Applications of Model Selection in Credit Models

Bob Stine & Dean Foster
Department of Statistics, The Wharton School
University of Pennsylvania, Philadelphia PA
www-stat.wharton.upenn.edu/∼bob

August 12, 1999

- Overview our application: predicting bankruptcy
- Four questions in model selection:
  structure, identification, validation, and searching
- Role for information theory (in particular, coding)
- Results for our application
- Next steps

Predicting Bankruptcy

Goal

Predictive model for personal bankruptcy.

That is...

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. (Too late?)

Data

- Historical data for about 250,000 bankcard accounts
- Short monthly time series for each customer
- Selected a two-year window 1996-1997
- Credit limits, spend, payments, bureau info, background
- No transaction data
- Merged data of several banks
- Wharton Financial Institutions Center

Needle in the haystack

About 2,500 bankrupt out of $24 \times 250{,}000 = 6$ million months in this particular collection of accounts.

Becoming a Typical Problem

Goals of modeling effort

- Exploratory, generating conjectures for subsequent study
- Accurate prediction of new observations

Don't know the real model

- Unsure of model form, much less which predictors to use in some "true" model.
- Consistency in usual sense not an issue.
- Not the traditional test of $H_0 : \beta_3 = 0$

Many potential predictors

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

Differing expectations for predictors

- Some *likely* to be useful ($|t| \approx 10$)
- Some *might* be useful ($|t| \approx 2$) $\Leftarrow$
- Others that somebody collects and are available.

Big Questions

Structure — What form of model?

- Linear regression model
- Additive models (smoothing), regression splines (MARS)
- Neural nets, projection pursuit
- Regression trees (CART), piecewise fits

Information theory handles this problem well, but the computing is onerous.

Identification — Which variables to use in the model?

- Informed combinations e.g. *limit − curr balance*
- Time lags, other ransformations e.g. logs
- Interactions "less-informed" combinations ⇐

Validation — What's the out-of-sample accuracy?

Dilemma for cross-validation is that if you use

- All the data to identify predictors, have little for validation.
- Little to identify predictors, with most for validation.

This problem may be going away?

Search — How do you find the right model?

We still rely on brute force. Getting cheaper!

Focus Attention on Variable Selection

Ubiquitous issue

- Smoothing splines = variable selection from special basis.
- Nonlinear patterns ≈ powers and interactions.

Many approaches

- Thresholding/testimator, with the threshold (for orthogonal)
  - $\overline{R}^2$ $|t| > 1$
  - AIC $|t| > \sqrt{2}$ Mallows C_p
  - BIC $|t| > \sqrt{\log n}$ Bayes factor, Rissanen's "MDL"
  - RIC $|t| > \sqrt{2 \log p}$ Bonferroni
- Shrinkage: Ridge regression, PLS, other Bayesian methods
- Dimension reduction: principal components
- Visual: grand tours

Information theory has been useful

- Unified perspective that gives common ground for analysis.
- Generates ideas for alternative selection criteria.
- Help understand differences among methods.

Coding and Statistics

Canonical question How to compress a file of text, say, into a smaller file without losing any information?

Two codes

- Code I: a fixed-length code (like ASCII, but with 2 bits each)
- Code II: a variable-length code, matching length to exponent

Symbol y	Code I	Code II	$p(y)$
a	00	0	$1/2 = 1/2^1$
b	01	10	$1/4 = 1/2^2$
c	10	110	$1/8 = 1/2^3$
d	11	111	$1/8 = 1/2^3$

Example

String	Code I	Code II	P(String)
baa	010000	1000	$\frac{1}{4}\frac{1}{2}^2 = \frac{1}{2^4}$
dad	110011	1110111	$\frac{1}{8}\frac{1}{2}\frac{1}{8} = \frac{1}{2^7}$

Connection to statistics

Each code implies a probability model for the data. If the code for Y has length $L(Y)$, then $P(Y) = 2^{-L(Y)}$.

Coding and Model Selection

Idea

"Good" models yield codes that compress the observed data into a shorter **message** than bad models.

