

# Applications of Adaptive Variable Selection in Credit Modeling

Bob Stine & Dean Foster

Department of Statistics, The Wharton School

University of Pennsylvania, Philadelphia PA

[www-stat.wharton.upenn.edu/~bob](http://www-stat.wharton.upenn.edu/~bob)

Edinburgh September 5, 2001

- Predict credit behavior (bankruptcy)
- Automated model construction
- Weighting: economics vs. efficiency.
- Identifying significant factors
  - Large pool of predictors (100,000)
  - Adaptive techniques
  - Variance estimation

# Predicting Bankruptcy

## Goal

Predictive model for personal bankruptcy...

- Monthly history of many accounts.
- Estimate probability of bankruptcy during the next billing cycle.

## Data

- “Large” data set: 250,000 bank-card accounts
- About 350 basic features
  - Credit limits, spend, payments, bureau info
  - Demographic background
  - Interactions are important

**100,000 predictors???**

## Bankruptcy is rare

2,244 bankruptcies in

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

## Trade-off

Ideal customers

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

# Which Accounts Matter?

## Whom to contact?

Need to use resources most efficiently.

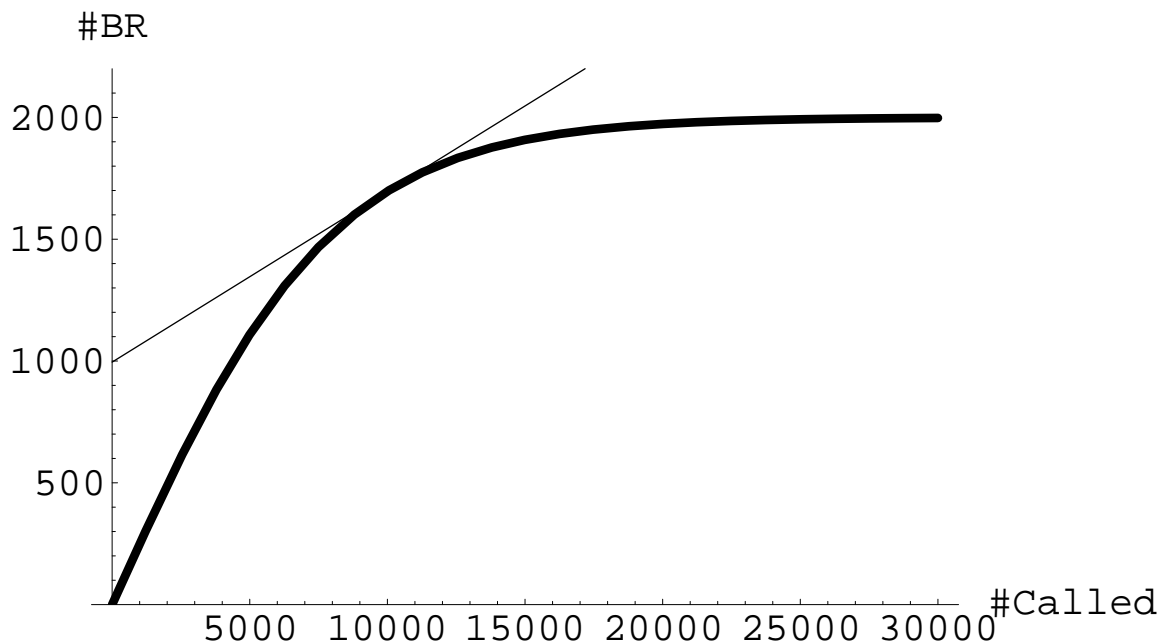
## Lift chart

Slope of tangent line in lift chart

$$\frac{\# \text{ Bankrupt Found}}{\text{Customer Called}}$$

## Which accounts are interesting?

- Extremes on left and right are “easy”.
- Those near break-even point are interesting.



