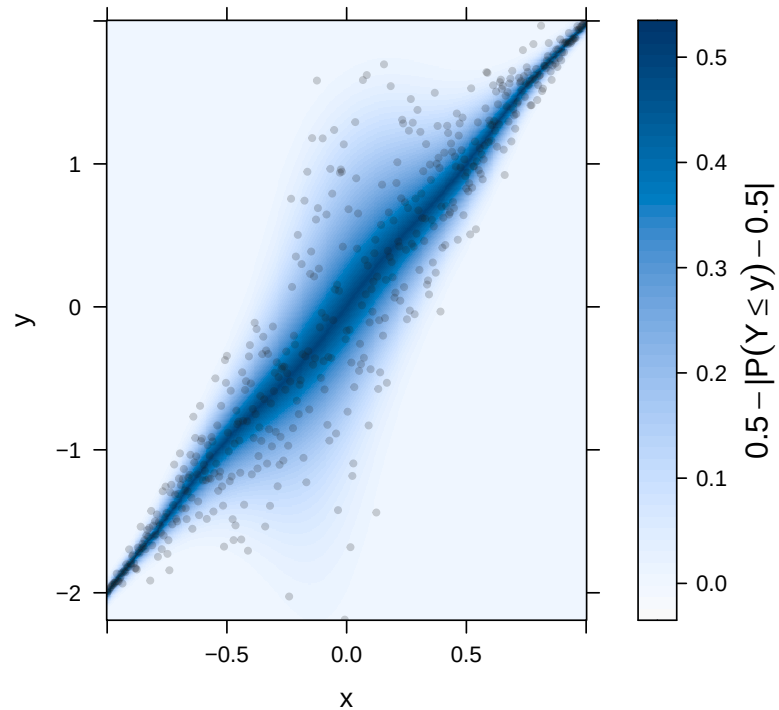
- Often (re-)discovered
- $\text{logit}(\mathbb{E}(I(Y \leq c)) = \alpha_c + \boldsymbol{x}^\top \boldsymbol{\beta}_c$
- Full likelihood possible
- Splines for $h(y)$ and $\boldsymbol{\beta}(y)$



Conditional Transformation Models

$$\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) \quad = \quad F_Z(h(y \mid \boldsymbol{x}))$$

- Practically no assumptions
- Tensor-product spline bases for y and $\boldsymbol{x}$
- Full likelihood possible
- 10.1111/rssb.12017



Transformation Likelihood

Log-likelihoods

Observed $(\underline{y}, \bar{y}] \subset \mathbb{R}$:

$$\log[F_Z\{h(\bar{y} \mid \boldsymbol{x})\} - F_Z\{h(\underline{y} \mid \boldsymbol{x})\}]$$

This includes discrete and censored observations and, via $(y_{(k)}, y_{(k+1)}]$, the nonparametric likelihood.

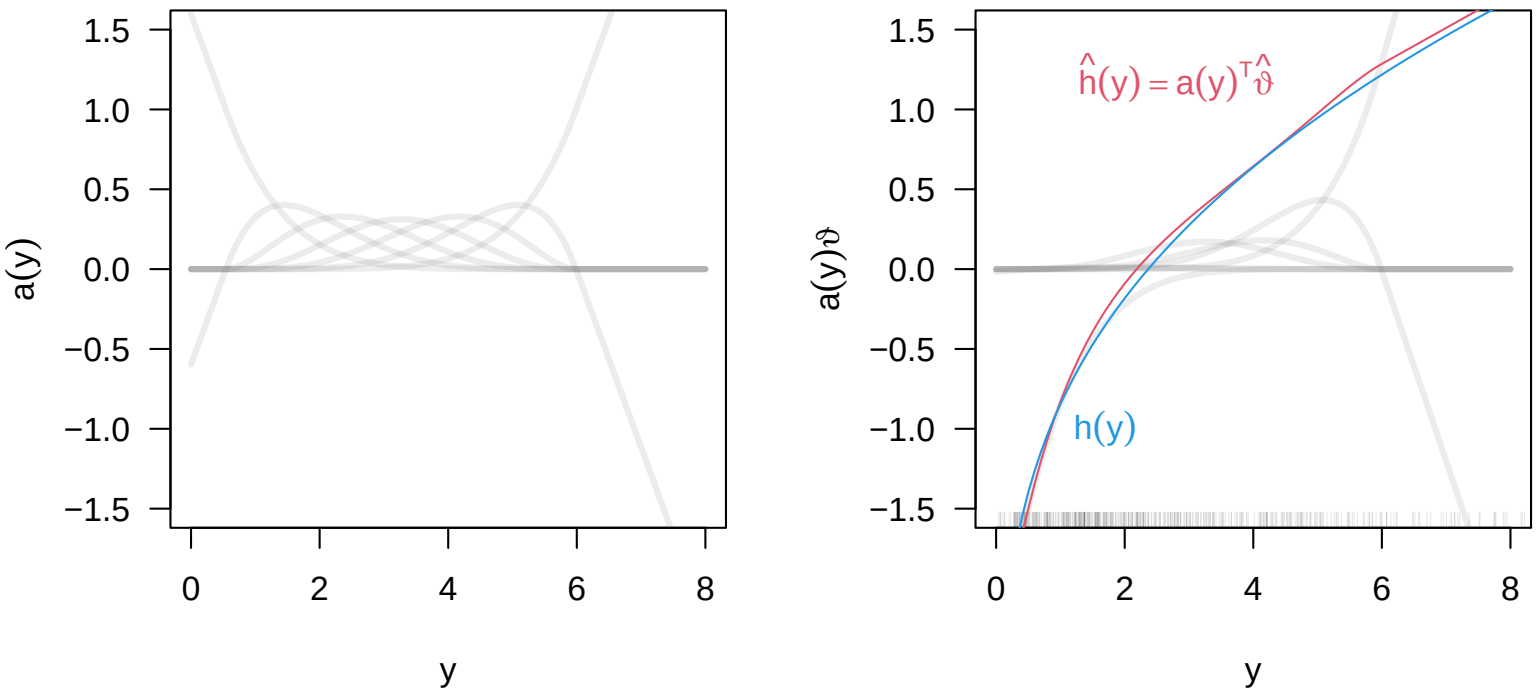
Observed $Y \in \mathbb{R}$:

