Tree-structured Regression Methods

Soo-Heang EO[1]

Joint work with Sungwan Bang[2], and HyungJun Cho[1]

[1]Department of Statistics
Korea University

[2]Department of Mathematics
Korea Military Academy

Oct 17, 2014

Table of Contents

Celebrating 50th anniversary

PROBLEMS IN THE ANALYSIS OF SURVEY DATA, AND A PROPOSAL

JAMES N. MORGAN AND JOHN A. SONQUIST*
University of Michigan

Most of the problems of analyzing survey data have been reasonably well handled, except those revolving around the existence of interaction effects. Indeed, increased efficiency in handling multivariate analyses even with non-numerical variables, has been achieved largely by assuming additivity. An approach to survey data is proposed which imposes no restrictions on interaction effects, focuses on importance in reducing predictive error, operates sequentially, and is independent of the extent of linearity in the classifications or the order in which the explanatory factors are introduced.

JASA 58(302) June 1963 (well, there were no early online versions, so most people read the paper in 1964 for the first time).

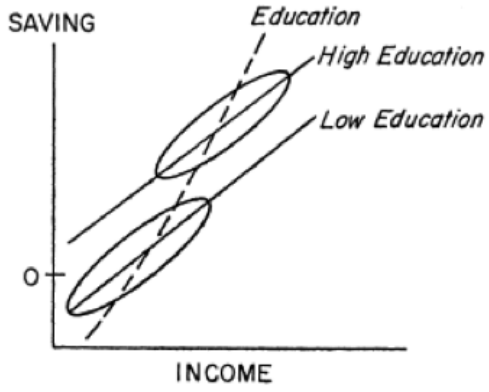
Celebrating 50th anniversary

Group	Spending unit average (1958) income	Number in unit	Number of cases
Nonwhite, did not finish high school	$ 2489	3.3	191
Nonwhite, did finish high school	5005	3.4	67
White, retired, did not finish high school	2217	1.7	272
White, retired, did finish high school	4520	1.7	72
White, nonretired farmers, did not finish high school	3950	3.6	87
White nonretired farmers, did finish high school	6750	3.6	24
The Remainder			
0–8 grades of school			
18–34 years old	4150	3.8	72
35–54 years old	4670	3.8	240
55 and older—not retired	4846	2.2	208
9–11 grades of school			
18–34 years old	5032	3.7	112
35–54 years old	6223	3.4	202
55 and older—not retired	4720	2.1	63
12 grades of school			
18–34 years old	5458	3.3	193
35–54 years old	7765	3.8	291
55 and older—not retired	6850	2.0	46
Some college			
18–34 years old	5378	3.0	102
35–54 years old	7930	3.8	112
55 and older—not retired	8530	2.0	36
College graduates			
18–34 years old	7520	3.8	80
35–54 years old	8866	2.9	150
55 and older—not retired	10879	1.8	34

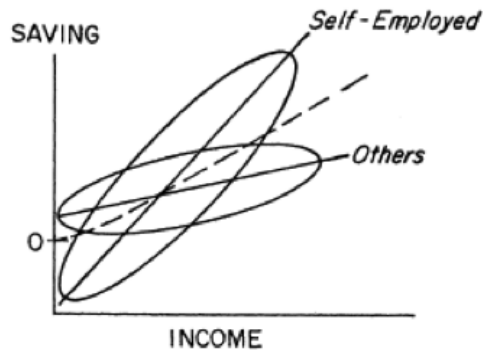
Source: 1959 Survey of Consumer Finances

Celebrating 50th anniversary

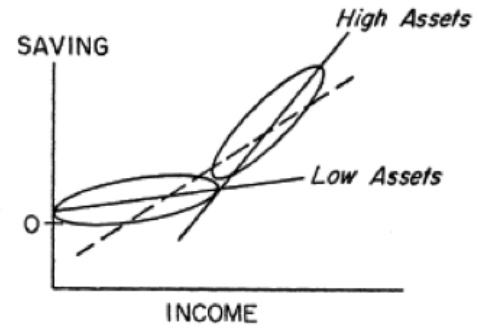
Muticollinearity, i.e., correlation between income and education but no interaction



Interaction, but no multicollinearity (no correlation between income and self-employment)