# Logistic Regression

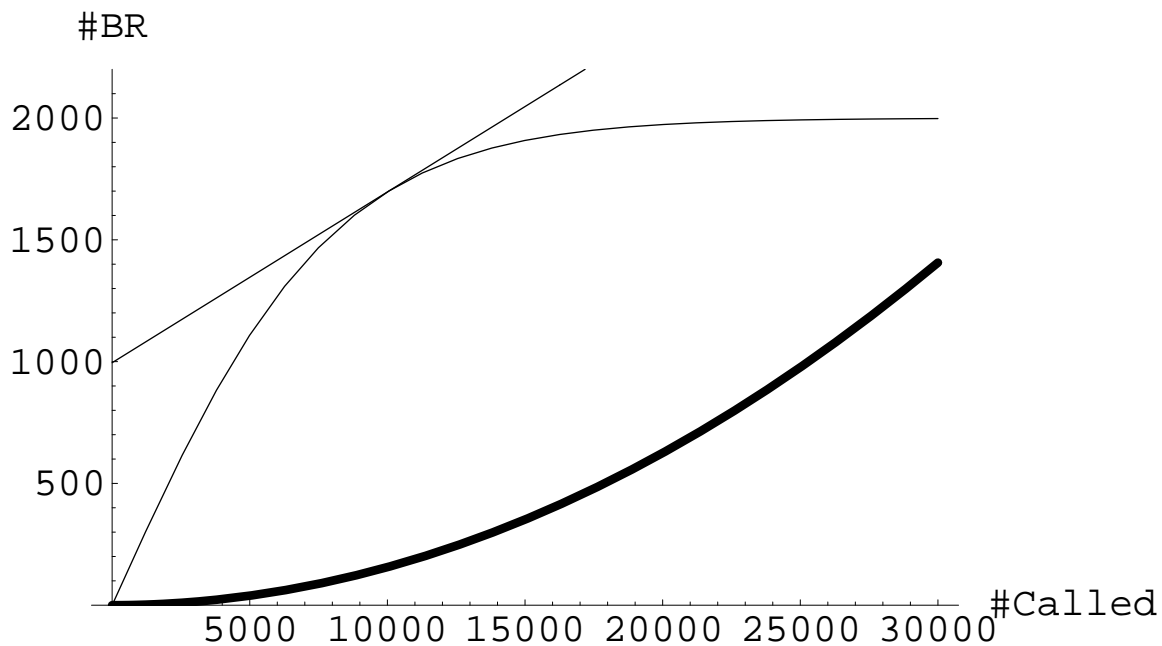
## Logistic regression = WLS

Estimate probability of bankruptcy  $p_i(\hat{\beta})$  in order to minimize weighted sum of squares

$$L(\hat{\beta}) = \sum_i \frac{(y_i - p_i(\hat{\beta}))^2}{p_i(\hat{\beta})(1 - p_i(\hat{\beta}))}$$

**Places most weight** on accounts with near zero chance for bankruptcy.

⇒ These are already most prevalent accounts.



