3 Ways to Improve Your Regression: Part 2

January 27, 2016

Outline

- Quick review of Part 1
  - Concepts covered
  - Results
- Nonlinear regression splines
  - Reading basis function code
  - Plotting
  - Including interactions
- Stochastic gradient boosting
  - Interaction control language
  - Partial dependency plots
  - Spline approximation
- Case studies and other applications
- Questions

Part 1 Recap

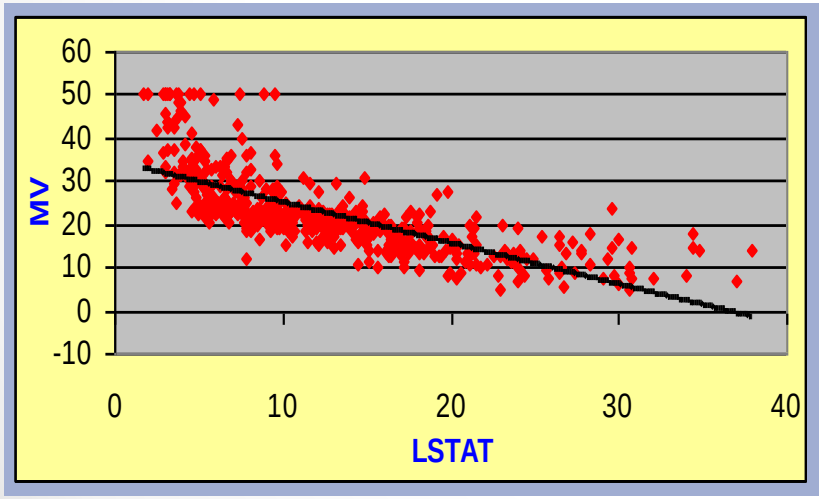
- Ordinary least squares (OLS)
  - Common issues: missing values, nonlinearities, interactions, variable selection, overfitting, instability due to collinearity
- Methods to overcome issues
  - Nonlinear regression splines (MARS)
  - Stochastic gradient boosting (TreeNet)
  - Random Forests
- Concrete strength prediction

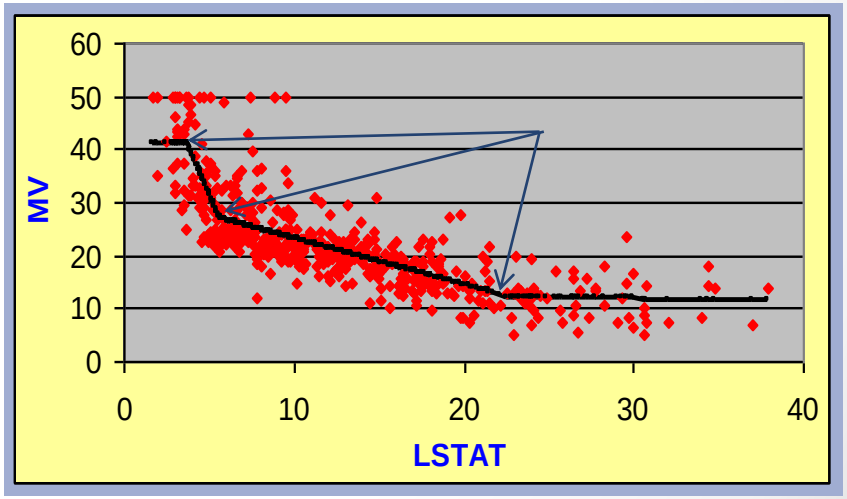
Method	MSE	R^2
OLS	107.21	61.55%
MARS	34.20	87.73%
TreeNet	37.47	86.56%
RandomForests	25.54	90.84%

- Feedback

Nonlinear Regression Splines

- Uses "knots" to impose local linearities
- These knots create "basis functions" to decompose the information in each variable individually
- Can also perform well on binary dependent variables
- MARS (Multivariate Adaptive Regression Splines)





