Variable Selection in Wide Data Sets

Bob Stine

Department of Statistics, The Wharton School

University of Pennsylvania, Philadelphia PA

www-stat.wharton.upenn.edu/∼stine

February 2, 2005

Collaboration with Dean Foster

- Main problem
  - Picking the features to use in a model
  - Wide data sets
- Main examples
  - Simulated idealized problems
  - Predicting credit risk
- Themes
  - Modifying familiar tools for data mining
  - Changing the approach to inference

Questions

Credit scoring

Can you predict who will declare bankruptcy?

Drug safety

Do reported adverse experiences suggest a systematic problem?

Identifying faces

Is this a picture of a face?

Genomics

Does a pattern of genes predict higher risk of a disease?

Text processing

Which references were left out of this paper?

These are great statistics problems, so...

Why not use our workhorse, regression?

- Different flavors (least squares, logistic, ...)
- Calculations well-understood.
- Results are familiar.
- Diagnostics are available.

Models

Simple structure

A linear or logistic regression ...

- n **independent** observations of m features
- q predictors in model with error variance σ^2:

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

Calibrate predictions Allow for various link functions ...

- Linear regression has identity link, logistic has logit.

$$E(Y|X) = h(\beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q)$$

- "One-step" solution at end of selection.

Rich feature space

Allow large, diverse set of features

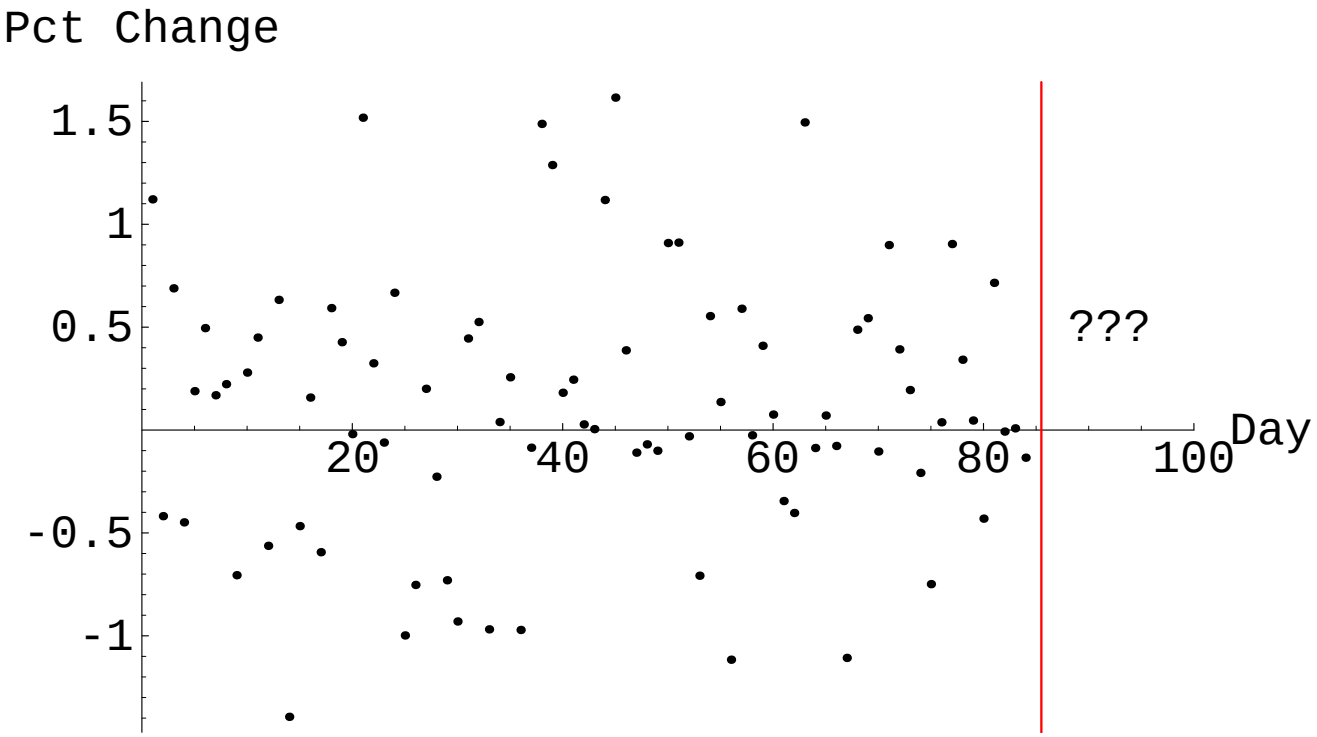
- Usual mix: continuous, categorical, dummy vars, ...
- Missing data indicators
- Combinations (principal components) and clusters
- Nonlinear terms, transformations (quadratics)
- **Interactions** of any of these

Which features generate best *predictions***?**

Stock Market

Where's the stock market headed in 2005?

Daily percentage changes in the closing price of the S&P 500 during the last 3 months of 2004 (85 trading days) ...



Predicting the Market

Problem

Predict returns on S&P 500 in 2005 using features built from a collection of 12 exogenous factors.

Regression model

$R^2 = 0.85$ using $q = 28$ predictors.