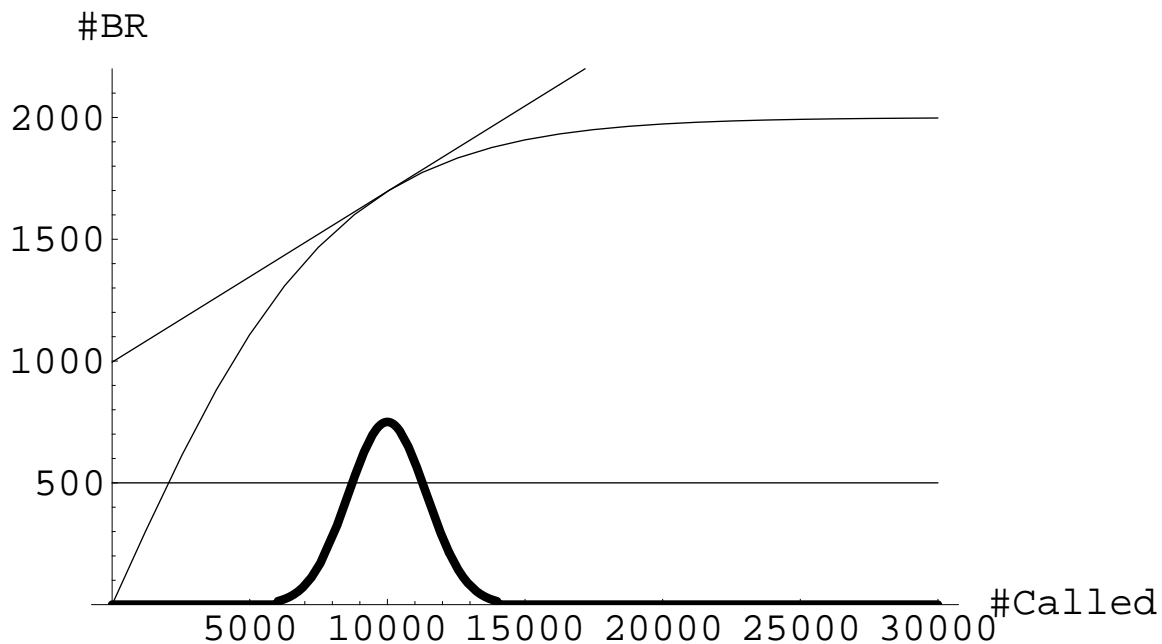
# Economic Weighting

**Economics** suggests putting more weight on accounts near the break-even point.

## Least squares

$$L(\hat{\beta}) = \sum_i \frac{(y_i - p_i(\hat{\beta}))^2}{\mathbf{1}}$$

- Does not require knowledge of where to locate the break-even point.
- Sacrifice efficiency *under utopian assumptions* (e.g., known model) for chance to find signal.



# Choices in Regression

## Structure — What type of model?

- Linear model
- Recent alternatives
  - Additive models
  - Neural nets and projection pursuit
  - Regression trees, piecewise fits (CART)

## Identification — Which predictors to use?

- Informed combinations e.g. *limit – curr balance*
- Time lags, other transformations e.g. logs, ratios
- Interactions “less-informed” combinations

## Search — How do you find them?

- Cannot try all possible solutions.
- We rely on greedy, brute force.
- Getting cheaper!

# Variable Selection

## Context

- $p$  potential predictors,  $n$  observations
- $q$  predictors in fitted model

$$\hat{Y}(q) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_q X_q$$

## Akaike Information Criterion – AIC

- Unbiased estimate of out-of-sample pred. MSE
- Threshold in an orthogonal regression

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

Picks too many predictors: 16% when no signal

- Originated in sequential comparison:  
Identify the order of an autoregression where one considers few models.
- Problem occurs because

$$\min(\text{unbiased estimates}) \neq \text{unbiased}$$

## $C_p$ , leave-one-out cross-validation

$AIC = C_p$  for large sample sizes.

# Variable Selection, *cntd*

## Bayesian Information Criterion – BIC

- Estimates the Bayes factor, ratio of posterior probabilities of different models  $\Pr\{H_k\}$
- Threshold in an orthogonal regression

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

- Parsimonious if  $n \gg p$ , promiscuous if  $n \ll p$ .

## Consistent model selection

- Suppose there is a “true” model for all  $n$ .
- Criterion must find this model with infinite data.
- *BIC* is consistent (*AIC* overfits).
- Note: Standard tests are not consistent in this sense (e.g., 5% error rate).

## True model?

Would you fit the same model with 1,000,000 observations as with 100?

# Variable Selection, *cntd*

## Minimax variable selection

- Which predictors min maximum prediction MSE?

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

- Answer: Pick them all! Not helpful.

## Hard thresholding

- Which predictors min *ratio* of prediction MSEs?

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

- Answer: (D&J, F&G 1994)

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

## Heuristic

- It's almost Bonferroni! ( $\sqrt{2 \log p}$  is a bit less strict)
- If  $Z_1, \dots, Z_p \sim N(0, 1)$  then

$$\Pr \{ \max(|Z_1|, \dots, |Z_p|) > \sqrt{2 \log p} \} \rightarrow 0.$$

# Variable Selection Criteria

In an *orthogonal* linear regression:

Criterion	Threshold $z$	Origin
Least squares	0	Gauss, minimax
$AIC$	$\sqrt{2}$	Unbiased PMSE (Akaike 73)
$C_p$	$\sqrt{2}$	Pred error (Mallows 73)
$BIC$	$\sqrt{\log n}$	Bayes (Schwarz 78)
$MDL$	$\sqrt{\log n}$	Coding (Rissanen 83)
$RIC$	$\sqrt{2 \log p}$	Relative risk (F & G 94)
hard threshold	$\sqrt{2 \log p}$	Minimax (D & J 94)

**Which one to use?**

# Adaptive Criteria

**Prior criteria** tuned for different problems:

Method	Optimized for
<i>AIC</i>	$p/2$ small slopes
<i>BIC</i>	Large slopes
<i>RIC</i>	1 significant slope

**Why pick one?**

**Adaptive solution** (Foster and Stine, 1997)

Information theory motivates the rule

$$\text{Add } X_j \iff |t_j| > \sqrt{2 \log p/q}$$

*Prediction error* of resulting model is within a constant factor of the best Bayesian model.

