



# Joint Colloquium of the IACA, PBSS and IAAHS Sections of the International Actuarial Association

Westin Copley Place Hotel, Boston, U.S.A. – 4-7 May 2008

## Data Mining and Predictive Modeling Applications for US P&C Insurance Industry

Cheng-sheng Peter Wu, FCAS, ASA, MAAA



ig LLP, 2008

Join



the IACA, PBSS at  
e Hotel, Boston, U.



s

3

# Theme

- Data mining and predictive modeling applications in US property and casualty insurance industry are – **HOT!**



- Why?

*“Because insurance is a “zero sum” game, and DM and PM will create “adversely selection” for the companies who are not doing it!”*

# Agenda

- What is Data Mining and Predictive Modeling
- A Success Story
- Applications in the US P&C Industry
- Introduction of Modeling Techniques

# What is Data Mining and Predictive Modeling

- Predictive modeling is an application of mathematical and statistical techniques and algorithms to produce a mathematical model that can effectively predict and segment future events
- Data mining is a process that utilizes predictive modeling techniques to analyze large quantities of internal and external data, in order to unlock previously unknown and meaningful business relationships

# What Drive DM and PM Hot Today?

- Rapid advancement of cheap computing power

## Moore's Law

Year	1980	1985	1990	1995	2000	2004
Storage Cost per Megabyte	\$190	\$ 70	\$ 10	\$0.90	\$0.05	\$0.001
Microprocessor Speed, MHz	5-8	16	33	75	200	400

# What Drive DM and PM Hot Today?

- DM and PM for insurance industry:
  - contains large amount of data
  - a under-explored area
- Development of new and powerful modeling and data exploration techniques
  - Examples: regression, GLM, neural networks, decision trees, clustering analysis, MARS, ...
  - Explore complicated patterns in data such as non-normality, non-linearity, interactions, etc.

# What Drive DM and PM Hot Today?

- True “multivariate” analysis with large amount of data and many variables
  - Analyze multiple variables “simultaneously” instead of one or two at a time.
  - Use large amounts of data
    - No need to use summarized data for actuarial analyses.
  - Create and analyze novel predictive variables.

# A Success Story

# A Case Study - Credit Score Revolution

## Progressive vs Industry



# Credit Score Revolution

- About credit score:
    - First important factor identified over the past 2 decades
    - Composite multivariate score vs. raw credit information
    - Introduced in late 80s and early 90s
    - Viewed at first as a “secret weapon”
    - Quiet, confidential, controversial, black box, ...etc
- “Early believers and users have gained significant competitive advantage!”*