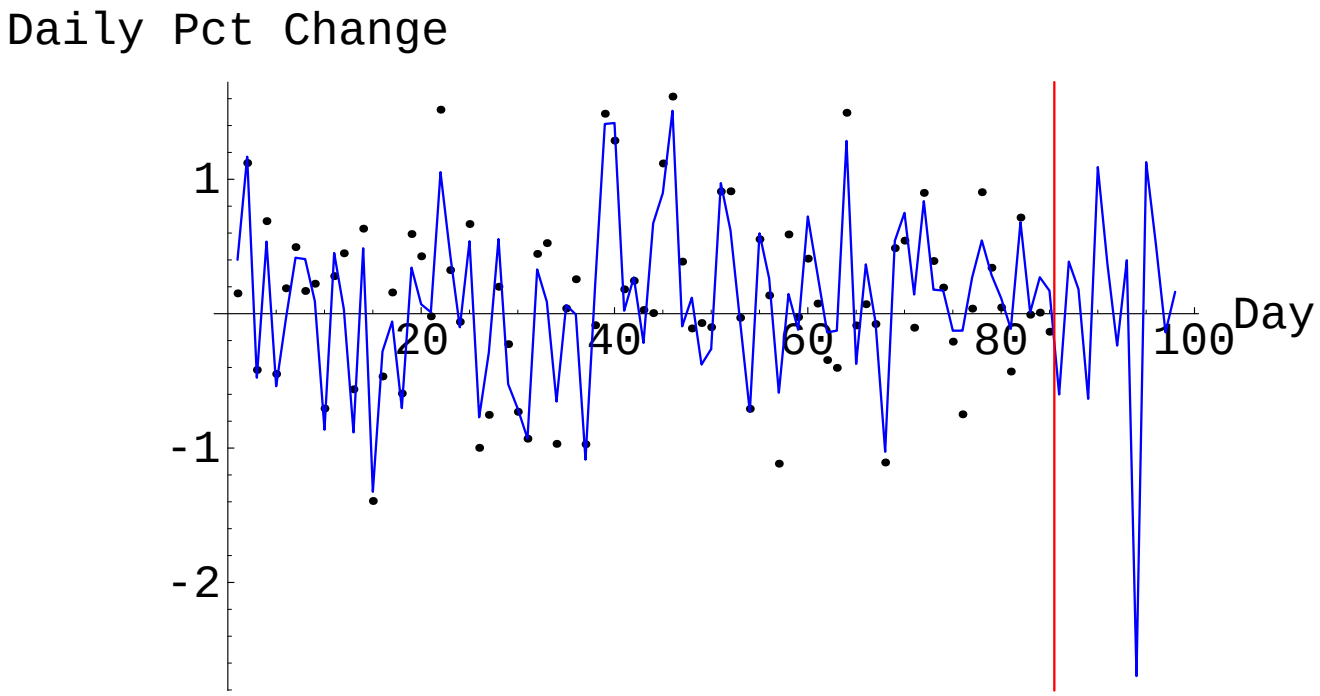
With $n = 85$, $F = 12$ with p-value < 0.00001.

Coefficients are impressive...

Term	Estimate	Std Error	t-Ratio	p-value
Intercept	0.323	0.078	4.14	0.0001
X3	0.095	0.039	2.45	0.0175
X4	0.172	0.040	4.34	0.0000
(X1)*(X1)	-0.202	0.039	-5.16	0.0000
(X1)*(X5)	0.256	0.048	5.34	0.0000
(X2)*(X6)	0.289	0.044	6.59	0.0000
(X4)*(X6)	-0.222	0.050	-4.43	0.0000
(X4)*(X7)	-0.213	0.047	-4.54	0.0000
(X5)*(X9)	-0.192	0.044	-4.35	0.0000
(X7)*(X9)	0.249	0.046	5.37	0.0000
...				

Model Fit and Predictions

Predictions from the model track the data closely, even matching turning points.

