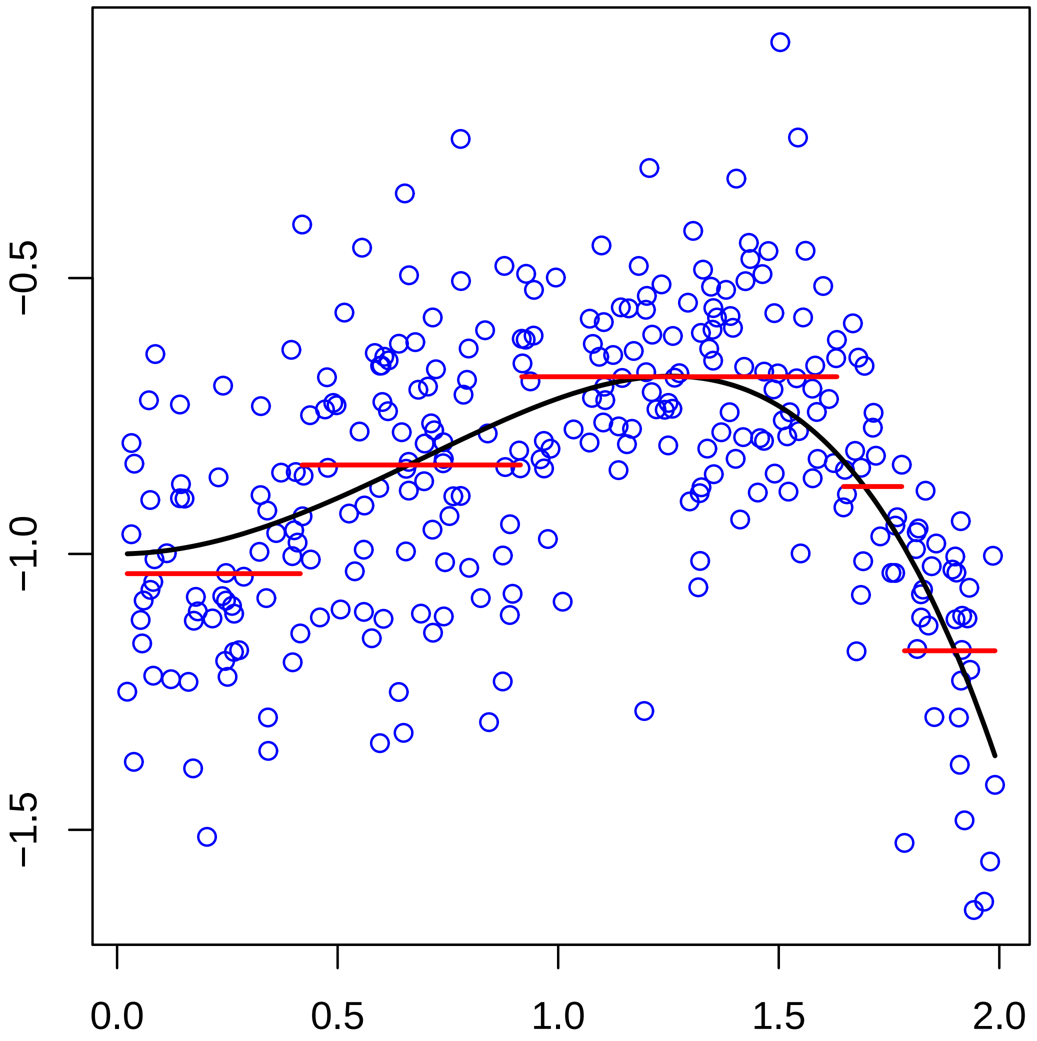
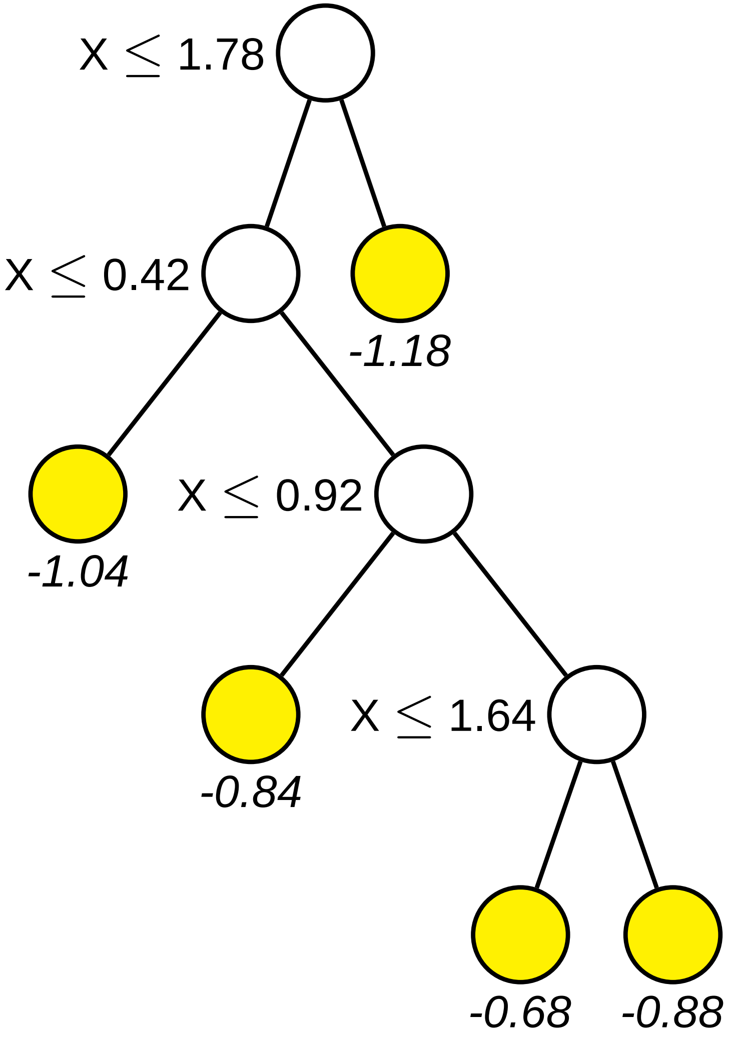
Regression tree models for multi-response and longitudinal data

Wei-Yin Loh
Department of Statistics
University of Wisconsin–Madison
http://www.stat.wisc.edu/∼loh/

Example of a piecewise-constant regression tree





CART approach for univariate response

1. Recursively partition the data:
   (a) Examine every allowable split on each predictor variable
   (b) Select and execute (create left and right daughter nodes) the best of these splits
   (c) Stop splitting a node if the sample size is too small
2. Prune the tree using cross-validation
3. Use surrogate splits to deal with missing values

Shortcomings of the CART approach

1. Biased toward selecting variables with more splits
2. Biased toward selecting variables with more (classification) or less (regression) missing values
3. Biased toward selecting surrogate variables with more missing values
4. Erroneous results if categorical variables have more than 32 values (RPART and commercial version of CART)

Extensions of CART to longitudinal data

Segal (JASA, 1992).

1. Assume AR(1) or compound symmetry structure in each node.
2. Use EM and multivariate normality to handle missing response values.
3. Assume compound symmetry if observation times are irregular.

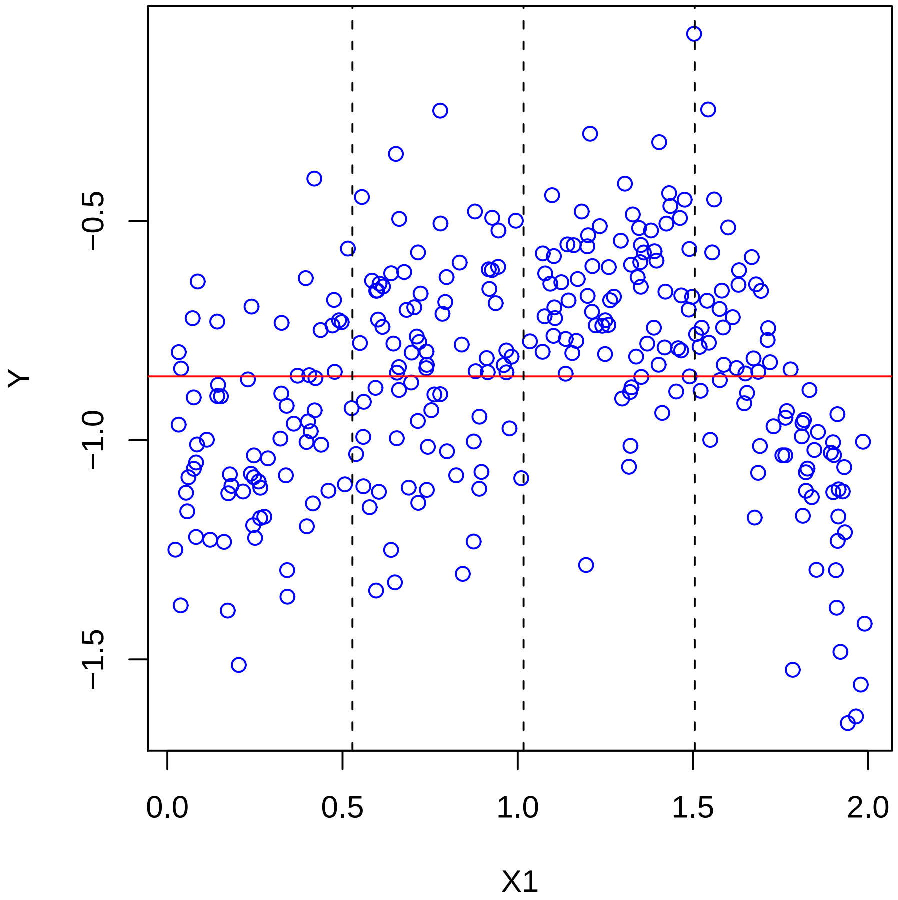
Zhang (JASA, 1998).

