Data Mining with Stepwise Regression

Bob Stine & Dean Foster

Department of Statistics, The Wharton School

University of Pennsylvania, Philadelphia PA

www-stat.wharton.upenn.edu/~stine

July 17, 2002

- Goals
  - Small squared prediction error
  - Small classification losses (asymmetric)
- Questions
  - Which model and estimator? Stepwise regression!
  - Which predictors to consider? Everything.
  - **Which predictors to use?**
- Examples
  - Smooth signal in presence of heteroscedasticity
  - Rough signal
  - Predicting bankruptcies

Some Modern Prediction Problems

Credit modeling, scoring

Can you predict who will declare bankruptcy?

Risk factors for a disease

Which factors indicate risk for osteoporosis?

Direct mail advertising

Who should receive a solicitation for a donation?

Internet/e-commerce

If you bought this CD, which others might you buy?

Financial forecasting

Which factors predict movement in stock returns?

These great statistics problems, so...

Why not use the workhorse, regression?

- Calculations well-understood.
- Results are familiar.
- Diagnostics possible.

An Application: Predicting Bankruptcy

Goal

Predictive model for personal bankruptcy...

Based on the recent history of an *individual* credit-card holder, estimate the probability that the card holder will declare bankruptcy during the next credit cycle.

Data

- Large data set: 250,000 bank card accounts
- About 350 “basic” predictors (aka, features)
  - Short monthly time series for each account
  - Credit limits, spend, payments, bureau info
  - Demographic background
  - Interactions are important (AC and cash adv.)

67,000 predictors???

Bankruptcy is rare

2,244 bankruptcies in

$$12 \times 250,000 = 3 \text{ million account-months}$$

Trade-off

Profitable customers look risky. Want to lose them?

“Borrow lots of money and pay it back slowly.”

Modeling Questions

Structure – What type of model?

A linear regression with least squares estimates.

- p potential predictors, n observations
- q non-zero predictors with error variance σ^2:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + \epsilon$$

Scope – Which X_j to consider?

Basically, everything...so p is very large.

- Demographics, time lags, seasonal effects
- Categorical factors, missing data indicators
- Nonlinear terms (quadratics)
- **Interactions** of any of these

Select – Which $q < p$ of the X_j go into the model?

Answering Modeling Questions

Structure – What type of model?

A linear regression with least squares estimates.

- p potential predictors, n observations
- q non-zero predictors with error variance σ^2:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + \epsilon$$

Scope – Which X_j to consider?

Basically, everything...so p is very large.

- Demographics, time lags, seasonal effects
- Categorical factors, missing data indicators
- Nonlinear terms (quadratics)
- **Interactions** of any of these

Select – Which $q < p$ of the X_j go into the model?

- Requires *conservative*, *robust* standard error

Conservative, Robust Standard Error

Conservative

Problem: Selection biases SE downward.

Solution: Estimate SE of contemplated predictor X_k using a model that *does not* include X_k. Use residuals from prior step to compute the SE for X_k.