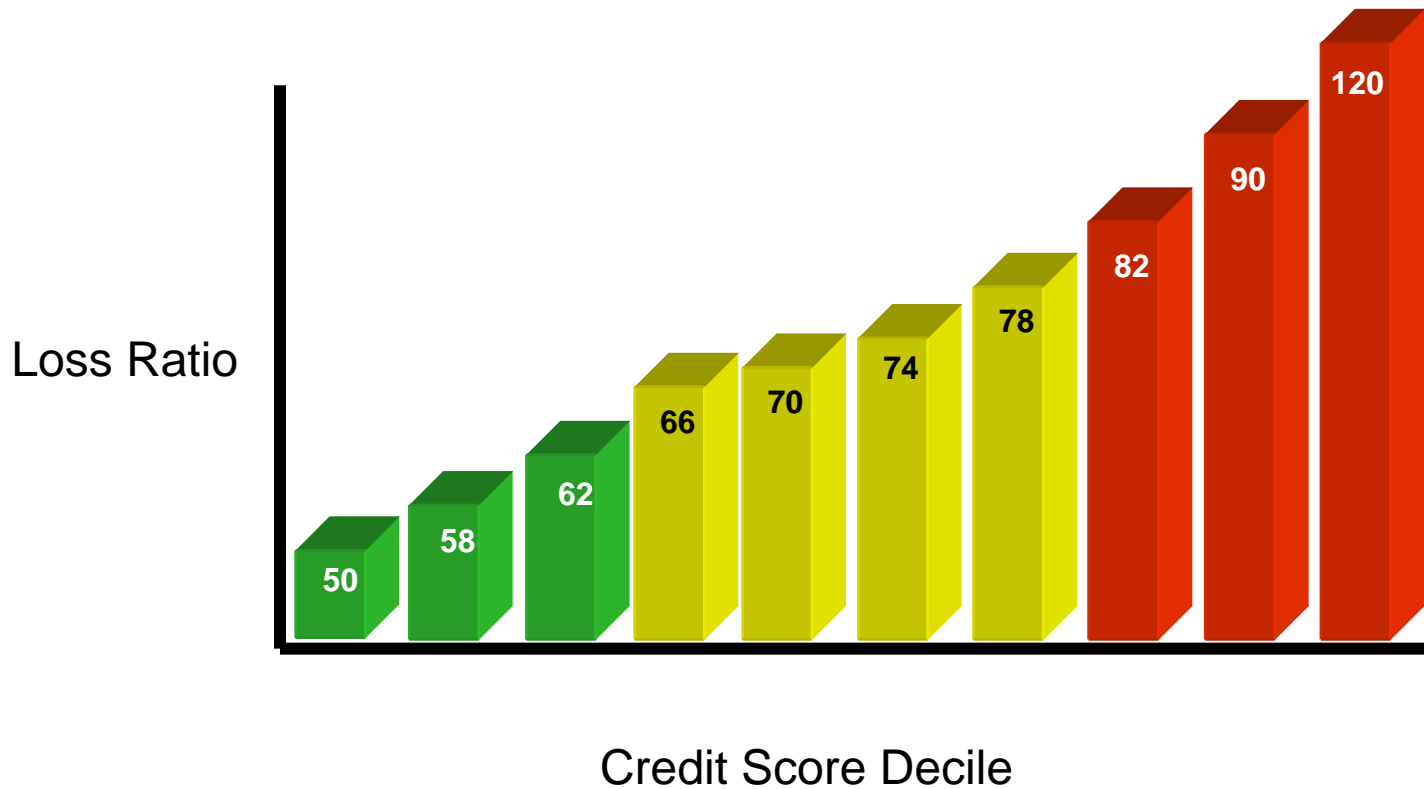
# Credit Score Revolutions

- What is “credit score”?
  - A composite score that usually contains 10 to 40 pieces of credit information
    - Payment pattern information, account history, bankruptcies/liens, collections, inquiries, bad debt/defaults...
    - Formula scoring or rule-based scoring
    - Industry scores and proprietary scores

# Credit Score Revolutions

- Why “credit score” is so successful?
  - “Large scale” “multivariate” scoring using “external data source”
  - Loss ratio lift is significant, a powerful class plan factor or rate tiering factor
  - “Brilliant” marketing approach for credit score:
    - Benefits/ROI are *measurable and* lift curve can be translated into bottom-line benefit
    - Blind test and independent validation can be done to verify the benefit

# Loss Ratio Lift Curve



# Credit Score Revolution

## 1997 NAIC/Tillinghast Study of 9 Companies' Data

### Loss Ratio Relativity of the Best and Worst 20% of Credit Score

	Co1	Co2	Co3	Co4	Co5	Co6	Co7	Co8	Co9	Avg
<b>Best 20%</b>	<b>-38%</b>	<b>-29%</b>	<b>-19%</b>	<b>-15%</b>	<b>-14%</b>	<b>-34%</b>	<b>-22%</b>	<b>-22%</b>	<b>-36%</b>	<b>-25%</b>
<b>Worst 20%</b>	<b>48%</b>	<b>20%</b>	<b>32%</b>	<b>30%</b>	<b>46%</b>	<b>59%</b>	<b>20%</b>	<b>22%</b>	<b>95%</b>	<b>41%</b>

# Applications in the US P&C Industry

# From Credit Scores to DM and PM

- A credit score is just “*one example*” of an insurance predictive model
- The same methods used to build credit scores are used in data mining to build insurance predictive models:
  - Fully utilize all sources of internal and external data sources
  - Fully utilize all available data
  - Same concepts to all lines of business

# Data Sources

- Company's internal data
  - Policy-level records
  - Loss & premium transactions
  - Billing
  - VIN.....
- Externally purchased data
  - Credit
  - CLUE
  - MVR
  - Census
  - ....