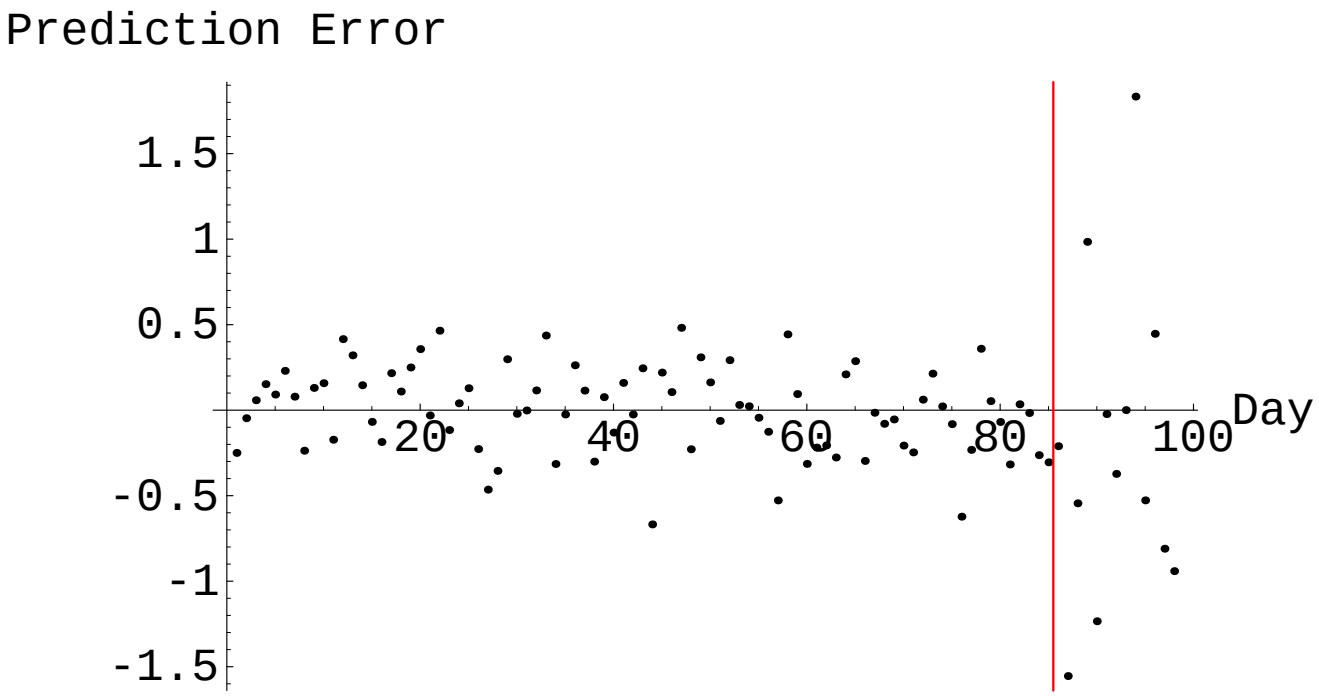


Looks like we should bet against the market on that day...

No guts, no glory!

Prediction Errors are LARGE

In-sample prediction errors are small, out-of-sample much larger.



How'd you lose the house?

How could this happen?

Significant fit, even by Bonferroni standards.

What was that model?

Exogenous base variables

12 columns of random Gaussian noise.

Null model is the right model.

Selecting predictors

Turn stepwise regression lose $m = 90$ constructed features

12 linear + 12 Squares + 66 Interactions

with "promiscuous" settings for adding variables,

prob-to-enter = 0.16 (AIC)

Then run stepwise backward to remove extraneous effects.

Result is an impressive-looking fit.

Why does this fool Bonferroni?

Once stepwise adds a predictor, s^2 becomes biased down, making the remaining predictors appear more significant.

Often "cascades" into a perfect fit.

Cannot fit saturated model to get unbiased estimate of σ^2 because $m > n$.

Easy fix

Control the **fitting process**. Use Bonferroni from the start, not after the fact.

Morals

Stepwise regression can make a silk purse from sow's ear.

Must control the process, not diagnose the final model.

Second Example: Predicting Bankruptcy

Predict onset of personal bankruptcy

Estimate probability use of credit card declares bankruptcy during the next billing cycle.

Challenge

Can **stepwise regression** predict as well as commercial "data-mining" tools or substantive models?

Many features

About 350 "basic" variables

- Short time series for each account
- Spending, utilization, payments, background
- Missing data and indicators
- Interactions are important (LV and cash adv)

$$m = \textbf{67,000 predictors!}$$

- **Transaction** history would vastly expand the problem.

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."

Approach

Stepwise search

Forward search with p-values to determine whether to add predictors rather than cross-validation.

Testimators rather than model averaging, MCMC.

Risk inflation criterion (RIC)

Select predictor if (m = number possible predictors)

$$\text{p-value} \approx 1/m$$

Obtains RIC bound (Foster & George 1994)

$$\min_{\hat{\beta}} \max_{\beta} E \frac{\|Y - X\hat{\beta}\|^2}{|\beta|\ \sigma^2} \leq 2 \log p$$

a.k.a.: hard thresholding, Bonferroni, Fisher's method

Adaptive thresholding

RIC best when "truth" is sparse, but lacks power if much signal. **False discovery rate** motivates an alternative.

If you've added q predictors, add the next if:

$$\text{p-value} \approx q/m$$

Further motivation...

- Half-normal plot (Cuthbert Daniel?)
- Generalized degrees of freedom (Ye 1998, 2002)
- Empirical Bayes (George & Foster 2000)
- Information theory (Foster, Stine & Wyner 2002)

Adaptive Variable Selection

Hard thresholding

Which predictors minimize max *ratio* of MSEs?

$$\min_{\hat{\beta}} \max_{\beta} \frac{E\,\|Y - X\hat{\beta}\|^2}{|\beta|\sigma^2}$$

Donoho&Johnstone, Foster&George 1994 answer (using t-stat)

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

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

Adaptive thresholding

Let π denote a symmetric, unimodal prior on for β and let $\hat{Y}(\pi)$ denote predictor based on π. Which predictors minimize ratio?