Modeling Process

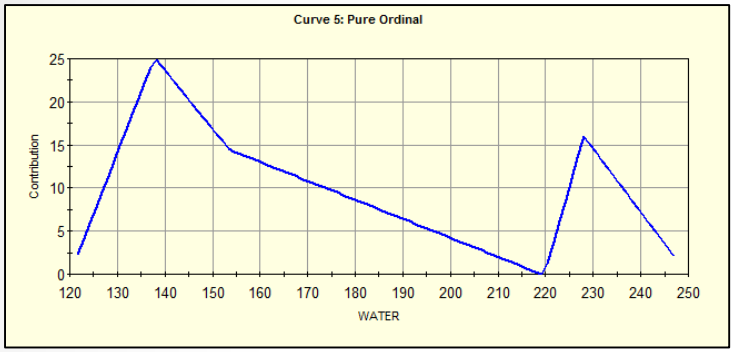
1. Forward stage:
   - Add pairs of BFs (direct and mirror, same knot) in a step-wise regression manner
   - The process stops once a user specified upper limit is reached
   - Possible linear dependency is handled automatically by discarding redundant BFs
2. Backward stage:
   - Remove BFs one at a time in a step-wise regression manner
   - This creates a sequence of candidate models of varying complexity
3. Selection stage:
   - Select optimal model based on the TEST performance (modern approach)
   - Select optimal model based on GCV criterion (legacy approach)

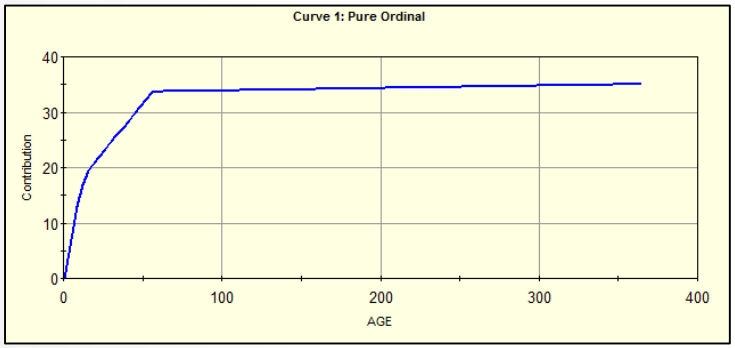
Basis Functions Code

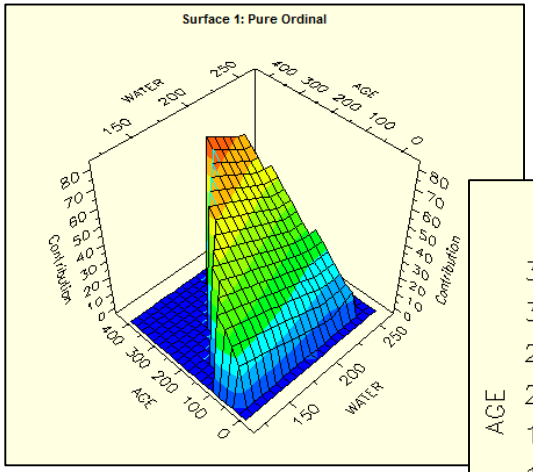
- **Basis Functions** (BF) provide analytical machinery to express knots:
  - Direct: max (0, X -c)
  - Mirror: max (0, c - X)
  - This is a continuous transformation of variable X into X*
  - Value 'c' defines the knot placement and constructed for any data value

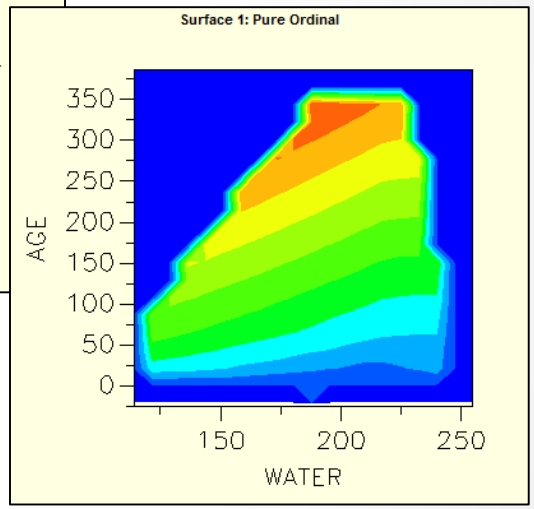
```
BF1 = max( 0, AGE - 56);
BF2 = max( 0, 56 - AGE);
BF3 = max( 0, CEMENT - 531.3);
BF4 = max( 0, 531.3 - CEMENT);
BF5 = max( 0, BLAST_FURNACE_SLAG - 19);

Y = 211.582 + 1.5734 * BF1 - 1.9215 * BF2 + 0.998637 * BF3
+ 0.078569 * BF4 + 0.26698 * BF5;
```