$$\approx \log[f_Z\{h(y \mid \boldsymbol{x})\}] + \log\{h'(y \mid \boldsymbol{x})\}$$

10.1111/sjos.12291

Parameterisation

- *h* typically swept under the carpet
- More fun: parameterise $h(y \mid \vartheta) = \boldsymbol{a}(y)^\top \vartheta$ in terms of $\vartheta \in \Theta$ (a la 10.1080/15598608.2013.772835)
- Estimate all parameters *simultaneously* (via ML)
- Does it hurt? Not really: 10.1002/sim.8425



Model Baking

- **Ingredients:** Take F_Z, parameterise $h(y \mid \vartheta)$, define impact of $\boldsymbol{x}$ via $\boldsymbol{x}^\top \beta$, $\boldsymbol{x}^\top \beta(y)$, $h(y \mid \boldsymbol{x})$
- **Mix:** Data defines likelihood function (handles discrete and continuous observations, censoring)
- **Oven:** Optimise, get $\hat{\vartheta}, \hat{\beta}$ + limiting distribution
- **Serve:** Interpret/interrogate fully specified model

And now for some new recipes...

Pushing the Boundaries

Correlated Observations

Mixed-effects transformation models via conditional distribution

$$\mathbb{P}\left(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{U} = \boldsymbol{u}, \xi\right) \quad = F_Z\left(h(y \mid \vartheta) - \boldsymbol{x}^\top \beta - \boldsymbol{u}^\top \xi\right)$$

with normal random effects $\xi \sim \mathrm{N}_q(\mathbf{0}, \Sigma)$

Integrate conditional likelihood wrt random effects, using the fabulous TMB package

10.32614/RJ-2021-075 10.1093/biostatistics/kxab045

High Dimensions

$$\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) \quad = \quad F_Z(h(y \mid \vartheta) - \boldsymbol{x}^\top \beta)$$