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

Pick q such that for $|t_1| \geq |t_2| \geq \cdots \geq |t_p|$,

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

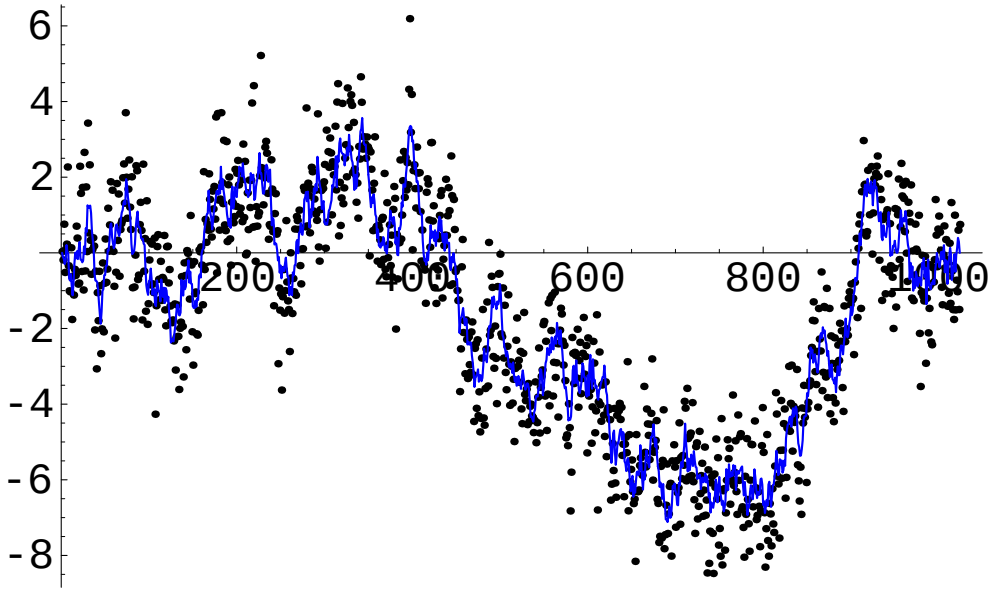
Details in Foster & Stine manuscript.

Example: Finding Subtle Signal

Signal is a Brownian bridge

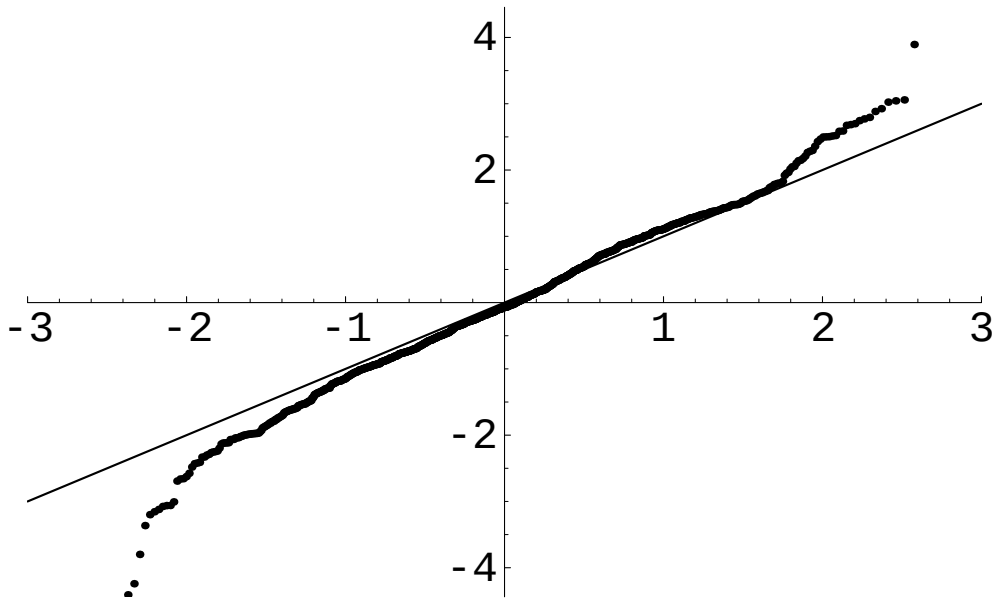
Stylized version of financial volatility.

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



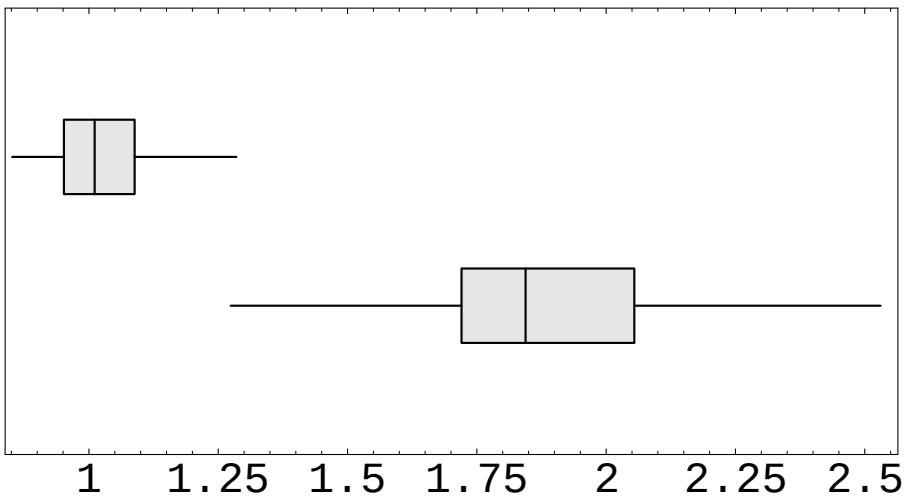
Example: Finding Subtle Signal

Wavelet transform has many coefficients



Comparison of MSEs

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



Does it really work?

Theory usually presumes...

- Normal distribution on error (thin tailed).
- Know "true" error variance σ^2.
- Error variance is constant.
- Predictors are orthogonal.

Combination of these mean that parameter estimates are independent with known sampling distributions.

$\Rightarrow$ know the p-values.

In practice...

- Is anything normally distributed?
- Don't believe model, much less error variance.
- Suspect measurements of varying accuracy.
- Collinear features, frequently with $m > n$.

