

Model Selection

Bob Stine

Dept of Statistics, Wharton School
University of Pennsylvania, Philadelphia PA
stine@stat.wharton.upenn.edu

May 11, 1998

- Regression modeling and data mining
- Example: financial time series
- Example: medical screening
- Plan for rest of day

Typical Problem

Goal of modeling effort

- *Ideal:*
Exploratory, generating conjectures for confirmatory study
- *Realistic:*
Accurate prediction of new observations

Don't know the real model

Unsure of which predictors are useful, much less the form of their role in some “true” model.

Many potential predictors

- Access to large database, data warehouse
- Potential nonlinearity, interactions, subsets/subgroups

Differing expectations for predictors

- Some *must* be in the model ($|t| \approx 10$),
- Some *might* be in the model ($|t| \approx 2$), \Leftarrow
- Variables that somebody thinks belong and are available.

Possible Analyses

Modeling techniques include (Rob Tibshirani's course)

- Stepwise regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- Additive models, smoothing, wavelets

$$Y = \beta_0 + S(X_1) + \cdots + S(X_p) + \epsilon$$

- Multivariate adaptive regression splines (MARS)

$$Y = \beta_0 + S_1(X_j) + \cdots + S_2(X_j, X_k) + \cdots + \epsilon$$

- Neural nets, projection pursuit

$$Y = \beta_0 + S(b_1 X_1 + \cdots + b_p X_p) + \cdots + S(c_1 X_1 + \cdots + c_p X_p) + \epsilon$$

- Regression trees (CART)

$$Y = \begin{cases} \begin{cases} \mu_{11}, & X_j \leq C_1 \\ \mu_{12}, & X_j > C_1 \end{cases} & X_k \leq C_2 \\ \beta_0 + \beta_1 X_\ell & X_k > C_2 \end{cases}$$

Question for all

Which predictors/categories/break-points belong in the model?

Stepwise as canonical example

Splines in MARS, for example, modeled as collection of piecewise functions which are chosen in a stepwise fashion. Just a different basis collection.

Data Mining

Features

- Goal: **predict** (out-of-sample error)
Software often does not reveal internal details.
- Many potential attributes in model (variable selection)
- Exploratory, with little potential to conflict theory
- Large data bases, data warehouse (relevance, quality)

Domains

- Financial markets
- Banking and credit industry
 - Fraud
 - Credit risk
 - Bankruptcy
- Chemistry
 - NIR spectroscopy ($p=757, n=42$)
 - High-throughput screening analysis
- “Science”
 - Weather patterns
 - Physics
 - Image classification, matching

Data Mining Example: Credit Risk

Problem

Should we give you credit? How much optimizes profits?

(It's not the card — it's the credit limit!)

Hard part

Creditors make the most money on those who carry a large balance, but these also pose the most risk.

Data

- Your application data, credit record, recent purchases
- Behavior of customers like you \Leftarrow
- Macro-economic, geographical, ... (merging databases)

Analysis issues

- Missing data
- Massive amount of data (250 million accts at one bank)
- What's comparable? (matching)

Common industry technique

- Discretize all data (missing data is just another level)
- Chaid analysis
- Proprietary methods (Fair-Isaacs)

Finance Example: McDonald's Stock

Problem

Predict movement of returns on McDonald's stock.

Some theory here

Finance theory indicates that returns on the overall market will be a good predictor of the returns on this particular stock.

Simple regression

Four years of monthly data over the years 1991-1994 show

$$\text{Return on McD} \approx 0.01 + \underbrace{1.0}_{\text{beta}} \text{ Market returns, } R^2 \approx 0.25$$

Other factors?

“Everyone” knows this model!