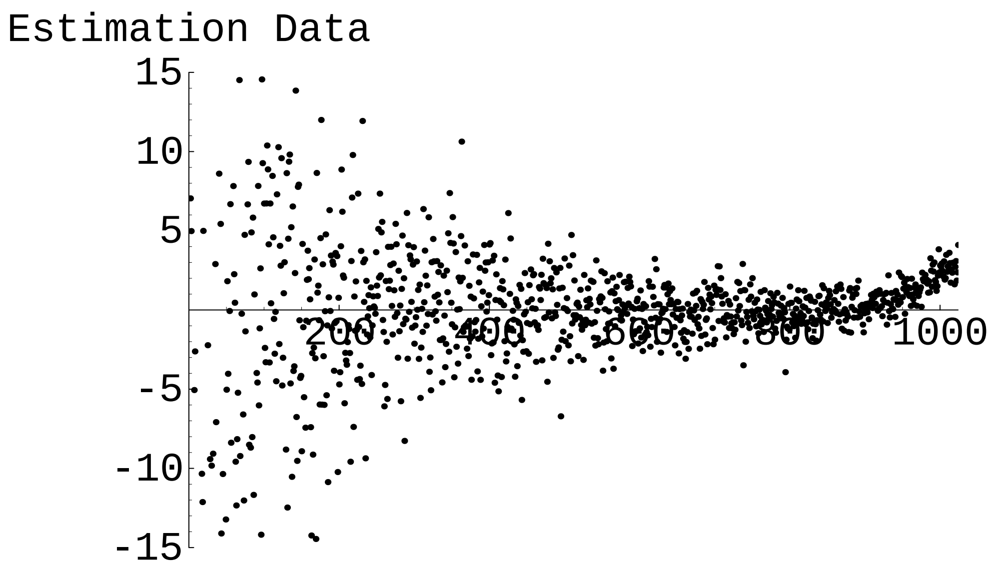
Robust

Problem: Heteroscedastic data lead to misleading SE's.

Example: Heteroscedasticity Can Fool You

Data

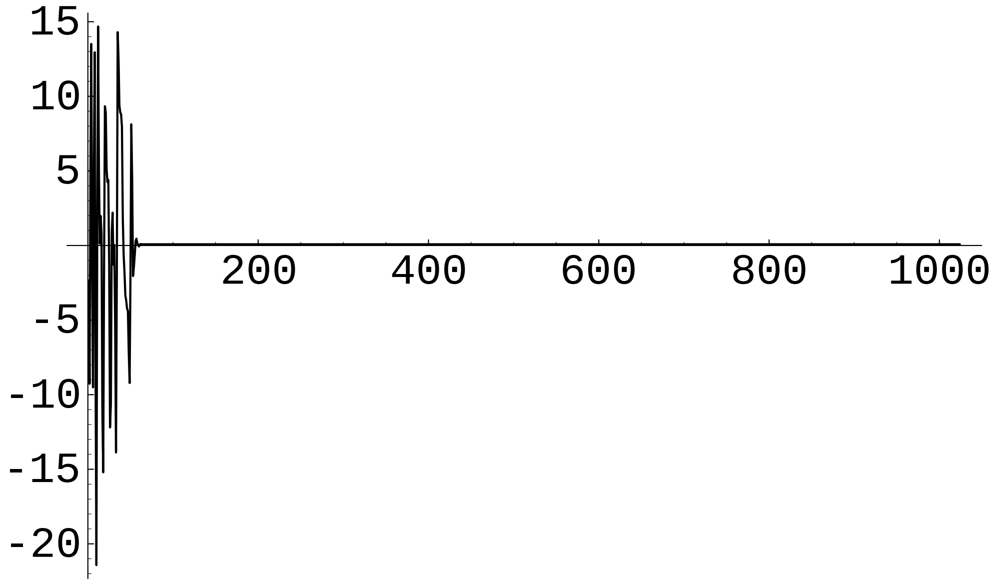
Do you see any “signal” in this data?