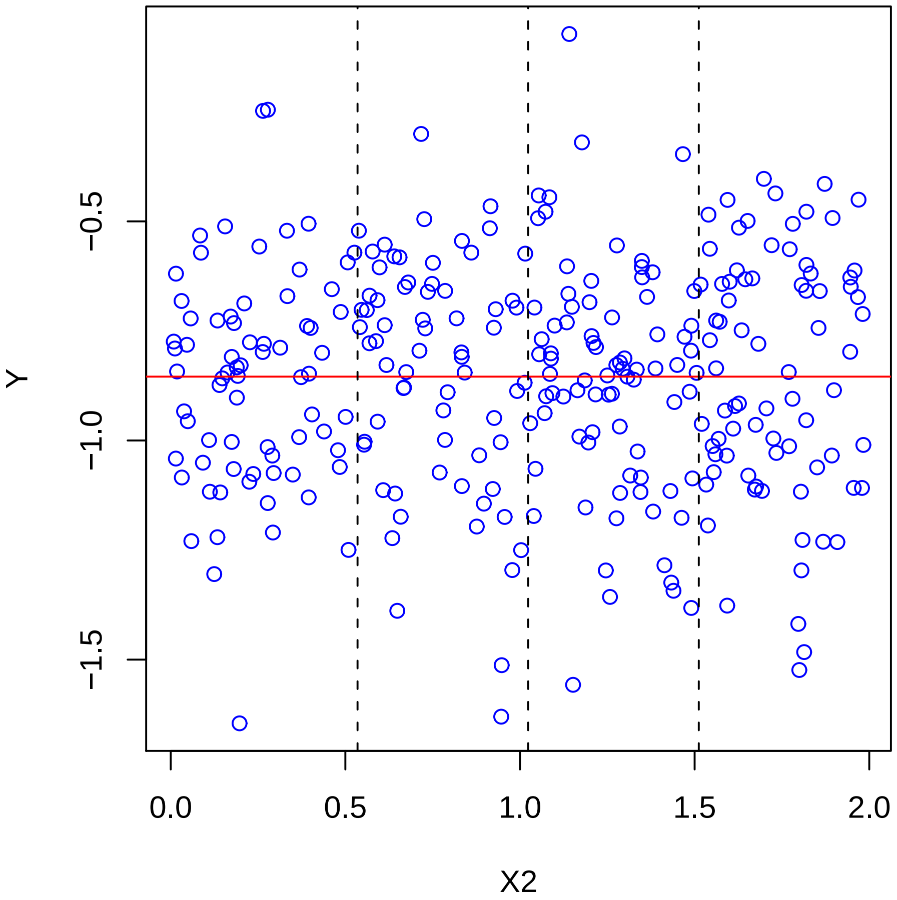
1. Assuming binary response variables, use log-likelihood of exponential family distribution as impurity criterion.

Yu and Lambert (JCGS, 1999).

1. Fit tree model with coefficients of a fitted spline function or a small number of the largest principal components.
2. Get predicted Y values in nodes from fitted spline functions or principal component scores.

Split variable selection based on residual patterns





Pos. res.	18	49	68	27
Neg. res.	52	31	10	45

χ^2_3 = 66.7, p = 2 × 10^{-14}

Pos. res.	37	41	45	39
Neg. res.	34	28	39	37

χ^2_3 = 1.14, p = 0.77

GUIDE (Loh 2002, 2009) split variable selection

1. Fit a model to the data in the node
2. Compute the residuals
3. For each ordered variable X (no grouping for categorical X):
   (a) Group its values into 3–4 intervals
   (b) Cross-tab the signs of the residuals vs. interval membership
   (c) Compute Pearson chi-squared statistic
4. Select the X with most significant chi-squared value

Four important consequences (vs. CART, C4.5, etc.)

1. Unbiased variable selection for piecewise-constant trees
2. Extensible to piecewise-linear and more complex models
3. Substantial computational savings if number of variables or samples is large
4. Chi-squared statistics form the basis for importance scoring of variables

Attempted extension of GUIDE to longitudinal data

Lee (CSDA, 2005).

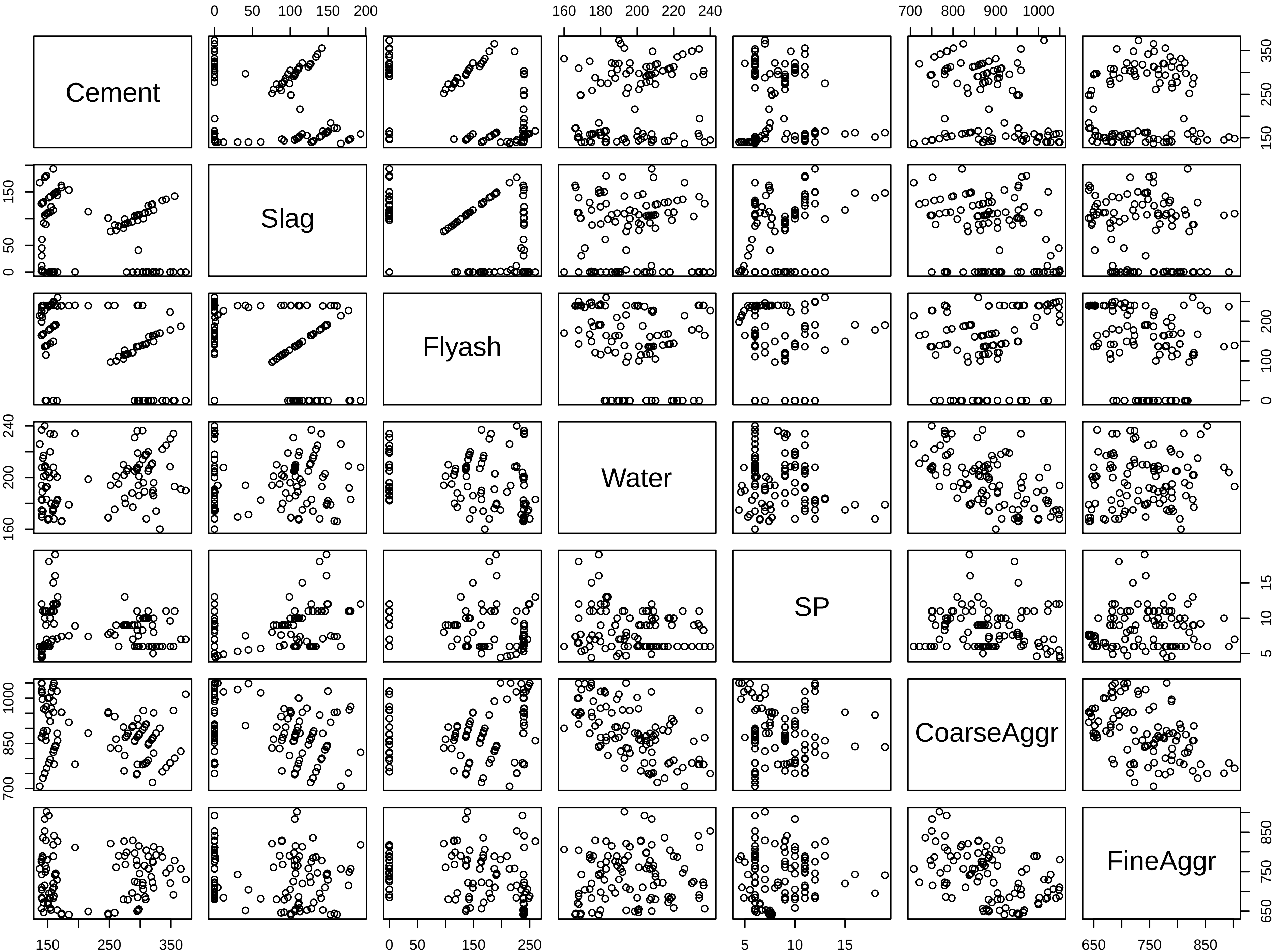
1. Fit a GEE model to the data in each node.
2. For each individual i, compute r_i, the **sum** of the standardized residuals over the time points.
3. Find p-value of t-test of two groups defined by signs of r_i for each X.
4. Split node with most significant X.
5. Use as split point a weighted average of the means of X in the two groups.
6. Stop splitting if p-value is insufficiently small.
7. Not applicable to categorical X variables.