# P&C Applications – Personal Line

- Almost all of the personal line companies are using credit scores or their proxy in marketing, pricing, and/or underwriting for personal auto and home business
- Using GLM to optimize the class plan factors has become a standard actuarial approach:
  - frequency vs. severity approach
  - Poisson, Gamma, Tweedie, and other long-tailed distribution assumptions
  - Sophisticated pricing and class plan structure
  - Large amount of pricing points (millions of pricing points/pricing cells)
- Using demographic information and geo-coding technique to redefine the rating territory

# P&C Applications – Small Commercial

- Copy the success from the personal line applications and rapidly deploy the applications for the small commercial lines
- Go beyond credit score models: non credit score models or mixed credit and non credit score models
  - 30% of companies have developed multivariate scoring models to score the small commercial book for business owners policies (BOP), auto, property, liability, and workers' compensation
  - The main application is in underwriting and profitability
  - Use business' financial information similar to personal financial information

# P&C Applications – Other Applications

- Retention model
- Marketing and mail solicitation model
- Claim model:
  - Fraud detection
  - Resource allocation
  - Early intervention
- Agent classification model
- Enterprise model: pricing, underwriting and supply/demand/elasticity
- Customer life time value model (LTV)

# Introduction of Modeling Techniques

# Some Definitions

- Target Variable:  $Y$ 
  - What we are trying to predict.
- Predictive Variables:  $\{X_1, X_2, \dots, X_N\}$ 
  - What we use to predict  $Y$
- Predictive Model :  $Y = f(X_1, X_2, \dots, X_N) + \varepsilon$  (error term)
- $f(x)$  represents the variation in  $Y$  that can be explained by the predictive variables, while  $\varepsilon$  (error term) represents the variation in  $Y$  that cannot be explained by the predictive variables
- The goal is to “maximize” the explainable variation and “minimize” the un-explainable variation

# Supervised vs Unsupervised Learning Models

- Supervised Learning: Attempt to predict  $Y$  in terms of  $X$ .
  - Ordinary regression, Analysis of Variance, Categorical Analysis
  - Generalized Linear Models (GLM)
  - Generalized Additive Models (GAM)
  - Neural Nets
  - Classification and Regression Trees (CART)
  - Multivariate Adaptive Regression Splines (MARS)
- Unsupervised Learning: Attempt to find interesting patterns amongst the characteristics where there is no outcome variable
  - Clustering
  - Principal Components / Factor Analysis
  - Neural Nets/KH Self-Organized Map

# Regression, AOV, and Categorical Analysis

$$Y = f(X_1, X_2, \dots, X_N) + \varepsilon \text{ (error term)}$$

- Error term is assume to be normally distributed
- Y – continuous for regression and AOV; categorical for Categorical Analysis
- X – continuous for regression; categorical for AOV and Categorical Analysis

# Generalized Linear Models

$$Y = f(X_1, X_2, \dots, X_N) + \varepsilon \text{ (error term)}$$

- $E(Y/\mu) = g(X_1, X_2, \dots, X_N)$ , link function
  - The link function is a linear form
  - When a log function is assumed log, it is a multiplicative formula
- Error term assumes various “exponential family” distributions:
  - Binomial; logistic regression; target is binomial (0 or 1, yes or no); used for retention models
  - Poisson; target is positive, integer value; used for claim count/frequency models
  - Gamma; target is continuous and long tailed – skewed to the right; used for severity model
  - Tweedie/Compound Poisson-Gamma; target is mixed, 0 then the rest continuous, used for loss cost or loss ratio models
- There are other exponential family distributions, such as Inversed Gaussian

# Generalized Adaptive Models

$$Y = f(X_1, X_2, \dots, X_N) + \varepsilon \text{ (error term)}$$

- GAM is very similar to GLM
- The only difference is in the link function:  $E(Y/\mu) = g(X_1, X_2, \dots, X_N)$ 
  - For GLM, the link function is a linear formula
  - For GAM, the link function can take “Taylor series/polynomial function” of higher order terms
  - In reality, GAM is a “non-linear” modeling technique
  - Obviously, more higher order terms for the predictive variables, more possibility of “over-fit”
  - Including a “penalty” calculation for goodness of fit associated with higher order terms for the predictive variables