Heteroscedasticity Example

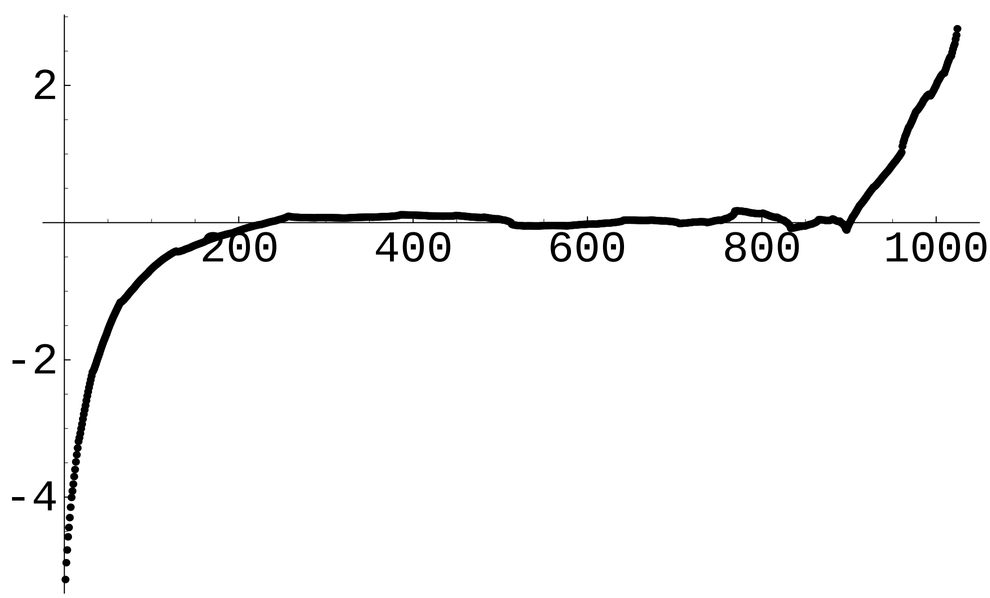
Wavelet regression

Standard wavelet regression with hard thresholding finds the following signal.



Wavelet regression, with corrected variances

Applied to standardized data, then rescaled.



Conservative, Robust Standard Error

Conservative

Problem: Selection biases SE downward.

Solution: Estimate SE of contemplated predictor X_k using a model that *does not* include X_k. Use residuals from prior step to compute the SE for X_k.

Robust

Problem: Heteroscedastic data lead to misleading SE's.

Solution: Adjust the data if you know weights that standardized the data (as the wavelet example or the BR application)

or

Use a SE that is robust to heteroscedasticity. eg. White's estimator.

Answering Modeling Questions

Structure – What type of model?

A linear regression with least squares estimates.

- p potential predictors, n observations
- q non-zero predictors with error variance σ^2:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + \epsilon$$

Scope – Which X_j to consider?

Basically, everything...so p is very large.

- Demographics, time lags, seasonal effects
- Categorical factors, missing data indicators
- Nonlinear terms (quadratics)
- **Interactions** of any of these

Select – Which $q < p$ of the X_j go into the model?

- Requires *conservative*, *robust* standard error
- **Measure significance without presuming CLT.**

Example: Sparse Data Can Fool You

Null model

Lots of data: $n = 10,000$

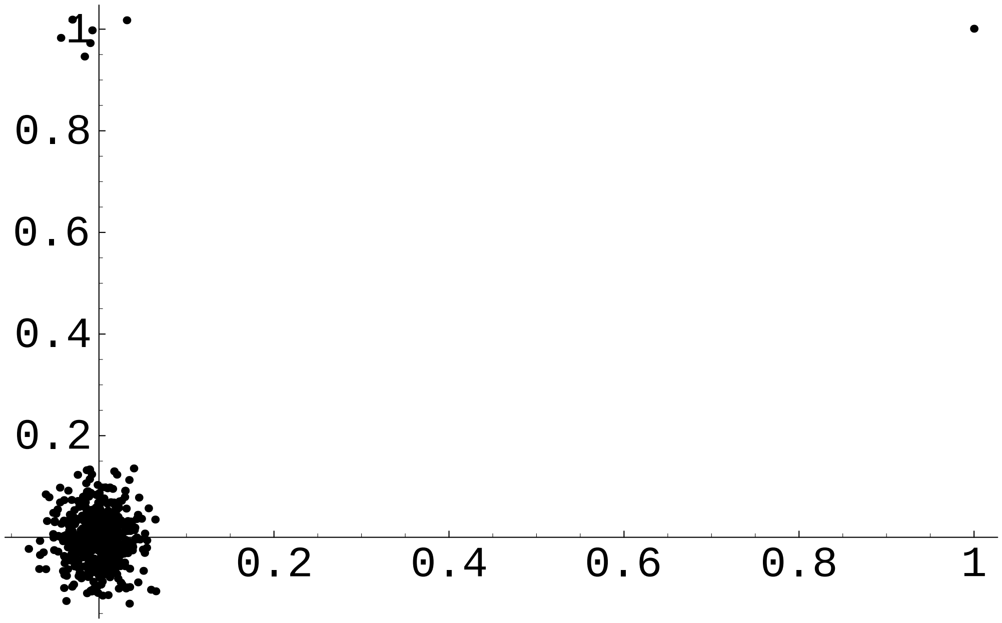
No signal: $Y_i \in \{0,1\}$ with $\mathrm{P}(Y_i = 1) = 1/1000$

Highly leveraged points

Get isolated, large X_{big} with $Y_{big} = 1$.