Multi-response: viscosity and strength of concrete

- 103 observations on seven input variables (kg per cubic meter):
  1. Cement
  2. Slag
  3. Fly ash
  4. Water
  5. Superplasticizer
  6. Coarse aggregate
  7. Fine aggregate
- Three output (dependent) variables:
  1. Slump (cm)
  2. Flow (cm)
  3. 28-day compressive strength (Mpa)
- Ref: Yeh, I-C (2007), *Cement and Concrete Composites*, vol 29, 474–480

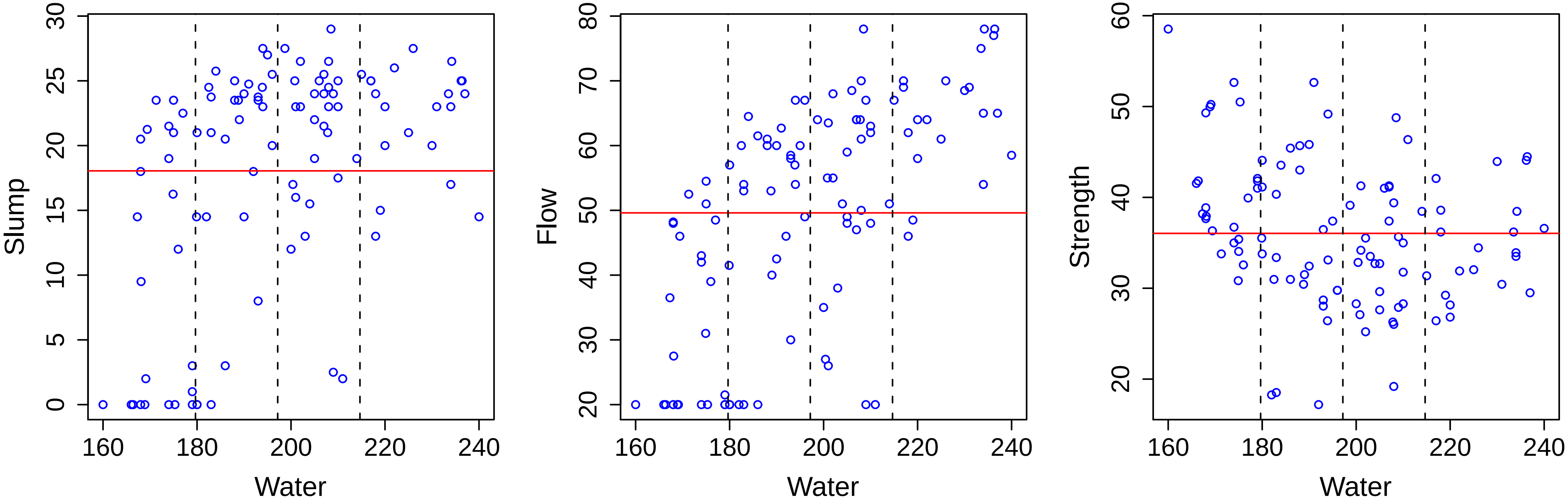
Separate linear models

	Slump		Flow		Strength	
	Estimate	P-value	Estimate	P-value	Estimate	P-value
(Intercept)	-88.53	0.66	-252.87	0.47	139.78	0.052
Cement	0.01	0.88	0.05	0.63	0.06	0.008**
Slag	-0.01	0.89	-0.01	0.97	-0.03	0.352
Flyash	0.01	0.93	0.06	0.59	0.05	0.032*
Water	0.26	0.21	0.73	0.04*	-0.23	0.002**
Superplasticizer	-0.18	0.63	0.30	0.65	0.10	0.445
Coarse aggregate	0.03	0.71	0.07	0.59	-0.06	0.045*
Fine aggregate	0.03	0.64	0.09	0.51	-0.04	0.178


Cement
Slag
Flyash
Water
SP
CoarseAggr
FineAggr
0
50
100
150
200
160
180
200
220
240
700
800
900
1000
350
250
150
0
100
200
15
10
5
850
750
650
150
50
0
240
200
160
1000
850
700
150
250
350
0
50
150
250
5
10
15
650
750
850

Patterns of residuals of

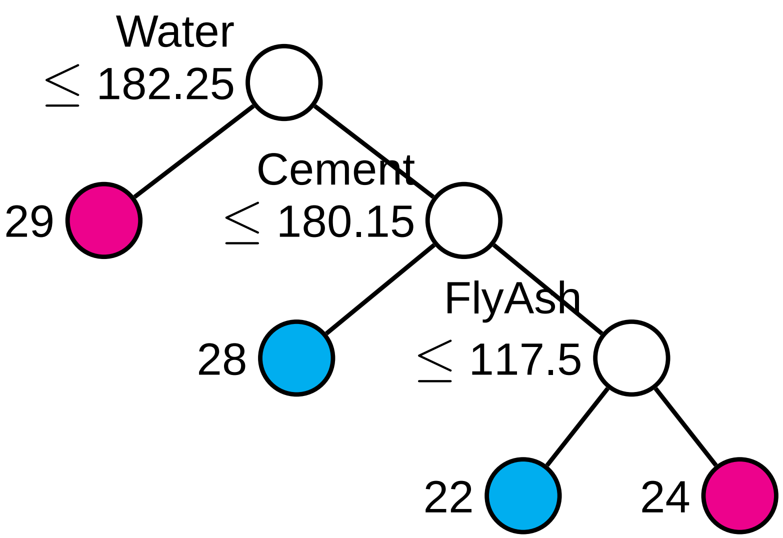
Slump, Flow and Strength vs. Water

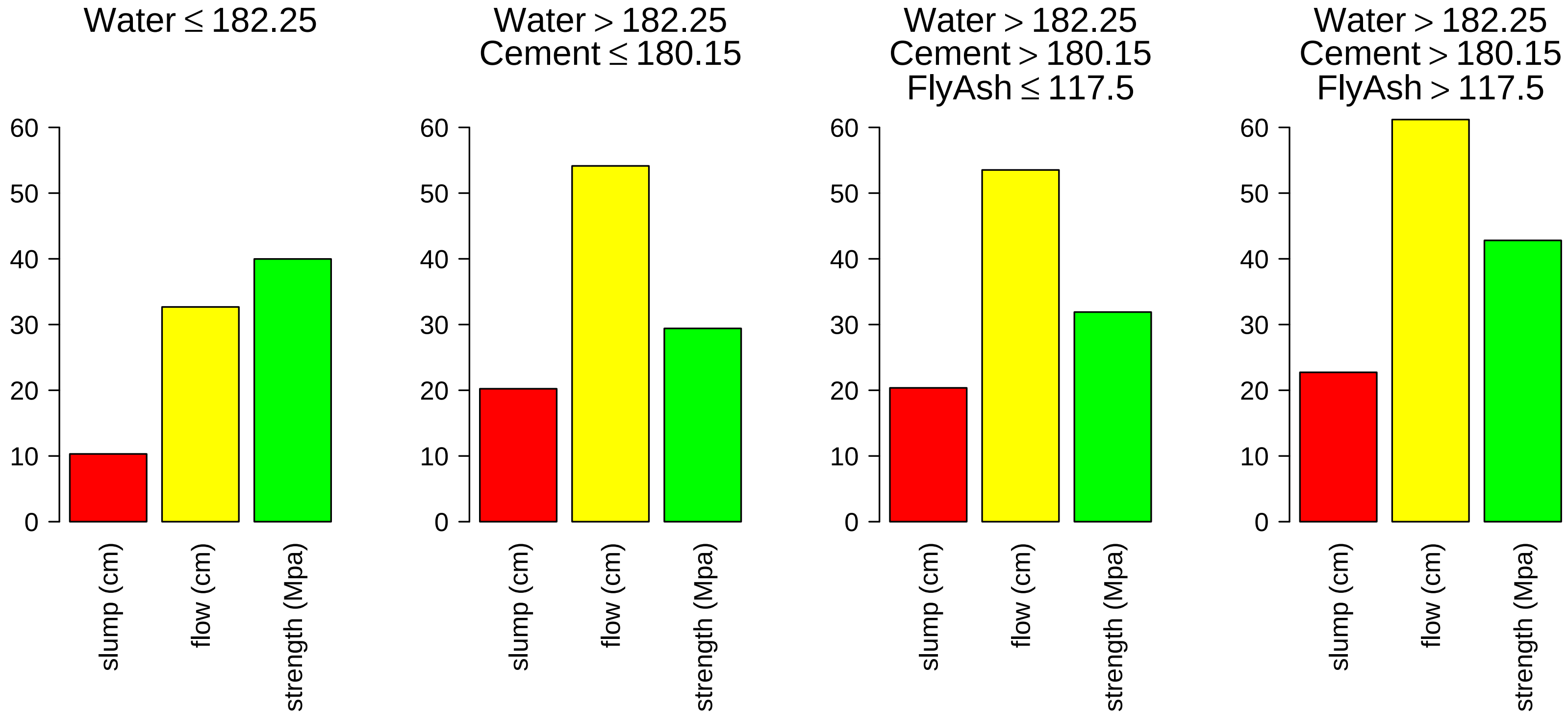


Residual sign patterns vs. Water

			Water			
Slump	Flow	Strength	$\leq$ 180	(180, 197]	(197, 215]	$>$ 215
$-$	$-$	$-$	2	6	5	1
$-$	$-$	$+$	14	3	2	1
$-$	$+$	$-$	0	0	1	1
$-$	$+$	$+$	0	0	0	1
$+$	$-$	$-$	1	2	2	0
$+$	$+$	$+$	4	0	1	0
$+$	$+$	$-$	3	9	11	10
$+$	$+$	$+$	0	9	7	7