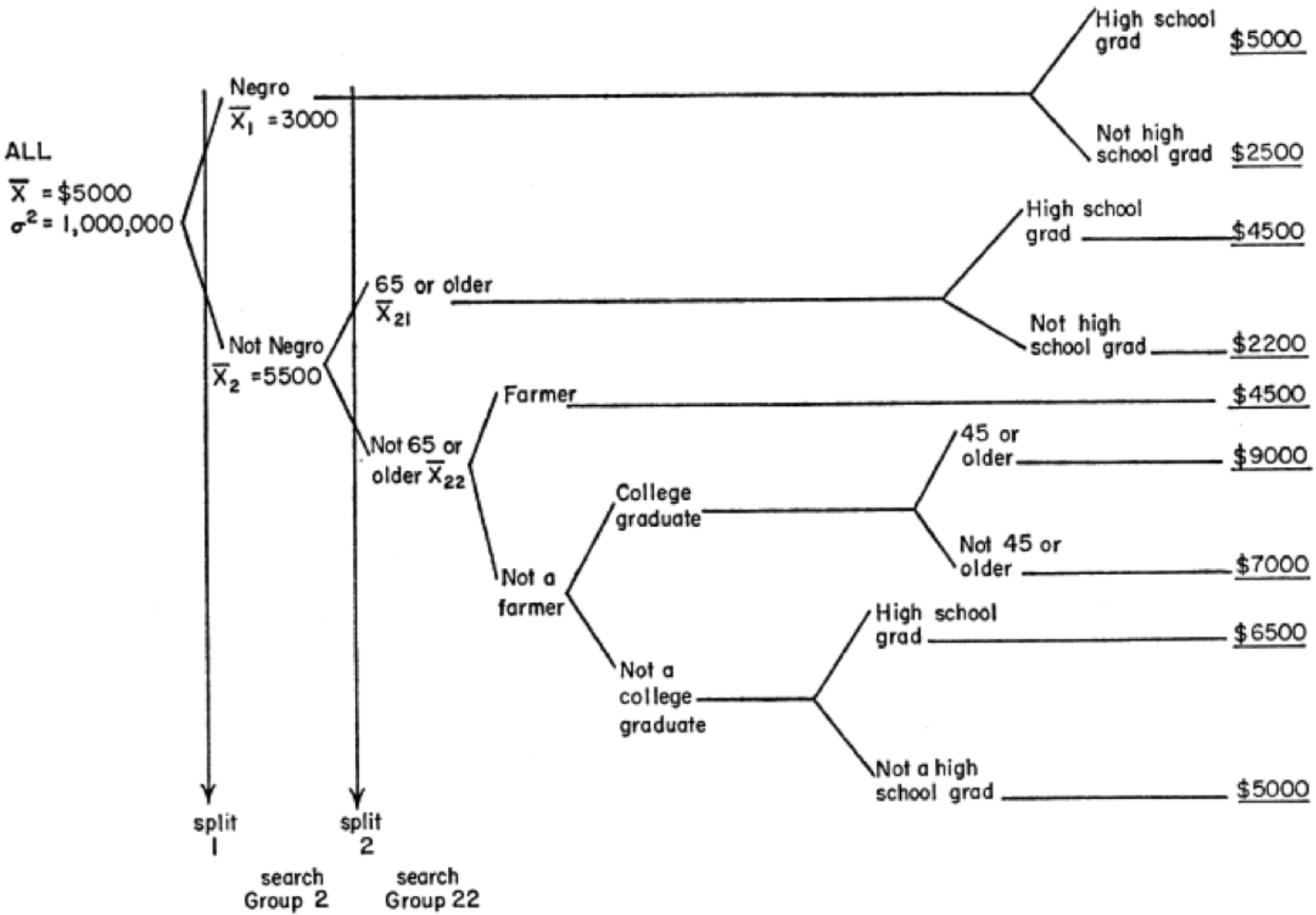


Both



Regression with pooled data
Separate regressions
Concentration of data

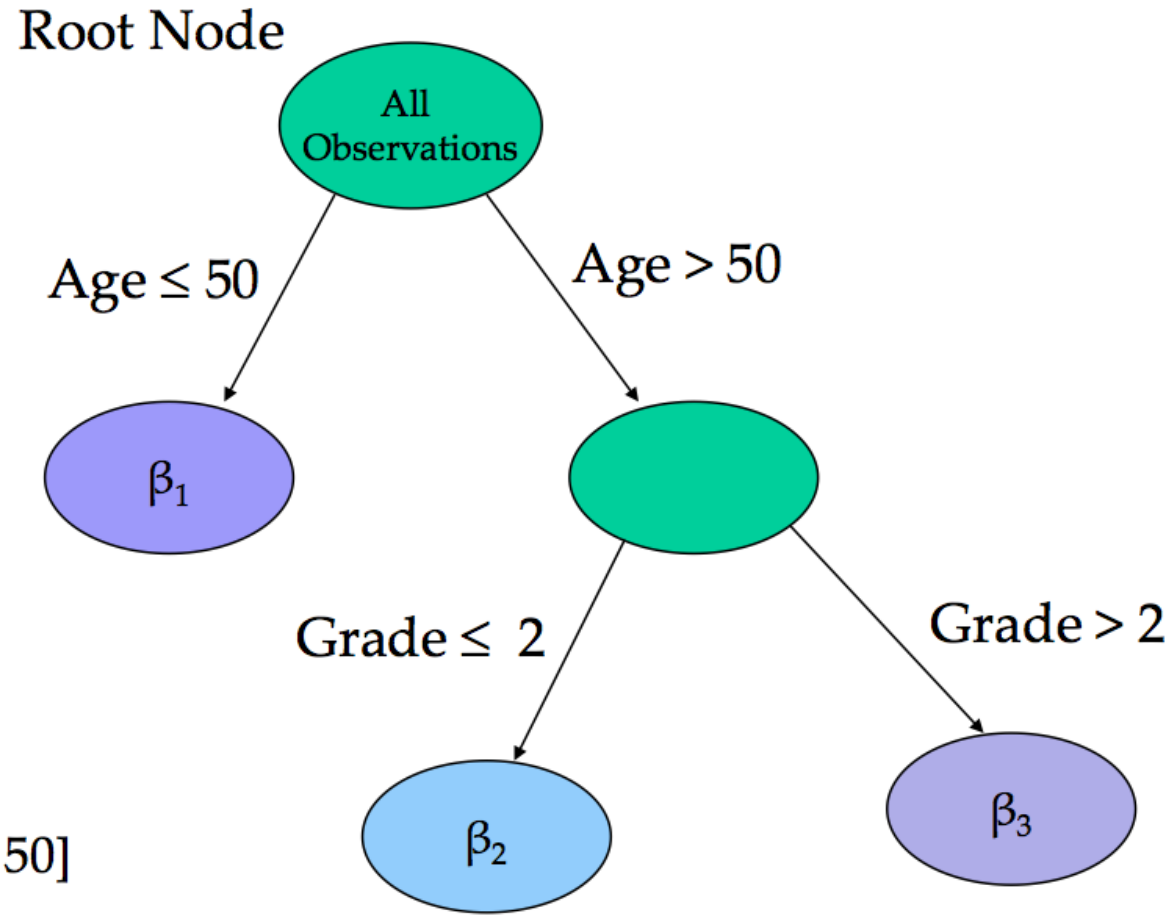
Celebrating 50th anniversary



Notations

- t and T: a node and a tree
- N and $N(t)$: sample size and number of samples in t
- Y: response variable
- p: number of predictor variables
- k: number of partitioning variables
- $X = (X_1, \ldots, X_p)$: vector of fitted variables
- $Z = (Z_1, \ldots, Z_k)$: vector of split variables

Classification and Regression Trees (CART)



$\beta_1 = E[T \mid Age \leq 50]$

$\beta_2 = E[T \mid Age > 50 \ \& \ Grade \leq 2]$

$\beta_3 = E[T \mid Age > 50 \ \& \ Grade > 2]$

Breiman, et al. (1984)

Classification and Regression Trees (CART)

1. Fit a constant, the node mean $\bar{y}$ at each node
2. Use sum of squared deviations $\sum_i (y_i - \bar{y})^2$ as node impurity
3. Choose the split that maximizes the decrease in node impurity
4. Use the sample $\bar{y}$ in node t as predicted value.
5. Prune tree using test sample or cross-validation
6. Use surrogate spilts to deal with missing values

Classification and Regression Trees (CART)

A *naive* approach is to evaluate the reduction of impurity, $\Delta(\cdot)$, over all possible splits, and select the split set with the greatest reduction of impurity,

$$\underset{A_{|Z}}{\operatorname{argmax}}\,\Delta(t) \quad = \quad R(t) - [R(t_L) + R(t_R)], \tag{1}$$
$$\forall A_{|Z} \in \{a_{1|Z}, a_{2|Z}, \ldots, a_{c|Z}\}, \forall Z \in \{Z_1, Z_2, \ldots, Z_r\},$$

where

- (t, t_L, t_R): a node t and its left and right child node
- Z: a candidate split variable
- $A_{|Z}$: a candidate split set given Z

We call this approach an **exhaustive search (ES)** algorithm (Breiman *et al.*, 1984).