Estimated significance

Chance that $Y_{big} = 1$ is 1/1000.



Regression gives $\hat{\beta}/SE(\hat{\beta}) = 13$.

Why so significant?

Leverage at outlier is $h_{big} = .14$.

Central limit theorem does not apply.

Measuring Significance

Large samples?

Problem: Data set has many observations, but certain combinations can be very sparse, giving the estimator a Poisson rather than normal character.

Solution: Compute a conservative p-value using an alternative bound on the distribution of the estimator.

Bennett's bound for tail probability (1962)

- Independent summands B_i, $\sup |B_i| \le M$.
- $E\, B_i = 0$, $\sum_i \text{Var}(B_i) = 1$.

$$P(\sum B_i \ge \tau) \le \exp\left(\frac{\tau}{M} - \left(\frac{\tau}{M} + \frac{1}{M^2}\right)\log(1 + M\,\tau)\right)$$

- If maximum is small relative to dispersion ($M\,\tau$ small)

$$P(\sum B_i \ge \tau) \le \exp(-\tau^2/2)$$

Example

Write the z-score for slope as the sum

$$\frac{\hat\beta}{SE(\hat\beta)} = \frac{\sum(X_i - \overline{X})Y_i}{\sigma\sqrt{SS_x}} = \sum B_i$$

Bennett's bound gives $P(\hat\beta/SE(\hat\beta) \ge 13) \le .011$.

Too conservative?

Only small part of variation is "Poisson" and we know which part this is.

Answering Modeling Questions

Structure – What type of model?

A linear regression with least squares estimates.

- p potential predictors, n observations
- q non-zero predictors with error variance σ^2:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + \epsilon$$

Scope – Which X_j to consider?

Basically, everything...so p is very large.

- Demographics, time lags, seasonal effects
- Categorical factors, missing data indicators
- Nonlinear terms (quadratics)
- **Interactions** of any of these

Select – Which $q < p$ of the X_j go into the model?

- Requires *conservative*, *robust* standard error
- Measure significance without presuming CLT.
- **Use an adaptive selection rule.**

Adaptive Variable Selection

Hard thresholding

- Which predictors minimize max *ratio* of MSEs?

$$\min_{\hat{q}} \max_{\beta} \frac{E \left\| Y - \hat{Y}(\hat{q}) \right\|^2}{q\sigma^2}$$

- Answer: (Donoho&Johnstone, Foster&George 1994)

$$\text{Pick } X_j \quad \Leftrightarrow |t_j| > \sqrt{2 \log p}$$

  Almost Bonferroni! ($\sqrt{2 \log p}$ is a bit less strict)

Adaptive thresholding

- Which predictors minimize max *ratio* of MSE's?

$$\min_{\hat{q}} \max_{\pi} \frac{E \left\| Y - \hat{Y}(\hat{q}) \right\|^2}{E \left\| Y - \hat{Y}(\pi) \right\|^2} \quad \text{for } \beta \sim \pi$$

- Answer: (Foster & Stine 2002, in preparation)
  Pick q such that for $|t_1| \geq |t_2| \geq \cdots \geq |t_p|$,

$$|t_q| \geq \sqrt{2 \log p/q} \quad \text{but} \quad |t_{q+1}| < \sqrt{2 \log p/(q+1)}$$

Other paths to similar criteria

Information theory (Foster & Stine)

Empirical Bayes (George & Foster)

Generalized degrees of freedom (Ye)