Maximum likelihood

Given a parametric model $P_\theta(Y)$ for the response data Y, (such as the Gaussian regression model)

$$\begin{aligned} \min_\theta(\text{code length for data}) &= \min_\theta \log 1/P_\theta(Y) = \max_\theta P_\theta(Y) \\ &= c \text{ Residual SS} \end{aligned}$$

How's this help? Isn't this just going to maximize R^2?

Rissanen's MDL criterion

Pick model obtaining shortest encoding of **itself** and **data**.

The message must encode *both* the slopes and the residuals (everyone knows all of the X's):

$$\begin{aligned} \text{Message length} &= L(\text{parameters}) + L(\text{data}) \\ &= L(\text{parameters}) + c \text{ Residual SS} \end{aligned}$$

$\Rightarrow$ penalized likelihood.

Thus, coding does not simply pick the model that maximizes R^2.

Coding and Models

MDL

Find the model giving the shortest *total message length.*

How to encode the parameters?

The representation of the parameters determines how much they contribute to the length of the message, and thus which to add:

$$\text{Add } \hat{\beta}_j \qquad \Longleftrightarrow \qquad (\downarrow \text{ Residual SS}) > \ (\uparrow\ L(\text{parameters}))$$

Code needs to answer two questions

Not enough to encode just the slope parameter, the code must also attach it to a predictor.

1. Which variables are in the model?
2. What are the coefficients of the chosen variables?

Bayesian tie

IT gives a nice way to get priors.

$$\text{Message length} \ = \ \text{Length parameters} \ + \ \text{Length data}$$

$$\Rightarrow \quad P(Y) = P(\theta)\ P(Y|\theta)$$

Representing the Regression Parameters

Context

1. Pick k from a collection of p possible predictors.
2. Pretend that predictors are orthogonal (e.g. wavelets).
3. Pretend error variance σ^2 is known.

Note: 2+3 → know $SE(\hat\beta_j)$ and $\hat z_j = \hat\beta_j / SE(\hat\beta_j)$.

Spike-and-slab, BIC (Rissanen 1983)

Suppose that you know the largest possible z score is z_{max}. Allocate enough bits to represent z_{max} for each coef regardless of realized value ("fixed format", Code I)

$$\underbrace{1\ 0\ 0\ 0\ 1\ 0\quad \ldots\quad 1}_{\text{p bits}} \mid \hat z_1 \mid \hat z_5 \mid \quad \ldots \quad \mid \quad \underbrace{\hat z_k}_{\text{1+log z-max bits}}$$

1. Identify chosen variables using p indicator bits.
2. Fixed length: Encode the k chosen coefficients by rounding **z-scores** to integers, each at a cost of $1 + \log_2 z_{max}$ bits.

AIC (Foster and Stine 1997, 1999)

With a different code, MDL also yields the AIC:

1. Identify chosen variables using p bits.
2. Variable length: Round to **z-scores**, and encode these using "log-Cauchy" at a cost of $1 + 2\log_2 z_j$ bits.

Representations for Many Predictors

Bankruptcy context

The number of coefficients p is very large, and we expect very few of these to enter our model. Why use that p bit prefix?

Problem with previous codes

We expect $k << p$, but prior AIC and BIC codes are designed for problems where **half** of the coefficients enter the model.

RIC (D&J, Foster and George 1993, Foster and Stine 1997)

Rather than identify coefficients using p bits, use $\log_2 p$ bits to identify the predictor.

$$(j, \hat{\beta}_j) \quad \Rightarrow \quad (\log_2 p, 1 + 2\log_2 z_j)\text{bits}$$

Optimal with about 1 coefficient to be coded (Poisson), and is equivalent to using the Bonferroni threshold $\sqrt{2 \log p}$.

EBIC (Foster and Stine 1997)

Rather than assume *a priori* that k/p is large or small, use a procedure that encodes the choice of predictors **adaptively**:

1. Compress the p bit indicator.
2. Variable length: Round to **z-scores**, encode these using "log-Cauchy" at a cost of $1 + 2\log_2 z_j$ bits.