Classification and Regression Trees (CART)

Shortcut algorithm for selecting split set (Breiman et al., 1984, Sec. 9)

- Label each observation in the node as belong to class 1 if the sign of residuals is positive and class 2 otherwise.
- Suppose there are L categories, $b_1, \ldots, b_L$, in the selected Z.

  Order the categories accodring to the class 1 proportions. That is,

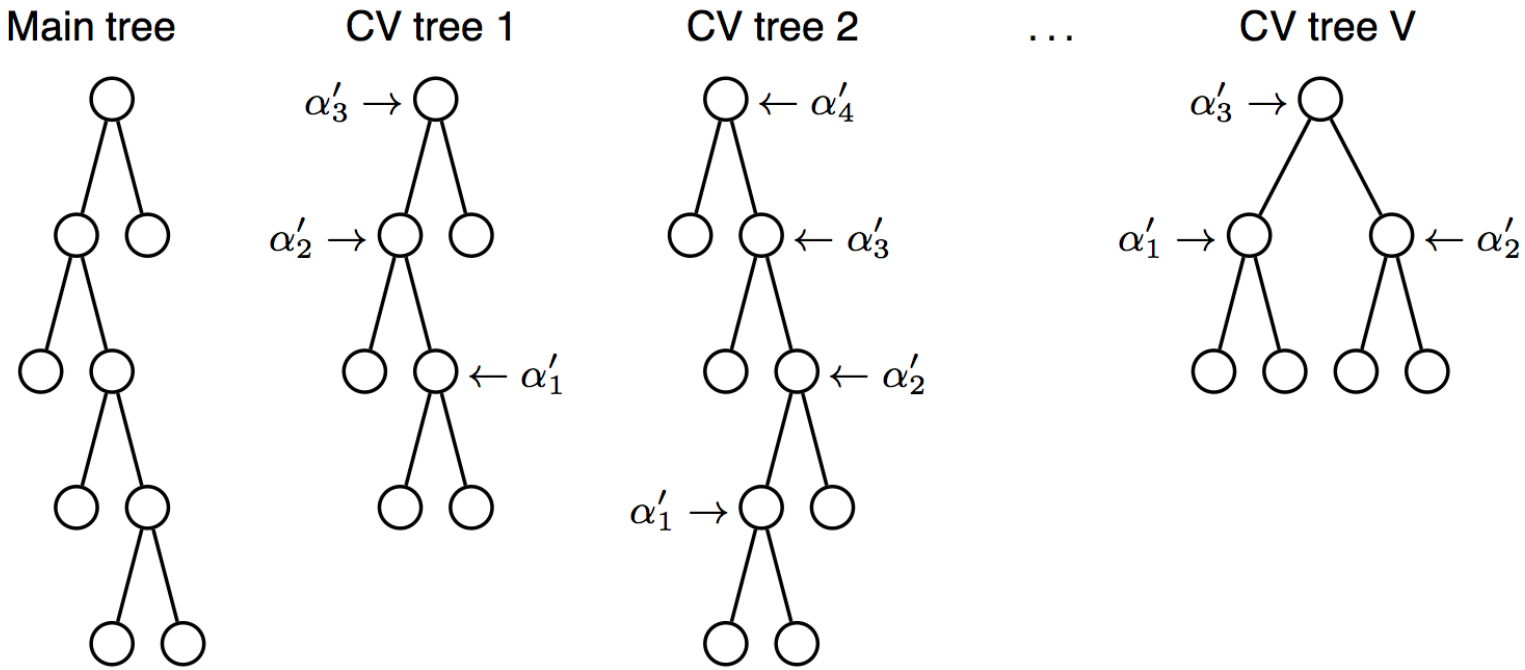
  $$Pr(Class1|Z = b_{l_1}) \leq Pr(Class1|Z = b_{l_2}) \leq \cdots \leq Pr(Class1|Z = b_{l_L}).$$
- The best split set belongs to one of the $L-1$ subsets

  $$\{b_{l_1}, \ldots, b_{l_k}\}, k = 1, \ldots, (L-1).$$

  Select the subset that minimizes the sum of impurities in the left and right subnodes.

Classification and Regression Trees (CART)

Cost-complexity pruning algorithm (Breiman *et al.*, 1984)



- Main tree is grown using all the data.
- Each CV tree is grown using $(V - 1)$ subsets.

Cost-complexity pruning algorithm (Breiman et al., 1984)

Find a tree T minimising the cost complexity:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|, \tag{2}$$

where

- $R(T)$: the sum of the resubstitution loss function over the terminal node of T
- α: complexity parameter satisfying $\alpha \geq 0$
- $\tilde{T}$: the number of terminal nodes in T

It consists of two parts:

1) Obtain candidate trees and **2)Select an optimal tree**.

Cost-complexity pruning algorithm (Breiman et al., 1984)

Step 1) Obtain candidate trees.

1) Let t be any node and T_t be the branch of a tree T with root node t. Then, $R_\alpha(\{t\}) = R(t) + \alpha$, and $R_\alpha(T_t) = R(T_t) + \alpha|\tilde{T}_t|$. Compute the critical value of α for which $R_\alpha(T_t) = R_\alpha(\{t\})$ is $\alpha = g(t)$, where $g(t) = (R(t) - R(T_t))/(|\tilde{T}_t| - 1)$.

2) Prune branches at nodes $t_{\min}$ for which $g(t_{\min}) = \min\{g(t) : t \in T - \tilde{T}\}$, where $T - \tilde{T}$ is the collection of intermediate nodes in T.

3) Repeat the preceding steps to obtain a nested sequence of trees $T_{\max} = T_0 \succ T_1 \succ T_2 \succ \cdots \succ T_{\text{root}}$, where $T_{\max}$ is the maximal tree, T_i is a tree with some branches of T_{i-1} pruned, and T_{root} is the trivial tree with only the root node.

Cost-complexity pruning algorithm (Breiman et al., 1984)

Step 2) Select an optimal tree.

1) Let $0 = \alpha_0 < \alpha_1 < \alpha_2 < \cdots$ be the α values associated with the pruned sequence of trees $T_{\max} = T_0 \succ T_1 \succ T_2 \succ \cdots \succ T_{\text{root}}$. Define $\alpha'_k = \sqrt{\alpha_k \alpha_{k+1}}$, where $k = 0, 1, 2, \ldots$.

2) Randomly divide the learning sample $\mathcal{L}$ into V subsets $\mathcal{L}_1, \ldots, \mathcal{L}_V$, so that every subset has approximately the same percentage of censored cases as $\mathcal{L}$.