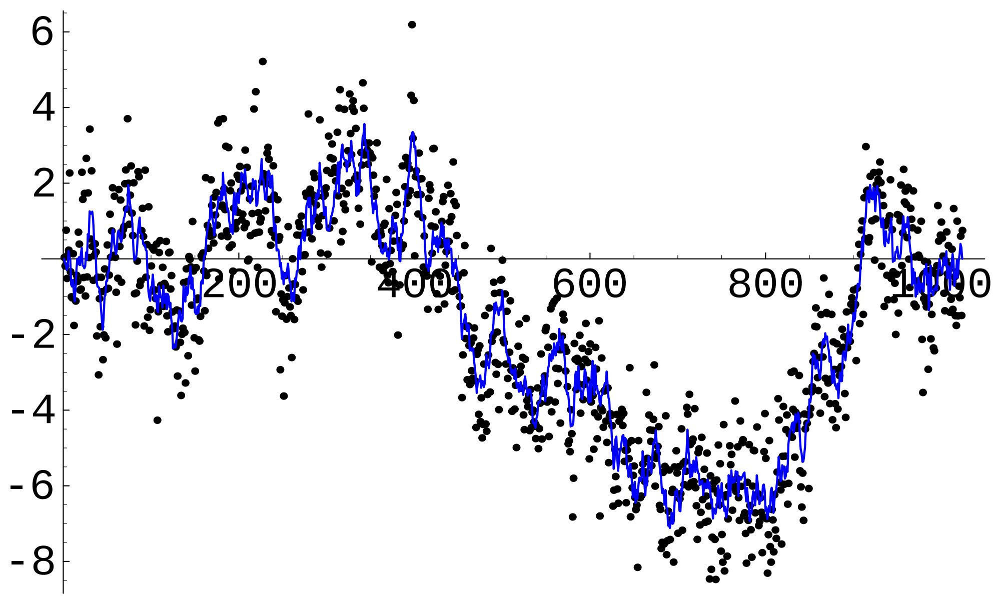
Simes method, step-up testing (Benjamini)

Example: Finding Subtle Signal

Signal is a Brownian bridge

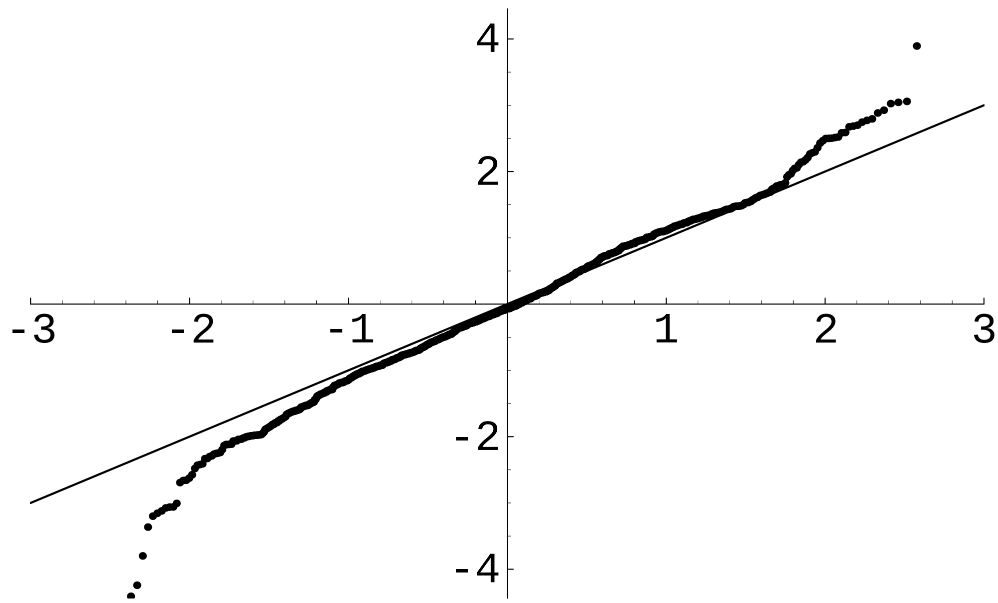
Stylized version of financial volatility.