χ^2_{21} = 57.1, p-value = 3.5 $\times$ 10^{-5}


Water
≤ 182.25
29
Cement
≤ 180.15
28
FlyAsh
≤ 117.5
22
24


Water ≤ 182.25
Water > 182.25
Cement ≤ 180.15
Water > 182.25
Cement > 180.15
FlyAsh ≤ 117.5
Water > 182.25
Cement > 180.15
FlyAsh > 117.5
0
10
20
30
40
50
60
slump (cm)
flow (cm)
strength (Mpa)

Longitudinal data example:

CD4 counts from an AIDS clinical trial

- Randomized, double-blind, study of 1309 AIDS patients with advanced immune suppression (Fitzmaurice, Laird and Ware, *Applied Longitudinal Analysis*)
- Four dual or triple combinations of HIV-1 reverse transcriptase inhibitors:

  1: 600mg *zidovudine* alternating monthly with 400mg *didanosine* (dual therapy)

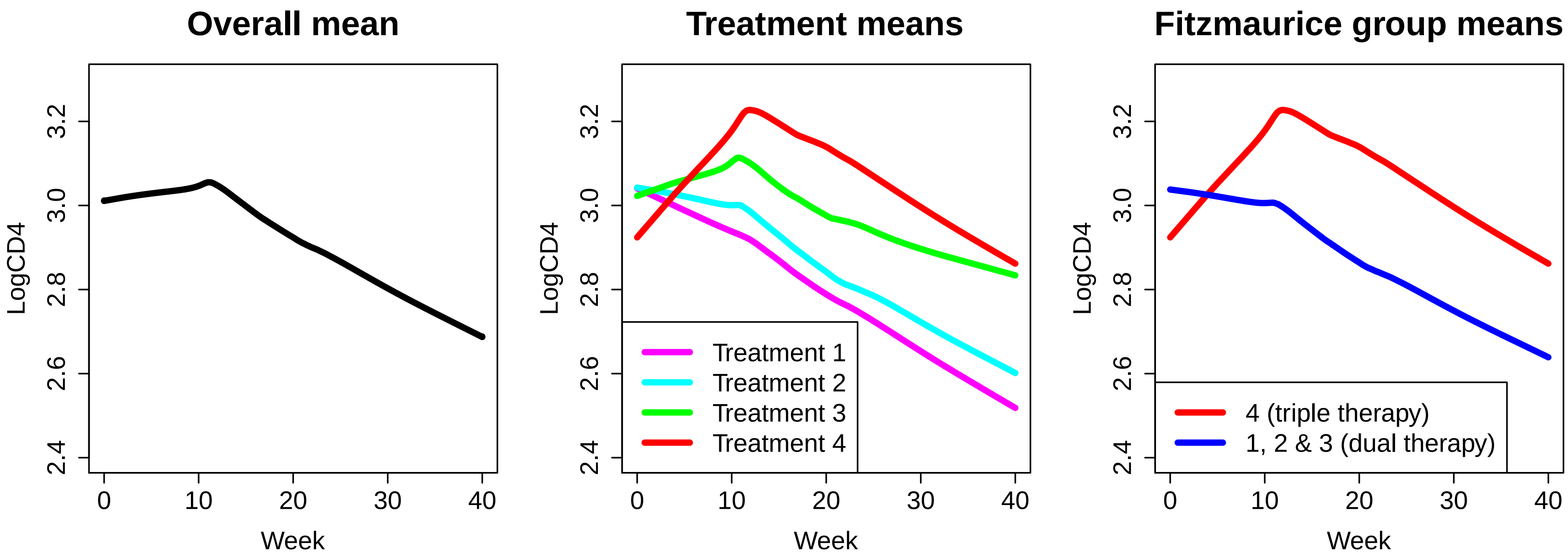
  2: 600mg *zidovudine* + 2.25mg *zalcitabine* (dual therapy)

  3: 600mg *zidovudine* + 400mg *didanosine* (dual therapy)

  4: 600mg *zidovudine* + 400mg *didanosine* + 400mg *nevirapine* (triple therapy)
- CD4 counts collected at baseline and at 8-week intervals during 40-week follow-up
- Patient observations during follow-up period varied from 1–9, with median of 4

  1. mistimed measurements
  2. missing measurements due to skipped visits and dropout
- Response variable is log(CD4 counts + 1)

Lowess smooths



Overall mean

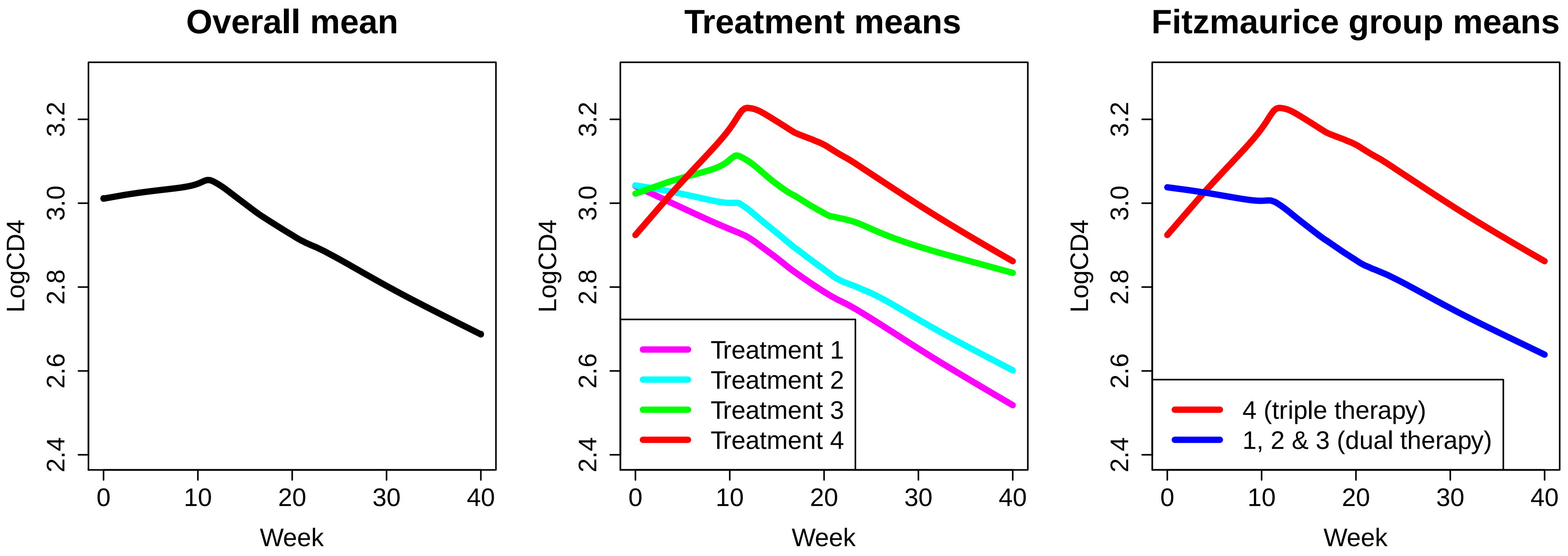
Treatment means

Fitzmaurice group means

Fitzmaurice et al. linear mixed effects model

$$
\begin{aligned}
E(Y_{ij} \mid b_i) &= \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij} - 16)_+ + \beta_4 I(\texttt{Trt} = 4) \times t_{ij} \\
&\quad + \beta_5 I(\texttt{Trt} = 4) \times (t_{ij} - 16)_+ + b_{1i} + b_{2i} t_{ij} + b_{3i} (t_{ij} - 16)_+
\end{aligned}
$$

Fitzmaurice et al. conclusions


Overall mean

Treatment means

Fitzmaurice group means

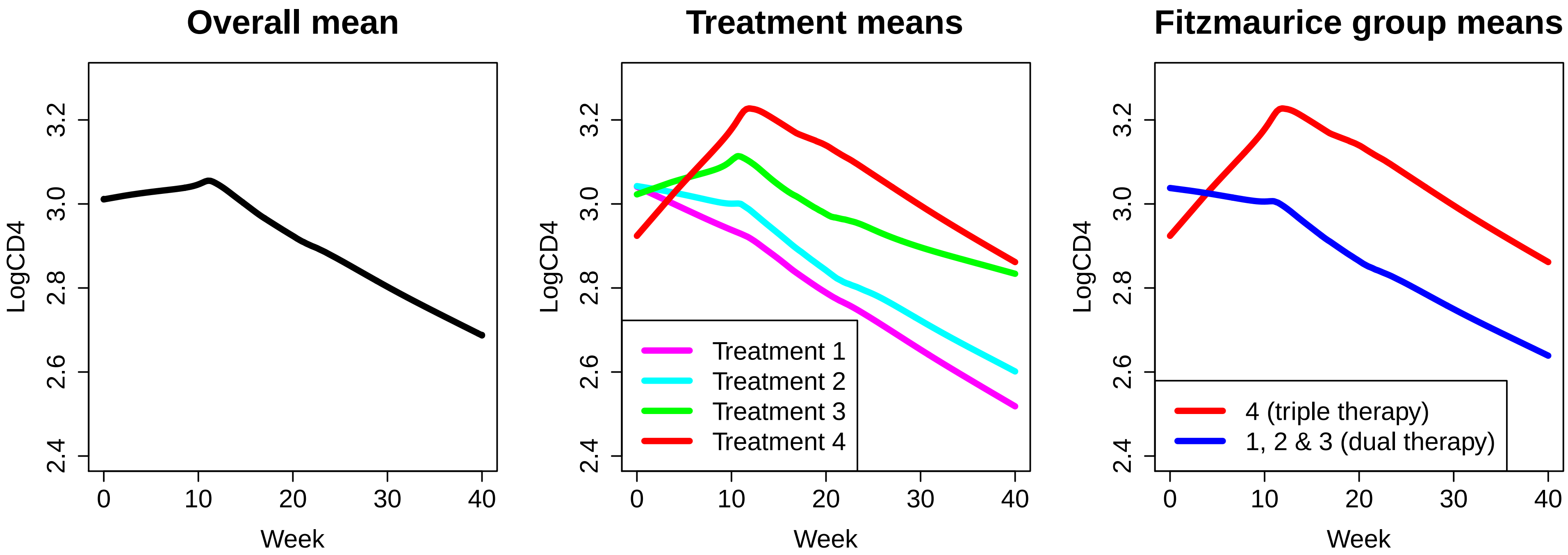
1. All fixed effects significant ($p < 0.005$)