3) For each α'_k, obtain the minimal-cost complexity tree $T^{(-v)}(\alpha'_k)$, which was constructed from $\mathcal{L} - \mathcal{L}_v$ for $v = 1, 2, \ldots, V$.

4) Estimate the cost of $T^{(-v)}(\alpha'_k)$ with the test sample $\mathcal{L}_v$. Let the cost be $R^v(T^{(-v)}(\alpha'_k))$.

5) Obtain the V-fold CV cost for tree T_k by setting $R^{\text{cv}}(T_k)$ to be equal to $\Sigma_{v=1}^{V} R^v(T^{(-v)}(\alpha'_k))$.

6) If $T_{\min}$ is the tree with the smallest value of $R^{\text{cv}}(T_i)$, T_{kSE} selected by the k-SE rule is the smallest tree T_i satisfying $R^{\text{cv}}(T_i) \leq R^{\text{cv}}(T_{\min}) + k\text{SE}(R^{\text{cv}}(T_{\min}))$.

Problems on the exhaustive search

ES suffers from

- undue preference to split variables with more possible splits (variable selection bias)
- end-cut preference, substantial computational cost, ...

To solve the problem,

- Loh (2002, 2009, 2014) have proposed a **residual analysis (RA)** algorithm.
- This approach called **GUIDE** is *selecting (1) the split variable and (2) its split set* **separately**.

Here, we adapt the RA philosophy to construct a tree framework.

Residual Analysis Tree (GUIDE)

- Proposed by Loh (2002, Sinica; 2009, AOAS) based on Loh and Shih (1996, Sinica) and Kim and Loh (2001, JASA)
- Define a residual from fitting a node model at node t as

$$r_i = y_i - \hat{y}_i, \quad i \in t. \tag{3}$$

- If the fitted model is correct, *the residuals should be randomly distributed* over each split variable. The randomness means that the fitted model is sufficient to explain the data so that no further split is needed.
- If the fitted model is not correct, the residuals have a systematic pattern against Z. Based on residuals, we can select a split variable to segment the whole data space into $\mathcal{B}$ subspaces (by default, $\mathcal{B} = 2$).

$$r_i^s = \text{sgn}(r_i) := \begin{cases} -1, & \text{if } r_i \leq 0, \\ +1, & \text{if } r_i > 0. \end{cases} \tag{4}$$

Residual Analysis Tree (GUIDE)

- Select a partitioning variable by performing main- and interaction- effects tests for **the randomness of residuals** against Z.
- Obtain a test statistic by a main-effects test.

  (a) Divide cases into three or four groups for ordered Z or number of categories for categorical Z.

  (b) Construct a table with signs of residuals as rows and groups as columns.

  (c) Compute the Wilson-Hilferty χ^2_1 approximation statistic, $W_M(Z)$.
- Do an interaction-effects test $W_I(Z_i, Z_j)$ if $\max_i W_M(X_i) \leq \chi^2_{1,\alpha}$

  where $\alpha = 0.05/K$, $\beta = 0.1/(K(K-1))$ and

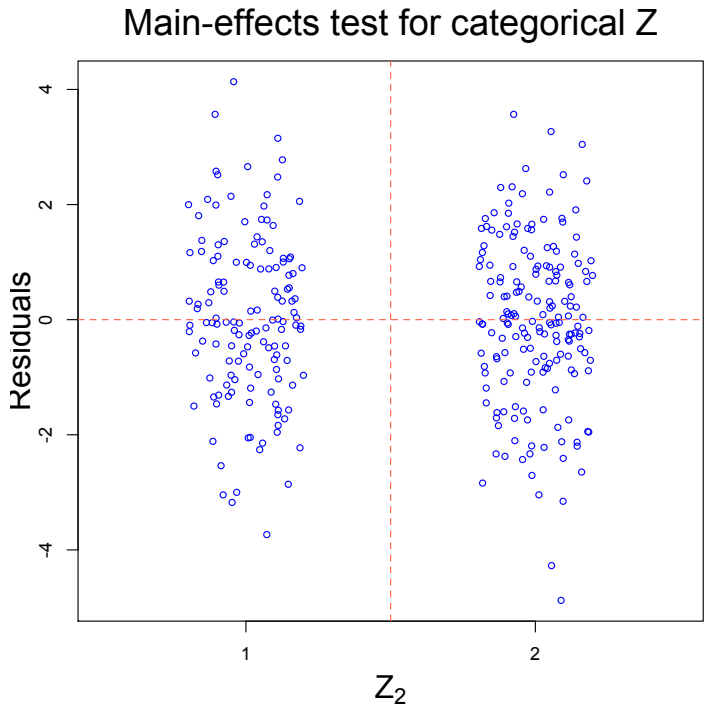
  K means the number of non-constant split variables

  (a) Find $W_I(Z_i, Z_j)$ for each pair of predictor variables

  (b) $\max_i W_I(Z_i, Z_j) > \chi^2_{1,\beta}$, select the pair with largest $W_I(Z_i, Z_j)$.

  (c) Otherwise, select variable with largest $W_M(Z_i)$ via a Two-level search.