**Other paths to adaptive rules**

- Empirical Bayes
- Half-normal plots
- Simes method, step-up testing

# Discussion of Adaptive Methods

## Sources of prediction error

- Include an extraneous predictor
- Omit a useful predictor
- Random estimation error

## Weakness of “Bonferroni”

Prediction error dominated by omitting predictors.

## Adaptive variable selection

- Bonferroni unpopular because of low power.
- Simes method – step-up/step-down tests:

$$|t_{(1)}| \geq |t_{(2)}| \geq \cdots \geq |t_{(p)}|$$

1. Compare  $t_{(1)}$  to  $\sqrt{2 \log p}$
2. Compare  $t_{(2)}$  to  $\sqrt{2 \log p/2}$
3. ... Compare  $t_{(q)}$  to  $\sqrt{2 \log p/q}$

## Idea

Once you find one variable, easier to add more.

# Predicting Personal Bankruptcy

## Goal

Identify accounts at “high” risk.

$$\Pr(\text{Bankrupt next month}) > 0.05 \quad (\textit{e.g.})$$

## Data

- Records for  $n = 250,000$  card holders
- Demographics (e.g., location, home ownership)
- Year of longitudinal data
  - Some monthly, others quarterly and annual
- Derived data
  - Interactions (regional differences, nonlinear)
  - Missing data

## Linear model

Consider selecting from (to begin!)

$$p = \mathbf{67,000} \quad \text{candidate predictors}$$

including interactions and missing indicators.

## Needle in the haystack

Bankruptcy is a rare event in our data:

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

# Validation and Estimation Samples

## Cross-validation

Split sample into two parts

- 20% for estimation ( $n \approx 600,000$ )  
458 bankruptcy events
- 80% for validation ( $n \approx 2,400,000$ )  
1,786 bankruptcy events

## Estimation

- Over-sample  
Use all 458 bankrupt cases, but only 2.5% of the rest, for an estimation sample of about 15,000.
- Adaptive selection  
Compare sequence of  $|t_j|$  to  $\sqrt{2 \log p/q}$ .

## Test of procedure

Predict the validation sample.

## Do you need a validation sample?

- To pick the model? No.
- To estimate the prediction MSE? Yes.

# Summary of Current Procedure

## Least squares estimator

Predictors chosen to minimize economically-motivated loss function

$$L(\hat{\beta}) = \sum_i (y_i - p_i(\hat{\beta}))^2$$

## Robust standard error

- Recognize heteroscedasticity when estimating SE.
- Could use Bernoulli weights  $\hat{p}_i(1 - \hat{p}_i)$
- Examples use non-parametric alternative

## Search procedure

Stepwise forward selection:

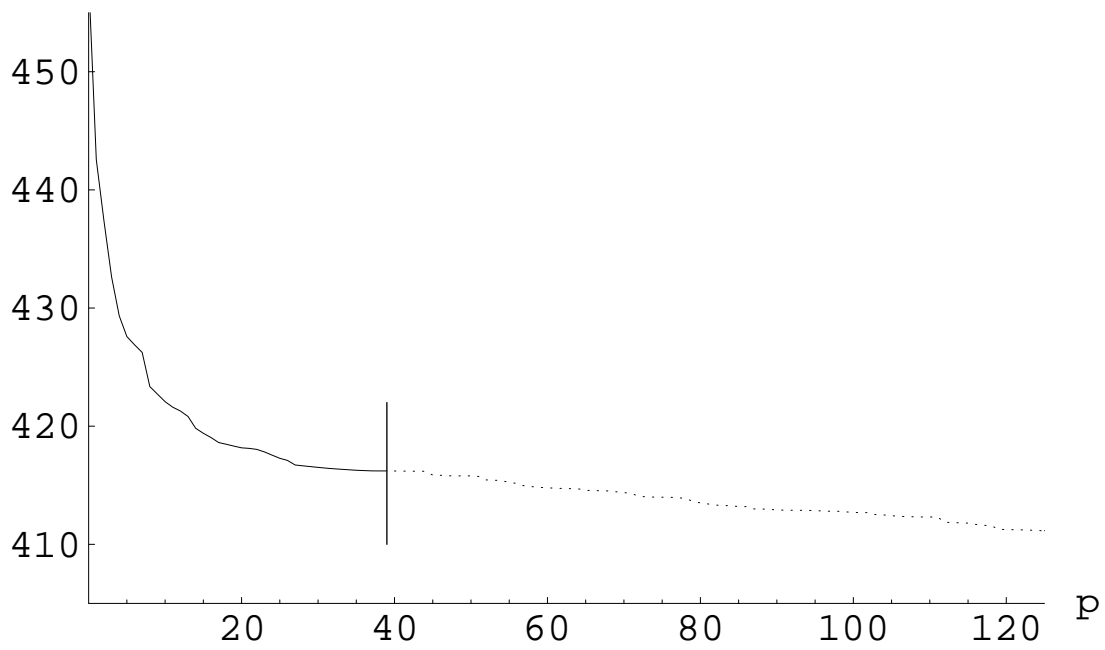
- Sort omitted predictors by change in residual SS and note the associated  $t$  ratio.
- Add variable with most explanatory power **if**  $|t|$  exceeds adaptive threshold.
- Use conservative SE based on current, not updated, residuals.

$$\text{Var}(\hat{\beta}_w) = (X'_k W X_k)^{-1} (X'_k W E_{k-1}^2 W X_k) (X'_k W X_k)^{-1}$$