Plotting Basis Functions









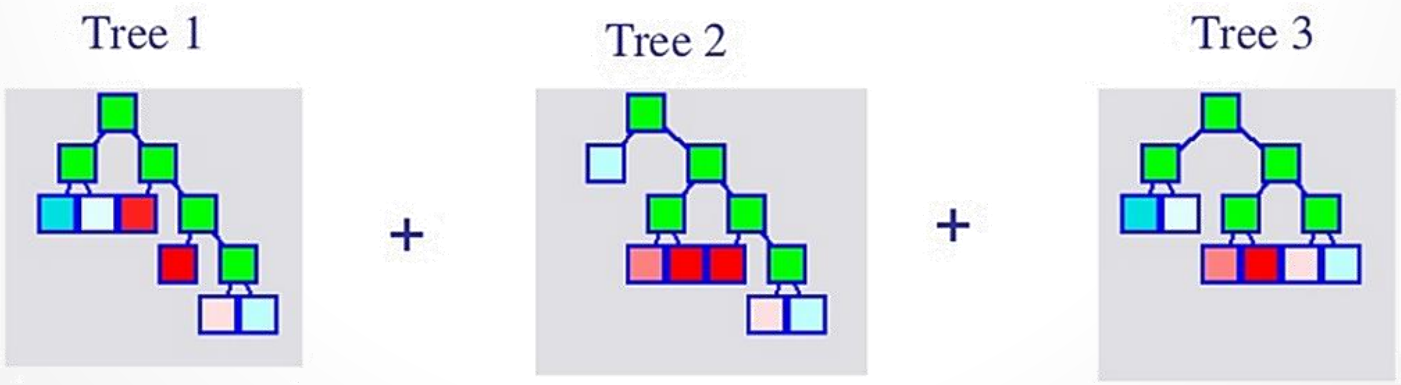
- Purely additive models produce 2D plots for each predictor
- Models with interaction effects allowed among all predictors produce 3D plots

Including Interactions

- Until now we have considered only ADDITIVE entry of basis functions
- Optionally, MARS will test an interaction with candidate basis function pair
  - Identify a candidate pair of basis functions
  - Test contribution when added to model as standalone
  - Test contribution when interacted with basis functions already in model
- Interactions are thus built by accretion
  - One of the members of the interaction must appear as a main effect
  - Then an interaction can be created involving this term
  - The second member of the interaction does NOT need to enter as a main effect
- Generally a MARS interaction is **region specific**
  - **i.e.** $(PT - 18.6)_+ * (RM - 6.431)_+$
- This is not the familiar interaction of PT*RM because the interaction is confined to the data region where RM>6.431 and PT>18.6
- MARS could construct a different interaction outside of this region
- Recommended that modeler try a series of models (AUTOMATE)
  - additive
  - 2-way interactions
  - 3-way interactions
  - 4-way interactions, etc.

Stochastic Gradient Boosting

- Small decision trees built in an error-correcting sequence
  1. Begin with small tree as initial model
  2. Compute residuals from this model for all records
  3. Grow a second small tree to predict these residuals
  4. And so on…
- Fast and efficient
- Data driven
- Immune to outliers
- Invariant to monotone transformations of variables