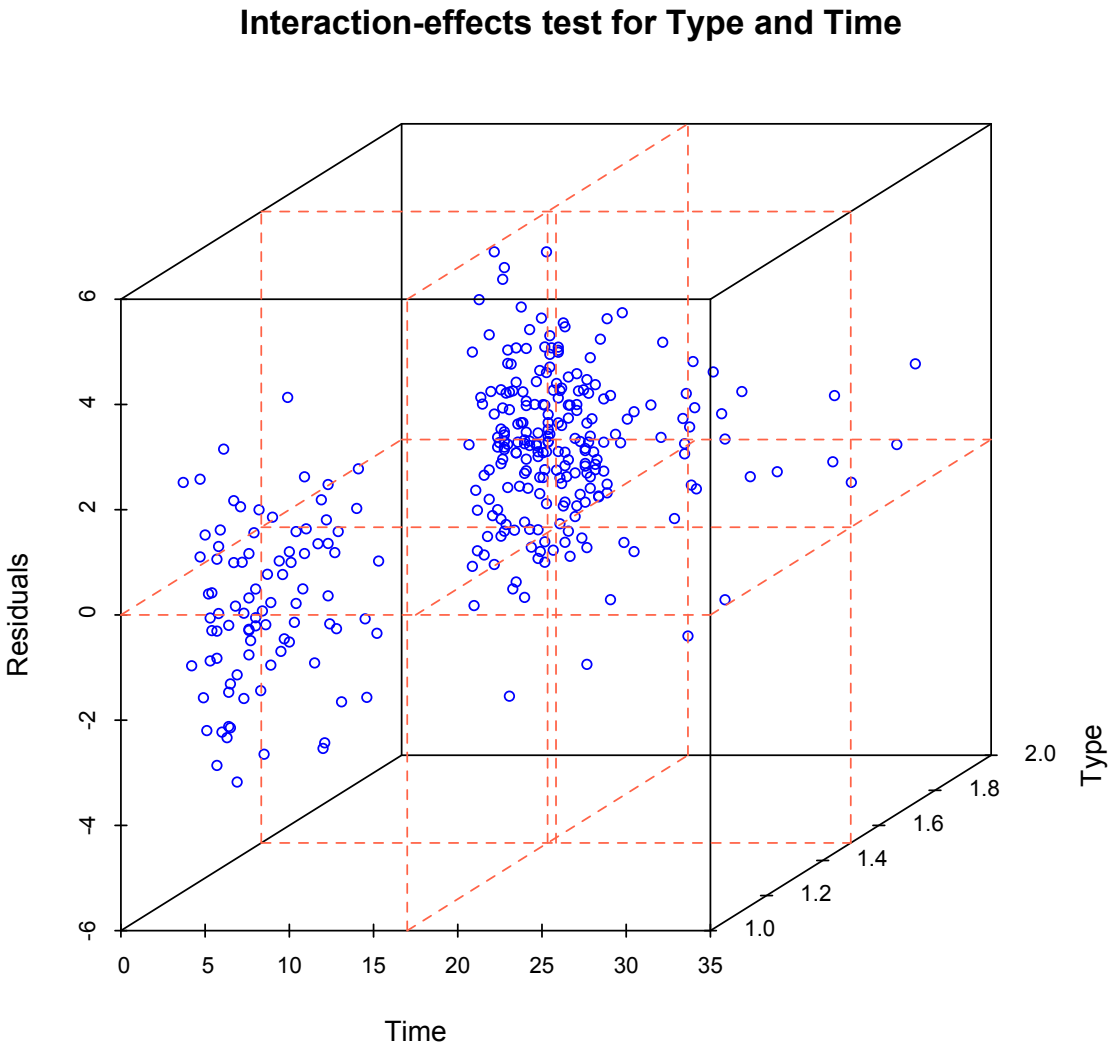
Main-effects test for selecting a split variable Z



Residuals	G_1	G_2	G_3
+	49	88	17
−	39	82	43
$W_M(Z) = 9.42$, p-value $= 0.002$			

Residuals	G_1	G_2
+	49	105
−	42	122
$W_M(Z) = 1.21$, p-value $= 0.271$		

Interaction-effects test for selecting a split variable Z



	Residuals	
	−	+
G_1	15	34
G_2	54	52
G_3	22	20
G_4	62	59
$W_I(Z_i, Z_j) = 4.481$		
p-value $= 0.034$		

Selecting split set for the selected ordered Z

We split a node t into two child nodes, t_L and t_R, by

- Searching all splits of the form $Z \leq c$ to maximize $\Delta(t)$.
- Spline-based point selection algorithm (Eo and Cho, 2013)

This split occurs when yielding the greatest reduction of impurity:

$$\Delta(t) = R(t) - [R(t_L) + R(t_R)], \tag{5}$$

where $R(t), R(t_L), R(t_R)$ are the impurity functions of node t, its left branch t_L, and its right branch t_R, respectively.

Selection of split set for categorical Z (Loh, 2009)

Let $\{a_1, a_2, \ldots, a_m\}$ be the set of distinct values of Z in node t,

- If $m \leq 11$, search all subsets S to find a split of the form $t_L = \{Z \in s : s \in S\}$.
- If $m > 20$, define the new categorical variable $Z' = \sum_i j_i I(Z = a_i)$ where j_i is the class that minimizes the impurity for the observations in $t \cap \{Z = a_i\}$. Search for the split based on Z' that minimizes the decrease in impurity.
- Otherwise, use linear discriminant analysis on the dummy vectors for Z.

Determining an optimal tree size

- Post pruning: Minimal cost-complexity pruning algorithm by Breiman et al. (1984) is used. Find a tree T minimizing the cost-complextiy

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|, \tag{6}$$

where $R(T)$ is the sum of the re-substititution loss functions over the terminal nodes of T, $\alpha \geq 0$ is a complexity parameter, and $|\tilde{T}|$ is the number of terminal nodes in T.

- Pre pruning: M-step stopping rule by Eo and Cho (2013) can be used.

Conditional Inference Tree (CTREE)

- Proposed by Hothorn, Hornik, and Zeilies (2006, JCGS)
- **Use conditional permutation tests to select split variables**
  - Requires computation of p-values, either by exact calculation, Monte Carlo simulation, or asymptotic approximation
- Use stopping rules controlled by Bonferroni adjustments without pruning
- Surrogate splits are used to deal with missing values
- Permutation tests (with subnode label as response variable) are used to find the surrogate variables.

Conditional Inference Tree (CTREE)