# In-sample Results

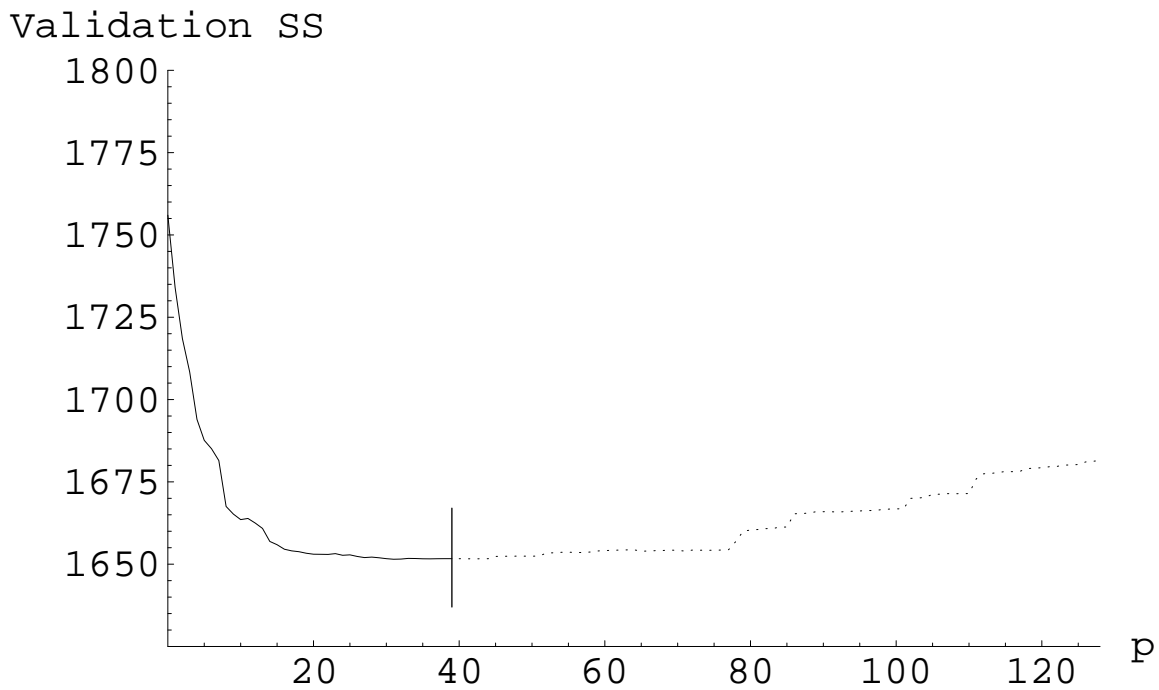
Sum of squared residuals, in-sample

Residual SS



# Validation Results

Sum of squared pred errors, validation sample

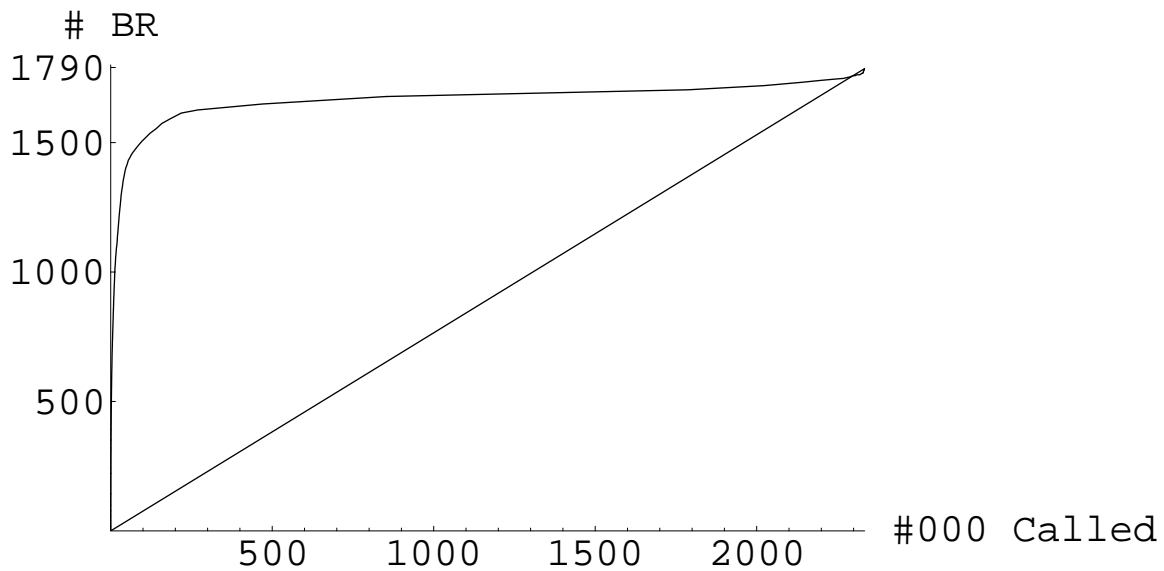


**Selected predictors improve accuracy.**

**Procedure does not overfit.**

# Validation Results

## “Lift” chart



**351 of bankruptcies in 999 largest predictions.  
60% of bankruptcies in largest 1% of predictions.**

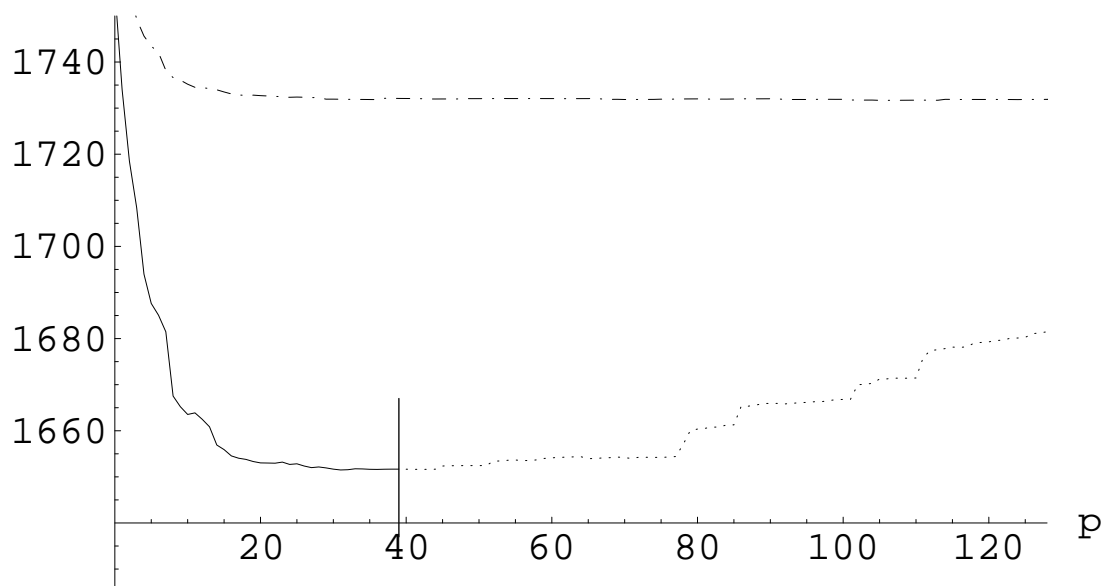
# Validation: Linear Model

Sum of squared pred errors, validation sample

## Linear model

Select predictors *without* considering possible interactions among the features.