2. Sig. diff. in rates of change from baseline to week 16 between dual and triple therapies

3. No sig. differences in rates of change from week 16 to 40 between the two groups

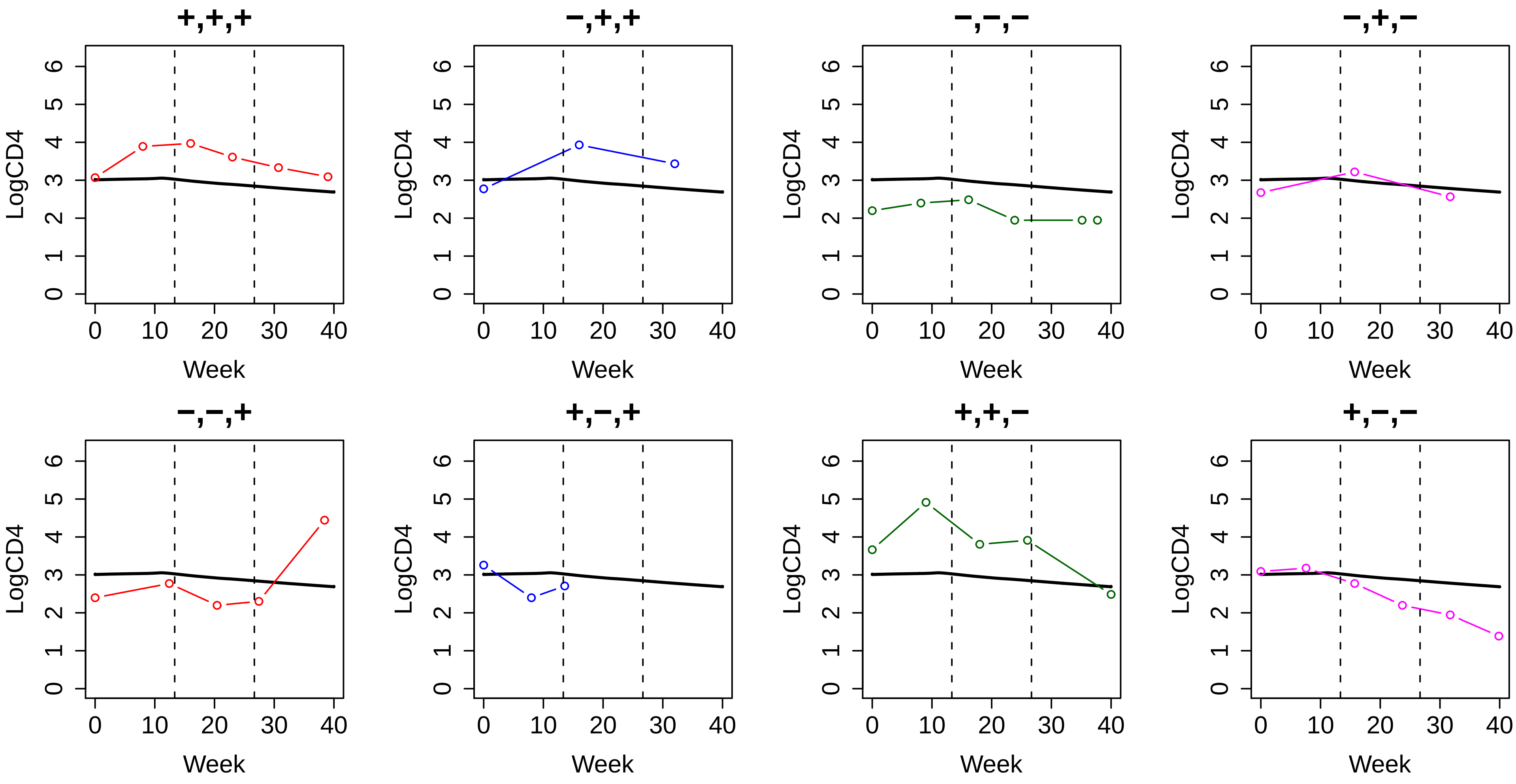
4. Substantial within and between-patient variability (large random effects)

Weaknesses in linear mixed model approach



1. Statistical inference is predicated on assumption that the parametric model is correct
2. Parametric model is subjective, often chosen after looking at the data (difficult to do if there are many predictor variables)
3. Different smoothers yield different models (assumed change point of 16 weeks is suspect)
4. Assumption of constant slopes after change point is similarly suspect

Examples of eight trajectory shapes



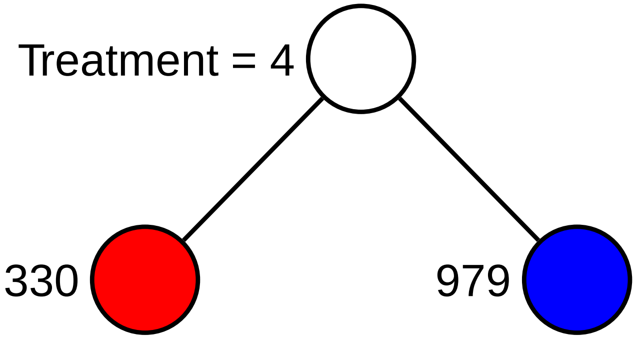
Chi-squared tests of trajectory patterns vs. X

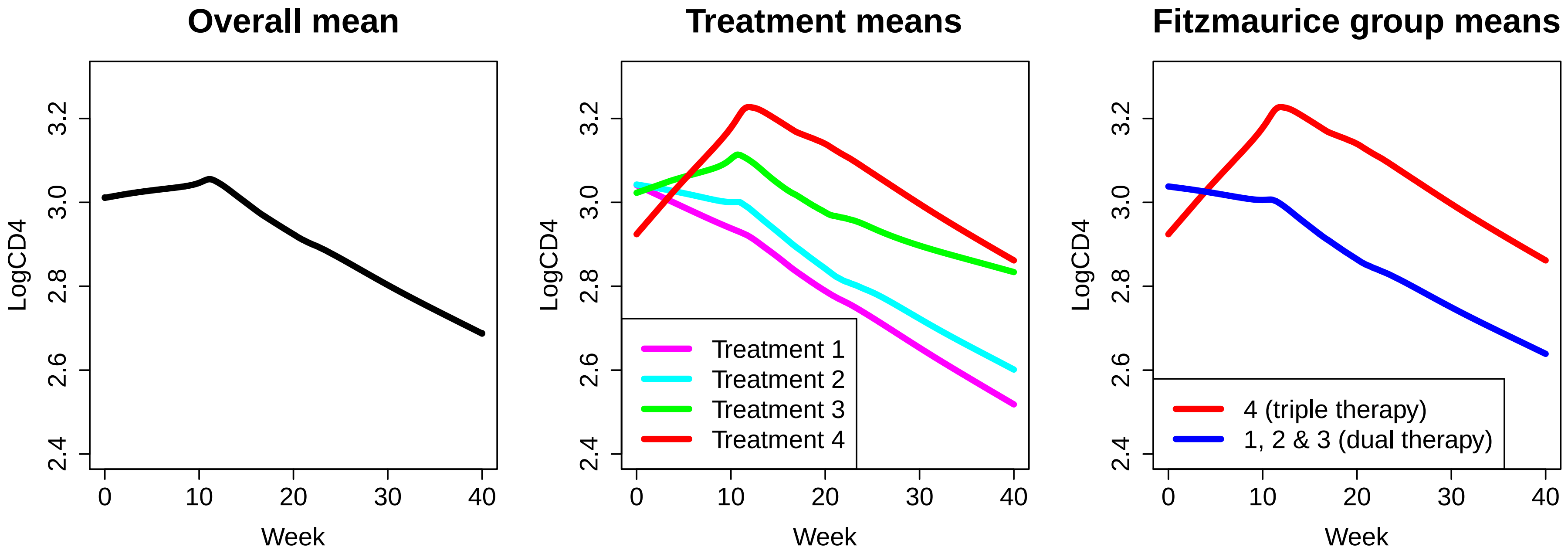
	Ordered X			Unordered X		
Pattern	$(-\infty, a]$	$(a, b]$	(b, ∞)	$X = c_1$	$\cdots$	$X = c_k$
(–,–,–)						
(+,–,–)						
(–,+,–)						
(+,+,–)						
(–,–,+)						
(+,–,+)						
(–,+,+)						
(+,+,+)						

Extension of GUIDE to longitudinal data

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 predictor variable X, find p-value of chi-squared 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 of trajectories from mean curves in the two child nodes
8. Stop splitting when sample size in node is too small
9. Prune the tree using cross-validation

GUIDE model (same grouping as Fitzmaurice et al.)