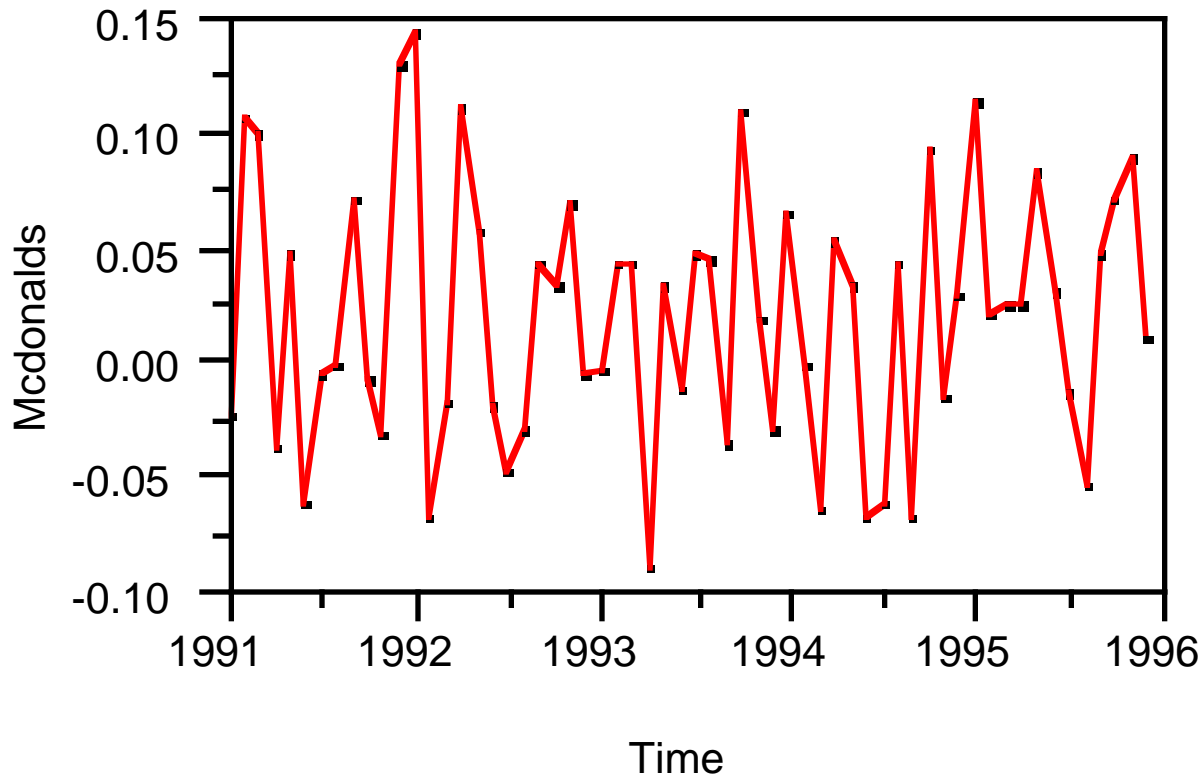
What other factors beat this standard model and lead to a real money-making scheme?

Model for McDonald's Stock

Stepwise results

- Extend model, selecting from 50 other factors.
- Forward stepwise ($p=0.25$), backward ($p=0.1$)
- Six p-values < 0.0001 (Bonferroni $0.05/51 = 0.001$)
- Overall $R^2 = 0.881$ ($F_{17,30} = 13.089$, $p < 0.0001$)
- Residual variance $s_{17}^2 = 0.0253^2$.

Time series plot



Term	Estimate	Std Error	<i>t</i> Ratio	p-value
Intercept	-0.0125	0.0058	-2.14	0.0402
Value-weighted	0.8572	0.1487	5.77	0.0000
X_2	0.0145	0.0054	2.67	0.0122
X_5	0.0206	0.0046	4.52	0.0001
X_7	-0.0269	0.0049	-5.44	0.0000
X_9	0.0119	0.0044	2.73	0.0104
X_{14}	-0.0098	0.0049	-2.00	0.0549
X_{16}	-0.0115	0.0043	-2.69	0.0116
X_{17}	-0.0155	0.0045	-3.47	0.0016
X_{20}	-0.0229	0.0044	-5.23	0.0000
X_{22}	-0.0327	0.0044	-7.39	0.0000
X_{26}	-0.0268	0.0053	-5.06	0.0000
X_{27}	-0.0150	0.0044	-3.42	0.0018
X_{31}	0.0111	0.0050	2.24	0.0327
X_{43}	0.0184	0.0045	4.07	0.0003
X_{44}	0.0193	0.0051	3.75	0.0007
X_{48}	0.0204	0.0045	4.50	0.0001
X_{49}	-0.0337	0.0046	-7.35	0.0000

Questions about Stock Model

What's a reliable predictor, what's not?

What are those other predictors?

What happened when the model was used for prediction?

Modeling Risk Factors

Osteoporosis

Weakening of bones due to loss of calcium. Particularly common in older, small women who often lose height as age.

Risks

Falls leading to complications, hospitalization.

Question

Which factors are predictive of osteoporosis in post-menopausal women, with the goal of using these as a *screening tool* in place of x-ray.

Subset of data

$n = 1252$ with $p = 120+$ predictors, with varying amounts of missing data.

Screening

Univariate vs. multivariate?

Subsets identified by predictors?

Role of missing data?

Approaches to Model Selection

Variable selection

- Want to be able to say which factors are important
Slope estimator is a “testimator.”
- Select terms via threshold from penalized likelihood
Add variable if: (gain in fit) > (penalty for adding)
- Key is understanding/interpreting penalty terms
- Caveats... selection leads to bias in
 - Chosen coefficients
 - Estimates of goodness-of-fit

Shrinkage alternatives

- Ridge regression, PLS
- Smaller MSE than OLS
- Other Bayesian methods

Yet others

- Dimension reduction: principal components
- Degree of freedom adjustments to standard calculations:
3 df for locating point where slope changes
- Exploratory: grand tours, graphical methods

Plan

Predictive risk

Select factors that appear to improve an estimate of the out-of-sample predictive power of the model.

Examples: C_p , AIC , cross-validation.

Bayesian methods

Emphasis on the notion of shrinking to a subspace using notion of a Bayes factor rather than classical shrinkage estimators (e.g., ridge regression).

Examples: BIC , classical MDL

Adaptive methods

Motivated from several points of view, conforming more to the problem at hand rather than an a priori specification of the model.

Example: $eBIC$

Information theory

Alternative perspective on the available choices for model selection.

From selection to evaluation

What model offers the best guaranteed predictive performance?