with $\beta \in \mathbb{R}^p$, with p being large

Add L_1, L_2, ... penalty for β to likelihood, using the fabulous CVXR package

Transformation ridge, lasso, elastic net, etc. 10.32614/RJ-2021-054

Additive Models

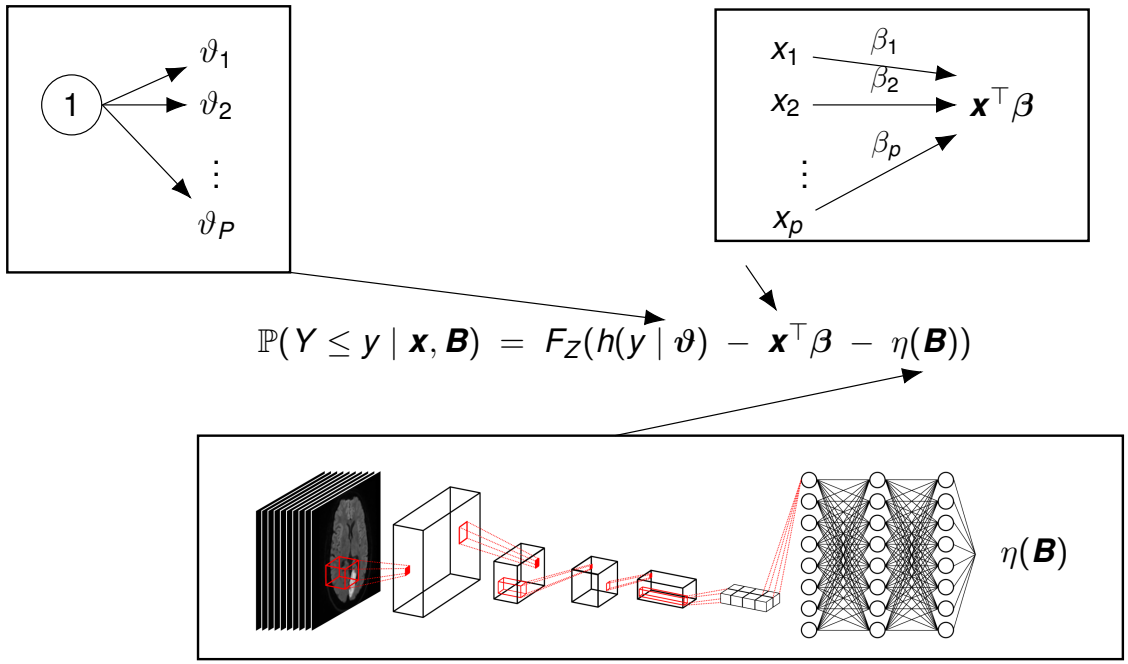
$$\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) \quad = \quad F_Z \left(h(y \mid \vartheta) - \sum_{j=1}^{J} f_j(\boldsymbol{x}) \right)$$

aka “transform-both-sides”

Use connection to mixed models and leverage mgcv infrastructure; in the making

Alternative: Boosting via mboost, 10.1007/s11222-019-09870-4

Unstructured Information



"Deep" transformations, 10.1016/j.patcog.2021.108263
10.1007/978-3-030-86523-8_1

Multivariate Transformation Models

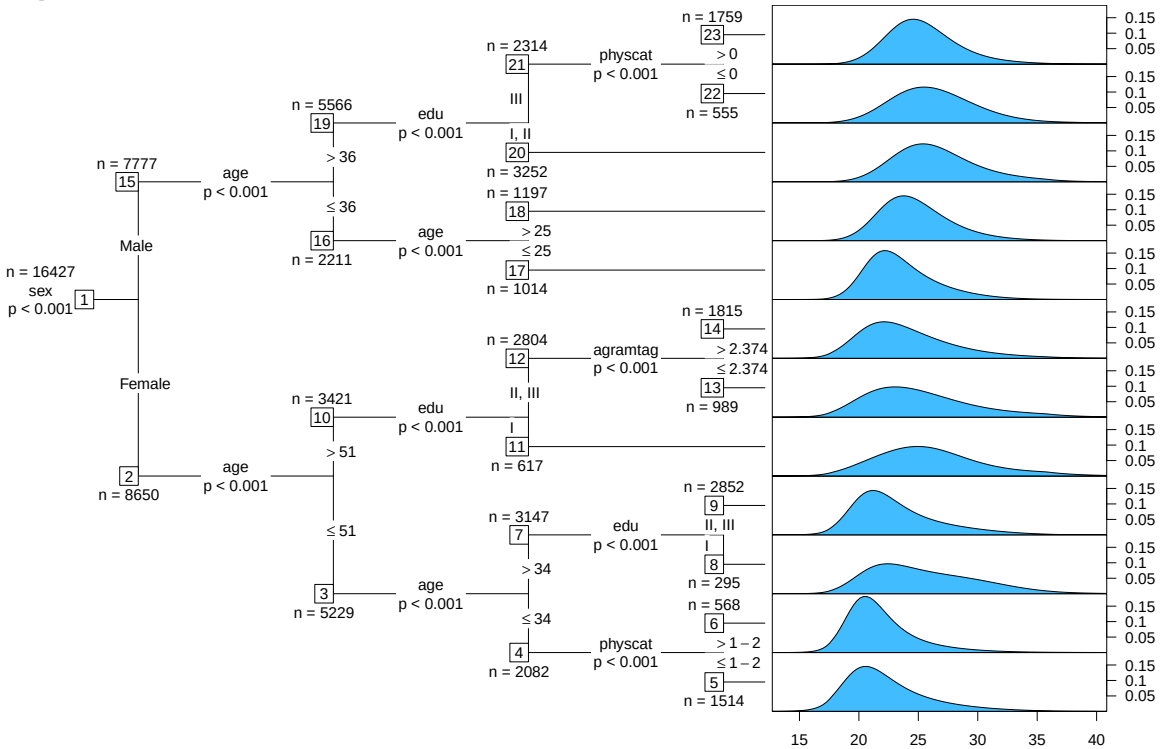
$$\mathbb{P}\left(\bigcap_{j=1}^{J} Y_j \leq y_j \mid \boldsymbol{X} = \boldsymbol{x}\right) = F_{\boldsymbol{Z}}(h_1(y_1 \mid \boldsymbol{x}) + h_2(y_2 \mid \boldsymbol{x}) + \lambda_{21}(\boldsymbol{x}) h_1(y_1 \mid \boldsymbol{x}) + \ldots + h_J(y_J \mid \boldsymbol{x}) + \sum_{j=1}^{J-1} \lambda_{Jj}(\boldsymbol{x}) h_j(y_j \mid \boldsymbol{x}))$$