$\Rightarrow$ don't necessarily have p-values.

Questions

How to handle the collinearity?

How to get p-values (or test statistics) to plug into adaptive selection?

Honest p-values for Testing $H_0 : \beta = 0$

Three methods

Each method makes some assumption about the data in order to produce reliable p-value.

Method	Requires
White estimator	Symmetry
Bennett bounds	Bounded Y
Robust estimator (e.g. ranks)	Homoscedastic

Concern from stock example

Have to avoid "false positive" that leads to inaccurate predictions, cascade of mistakes.

Need to be sure that if select a feature as a predictor, it genuinely improves accuracy of fit.

Initial approach to bankruptcy

Thresholding with OLS worked fine there, lets use it here. With very large n, CLT surely protects us.

Test Problem

Motivation

Used 5-fold "reversed" cross-validation to confirm results,
fit on 600,000, predict 2,400,000.

As fitting proceeds, monitor out-of-sample error.

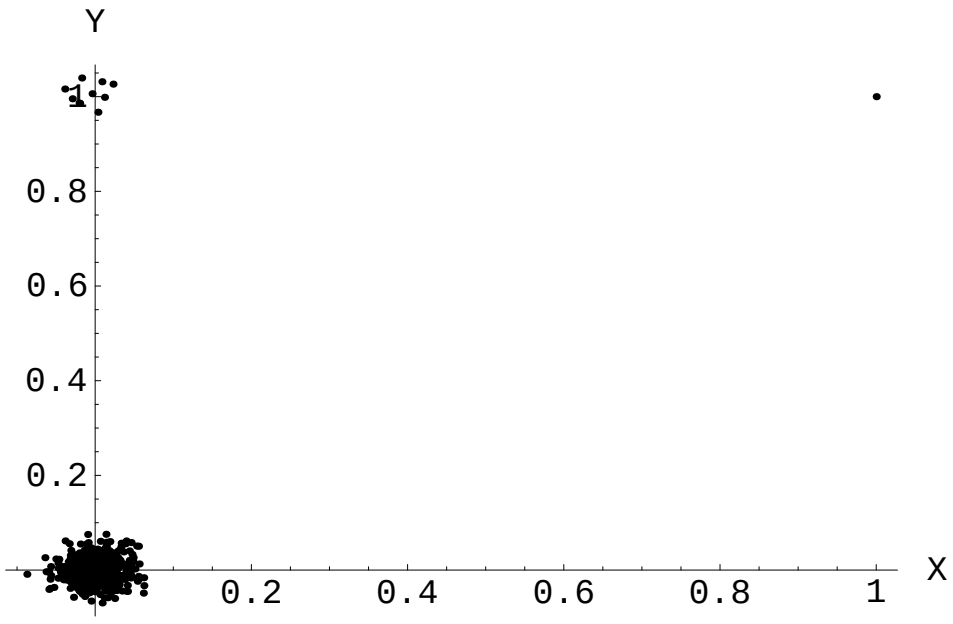
Sudden spike in prediction error

Diagnostic **plots** reveal problem due to sparse nature of interactions, even though n=600,000.

Stylized problem

$n = 10,000$ and $X_1 = 1$, $X_2, \ldots, X_{10,000} \sim N(0, 0.025)$

$P(Y = 1) = 1/1000$, independent of X. (dithered in plot)



Usual t-statistic

$se = 0.016 \Rightarrow t = 14$ for a p-value of 0.0000000...

Common sense suggests p-value $\approx 1/1000$.

White Estimator

Least squares estimator

$$Y = \hat{\beta}_0 + \hat{\beta}_{q,1} X_{q,1} + \cdots + \hat{\beta}_{q,q} X_{q,q} + \epsilon$$

$\Rightarrow$ estimator $\hat{\beta}_q$ with design matrix X_q.

Sandwich formula (H. White, 1980, *Econometrica*)

$$\mathrm{Var}(\hat{\beta}_q) = (X_q' X_q)^{-1} X_q' \underbrace{\mathrm{Var}(\epsilon)} X_q (X_q' X_q)^{-1}$$

Estimate variance using the residuals from **prior** step:

$$\mathrm{Var}(\hat{\beta}_q) = (X_q' X_q)^{-1} X_q' \underbrace{\mathrm{Diag}(e_{q-1}^2)} X_q (X_q' X_q)^{-1}$$

Result in stylized problem

Estimated std error is 10 times larger, $\widehat{se} = 0.16$, giving a more modest (and appropriate) p-value.

Working under null

Residuals from prior step e_{q-1} computes SE under null, not under the alternative.