Selection of split set or points

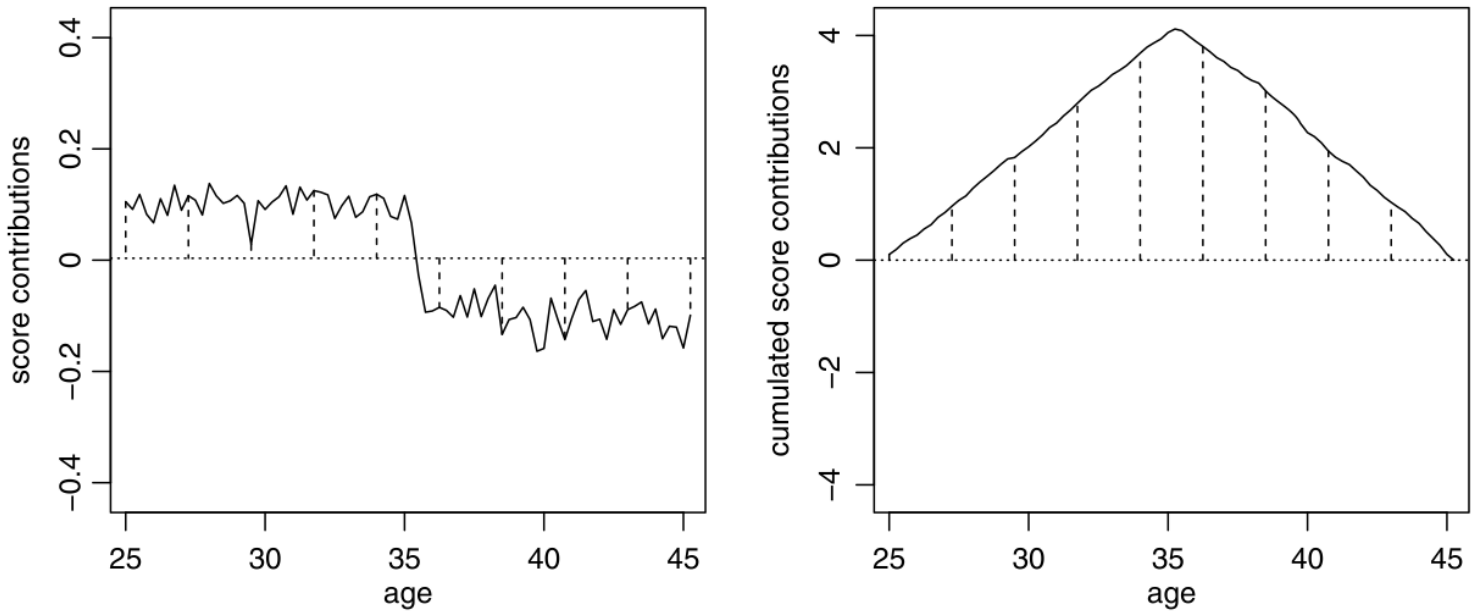


FIGURE 2.
Structural change in the variable age (artificial data for illustration purposes). In the *left plot*, the individual score contributions are ordered with respect to the variable. The *dashed lines* indicate deviations from the overall mean zero, which are positive before the structural change and negative afterward. In the *right plot*, the positive and negative deviations are cumulated and the structural change is now noticeable from the peak in the cumulative sum process.

Figure: Concept of an instability test (Strobl et al., 2013)

Model-based CART

Alexander and Grimshaw (1996, JCGS) proposed **treed regression algorithm** that combines the merits of CART and OLS.

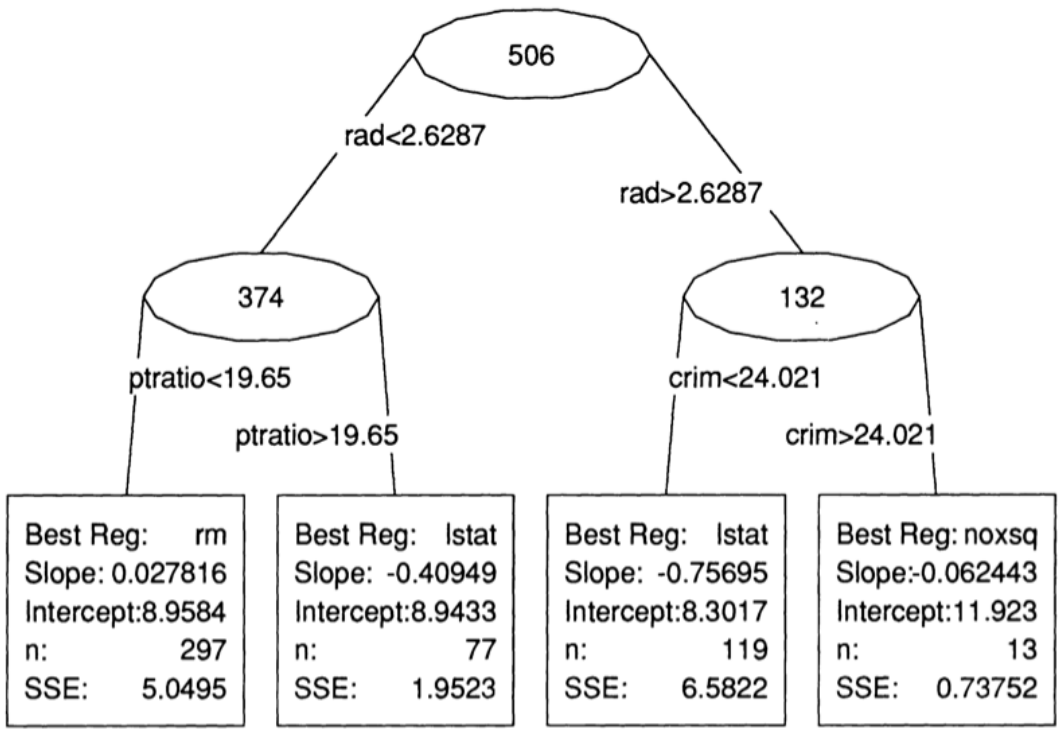


Figure 9. Monotone treed regression model of level two (four terminal nodes) of the log of the median value of owner-occupied homes using 13 explanatory variables measured on Boston area census tracts in 1970 from an example in Belsley, Kuh, and Welsch (1980).

Previous works on model-based regression trees

- Alexander and Grimshaw (1996, JCGS): CART with simple regression
- Chan and Loh (2003, JCGS): GUIDE with logistic regression
- **Zeilies et al. (2008, JCGS) and Rusch et al. (2013): CTREE with GLM**
- **Sela and Simonoff (2012) and Hajjem et al. (2011): CART with LME**
- Brandmaier *et al.* (2013): CART with SEMs
- **Loh and Zheng (2013, AOAS) and Eo and Cho (2014, JCGS): GUIDE with LME**
- Simonoff (2013): CART with linear multilevel model
- Strobl *et al.* (2013): CTREE with Rasch model
- Rusch *et al.* (2013b, AOAS): CTREE with logistic regression
- Rusch *et al.* (2013c, AOAS): CTREE with latent dirichlet allocation
- Tutz and Berger (2014): CTREE with GLM
- ...

Longitudinal CART (Segal, 1992)

		CART algorithm	SEGAL algorithm
Model		$y_i = \beta_0 + \varepsilon_{ij}$	$y_{ij} = \beta_{0j} + \varepsilon_{ij}$
Split	Split Mtd.	Binary Split	
	Split Ft.	$\phi_m(s, g) = SS(g) - SS(g_L) - SS(g_R)$.	
	R(t)	$SS(g) = \Sigma(y_i - \bar{y}(g))^2$	$SS(g) = \sum_{i \in g} (y_i - \mu(g))' V(\theta, g)^{-1} (y_i - \mu(g))$.
	Cov Str.		$V(\theta, g) = V(\theta_L, g_L) = V(\theta_R, g_R)$
Stopping		Cost-Complexity Pruning	
Missing		Surrogate rule	

Figure: Comparison between CART and Longitudinal CART (Eo, 2009)

RE-EM tree (Sela and Simonoff, 2012)

Hajjem et al. (2011, SPL) and Sela and Simonoff (2012) independently proposed **random effect with EM algorithm tree model**.

$$y_{it} = f(X_i) + Z_{it} b_i + \epsilon_{it},$$

$$b_i \sim N(0, D), \quad \epsilon_{it} \sim N(0, \sigma^2 I),$$

$$i = 1, 2, \ldots, n.$$

- If the random effects, b_i, were known, the model implies that we could fit a regression tree $y_{it} - Z_{it} b_i$ to estimate f.
- If the fixed effects, f, were known and can be represented as a linear function, then we could estimate the random effects using a traditional mixed effects linear model with $f(X_i)$.