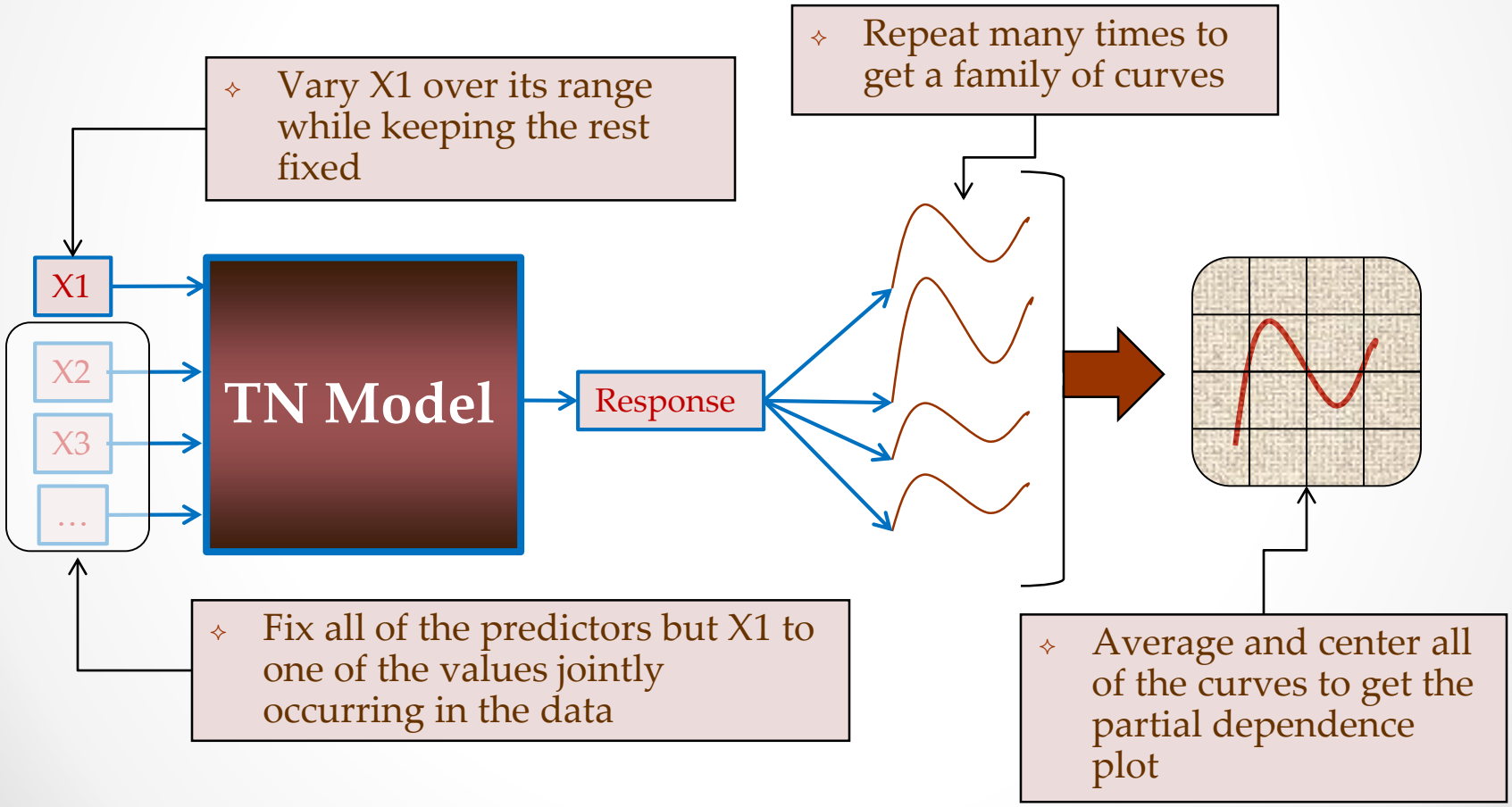


Interaction Control

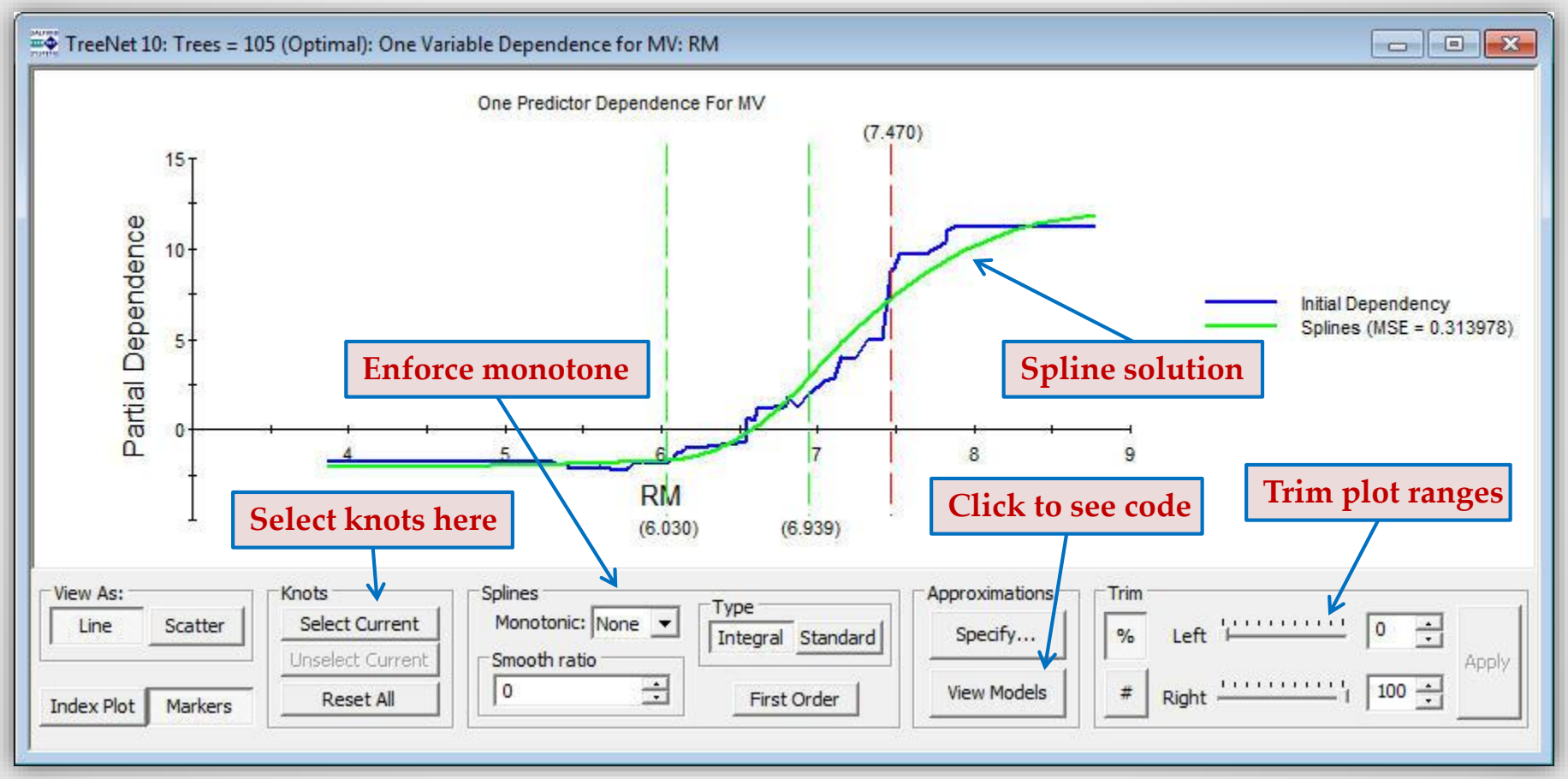
- TreeNet models are always additive in trees
- Interactions may only enter at the individual tree level
- Larger trees allow more opportunities for interactions
  - 2-node trees force additive models
  - 3-node trees allow pair-wise interactions only
  - The degree of interactions relates to the number of levels in the tree
- TN can generate a special report with estimates of interactions in the model
- TN offers additional model building flexibility by allowing direct control over which variables are allowed to interact and to what degree

Partial Dependence Plots

- This procedure minimizes the effect of correlated inputs and gives the exact result when *X* enters additively or strictly multiplicatively



Spline Approximations



Model Automation

- Automates variable selection process
  - Varying engine parameters
  - Univariate dependencies
  - Cross-validation
  - Stratified models
  - Sampling strategies
  - Variable selection
  - Hot spot detection
  - Missing value imputation
  - Granularity reduction
  - Transformation discovery
  - Interaction discovery
  - Bootstrap aggregation
  - Variable binning
  - 2-stage model building
  - Rolling data window
  - Outlier detection
- Top Shaving: each step the most important variable is eliminated
  - Capable to detect and eliminate "model hijackers" – variables that appear to be important on the learn sample but in general hurt model performance (for example, ID variable)