# Neural Networks

$$Y = f(X_1, X_2, \dots, X_N) + \varepsilon \text{ (error term)}$$

- No assumption for the error terms
- Focus is on  $f(x)$ :
  - Essentially is a curve fitting technique or non-parametric regression
  - Assume  $f(x)$  to be “non-linear”
  - Can fit any nonlinear patterns; universal approximator; similar to Taylor series
  - The function form of  $f(x)$  is a series of “sigmoid functions”:

$$Y = \frac{1}{1 + e^{b_0 + b_1 z_1 + b_2 z_2}}$$

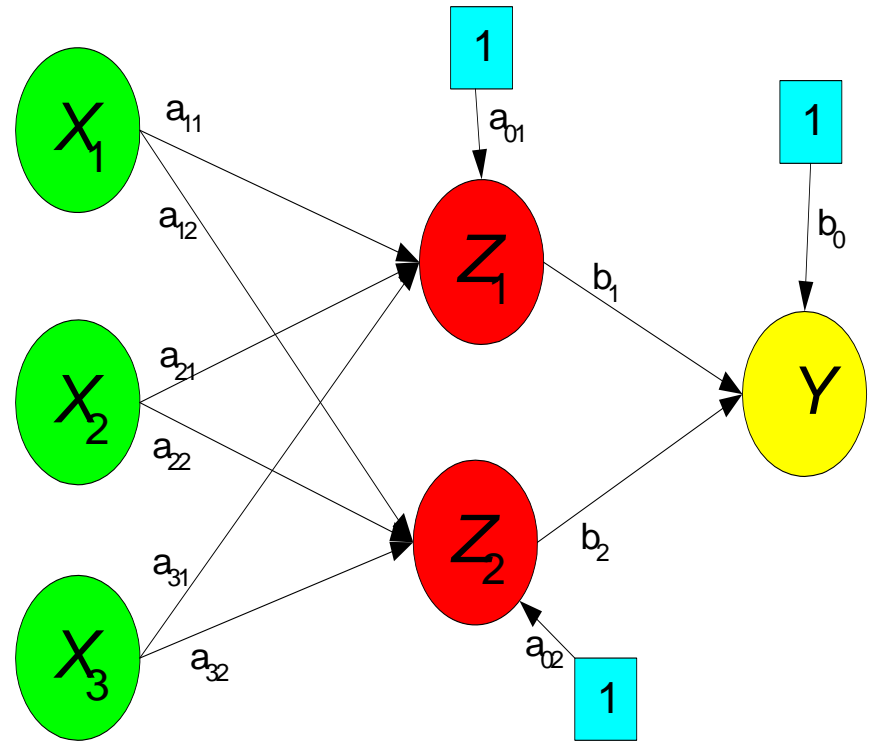
- No need to assume the non-linear form for predictive variables; Hard to interpret the results/black box; computational intensive