with $\boldsymbol{Z} \sim \mathrm{N}_J(\boldsymbol{0}, \boldsymbol{I})$. Correlations depend on $\boldsymbol{x}$ through $\lambda_{jj'}(\boldsymbol{x})$

Connection to Gaussian copulas and normalising flows, 10.1111/sjos.12501

Trees and Forests

Model-based recursive partitioning (MOB) based on transformation models, trees and forests for distributional regression



10.1515/ijb-2019-0063, 10.1177/0962280219862586

Random Forest

What is a random forest (in general, not only B&C)?

Classical:

$$\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \chi$$

Here:

$$\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = f(y \mid \boldsymbol{x}), \quad \forall \boldsymbol{x} \in \chi$$

RF often understood as being “nonparametric” but life is easier when we have a *model*.

Parametric (!) Setup

Unconditional model for response

$$\mathbb{P}_{Y,\Theta} = \{\mathbb{P}_{Y,\boldsymbol{\vartheta}} \mid \boldsymbol{\vartheta} \in \Theta\}$$

Assumption: Regression model belongs to this family:

$$\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}} = \mathbb{P}_{Y,\boldsymbol{\vartheta}(\boldsymbol{x})}$$

Task: Estimate $\boldsymbol{\vartheta}$ function

Likelihood Contributions

"Learning" data $(y_i, \boldsymbol{x}_i), i = 1, \ldots, N$ plus family $\mathbb{P}_{Y,\Theta}$ defines likelihood function

$$\ell_i : \Theta \to \mathbb{R}$$

$\ell_i(\boldsymbol{\vartheta}(\boldsymbol{x}_i))$ gives the likelihood for observation i with candidate parameters $\boldsymbol{\vartheta}(\boldsymbol{x}_i)$

Handle censoring and truncation appropriately here

Adaptive Local Likelihood Estimators

$$\hat{\boldsymbol{\vartheta}}^N(\boldsymbol{x}) := \arg\max_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} w_i^N(\boldsymbol{x}) \ell_i(\boldsymbol{\vartheta})$$

Conditioning works via weight functions $w_i^N(\boldsymbol{x})$ only

Unconditional Maximum Likelihood

$$\hat{\boldsymbol{\vartheta}}_{\text{ML}}^{N} := \arg\max_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} \ell_i(\boldsymbol{\vartheta})$$

Trees

$$\chi = \dot{\bigcup_{b=1,\ldots,B}} \mathcal{B}_b$$

$$\begin{aligned} w^N_{\text{Tree},i}(\boldsymbol{x}) &:= \sum_{b=1}^{B} I(\boldsymbol{x} \in \mathcal{B}_b \wedge \boldsymbol{x}_i \in \mathcal{B}_b) \\ \hat{\boldsymbol{\vartheta}}^N_{\text{Tree}}(\boldsymbol{x}) &:= \arg\max_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} w^N_{\text{Tree},i}(\boldsymbol{x}) \ell_i(\boldsymbol{\vartheta}) \end{aligned}$$

Forests

$$\chi = \dot{\bigcup_{b=1,\ldots,B_t}} \mathcal{B}_{tb} \text{ for } t = 1, \ldots, T \text{ trees}$$

$$w^N_{\text{Forest},i}(\boldsymbol{x}) := \sum_{t=1}^{T} \sum_{b=1}^{B_t} I(\boldsymbol{x} \in \mathcal{B}_{tb} \wedge \boldsymbol{x}_i \in \mathcal{B}_{tb})$$

$$\hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\boldsymbol{x}) := \arg\max_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} w^N_{\text{Forest},i}(\boldsymbol{x}) \ell_i(\boldsymbol{\vartheta})$$

OK, Done! Really?