Back to the Application

Data preparation

- Merge data from multiple files.
- Convert categorical data into indicators.
- Handle missing data.

Result is a numerical array for modeling.

Variable selection

- Choose predictors from collection of interactions, including missing data indicators, nonlinear, and interaction terms.
- We'll use a straight-forward greedy stepwise method, using an information theoretic criterion to evaluate the selected model.

Estimate out-of-sample performance

- Ideally, use a method unified with the choice of predictors.
- Currently, use an 'honest' confidence interval for scale of prediction error

Data Preparation

Merging data

- Data in 35 files:

  24 monthly, 8 quarterly, 2 annual, and descriptive data
- Convert incoherent formats

Sample data

- Find all cards where bankruptcy appears
- Record several prior months of data; prior quarter, curr year.
- Sample random month for some fraction of others

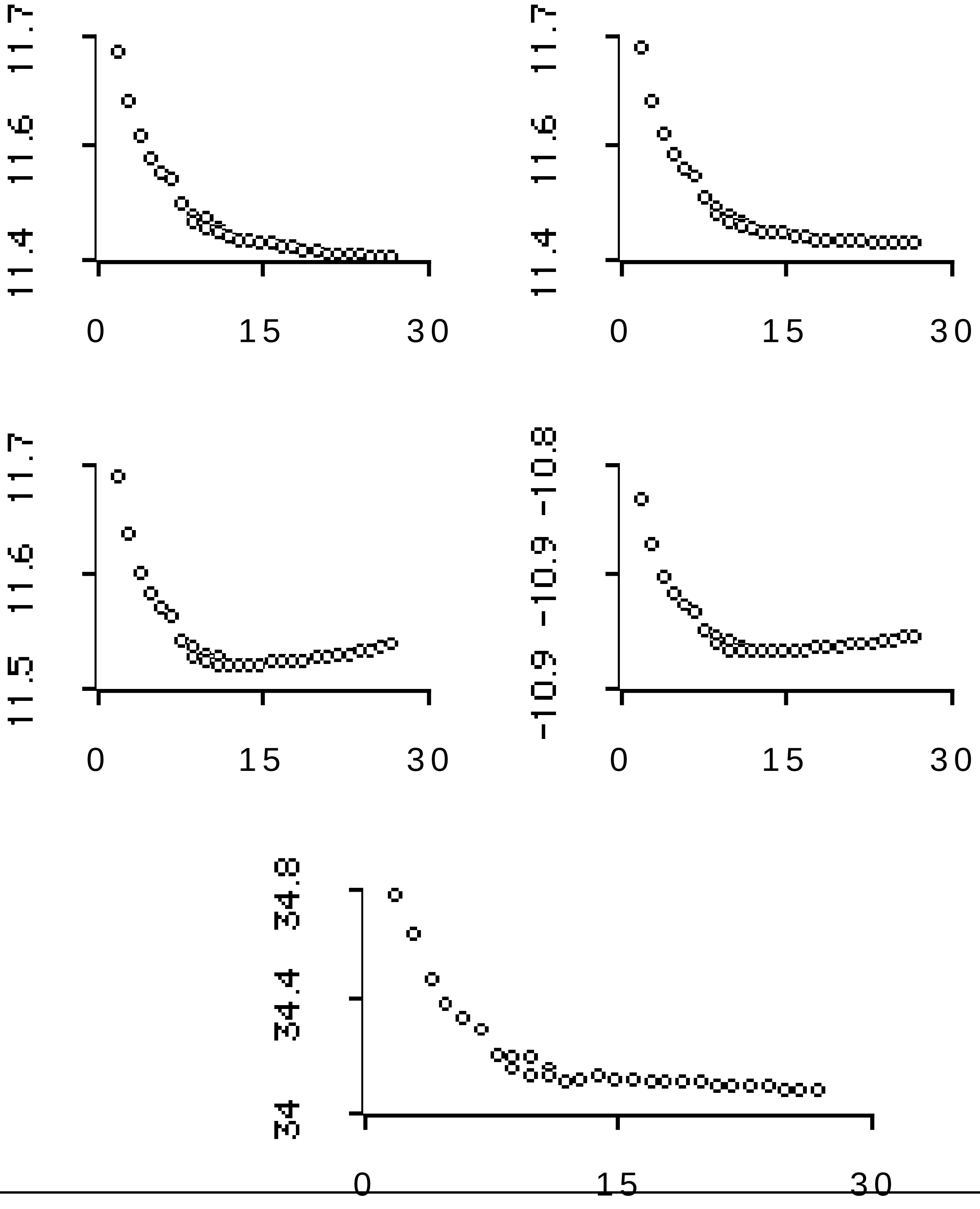
Missing data

- Most predictors have missing cases.
  - Structural: Subsets with more extensive data.
  - Temporal: “We didn’t collect that back then.”
  - Some are informative?
- Treat observed X_j as the product of missing indicator with ‘true’ covariate,

  $$X_j = I_j \, X_j^*$$
- Include both X_j and I_j as possible predictors — roughly doubling the number of predictors.