RE-EM tree (Sela and Simonoff, 2012)

For a known cluster: Prediction $= \hat{f}(x_{ij}) + Z_i \hat{b}_i$.

For a new cluster: Prediction $= \hat{f}(x_{ij})$.

Step 0. Set $r = 0$. Let $\hat{b}_{i(0)} = 0$, $\hat{\sigma}^2_{(0)} = 1$, and $\hat{D}_{(0)} = I_q$.

Step 1. Set $r = r + 1$. Update $y^*_{i(r)}$, $\hat{f}(X_i)_{(r)}$, and $\hat{b}_{i(r)}$

i) $y^*_{i(r)} = y_i - Z_i \hat{b}_{i(r-1)}$, $i = 1, \ldots, n$,

ii) Let $\hat{f}(X_i)_{(r)}$ be estimated from a tree (or forest) algorithm with $y^*_{i(r)}$ as responses and X_i as covariates,

iii) $\hat{b}_{i(r)} = \hat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} \left(y_i - \hat{f}(X_i)_{(r)} \right)$, $i = 1, 2, \ldots, n$,

where $\hat{V}_{i(r-1)} = Z_i \hat{D}_{(r-1)} Z_i^T + \hat{\sigma}^2_{r-1} I_{n_i}$, $i = 1, 2, \ldots, n$.

Step 2. Update $\hat{\sigma}^2_{(r)}$, and $\hat{D}_{(r)}$ using

$$\hat{\sigma}^2_{(r)} = N^{-1} \sum_{i=1}^{n} \left\{ \hat{\epsilon}^T_{i(r)} \hat{\epsilon}_{i(r)} + \hat{\sigma}^2_{(r-1)} [n_i - \hat{\sigma}^2_{(r-1)} tr(\hat{V}_{i(r-1)})] \right\}$$

$$\hat{D}_{(r)} = n^{-1} \sum_{i=1}^{n} \left\{ \hat{b}_{i(r)} \hat{b}^T_{i(r)} + [\hat{D}_{(r-1)} - \hat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} Z_i \hat{D}_{(r-1)}] \right\},$$

Figure: Main algorithm of RE-EM tree

Model-based CTREE (Zeileis et al., 2008)

1. **Fit a parametric model in each node, e.g., based on maximum likelihood (Zeileis et al, 2008, JCGS) or GLM (Rusch et al., 2013).**
2. Assess whether parameter estimates are stable with respect to each split variable, using Bonferroni-adjusted p-values (instability tests).
3. If minimum p-value is sufficiently small, select the most unstable variable and split the node into two. Otherwise stop.

Model-based CTREE (Zeileis et al., 2008)

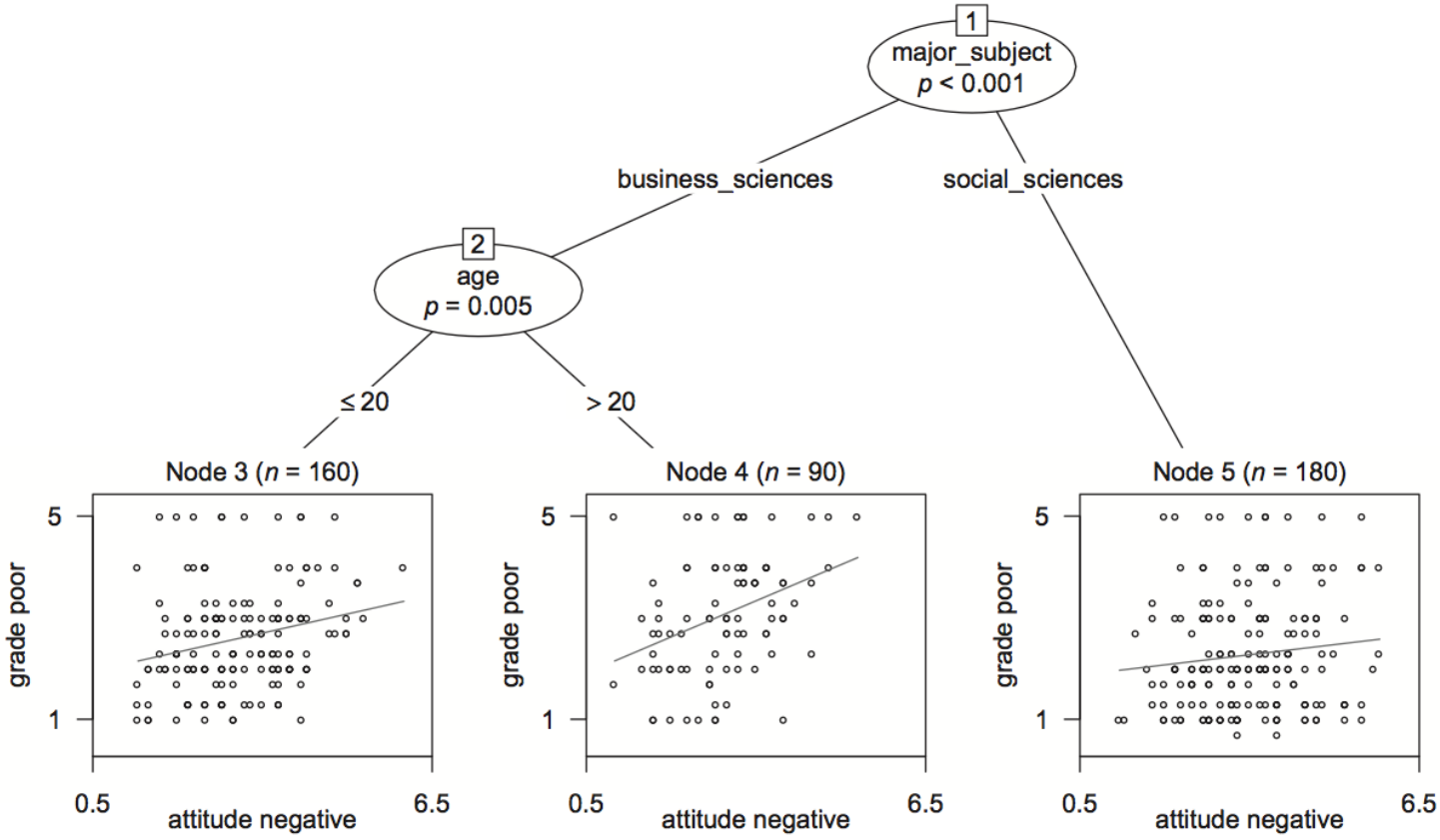


Figure 9. Model-based partition for the attitude toward statistics data. The model of interest relates the final statistics grade to the Cognitive Competence score obtained in the first week. (For the interpretation, note that the Cognitive Competence score was recoded such that high values correspond to a negative attitude, i.e., agreement with items such as "I will find it difficult to understand statistical concepts," and numerically high grades indicate poor performance.)