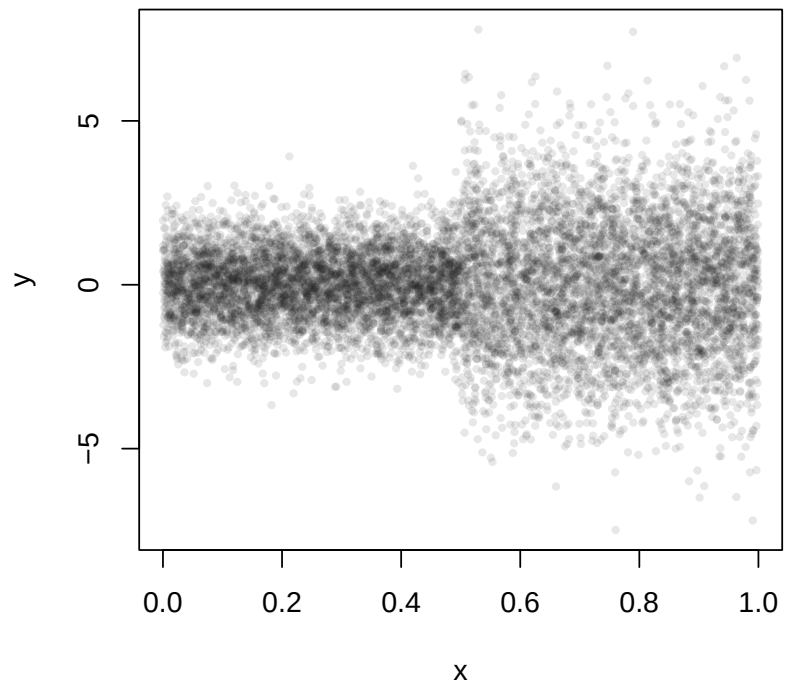
These “nearest neighbor weights” have been used before, first in

- “bagging survival trees” (2004), in
- “conditional inference forests” (**party(kit)**, since 2005) and in
- “quantile regression forests” (**quantregForest**, since 2006)

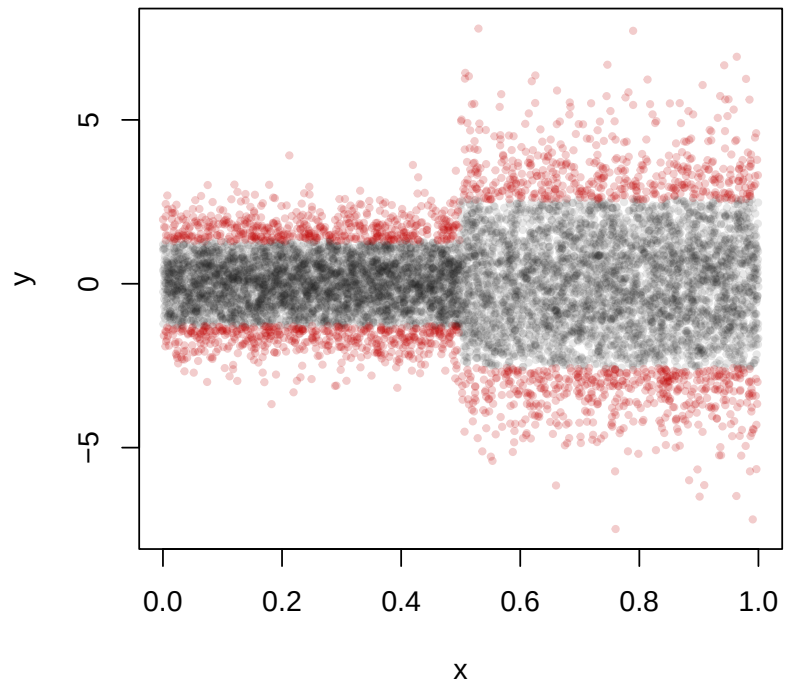
with *standard* trees (CART- or CTree-like).

Unfortunately, there is a catch.