$$Y_t = BB_t + \sigma\ \epsilon_t$$



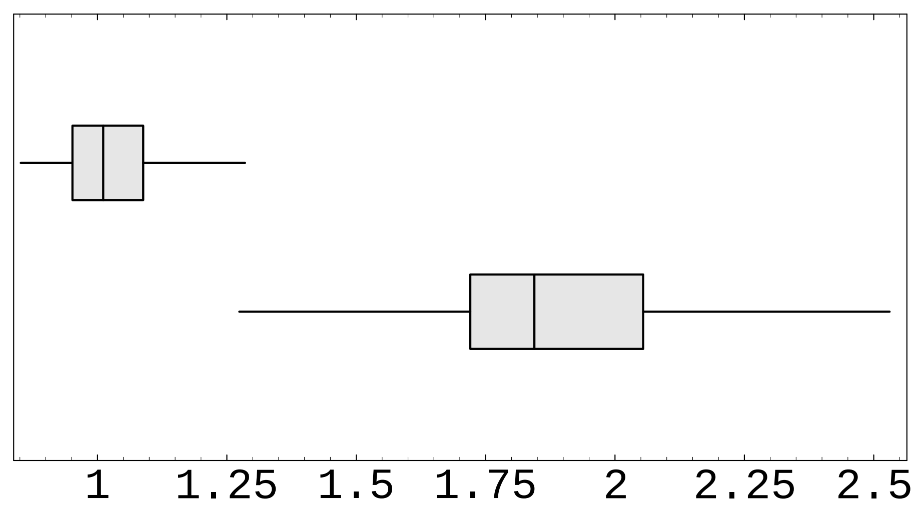
Example: Finding Subtle Signal

Wavelet transform has many coefficients



Comparison of MSEs

Boxplots show MSE of reconstructions using adaptive (top) vs. hard (bottom)



Modeling Approach

Structure

A linear regression with least squares estimates.

- p potential predictors, n observations
- q non-zero predictors with error variance σ^2:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + \epsilon$$

Scope

Basically, everything...so p is very large.

- Demographics, time lags, seasonal effects
- Categorical factors, missing data indicators
- Nonlinear terms (quadratics)
- **Interactions** of any of these

Select

- Requires *conservative*, *robust* standard error
- Measure significance without presuming CLT.
- Use an adaptive selection rule.

Test Case Study: Predicting Bankruptcy

Goal

Identify customers at "high" risk of declaring bankruptcy.

Rare event

Bankruptcy is a rare event in our data:

2,244 events in 3,000,000 months of data

Possible predictors

Collection of more than **67,000** possible predictors include

- Demographics
- Credit scores
- Payment history
- Interactions
- Missing data

Need all three aspects of our approach

- Robust SE

  Heteroscedastic because of 0/1 response variable.

- Bennett bound

  Sparse response *and* predictors like interactions.

- Diffuse, weak signal

  No one predictor will explain much variation alone.

Split-Sample Comparison

Reversed 5-fold cross-validation

- 20% for estimation ($n = 600,000$)
  about 450 bankruptcy events
- 80% for validation ($n = 2,400,000$)
  about 1,800 remaining bankruptcy events

Goal

Two ways to assess the models:

1. Predictive accuracy (squared error) and
2. Minimal costs, assigning differential costs to
   - Missing a bankruptcy (expensive)
   - Aggravating a customer (smaller cost)