A somewhat harder example:

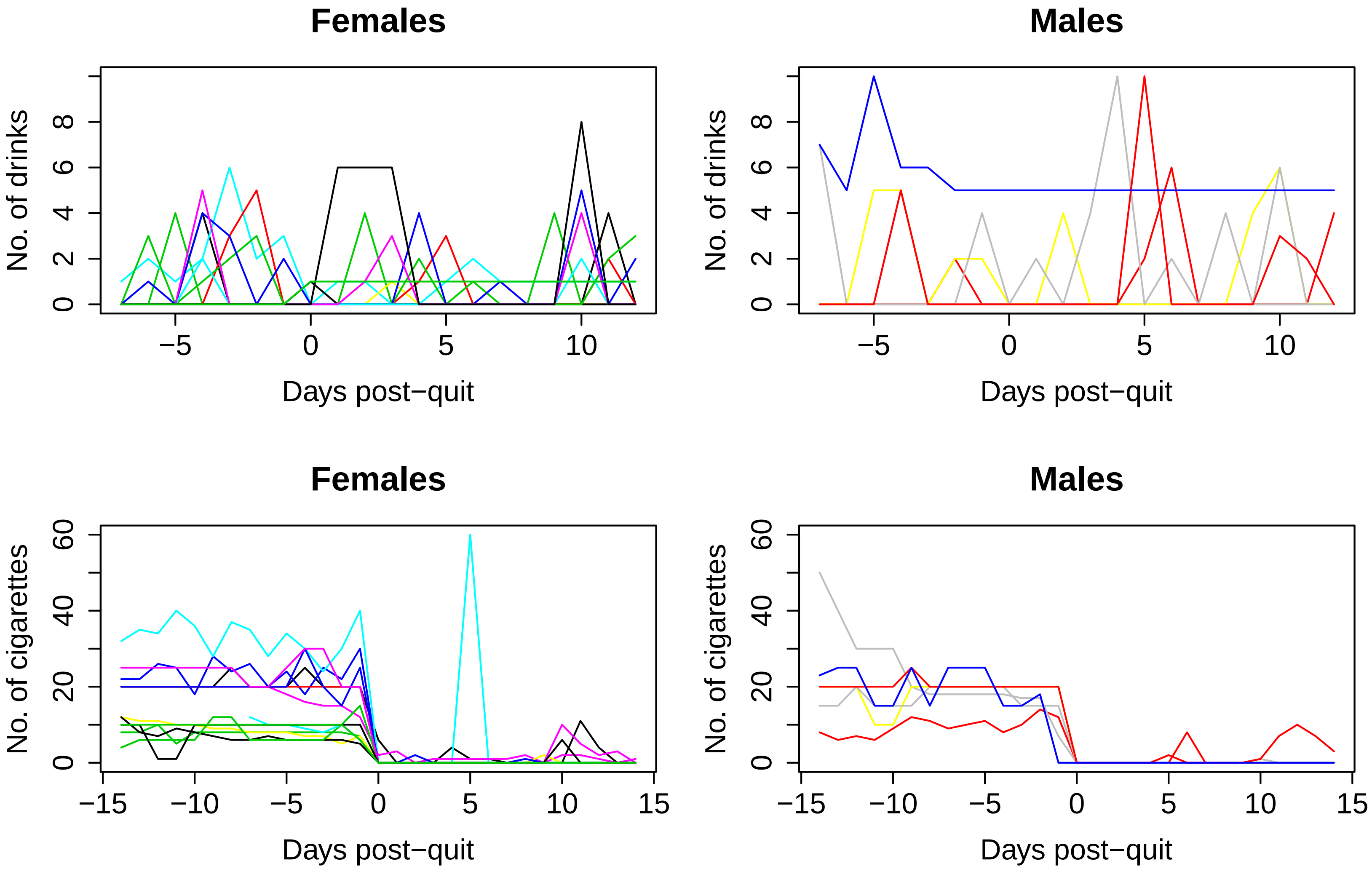
smoking cessation clinical trial

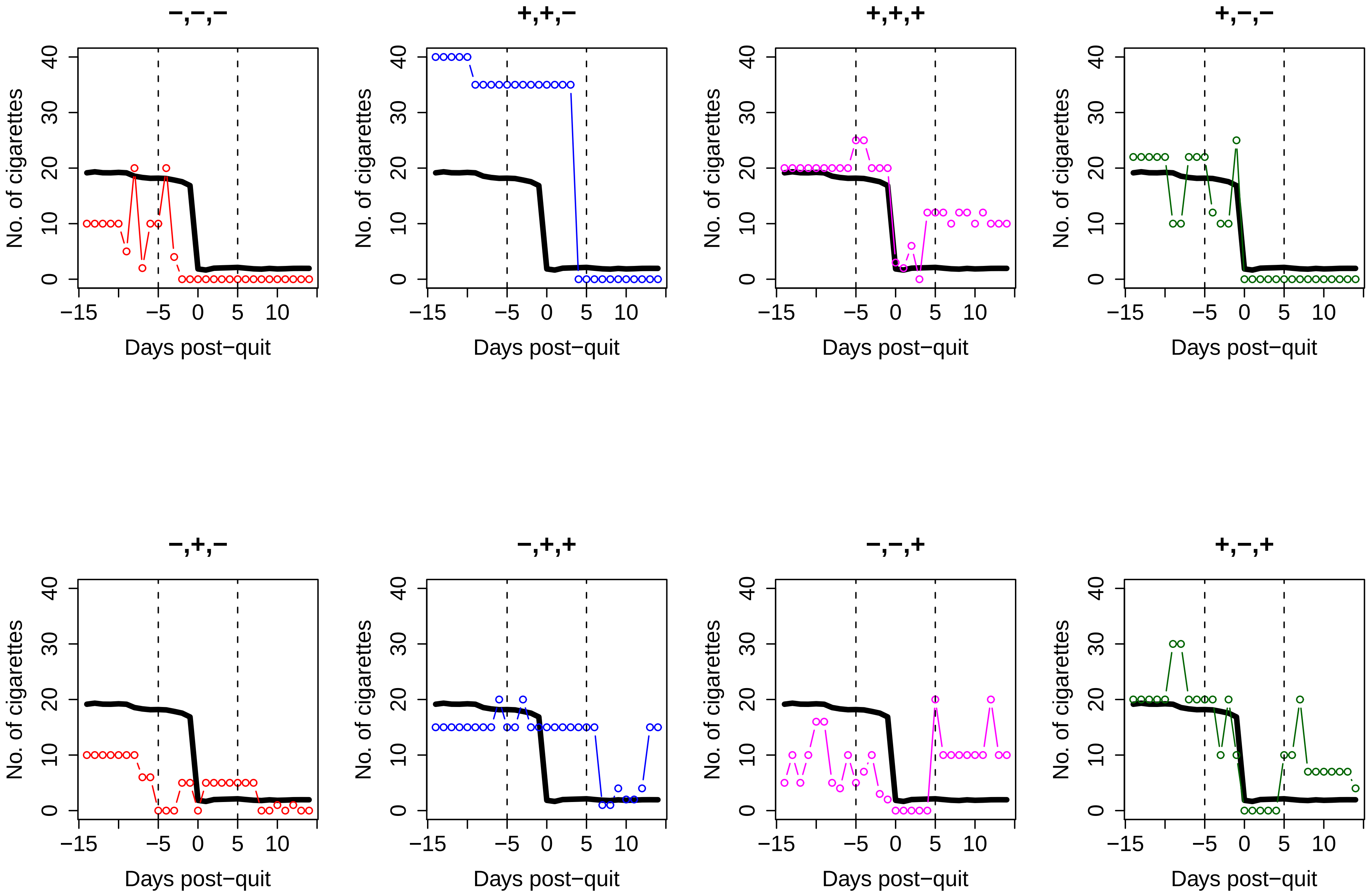
- Responses are number of drinks and cigarettes consumed daily for two weeks before and after quit day for 1470 persons
- 135 explanatory variables with 0–1308 missing values
- 32–63% persons missing drink responses in 8–14 days pre-quit and 13–14 days post-quit
- Goal: model drinking and smoking responses *jointly*