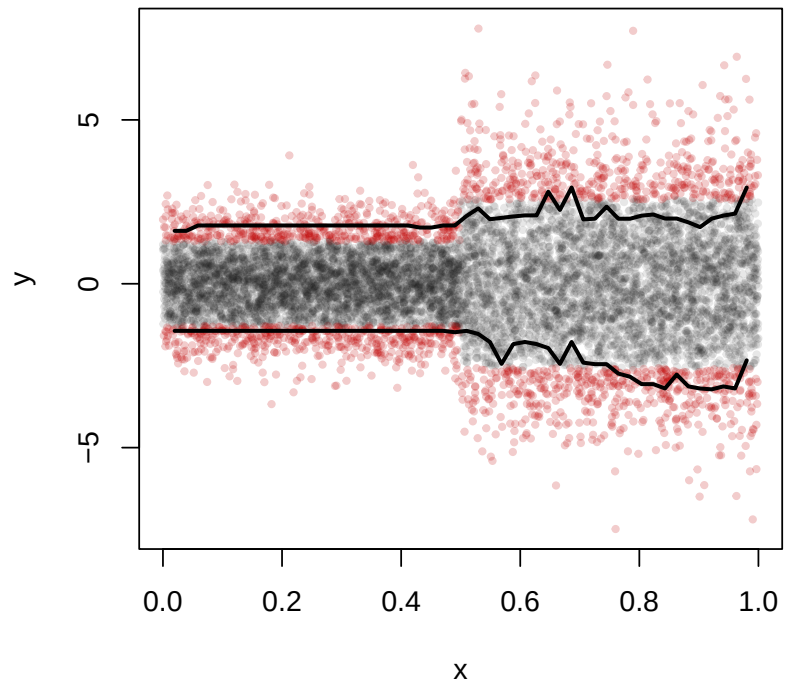
The Problem



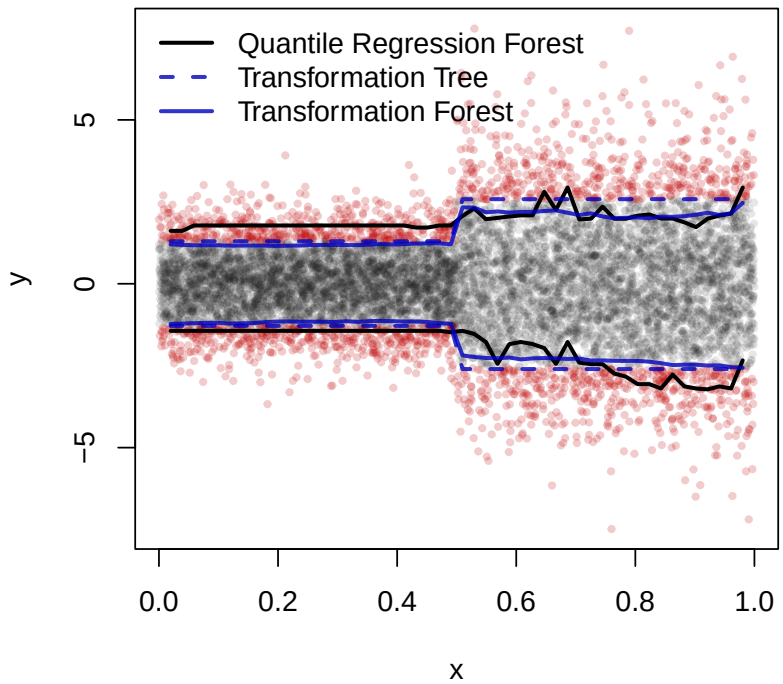
The Problem



The Problem



The Problem


Quantile Regression Forest
Transformation Tree
Transformation Forest
y
5
0
−5
0.0
0.2
0.4
0.6
0.8
1.0
x

The Solution

We need splits sensitive to *distributional* and not just *mean* changes.

Generic approach ("Distribution trees and forests"):

$$\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}_{Y, \boldsymbol{\vartheta}(\boldsymbol{x})}(y)$$

Here: Use transformation model

$$\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = F_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x}))$$

10.1080/10618600.2021.1872581

Parameterisation

Transformation trees and forests based on parameterisation

$$\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = F_Z(\boldsymbol{a}_{\text{Bs},d}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x}))$$

- $\boldsymbol{a}_{\text{Bs},d}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x})$ is a smooth, monotonic Bernstein polynomial of degree d
- $d = 1$ with $F_Z = \Phi$ means $\mathbb{P}_{Y|\boldsymbol{X}=\boldsymbol{x}} = \mathcal{N}(\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x}))$
- $d = 5$ is surprisingly flexible

Model-based Recursive Partitioning (MOB)

Core idea

- Fit parameters $\hat{\vartheta}_{\text{ML}}$ in *unconditional* model $\mathbb{P}_{Y,\vartheta}$
- Compute individual gradient contributions ("scores")

$$\boldsymbol{s}_i = \left. \frac{\partial \ell_i(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} \right|_{\boldsymbol{\vartheta} = \hat{\vartheta}_{\text{ML}}}$$

- Select predictor from $\boldsymbol{x}$ with strongest parameter instability as indicated by highest association to $\boldsymbol{s}_i$, $i = 1, \ldots, N$
- Find "best" binary split; repeat recursively

Implemented for many models, including (G)LM(M)s, parametric survival, β-regression, spatial lag, Bradley-Terry-Luce, various Item Response Theory models, subgroup analyses, etc.