Figure: Example of model-based CTREE (Strobl et al., 2009)

Longitudinal GUIDE (Loh and Zheng, 2013)

1. Treat each data point as a curve (trajectory).
2. **Fit a mean curve (lowess or smoothing spline) to data in the node.**
3. Group trajectories into classes according to shapes relative to mean curve.
4. For each X variable, find p-value of χ^2 test of class vs. X.
5. Select X with smallest p-value to split node.
6. For each split point, fit a mean curve to each child node.
7. Select the split that minimizes sum of squared deviations or trajectories from mean curves in two child nodes.
8. Stop splitting when sample size in node is too small.
9. Prune the tree using cross-validation.

Longitudinal GUIDE (Loh and Zheng, 2013)

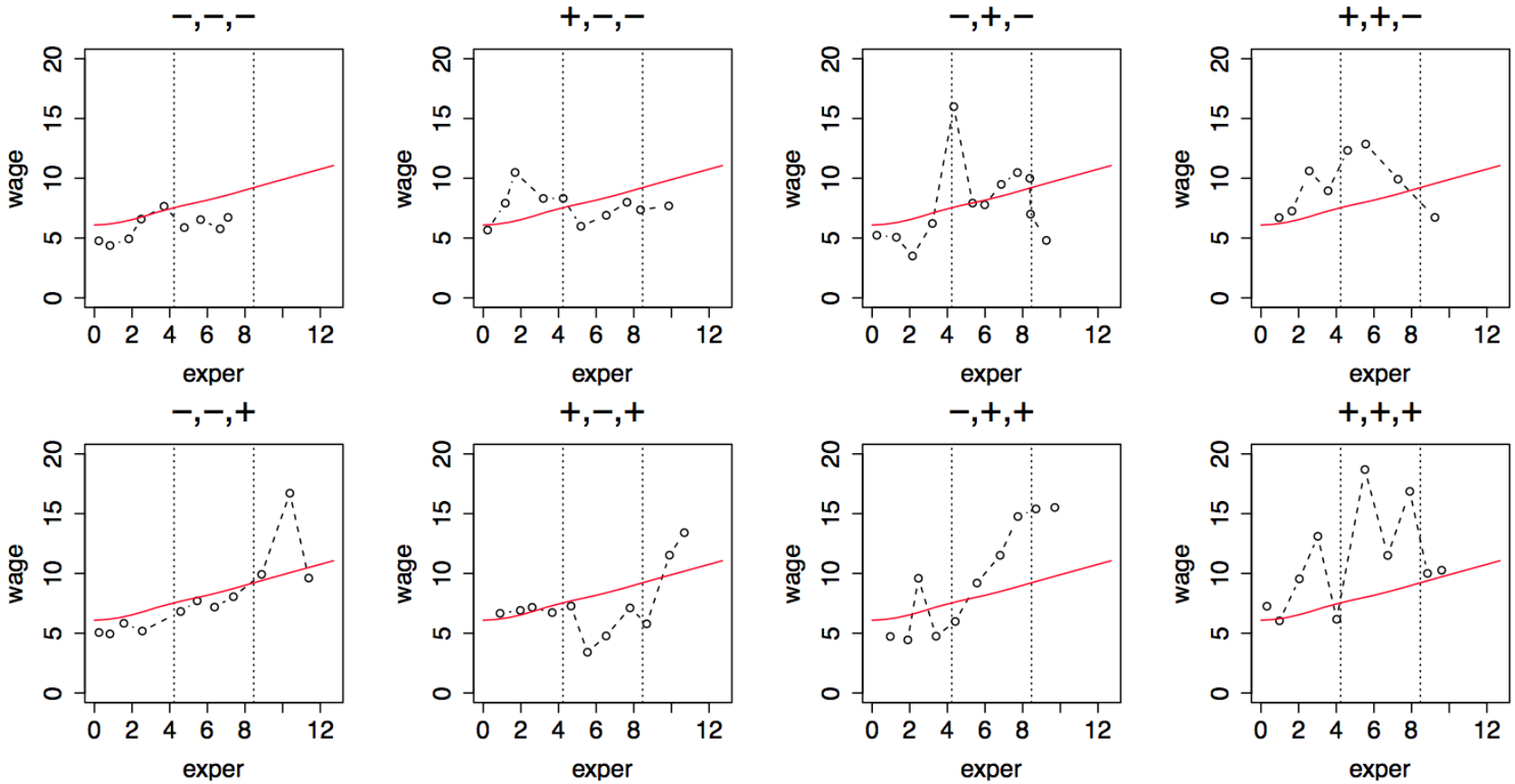


Figure: Some individual trajectories

Main References

- Alexander, W. P., and Grimshaw, S. D. (1996) Treed regression, *JCGS*, **5**, 156-175.
- Boulesteix, A-L., and Strimmer, K. (2006). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data, *Brief Bioinform*, **8**, 32-44.
- Eo, S-H and Cho, H (2014). **Tree-structured mixed-effects regression modeling for longitudinal data**,*JCGS*, .
- Erriksson, L., Trygg, J., and Wold, S. (2009). **PLS trees, a top-down clustering approach**, *J.Chem*, **23**, 569-580.
- Kettaneh, N., Berglund A., and Wold, S. (2005). PCA and PLS with very large data sets, *CSDA*, **48**, 69-85.
- Loh, W-Y. (2002). Regression trees with unbiased variable selection and interaction detection, *Stat Sinicai*, **12**, 361-386.
- Loh, W-Y. (2009).**Improving the precision of classification trees**, *AOAS*, **3**, 1710-1737.
- Loh, W-Y., and Zheng, W (2013). Regression trees for longitudinal and multi-response data, *AOAS*, **7**, 495-522.
- Suh, H-Y. *et al.* (2012). A more accurate method of predicting soft-tissue changes after mandibular setback surgery, *JOMS*, **70**, 553-562.
- Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning, *JCGS*, **17**, 492-514.