

Beyond Semi-parametric Models: Trees and CARTscans

.....

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics,
University of Washington

February 8, 2008

1

© 2002, 2003, 2004 Scott S. Emerson, M.D., Ph.D.

Outline

.....

Standard Regression Models
Flexible, Highly Predictive Models

- Classification and Regression Trees

Descriptive Statistics

- CARTscans

Inferential Statistics

Acknowledgements:

- Martha Nason (CARTscans)
- Kyle Rudser (Inference for variable importance)
- Michael Leblanc (CARTscans and inference)

2

Regression Models

.....

According to the parameter compared across groups

- Means → Linear regression
- Geom Means → Linear regression on logs
- Odds → Logistic regression
- Rates → Poisson regression
- Hazards → Proportional Hazards regr
- Quantiles → Parametric survival regr

3

General Regression

.....

General notation for variables and parameter

Y_i	Response measured on the i th subject
X_i	Value of the POI for the i th subject
W_{1i}, W_{2i}, \dots	Value of adjustment variables for the i th subject
θ_i	Parameter of distribution of Y_i

- The parameter might be the mean, geometric mean, odds, rate, instantaneous risk of an event (hazard), etc.

4

Multiple Regression

General notation for simple regression model

$$g(\theta_i) = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_{1i} + \beta_3 \times W_{2i} + \dots$$

$g(\)$ "link" function used for modeling

β_0 "Intercept"

β_1 "Slope for Pred of Interest X"

β_j "Slope for covariate W_{j-1} "

- The link function is usually either none (means) or log (geom mean, odds, hazard)

5

Common Uses of Regression

We have talked about

- Prediction – estimating θ for a particular combination of modeled covariates
- Inference about associations – determining whether the value of θ is markedly (and statistically significantly) different across groups differing in their value for the predictor of interest

6

Borrowing Information

Use other groups to make estimates in groups with sparse data

- Intuitively: 67 and 69 year olds would provide some relevant information about 68 year olds
- Assuming straight line relationship tells us how to adjust data from other (even more distant) age groups
 - If we do not know about the exact functional relationship, we might want to borrow information only close to each group

7

Defining "Contrasts"

Define a comparison across groups to use when answering scientific question

- If straight line relationship in parameter, slope for POI is difference in parameter between groups differing by 1 unit in X when all other covariates in model are equal
- If nonlinear relationship in parameter, slope is average difference in parameter between groups differing by 1 unit in X "holding covariates constant"
 - Statistical jargon: a "contrast" across the groups

8

Drawbacks of Standard Regression

Need to pre-specify

- The covariates to include
- Any transformations of those covariates
- Any terms modeling effect modification and the form of the effect modification

Use in decision rules can be complicated

- Need to compute the “linear predictor”
- Not easily implemented as diagnostic criteria

$$g(\theta_i) = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_{1i} + \beta_3 \times W_{2i} + \dots$$

9

Highly Predictive Models

Statistical modeling techniques that are more “adaptive” (driven by) the data

- Stepwise model building to select which variables to include
- Splines and smoothers to select nonlinear effects
- “Automated interaction detection” to automatically look for interactions
 - “Classification and Regression Trees (CART)”
- Many others

10

Regression Trees

Recursively split the data into two groups that differ the most from each other with respect to some summary measure

- Start with one sample with response Y and p covariates X_1, X_2, \dots
- Perform many, many two sample tests
 - For each covariate and every possible place that covariate can be split
- Split the sample into two based on the biggest difference
- Repeat for each subsample...

11

Example: FEV and Smoking

Start with 654 cases

- Consider every place you can split height, every place you can split age, every place you can split sex, every place you can split smoking
- For each such split perform a t test
- Find the split that gives the smallest p value
 - (It will not have the right type I error, but we do not care about that at this stage)

12

Regression Tree on FEV Data

Is Height less than 61.75 inches?

- YES: Is Height less than 58.75 inches?
 - YES: Is Height less than 55.25 inches?
 - ... (until we decide to stop splitting, then give mean FEV)
 - NO: Is Age < 9.5 years?
 - ... (until we decide to stop splitting, then give mean FEV)
- NO: Is Height < 66.75 inches?
 - YES: Is Height < 64.25 inches?
 - ... (until we decide to stop splitting, then give mean FEV)
 - NO: Is Male < 0.5?
 - ... (until we decide to stop splitting, then give mean FEV)

13

Regression Tree on FEV Data

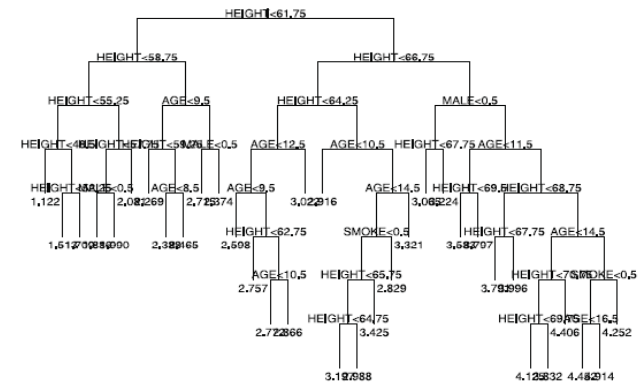


Figure 1. Hierarchical view of tree.