- Bottom Shaving: each step the least important variable is eliminated
  - May drastically reduce the number of variables used by the model
- Error Shaving: at each iteration all current variables are tried for elimination one at a time to determine which predictor contributes the least
  - May result to a very long sequence of runs (quadratic complexity)

Other Applications

- Dental Research
  - Predicting risk of obesity and type 2 diabetes in children using saliva samples
    - J. Max Goodson (DDS, PhD) at The Forsyth Institute
- Economics
  - "Rules of Thumb" for Sovereign Debt Crises by Paolo Manasse and Nouriel Roubini
  - Forecasting Recessions
- Ecology
  - Paradigm Shifts in Wildlife and Biodiversity Management
- Business
  - "Techniques for Business Failure Prediction"
  - "Business Analytics in Financial Services"

More Applications

- Epidemiology
  - Relationship between vitamin E levels and myocardial infarction
  - Is coronary calcification associated with regional left ventricular dysfunction?
- Healthcare
  - "Identifying Key Determinates of Quality in Health Care"
  - "Applied Multivariable Modeling in Public Health"
- Social Sciences
  - Analyzing RMV data to determine the extent of racial and gender profiling
  - Profiling Poverty with Multivariate Adaptive Regression Splines

Questions?

- support@salford-systems.com

- Follow-up email:
  - Recording of webinar
  - PowerPoint slides
  - SPM 30-day trial download instructions
  - Tutorial with concrete dataset