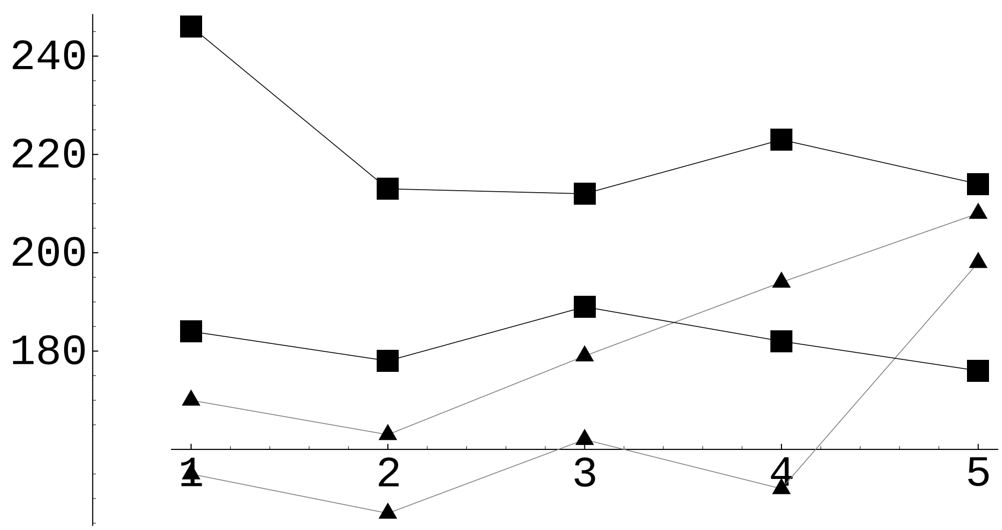
Machine-learning competitor

Two classification algorithms developed in the computer learning community (Quinlan):

C4.5 and C5.0 (with boosting)

Stepwise Has Better Brier Scores

Plot shows the *reduction* in the MSE of prediction over the null model for the five replications. Larger values are better.



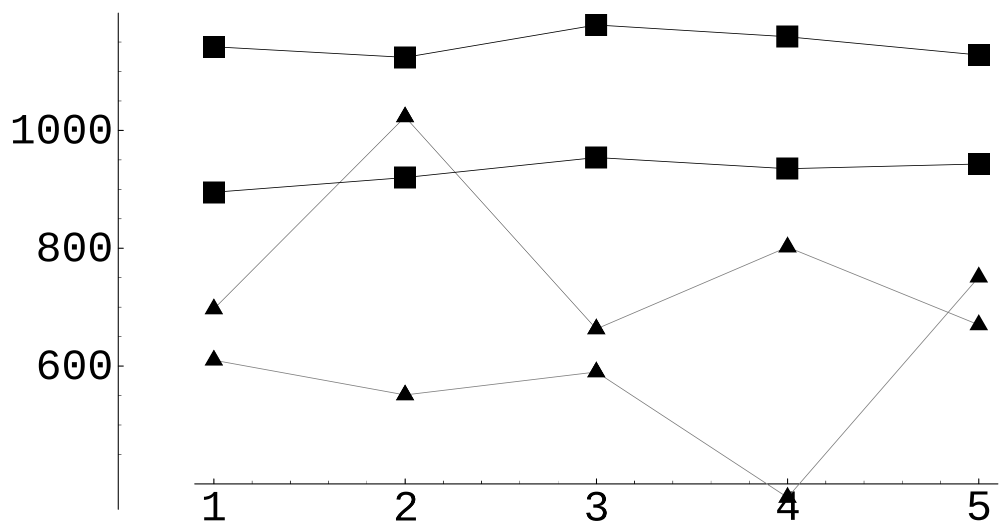
Boxes: Stepwise, with and without *calibration*