135 explanatory variables

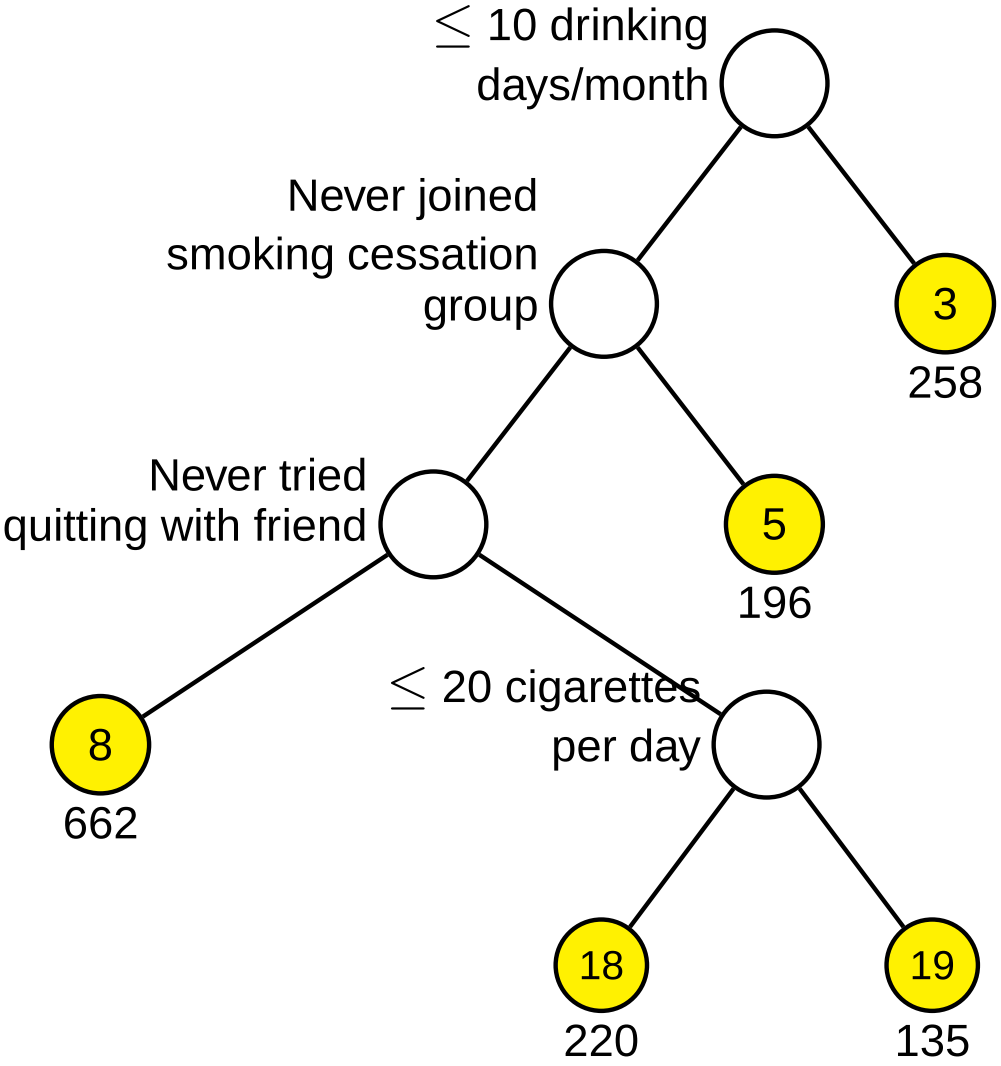
- Age, gender, marital status, education, income, race
- Age 1st cigarette, years smoked, cigarette type
- Various measures of emotional attachment to smoking
- Living and working environments w.r.t. smoking
- Number of past attempts at quitting and quitting methods
- Tobacco dependence scores (FTND, PRISM, WISDM)
- Baseline health and physical measurements (blood pressure, BMI, etc.)
- Treatment type
- Past drinking frequency

20 drinking and smoking profiles by gender

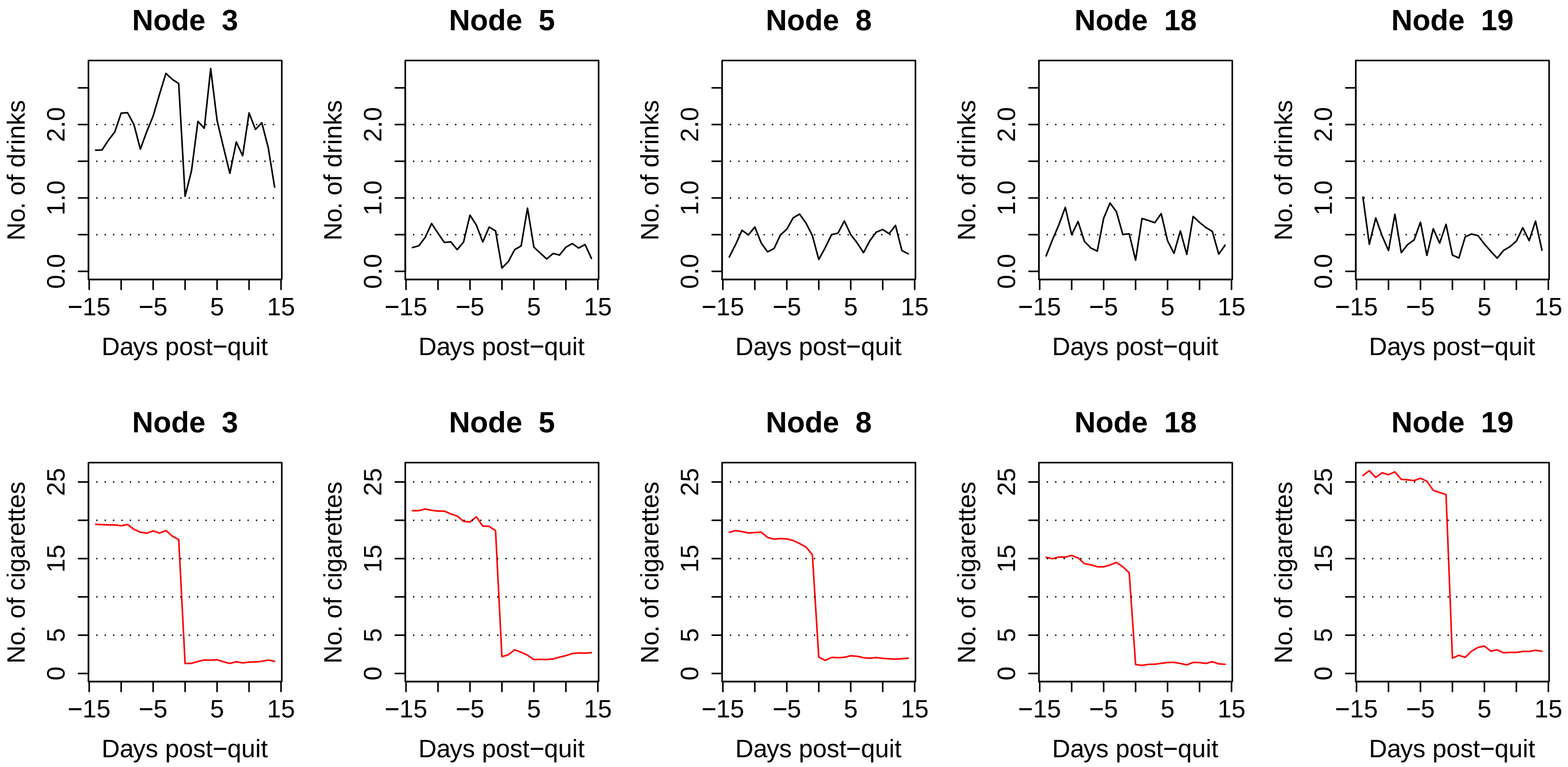



−,−,−
+,+,−
+,+,+
+,−,−
−,+,−
−,+,+
−,−,+
+,−,+
No. of cigarettes
0
10
20
30
40
Days post−quit
−15
−5
0
5
10

Longitudinal regression tree for drinking & smoking


$\leq$ 10 drinking days/month
Never joined smoking cessation group
3
258
Never tried quitting with friend
5
196
8
662
$\leq$ 20 cigarettes per day
18
220
19
135

Mean drinking and smoking profiles in leaf nodes



How to include trajectory fluctuations?

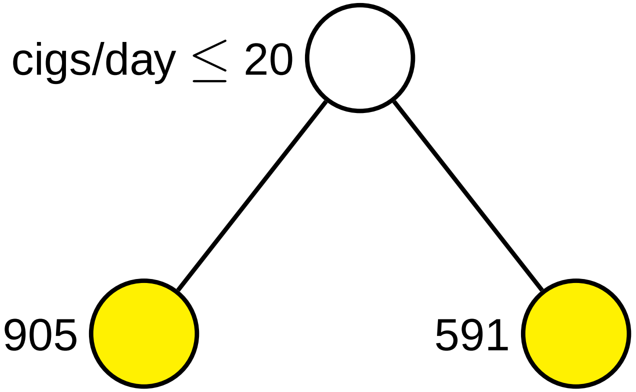
- Let t_0 denote the quit day. Define the absolute deviation z_t at time t to be

$$z_t = \begin{cases} |y_t - y_{t-1}|, & t = t_0 - 1 \\ |y_t - y_{t+1}|, & t = t_0 \\ (|y_t - y_{t-1}| + |y_t - y_{t+1}|)/2, & \text{otherwise} \end{cases}$$