Idea is to test only

$$H_0 : \beta_{q,q} = 0$$

rather than

$$H_0 : \beta_{q,q} = 0 \quad \& \quad \mathrm{Var}(\epsilon_i) = \sigma^2$$

Example: Finding Missed Signal

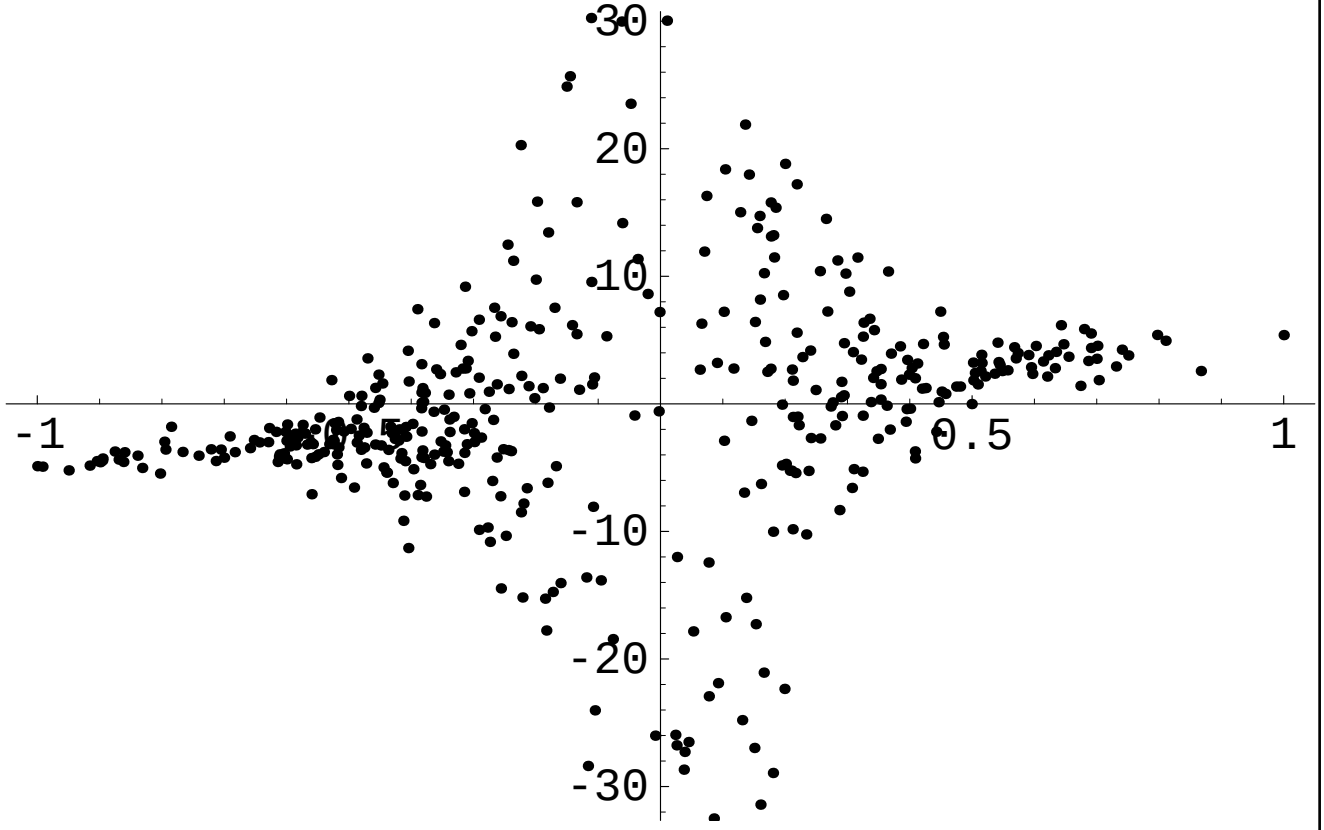
Question

Does White estimator ever produce significant effect that OLS misses?

In most cases, $\mathrm{Var}(\epsilon_i)$ increases with $\hat{y}_i$, and OLS underestimates SE.

Heteroscedastic data

High variance at 0 obscures the differences at extremes.



Least squares

Standard OLS reports SE = 3.6 giving t = 1.2.

White SE is 0.9, so t is 4 times larger and finds the underlying effect.

Honest p-values for Testing $H_0 : \beta = 0$

Three methods

Each method makes some assumption about the data in order to produce reliable p-value.

Method	Requires
White estimator	Symmetry
Bennett bounds	**Bounded Y**
Robust estimator (e.g. ranks)	Homoscedastic

Concerns

Avoid "false positive" that leads to inaccurate predictions, cascade of mistakes.

Need to be sure that if select a feature as a predictor, it genuinely improves accuracy of fit.

Aside...

Use OLS with 0/1 response?

Yes!

Weighted with $\hat{Y} \times (1 - \hat{Y})$ to compute SE, but not to pick variables or estimate parameters.

Context

Handful of bankrupt cases ($y_i = 1$).

Most $\hat{y}_i \approx 0$, the regular people.

Do you really want to minimize

$$\sum_i \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i(1 - \hat{y}_i)}$$

and put yet more weight on the least interesting cases?

Compromise

Minimize $\sum(y_i - \hat{y}_i)^2$, but recognize heteroscedasticity when computing SE.

If you're going to be judged by squared error, minimize it!

Ideal

Choose economically correct loss function (from cost of BR and cost of annoying others).

Lenders loath to reveal such details.

Bennett Inequality

Think differently

Sampling distribution of $\hat{\beta}$ is not normal because huge leverage in one point contributes "Poisson-like" variation to the estimator.

Bennett inequality (Bennett, 1962, *JASA*)

Bounded independent r.v. $U_1, \ldots, U_n$ with $\max |U_i| < 1$, $E\, U_i = 0$, and $\sum_i E\, U_i^2 = 1$,

$$P(\sum_i U_i \geq \tau) \leq \underbrace{\exp\left(\frac{\tau}{M} - \left(\frac{\tau}{M} + \frac{1}{M^2}\right)\log(1 + M\tau)\right)}$$

$$P(\sum_i U_i \geq \tau) \leq \exp(-\tau^2/2)$$

If $M\,\tau$ is small, $\log(1 + M\tau) \approx M\tau - M^2\tau^2/2$

Allows heteroscedastic data

Free to divy up the variances as you choose, albeit only for bounded random variables.

