## Mη παραμετρική παλινδρόμηση ΙV: Πολλαπλή μη παραμετρική παλινδρόμηση, GAMs και Regression Trees

### Δένδρα παλινδόμησης (regression trees)

**Iδέα:** αν , βρές διαμέριση του  σε ξένα μεταξύ τους «παραλληλεπίπεδα»  με πλευρές παράλληλες στους άξονες , που να δίνουν τη καλύτερη δυνατή προσαρμογή της μορφής 

π.χ. 



**Πως** (π.χ. )

* Αρχή: για κάθε συντεταγμένη (μεταβλητή)  ,
	+ εξέτασε όλες τις διαμερίσεις του R, σε υποσύνολα που ορίζονται από  και , για όλα τα πιθανά s (αρκεί να πάρουμε για s όλες τις συντεταγμένες των  παρατηρήσεων μιας )
	+ για κάθε διαμέριση  θέτουμε μέσος όρος των  για τα οποία .
	+ Βέλτιστη διαμέριση είναι εκείνη που ελαχιστοποιεί την 
* Συνεχίζουμε διαδοχικά διαμερίζοντας «βέλτιστα» κάθε μια από τις υποπεριοχές που έχουμε λάβει στα προηγούμενα βήματα

Το αποτέλεσμα αυτής της διαδικασίας μπορεί να παρασταθεί ως ένα «**δυαδικό δένδρο»,** όπου κάθε τελικό κόμβο έχουμε μια υποπεριοχή , ενώ σε κάθε εσωτερικό κόμβο έχουμε μια διακλάδωση με βάση μια συντεταγμένη  και μια τιμή s,



From Patrick Breheny’s course at the university of Kentucky



****

#### Πόσο «μεγάλο» να διαλέξουμε το δένδρο Τ;

Πρώτα βρες μέγιστο δένδρο.

Μετά: Έστω

 , όπου .

 Θέτουμε

=

όπου  παράμετρος εξομάλυνσης και  ο αριθμός των τελικών κόμβων του Τ

* Για ένα πλέγμα τιμών του α ,
	+ Για κάθε τιμή του α στο  βρες , το υπο-δενδρο που ελαχιστοποιεί την , κόβοντας κατ΄εξακολούθηση από ένα μεγαλύτερο δένδρο το «ασθενέστερο» κλαδί, μέχρι να ελαχιστοποιηθεί η .
* Έτσι κατασκευάζεατι μια ακολουθία «βέλτιστων» δένδρων 

**Το  μπορεί να επιλεγεί με cross-validation:**

* Σπάσε το δείμα σε π.χ 10 μέρη (folds)
* Για κάθε ένα από αυτά (k=1,…,10):
	+ Εξαίρεσε τα δεδομένα του k-fold (test-fold) και με όλα τα υπόλοιπα (training sample) ξαναφτειάξε εξ αρχής μια ακολουθία δένδρων  που να ελαχιστοποιούν την  για ένα πλέγμα τιμών του α (, όπως παραπάνω).
	+ Για κάθε δένδρο  (ξανα-) εκτίμησε το μέσο τετραγωνικό σφάλμα πρόγνωσης  χρησιμοποιώντας μόνο τα δεδεμένα του k-fold (test-fold) . Αυτό δίνει  ως συνάρτηση των .
	+ Για κάθε  Σχημάτισε το μέσο όρο των 
	+ Διάλεξε εκείνο το  που ελαχιστοποιεί την .







**Node number 1: 462 observations, complexity param=0.125**

 **Primary splits:**

 age < 50.5 to the left, improve=24.58856, (0 missing)

 tobacco < 0.49 to the left, improve=19.42366, (0 missing)

 famhist splits as LR, improve=15.51823, (0 missing)

 ldl < 4.315 to the left, improve=12.58910, (0 missing)

 adiposity < 25.16 to the left, improve=10.38739, (0 missing)

**Node number 2: 290 observations, complexity param=0.025**

 **Primary splits:**

 age < 30.5 to the left, improve=7.398448, (0 missing)

 tobacco < 0.49 to the left, improve=6.476421, (0 missing)

 typea < 68.5 to the left, improve=5.850580, (0 missing)

 ldl < 4.155 to the left, improve=5.072853, (0 missing)

 adiposity < 25.16 to the left, improve=3.814103, (0 missing)