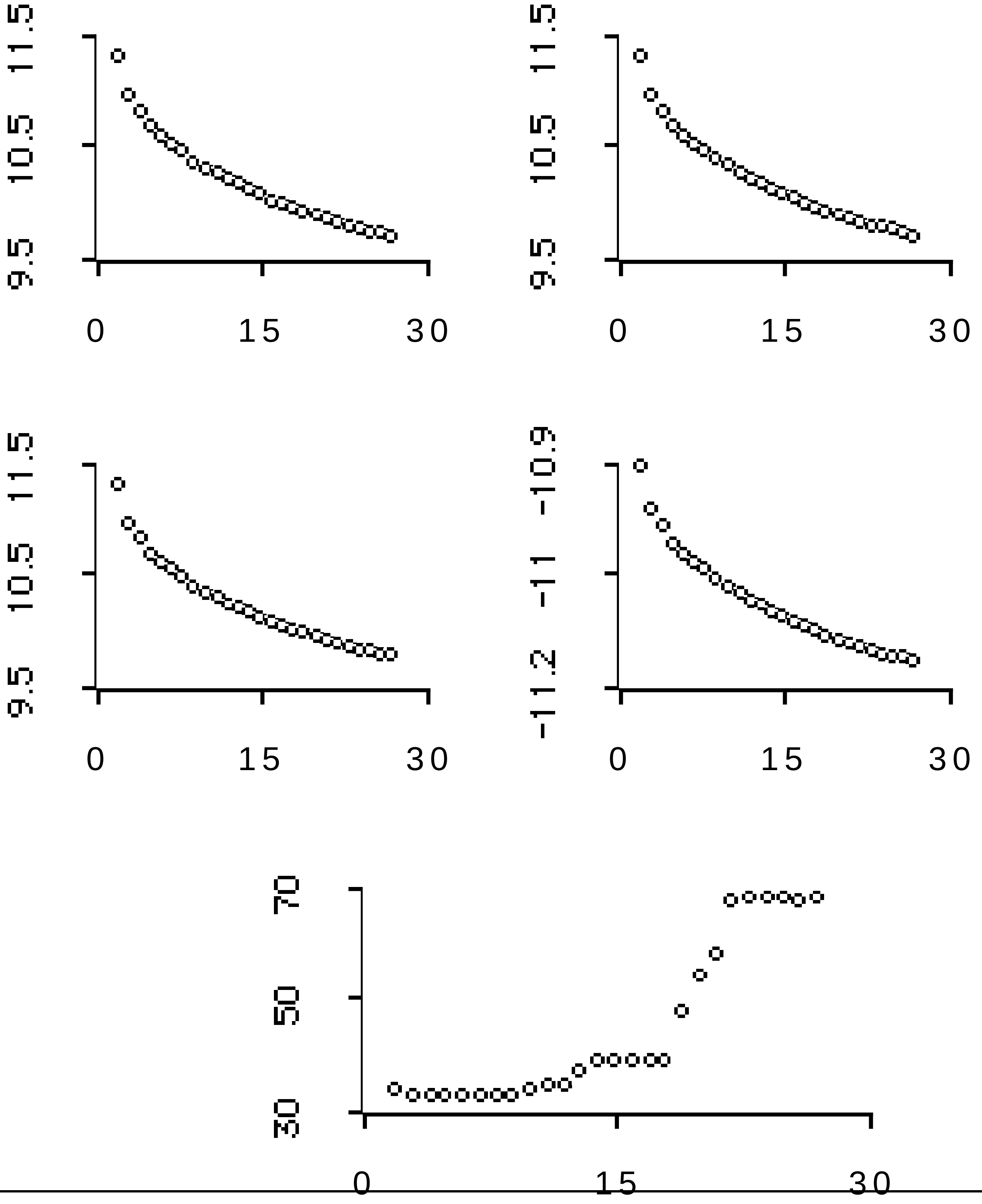
Some Results with No Interactions

365 predictors, 15,000 obs fit, 45,000 validation



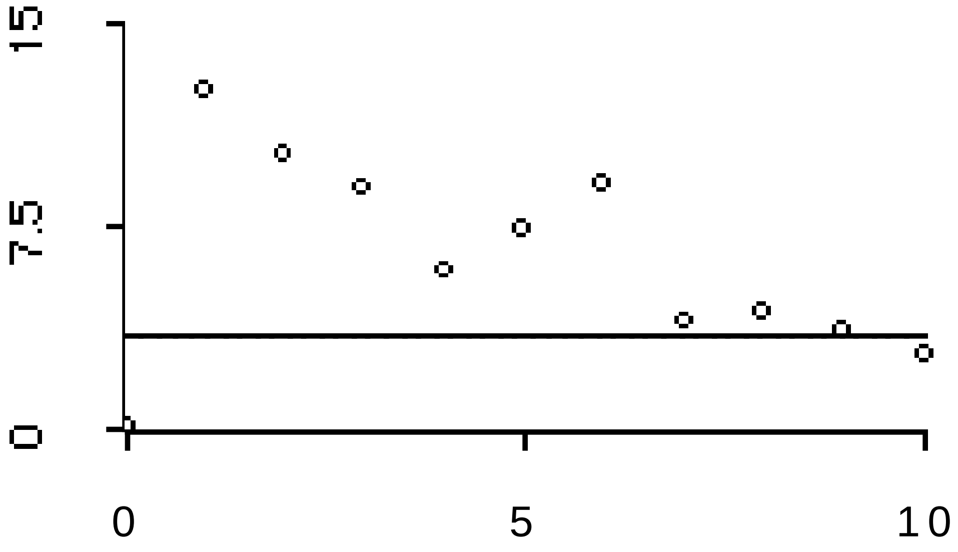
Same Criteria, With Interactions

67,000 predictors, 15,000 obs fit, 45,000 validation

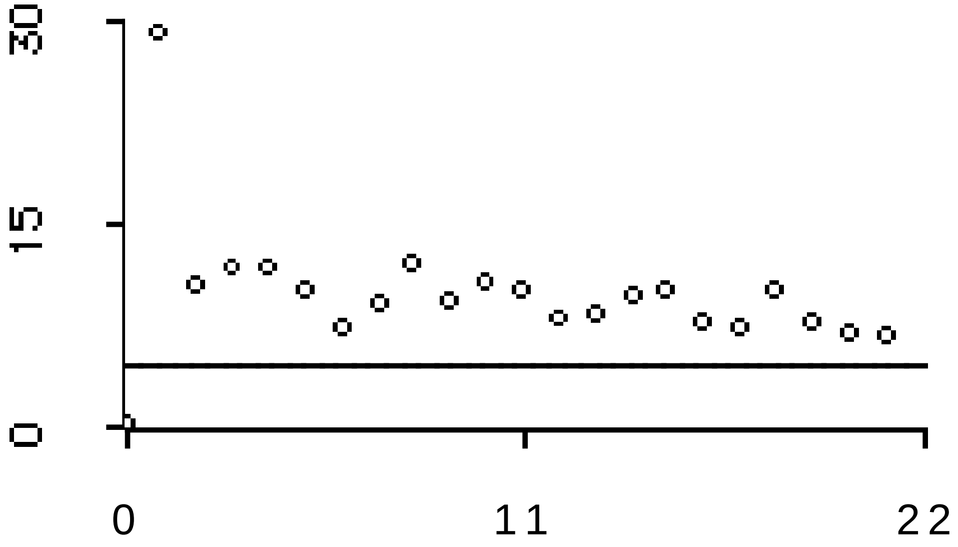


Fitted Models and Bonferroni Threshold

t-statistics in order of entry for 10 predictors, chosen from $p = 365$.



With 21 predictors chosen from $p = 67,000$.



Horizontal lines gives Bonferroni bounds $\sqrt{2 \log p} = 3.4$, 4.7

What are some of those factors?

t-statistic	Interaction Terms	
29.25	NUM-CARDS-CLOSED	NUM-CARDS-EVER-60
10.20	LATE-CHG-m0	NUM-CARDS-EVER-60
11.67	LATE-CHG-m0	NUM-CARDS-CLOSED
-11.83	EXT-STAT-E-m0	NUM-OURS
10.20	CASH-INT-m0	EXT-STAT-I-m3
7.60	EXT-STAT-E-m0	EXT-STAT-NA-m3
-8.94	NUM-CARDS-EVER-60	NUM-OURS
-12.03	EXT-STAT-B-m2	NUM-CARDS-CLOSED
9.47	EXT-STAT-E-m0	MAIL-IND-M
-10.87	CYC-CA-m0	EXT-STAT-I-m3
10.28	MRCH-INT-m0	NUM-CARDS-EVER-60
-9.65	NUM-CARDS-POS	NUM-CARDS-EVER-60
-10.12	EXT-STAT-E-m3	NUM-CARDS-EVER-60
-7.67	INT-STAT-D-m0	OVR-CHG-m0
-7.27	EXT-STAT-C-m2	NUM-CARDS-EVER-60

Need to be really careful at this point!!!

So What's Going On?

Confused state

Selection tools appear effective with $p = 350$ (stop selecting more when valication error goes up, without looking at validation error), but the fits are poor.

When we add more predictors we find that...

Same criteria fail (i.e., suggest factors that are useful, but they aren't) when the number of predictors soars to $n << p = 65,000$.

Dilemma

- Good est of out-of-sample performance of poor estimator, vs
- Poor est of out-of-sample performance of good estimator.