In regression (Foster & Stine, 2004, *JASA*)

Conditional on prior $q - 1$ variables,

$$\hat{\beta}_{q,q} = \sum h_i e_{q-1}$$

a weighted sum of **prior residuals**.

In stylized example assigns p-value $\approx 1/100$.

Honest p-values for Testing $H_0 : \beta = 0$

Three methods

Each method makes some assumption about the data in order to produce reliable p-value.

Method	Requires
White estimator	Symmetry
Bennett bounds	Bounded Y
Robust estimator	**Homoscedastic**

Concerns

Avoid "false positive" that leads to inaccurate predictions, cascade of mistakes. Selected features genuinely improve accuracy of fit.

Robustness of validity

Rank regression would also protect you to some extent, but has problems dealing with extreme heteroscedasticity.

Bounded influence estimators

Another possibility, but can it be computed fast enough?

Calibration

Low-hanging fruit

Model is calibrated if

$$E\left(Y \mid \widehat{Y}\right) = \widehat{Y}$$

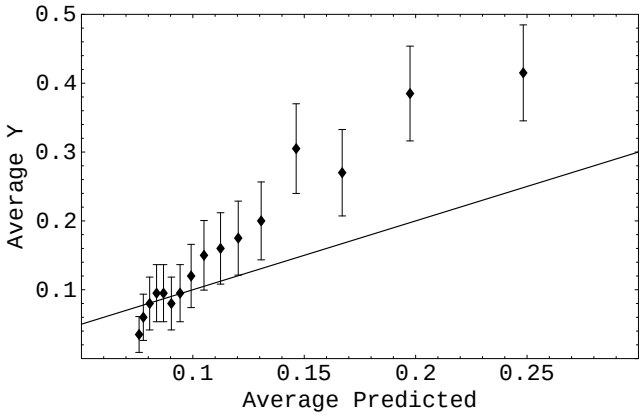
Obtain with scatterplot smoothing

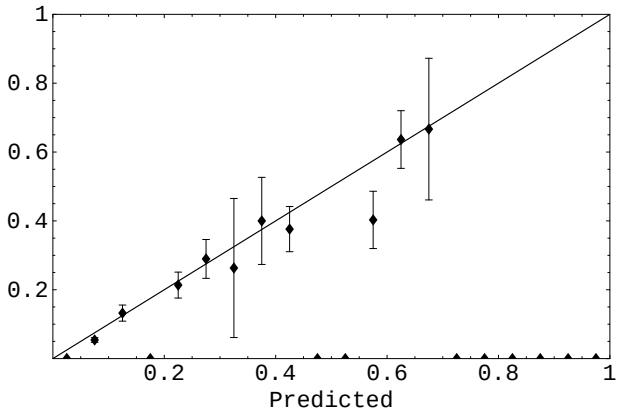
Find smooth/monotone function (inverse link) so that

$$E\left(Y \mid \widehat{Y}\right) = h(\widehat{Y}) = \widehat{\widehat{Y}}$$

Example in bankruptcy application

We used simple implementation of pool-adjacent-violators to obtain satisfactory results.





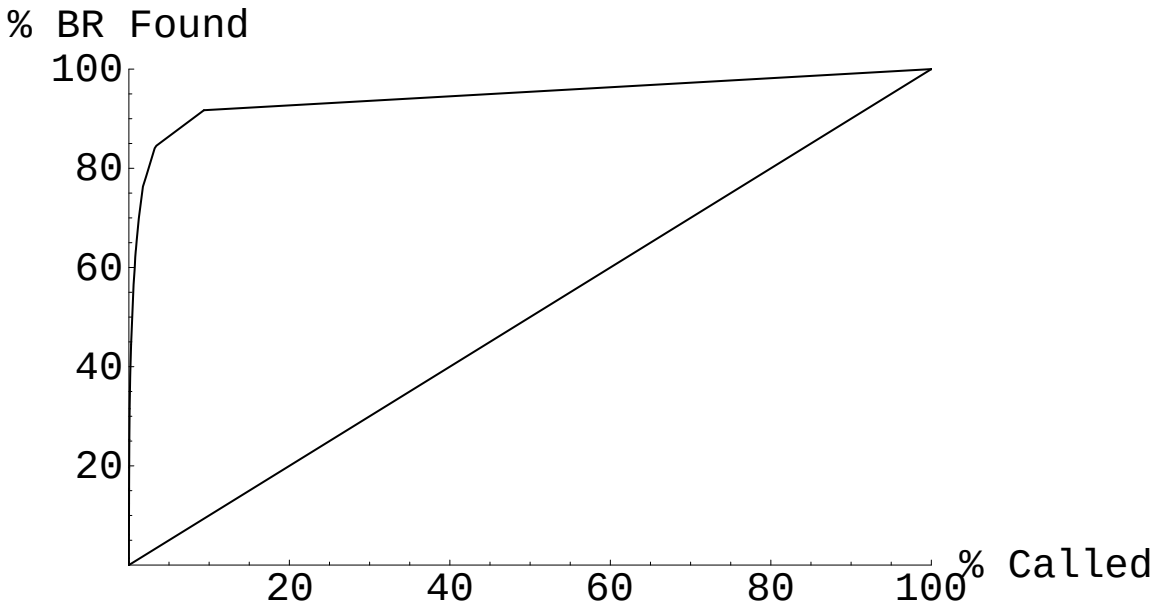
Ideally ... Iterative, "self-consistent" search.