- This yields a “deviation” trajectory for each individual

- Join the smoking and deviation trajectories and fit a regression tree to them

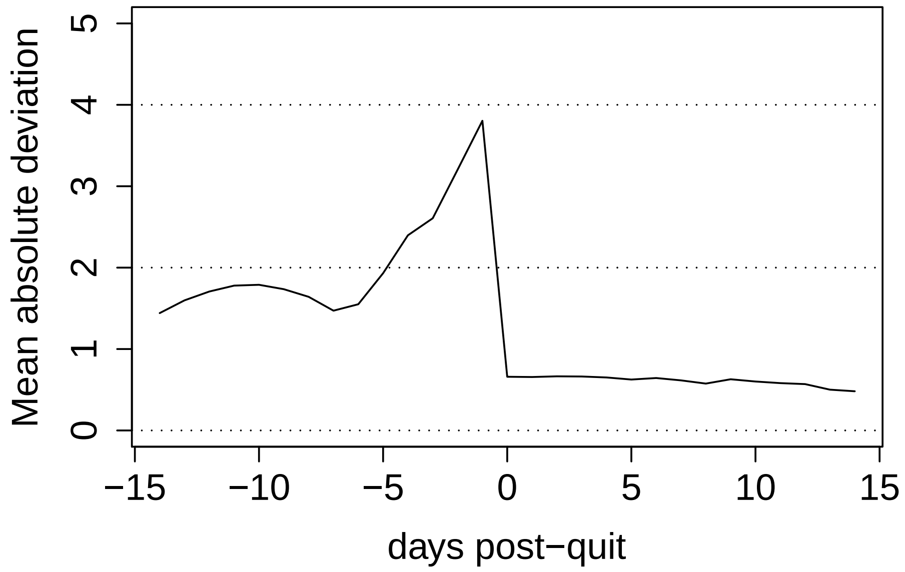
Longitudinal regression tree for number and deviation in cigarettes smoked



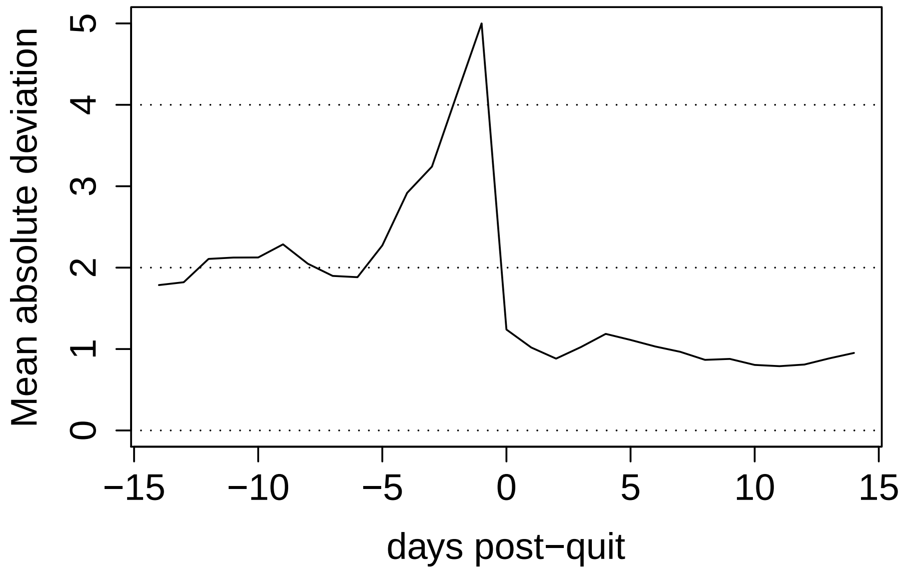
An observation goes to the left branch if and only if it satisfies the stated condition

Sample sizes given on the left of leaf nodes

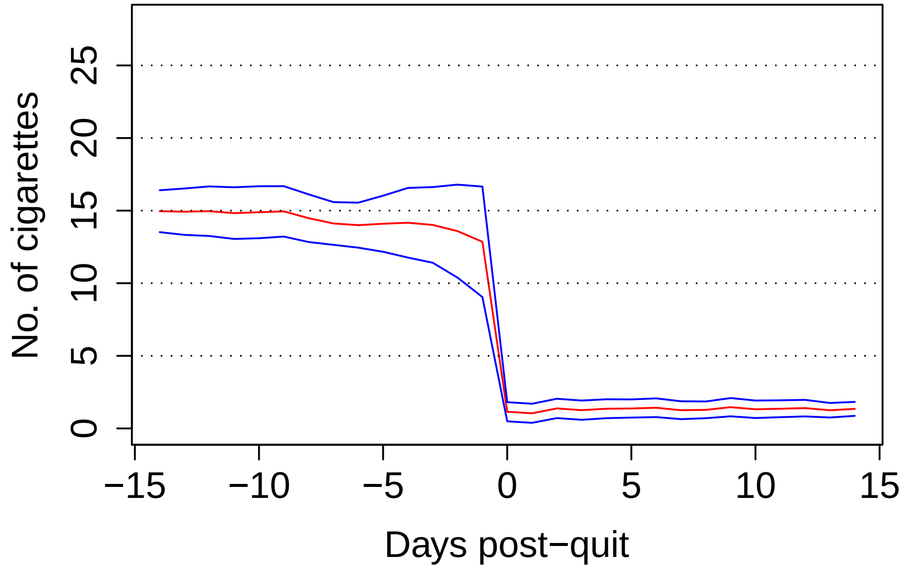
Cigarettes per day ≤ 20



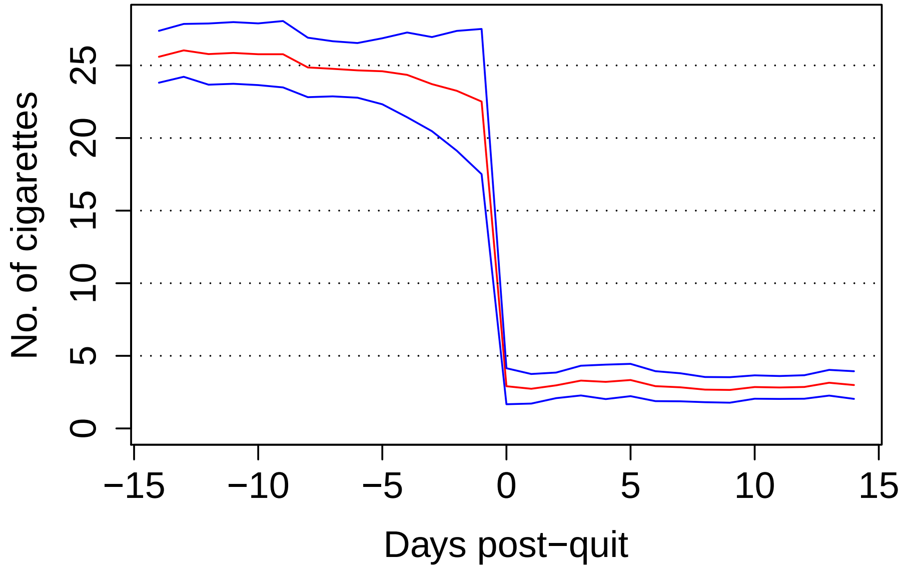
Cigarettes per day > 20



Cigarettes per day ≤ 20



Cigarettes per day > 20



Bayes risk consistency of piecewise-constant models (Breiman et al. 1984)

Theorem. Suppose that $E|Y|^q < \infty$ for some $1 \leq q < \infty$. Let $p_N(t)$ be the proportion of observations in node t such that $p_N(t) \geq k_N \log(N)/N$ for some k_N. Let $D_N(x)$ denote the diameter of the node containing x. Assume

$$k_N \to \infty \text{ and } D_N(X) \xrightarrow{P} 0 \text{ as } N \to \infty. \qquad (1)$$

Let $d_B(x) = E(Y \mid X = x)$ and $d_N(x)$ be the piecewise constant regression tree estimate of $d_B(x)$. Then $E|d_N(X) - d_B(X)|^q \to 0$.

Theorem. Given any function d on $\mathcal{X}$, let $R(d) = E[Y - d(X)]^2$. Suppose that $EY^2 < \infty$ and that condition (1) holds. Then $\{d_N\}$ is *risk consistent*, i.e., $ER(d_N) \to R(d_B)$ as $N \to \infty$.

Asymptotic uniform consistency

(Kim, Loh, Shih & Chaudhuri 2007)

Let $f(x) = E(Y|x)$ be continuous in a compact rectangle C. Suppose there is $a > 0$ such that

$$\sup_{x \in C} E\{\exp(a|Y - f(x)|) \mid X = x\} < \infty.$$

Let T_N be the regression tree based on training sample size N, m_N = minimum node sample size, and $\delta(t) = \sup_{x,z \in t} \|x - z\|$ be the diameter of node t.

Assume that as $N \to \infty$,

1. $(\log N)/m_N \xrightarrow{P} 0$
2. $\sup_{t \in T_N} \delta(t) \xrightarrow{P} 0$
3. Minimum eigenvalue of node design matrices is bounded from 0 in probability

Let $\hat{f}(x)$ be the regression estimate at x. Then

$$\sup_{x \in C} |\hat{f}(x) - f(x)| \xrightarrow{P} 0.$$

Conclusion

- GUIDE extension to multi-response and longitudinal data is applicable to irregularly observed and missing data
- Does not use mixed effect models — no need to estimate covariance matrices
- Dependence in longitudinal data handled by treating data as curves — shapes of the curves are used to select variables for splitting
- Curves are clustered according to the predictor variables — traditional cluster analysis does not use predictor variables
- Purpose is exploratory analysis; objective is prediction

Software availability

Mac, Linux and Windows executables for GUIDE may be obtained from:

http://www.stat.wisc.edu/∼loh/guide.html