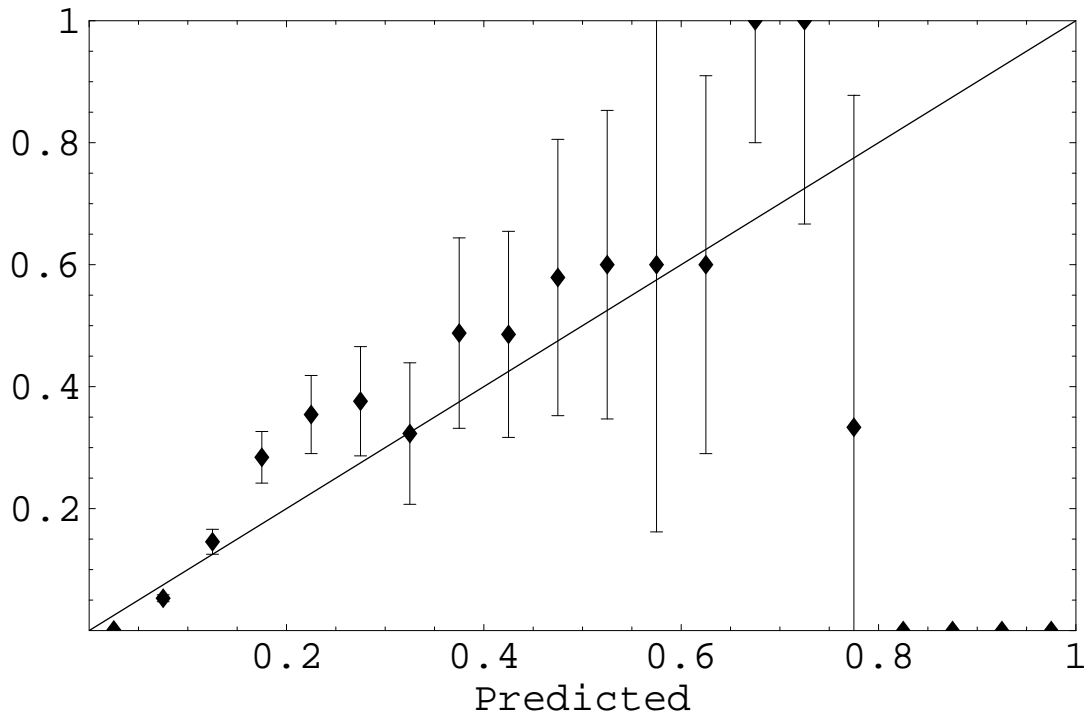
Validation SS



**Linear model cannot achieve comparable accuracy.**

# Validation Results

## Calibration chart



Some calibration error near  $\Pr(\text{bankrupt}) \approx 0.2$ .  
Further nonlinearity?