Transformation Trees (TTree)

- Start with $\hat{\vartheta}^N_{\text{ML}}$
- Search for parameter instabilities in $\hat{\vartheta}^N_{\text{ML}}$ as a function of $\boldsymbol{x}$ using (a beefed-up version) of MOB
- Potentially find changes in the mean AND higher moments
- Forests: Aggregate these trees via adaptive local likelihood estimation

Transformation Forests (TForest)

$$\hat{\mathbb{P}}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = F_Z(\boldsymbol{a}_{\text{Bs},d}(y)^\top \hat{\boldsymbol{\vartheta}}^N_{\text{Forest}}(\boldsymbol{x}))$$

makes the forest "parametric" (one model for each $\boldsymbol{x}$) with

- Forest likelihood
- Prediction intervals
- Likelihood-based variable importance
- Parametric bootstrap
- ...

and applicable to censored and truncated data.

More Complex Models

For example: Subgroup analysis, stratified / personalised medicine, ...

Conditional transformation model

$$\mathbb{P}(Y \leq y \mid \text{treatment}, \boldsymbol{X} = \boldsymbol{x}) = F_Z(\boldsymbol{a}_{\text{Bs},d}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x}) - \beta(\boldsymbol{x}) I(\text{treated}))$$

- Both the "intercept function" $\boldsymbol{a}_{\text{Bs},d}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x})$ and
- the treatment effect $\beta(\boldsymbol{x})$ may depend on $\boldsymbol{x}$
- $F_Z() = 1 - \exp(-\exp())$ makes β a log-hazard ratio
- Include $\hat{\beta}$ in search for parameter instabilities

Stratified Medicine

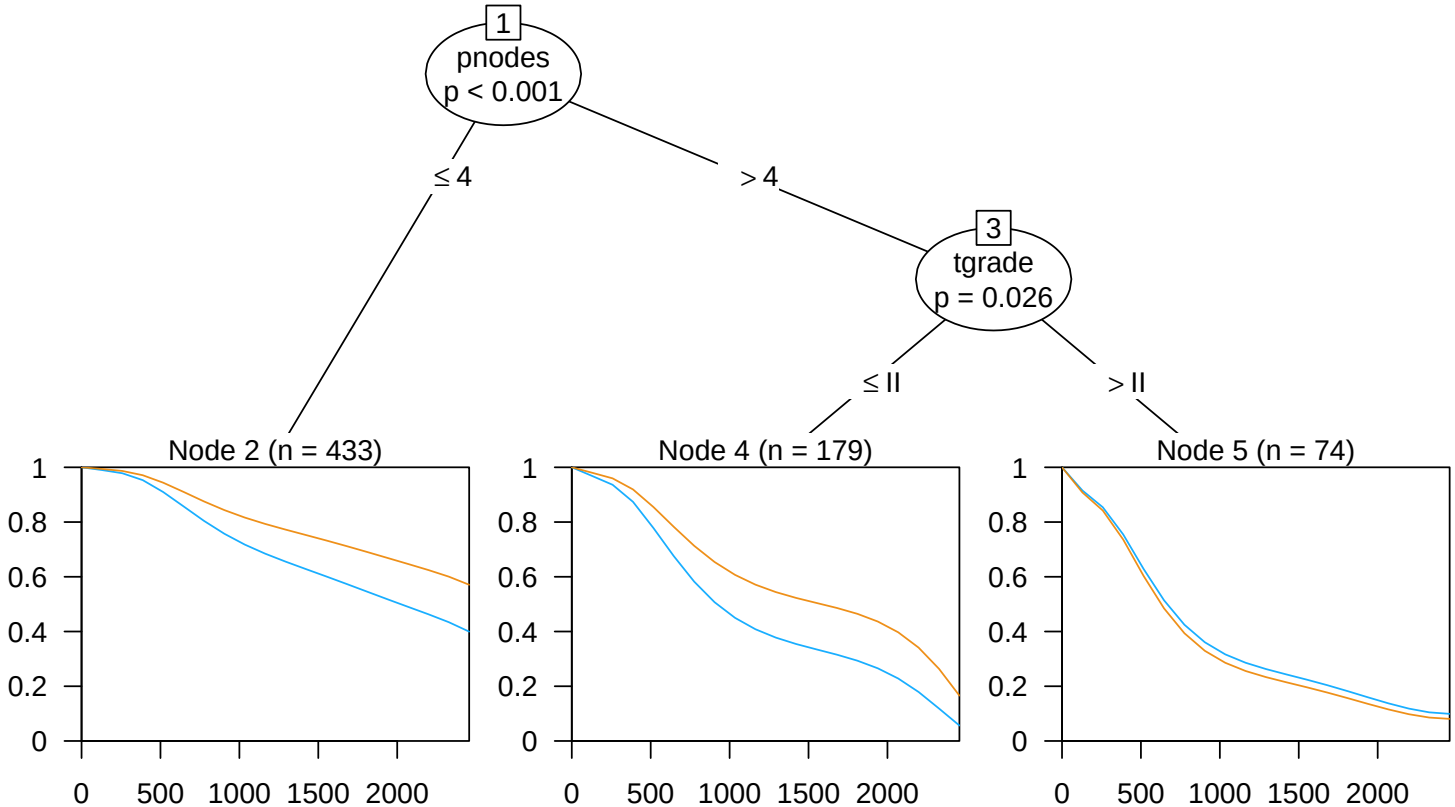
Partition log-hazard ratio β from a fully parametric Cox model