# Neural Nets

$$Z_1 = \frac{1}{1 + e^{a_{01} + b_{11}x_1 + b_{21}x_2 + b_{31}x_3}}$$

$$Z_2 = \frac{1}{1 + e^{a_{02} + b_{12}x_1 + b_{22}x_2 + b_{32}x_3}}$$

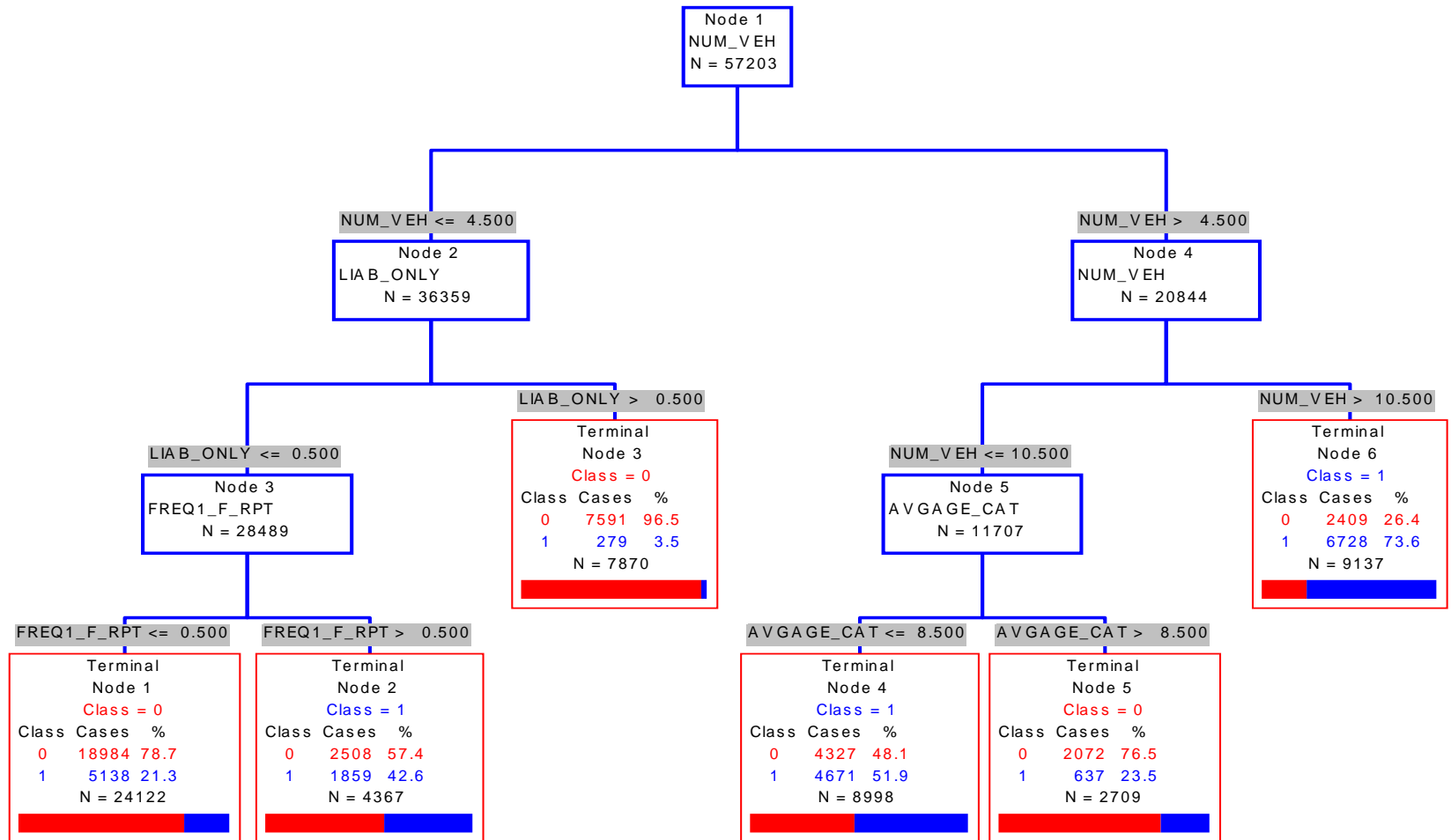
$$Y = \frac{1}{1 + e^{b_0 + b_1z_1 + b_2z_2}}$$



# CART

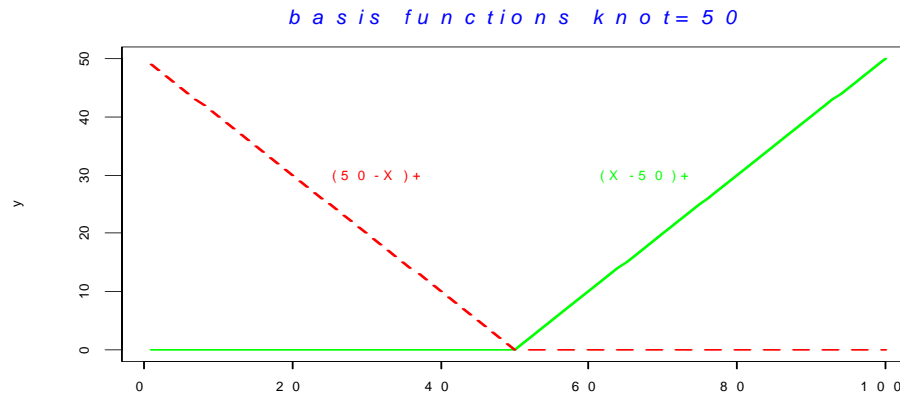
- No assumption for the error terms
- Focus is on  $f(x)$ :
  - Find out the segment/cut that maximize the separation of the target variable
  - Try one variable at a time and continue to find segmentation/grow the tree
  - End result is similar to “if then” rules
  - Target can be either categorical or continuous
  - Simple to understand; less sensitive to outliers; no need to transform the predictive variables; less powerful; less optimal if the target or predictive variables are continuous; too big a tree can be hard to handle
  - One potential application is “transforming or grouping” predictive variables