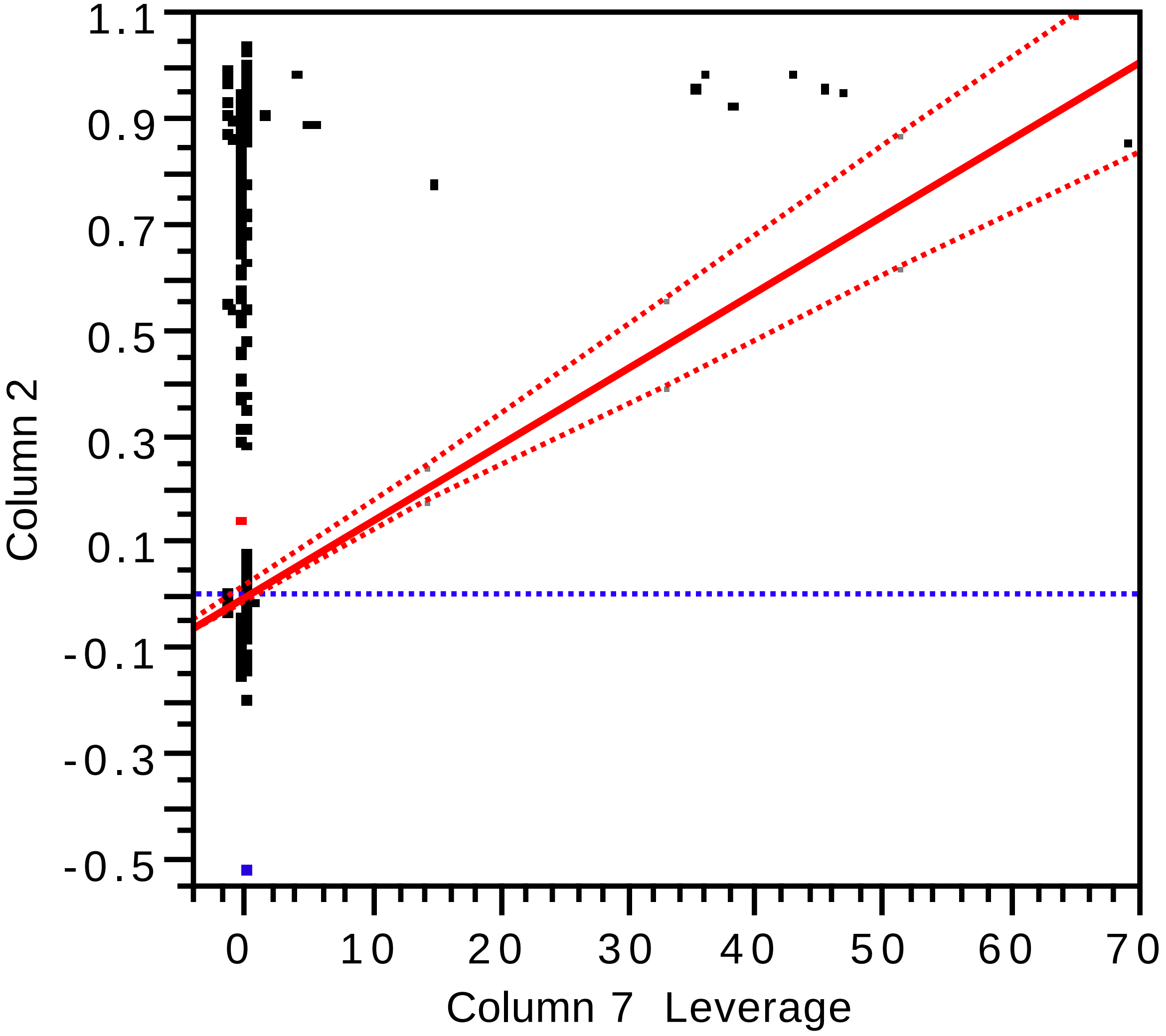
Why are selection tools failing?

- Weighted sampling method (oversampling)
- Outliers (see leverage plot)
- Linear regression rather than logistic
- Practical vs statistical significance

Still a Role for Pictures

Plots remain useful

Partial regression leverage plot for predictor with interactions, (cross-product of internal status code with number of cards).



Honest Intervals for σ^2

Bias in $\mathbf{s^2}$