$$\mathbb{P}(T > t \mid \text{treatment}) = \exp(-\exp(\boldsymbol{a}_{\text{Bs},d}(t)^\top \boldsymbol{\vartheta} - \beta I(\text{treated}))$$

for a randomised controlled clinical trial on hormonal treatment of breast-cancer patients

```
library("tram")
cmod <- Coxph(ctime ~ horTh, data = GBSG2)
library("trtf")
tmod <- trafotree(cmod,
                  formula = ctime ~ horTh | .,
                  data = GBSG2)
```

Stratified Medicine



Survival Forests

Log-rank splitting implicitly assumes proportional hazards model

$$\mathbb{P}(T > t \mid \boldsymbol{X} = \boldsymbol{x}) = \exp(-\exp(h(y) - \beta(\boldsymbol{x})))$$

$\Rightarrow$ `cforest`, `ranger`, `randomForestSRF` are insensitive to non-proportional hazards effects.

Switching to transformation forests based on

$$\mathbb{P}(T > t \mid \boldsymbol{X} = \boldsymbol{x}) = \exp(-\exp(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}(\boldsymbol{x})))$$

relaxes this restriction, 10.1177/0962280219862586.

Ordinal Transformation Forests

Proportional-odds models assume

$$\mathbb{P}(Y \leq y_k \mid \boldsymbol{X} = \boldsymbol{x}) = \text{expit}(\vartheta_k - \beta(\boldsymbol{x}))$$

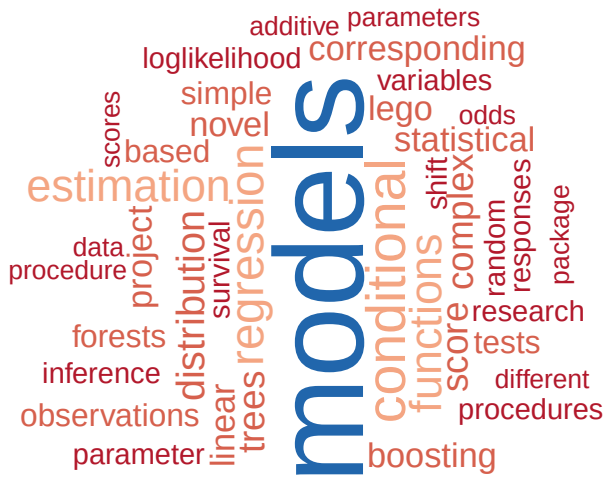
Transformation models can be set-up as estimators for $\beta(\boldsymbol{x})$ or, without assuming proportional odds, for more general models

$$\mathbb{P}(Y \leq y_k \mid \boldsymbol{X} = \boldsymbol{x}) = \text{expit}(\vartheta_k(\boldsymbol{x}))$$

10.1089/neu.2020.7407

R Add-on Packages

- mlt: Basic infrastructure
- tram: Model interfaces, multivariate models
- cotram: Count models
- tramnet: Penalisation
- tramvs: Variable selection
- tramME: Mixed-effects
- tbm: Boosting
- trtf: Trees and Forests



https://ctm.R-forge.R-project.org/