Triangles: C4.5, C5.0

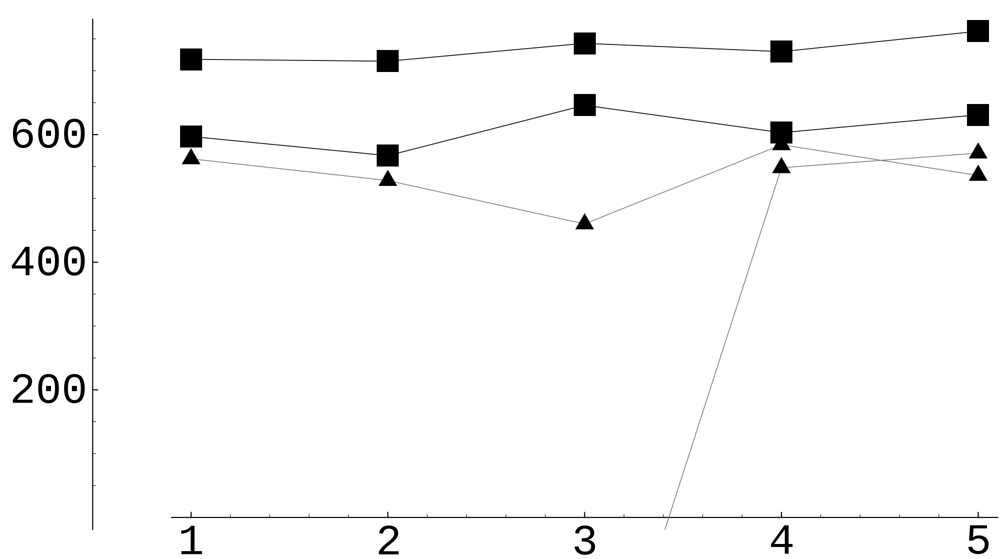
Stepwise Generates Larger Savings

Plot show the *savings* in accumulated losses over the null model for the five replications. Larger values are better. (Boxes–stepwise, Triangles–classifier).

Savings at a trade-off of 995 to 5.

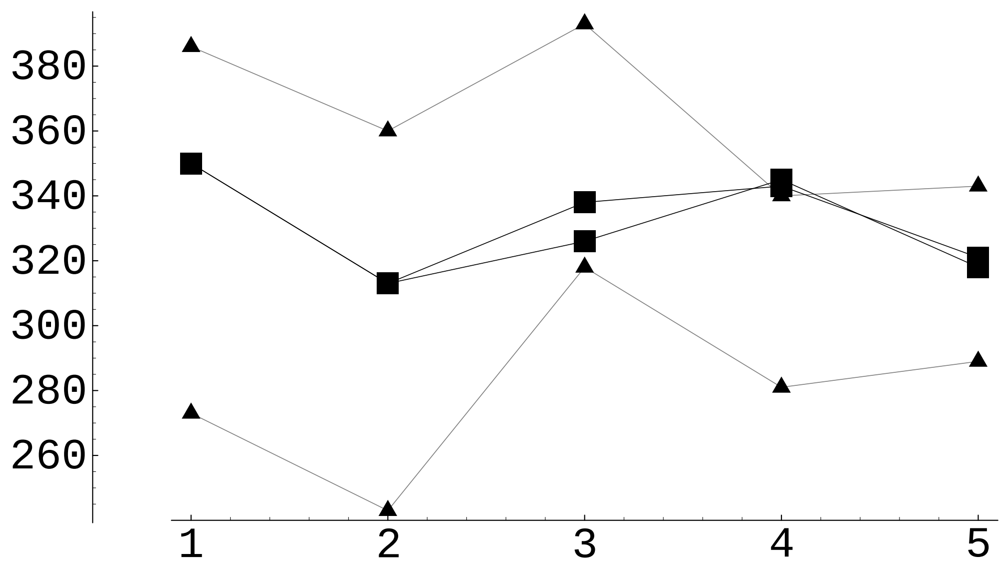


Savings at a trade-off of 980 to 20.



Well, Not Always

This plot shows the *savings* in accumulated losses over the null model for the five replications at a less extreme trade-off of 900 to 100.
(Boxes–stepwise, Triangles–classifier).



Notice that the differences are not so large as those in prior plots.

Calibration was not so helpful here as we expected.

Discussion

Adaptive variable selection

Powerful technique, strong theoretical basis

- Crucial role of standard error estimates
- Avoids "patterns" introduced by sparse data
- Adaptive cut-off finds structure Bonferroni misses
  Significant terms shown to help in validation

Implications for practice

- Automated search with good validation properties
  - Use more to estimate
- Supplement to "manual" analysis

Next steps Better searching ...

- More efficient search strategies
- Use of "expert" information
- Open vs. closed view of space of predictors