Since $p > n$, no simple unbiased estimate exists. Once a "noise" variable is selected, resulting selection bias makes it easier for other noise terms to enter and appear significant.

⇒ Inflated nominal significance.

Goal

Offer a guaranteed level of performance in prediction.

Honest interval for σ^2

Bonferroni adjust *one-sided* interval for σ^2, with $m = \binom{p}{k}$,

$$\left[0,\ \frac{(n-k-1)s_k^2}{\chi^2_{n-k-1,0.05/m}}\right]$$

Need to compute extreme left quantile of χ^2 distribution.

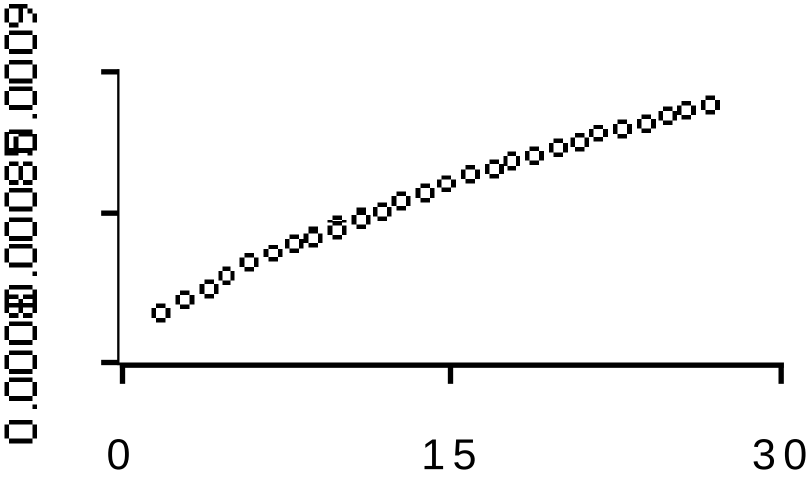
Use in model selection?