Properties of Trees

Flexible

- Approximates nonlinear relationships as a “step function”
- Able to detect interactions
 - In fact, every “leaf” tends to be a quite complex interaction

Mimics common decision rules

Overfits data

- Perhaps some leaves are quite similar to each other

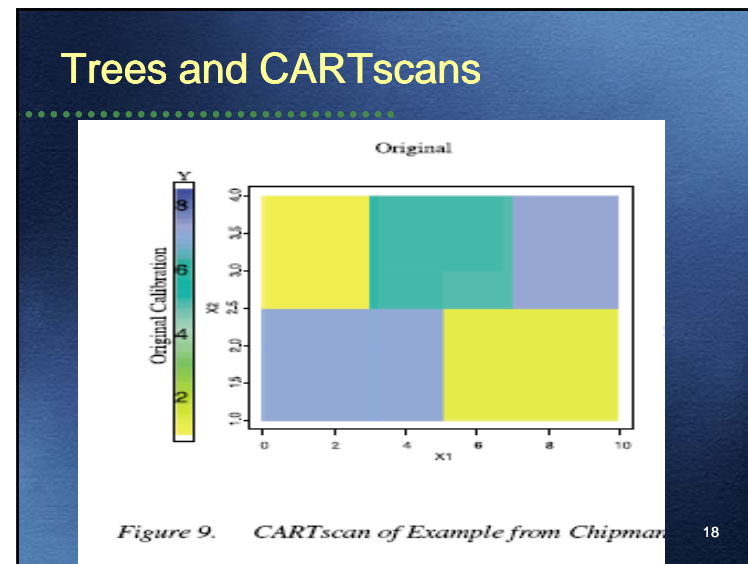
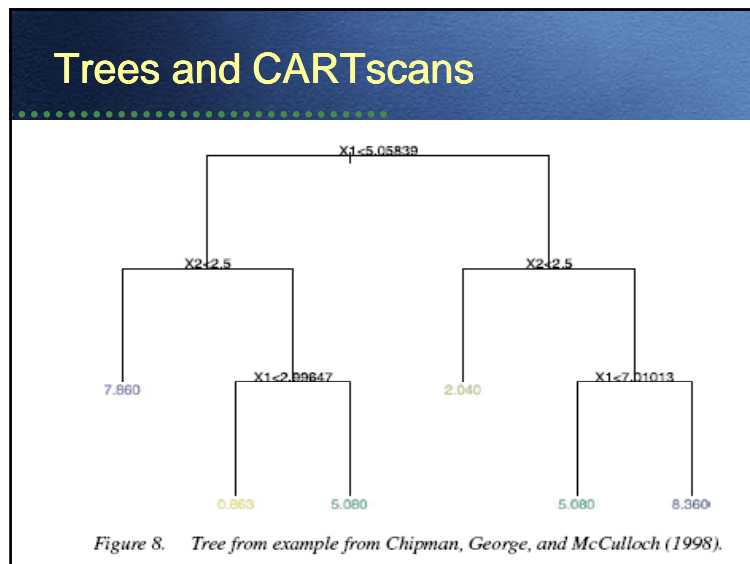
15

CARTscans

Graphical presentation of complex predictive models

- Goal: Visual assessment of variable importance
- Nason, et al. (JCGS, 2004)

16

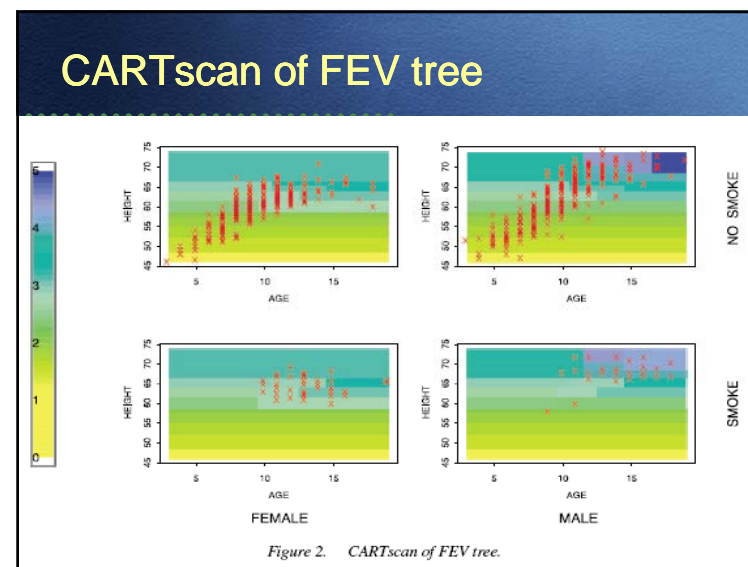


More Than Two Covariates

Create a CARTscan array to mentally “stack” images much like in CAT scans

- Two “inner variables”
- Two “outer variables”
 - Categorized, perhaps with overlapping intervals
- Handling of other variables
 - Averaging, restrictions, more pages