For example, model with many interactions might not be needed with a logistic link that describes multiplicative structure more simply.

Classification Results

Lots of lift

Not only did the model generate smaller costs than C4.5 (w/wo boosting), it also had huge lift:



But...

- Most predictors were interactions.
- Slowwwwww.
- Know that you missed other things.

Who to blame?

Attribute many problems to the greedy, "breadth-first" search of the space of predictors.

Alternative?

Sequential Selection

Search features in order rather than all at once.

Goal

Better predictions...

- Incorporate substantive knowledge
- Sequential selection of features based on current status
- Open vs. closed view of space of predictors
- Run faster

Heuristics

- Theory from adaptive selection suggests that if you can order predictors, then little to be gained from knowing β.
- Alpha spending rules in clinical trials.
- Depth-first rather than breath-first search.

Multiple Hypothesis Testing

How to test a finite collection of hypotheses?

Notation

m null hypotheses $\{H_1, \ldots, H_m\}$

Test results

$$R_j = 1 \text{ if reject } H_j, \text{ 0 otherwise}$$

and

$$V_j = 1 \text{ if } \textbf{falsely} \text{ reject } H_j, \text{ 0 otherwise}$$

Accumulated counts

$$R(m) = \sum_{j=1}^{m} R_j \quad V(m) = \sum_{j=1}^{m} V_j$$

False discovery rate (criterion)

Proportion of false positives among rejects

$$\text{FDR}(m) = E\left(\frac{V(m)}{R(m)} \mid R(m) > 0\right) P(R(m) > 0) \, .$$

Step-up testing (procedure)

Order p-values of **independent** tests

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$$

Reject $H_{(1)}, \ldots, H_{(j^*)}$ for $j^* = \max\{j : p_{(j)} < \alpha \, j/m\}$.

Benjamini & Hochberg (1995, *JRSSB*) show that this procedure satisfies $FDR(m) < \alpha$.

Sequential Discovery Rate

How to test an infinite sequence of hypotheses?

Same notation

$R(m)$ total rejections with $V(m)$ false positives.

Sequential discovery rate (criterion)

When testing a **sequence** of null hypotheses, difference in counts rather than ratio

$$\text{SDR}(m) = E\left[V(m) - \alpha\, R(m)\right]$$

and for all stopping times M,

$$\text{SDR} = \sup_{M}\ E\left[V(M) - \alpha\, R(M)\right]\ .$$

Alpha-investing rule (procedure)

Start with an initial "wealth" $W(0)$. The rule "invests" this allowance for Type 1 errors.

To test H_j ...

Before test, rule announces α-level $\alpha_j \leq W(j-1)$.

After test, wealth is

$$W(j) = W(j-1) + \begin{cases} -\alpha_j & , \quad p_j > \alpha_j \\ \omega - p_j & , \quad p_j \leq \alpha_j \end{cases}$$

If the test rejects, the rule "earns" payout ω for future Type 1 errors. Otherwise, its wealth decreases.

SDR, continued

Theorem

If the payoff for rejecting a null $\omega < \alpha/2$, then this procedure meets the SDR criterion.

Generality

The tests need not be independent, just conditionally correct in the sense that

$$E\left(V_j | R_1, \ldots, R_{j-1}\right) = \alpha_j \quad \text{under } H_j$$

Proof of theorem by showing super-martingale.

Examples

FDR setting: Testing a fixed set of m hypotheses with m independent p-values $p_1,\ p_2, \ldots, p_m$, $p_{j^*} < \alpha\, j^*/m$.

Knowledgeable:

Order hypotheses in increasing p-value. Earns $j^*\ \alpha/2$ rejecting those that FDR rejects, so has "money to burn" for more.

Random order:

Start testing at Bonferroni level α/m.

Expect to find $p_{(1)} < \alpha/m$ after $m/2$ tests.

"Spend" $\alpha/m \times m/2 = \alpha/2$, "earn" $\alpha/2$. So, break-even.

Rejects those FDR rejects.

Auction Strategies

Proliferation of interactions

So many among the choices, so inevitable to find them.

Sequential search Gradually search interactions, with more emphasis to main effects.

- Start with raw variables.
- Once select X_1 and X_2, try interaction $X_1 * X_2$.

Multiple strategies

Use an **auction** to select the strategy.

Two strategies, one for X's and second for interactions.

Each starts with wealth $\alpha/2$.

First strategy can wager more because it has to explore fewer possible effects.

Generalizes to many strategies. Strategy that finds the most significant effects dominates choices.

Computing

$$\text{Strategies} \iff \text{Auction} \iff \text{Model}$$

Each step very fast since basically a simple regression/one-dimensional calculation.

Discussion

Adaptive variable selection

Powerful technique, strong theoretical basis

- Crucial role of standard error estimates
- Scales well to large predictor sets
- Fast

Strategies

Key becomes strategies for generating features.

- Substantive expert can order features (chemist)
- Allows narrow search of very-high order interactions
- Compete with SVM, others from machine learning