Pick the model having the smallest adjusted upper endpoint, noting that ...

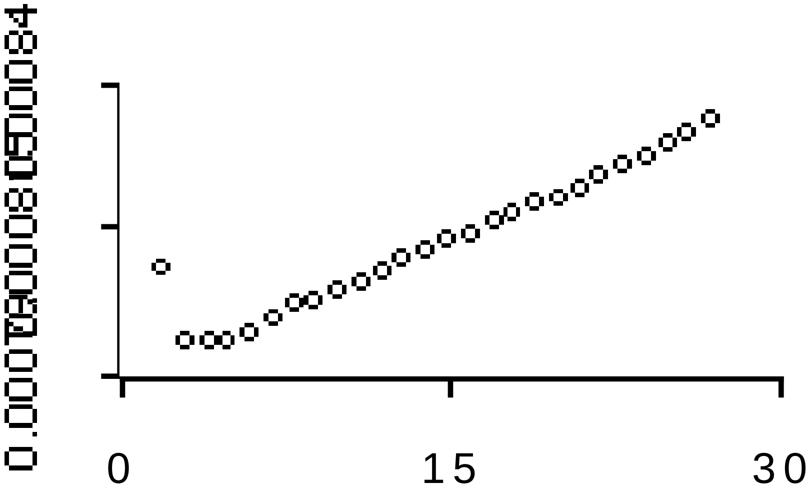
The method tends to be conservative since it does not really adjust for the presence of weak signal. Hard to see how without many assumptions.

Plot of Honest Estimates

For the model with $p = 365$, the honest s^2 likes the simplest model.



With $p = 67,000$, it stops the selection process.



Next Steps

Our application

- Transaction data (as used in fraud models)
- Different forecast horizons, profitability analysis
- Other models: logistic, survival, ...
- More insight into industry experience, 'priors'

Theory Allow richer model spaces, using information theory to choose class of models and combine

- Selection of model form
  - Hybrid combination of regression and CART
    - Split the sample vs the interaction approach
  - Context trees
- Selection of predictors
  - Stochastic complexity criterion
- Estimate of prediction MSE (reproduce honest s^2 behavior)

Computing

- Which subsets of predictors? Can't do all...
- Which splits of the data?
- Is brute force the only way?

A Different Formulation for IT

Experts and Oracles

- Way of thinking about side-information.
- Provide a benchmark for comparison, competitor.

What could an oracle supply?

Coefficient?

Coefficient value / level of significance.

Fit of model? $\Leftarrow$

Goodness-of-fit (e.g., R^2) with one predictor, two, three,...

Advantages

- Easier to elicit.
- Easier to compete against.
- Coordinate free.

Status

A work in progress. Seems to solve some technical problems with stochastic complexity, but have to tie to ideas that underlie honest s^2.

Overfitting

What does it mean to overfit?

- Does it mean using a poor estimator, such as a procedure that gives an inefficient estimate?

  This occurs in our setting if we use a procedure that expects $\frac{1}{2}$ the predictors to enter the model. Like fixing the probability of heads at $\frac{1}{2}$, ignoring the data.

  Works well if $\frac{1}{2}$ is the right answer!

- Adaptive methods handle this one pretty well.

- Does it mean that we claim that our model predicts better than it actually will?

  Harder question to answer, and one that is not so obvious to answer from information theory.

  "Arbitrary" attribution of message length to parameters and data in *tight* codes. Previous illustrative codes are not tight. Analogous to integrating over the model space in a Bayesian analysis — Where'd the model go?