19



Observations From Tree

- Height is strongest predictor
- Age – sex interaction after puberty
- Difference between smokers and nonsmokers best seen among older boys
 - Note that areas with no data sometimes appear to be arbitrary in their inclusion with other groups

21

CARTscans with Regression

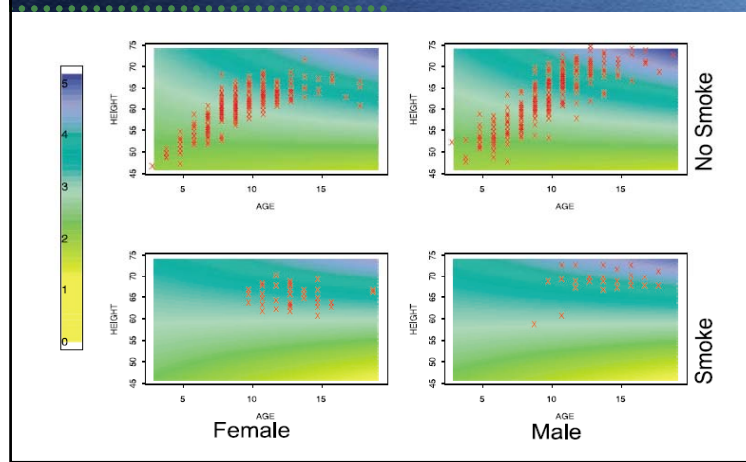
We can actually use CARTscans to help us understand complex regression models

- Regression model with two way interactions
- Regression model with four way interactions

(Compare this type of display to the scatterplots with stratified smooths used with SEP data set)

22

Regression with 2-way Interactions



Regression with 4-way Interactions

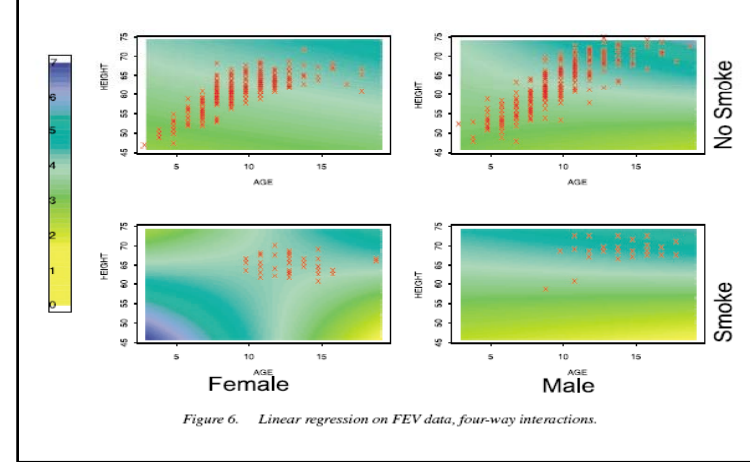


Figure 6. Linear regression on FEV data, four-way interactions.

CARTscans: Variability

It is difficult to judge the “statistical significance” of a predictive model due to the multiple comparisons

- Can use bootstrapping to judge the reproducibility of the predictive model across pseudo-replicates
 - Present quantiles of predictions across bootstrapped samples
 - Do the contrasts among regions remain?

25

Trees and CARTscans

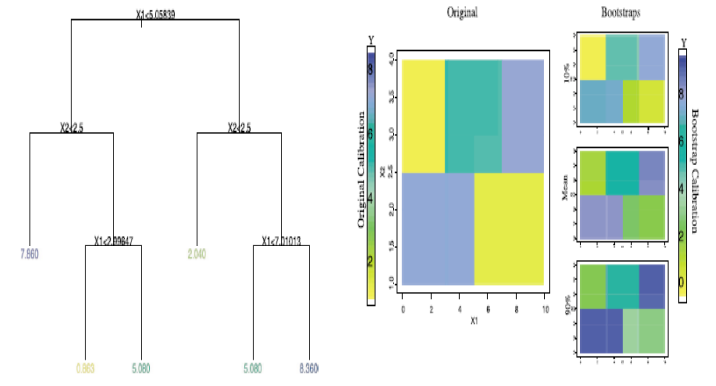


Figure 8. Tree from example from Chipman, George, and McCulloch (1998).

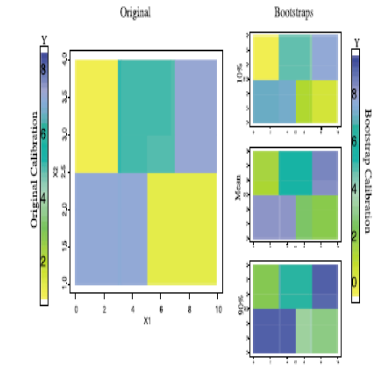


Figure 9. CARTscan of Example from Chipman, George, and McCulloch (1998).

Inference: Variable Importance

General approach (Rudser)

- Highly predictive model to identify homogeneous groups
 - Adaptively select variables, transformations, interactions
- Estimate distribution within each leaf and derive arbitrary summary measures
 - Works with quantiles, censored data means, ...
- Other regression models to form contrasts
- Standard errors by bootstrapping
 - Must account for overfitting of data

27