# Growing The Tree



# MARS

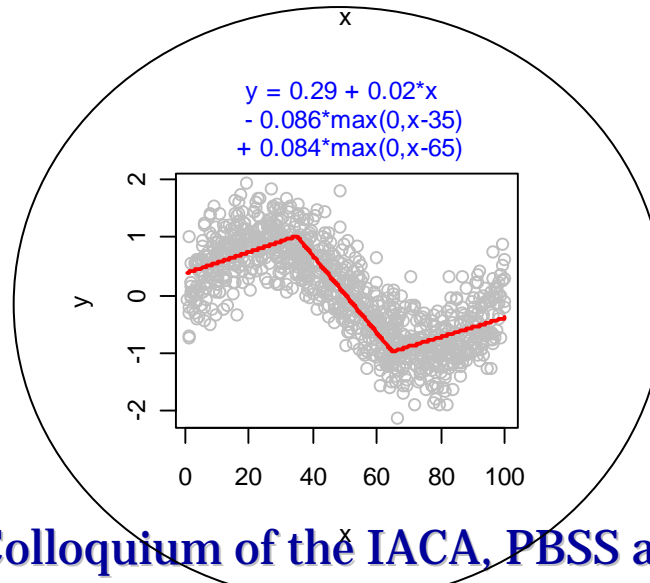
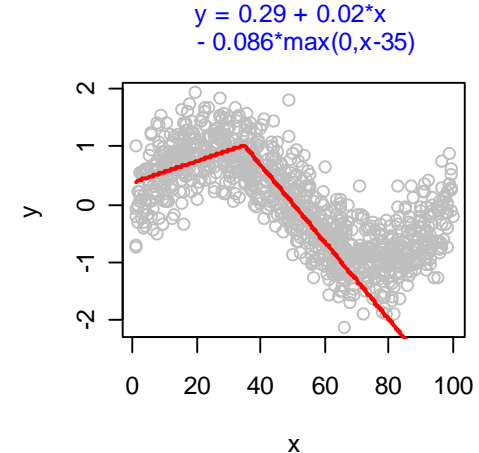
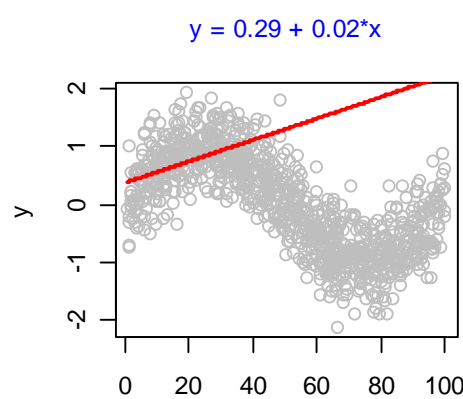
- No assumption for the error term
- Focus is on  $f(x)$ :
  - Similar to neural networks, essentially a nonlinear modeling technique; can also include interaction terms
  - First transform variables via “basis functions”, and then perform stepwise procedure on the transformed variables:



- Modelers do not need to assume the function form for predictive variables; model output will advise the optimal predictive variables transformation; computational intensive

# MARS

- After the forward stepwise search and pruning back the final MARS model is:
- $\hat{y} = 0.29 + 0.02*x$   
-  $0.086*\max(0,x-35)$   
+  $0.084*\max(0,x-65)$



# Unsupervised Learning Techniques

- No target variables
- Given a number of characteristics variables, naturally form several “profiles” that will group similar data points together
  - For example, given 1000 data points for people and 20 characteristics, can these 1000 people be grouped into several similar groups/profiles?
  - *Clustering technique*: an iterative algorithm; need to specify the number of profiles (such as 5 groups); mathematically, a person is placed into one of the 5 groups is based on the “deviation of the person’s characteristics and the group’s average characteristics
  - *Neural networks/KH map*: very similar to the clustering technique; an iterative algorithm; place data points on a “geometric framework”, like 10 by 10 grids (high income-young profession-single placed at upper left corner while low income –married- mature placed at lower right corner, etc) .
  - *Principal component analysis*: create another 20 principal component variables to replace the original characteristics; use “Eigen function transformation” to derive the pc; each pc represent a profile