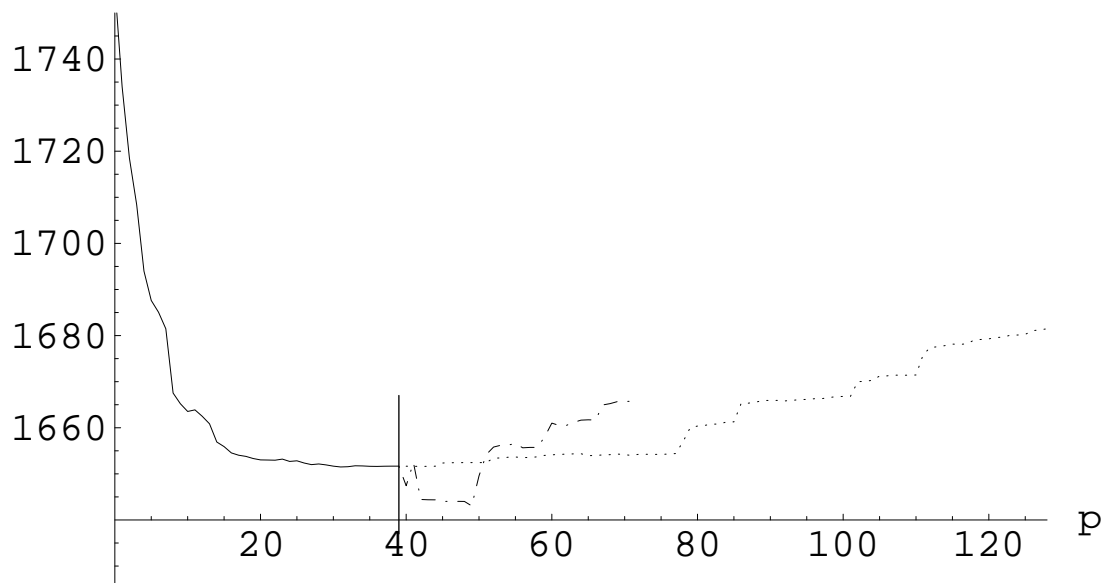
# Validation: Higher-Order Model

Sum of squared pred errors, validation sample

## Higher-order model

Expand search from 39-predictor model by considering *interactions* of these 39 with other features

Validation SS

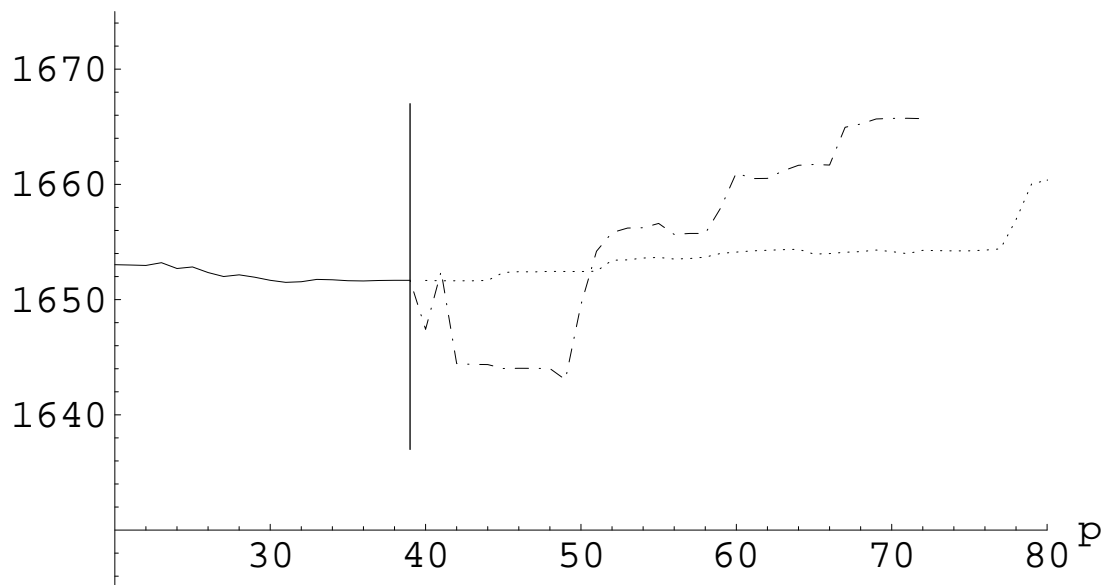


**Added features significantly improve accuracy.**

# Zoom in on Higher-Order Model

Sum of squared pred errors, validation sample

Validation SS



**Added features significantly improve accuracy.**

# Conclusions

## Adaptive variable selection

- Powerful technique, strong theoretical basis
- Crucial role of standard error estimates
- Significant terms lower validation SS.
- Other successful applications
  - Other credit modeling
  - Clinical data

## Implications for practice

- Automatic search as a baseline
- Supplement to “manual” analysis

## Next steps

- Formalize “expert knowledge”
  - Leverage ordering of predictors
- Dynamically expand interactions set
  - Interact found predictor with excluded.
- Categorize predictors
- Estimation from expanding data set

# Appendix: Over-sampling

## Over-sample bankrupt events

- Use *all* bankrupt events, 2.5% of rest.
- Unlike logistic regression, linear regression slopes are *biased unless we adjust* for sampling wts  $w_i$ .

## Weighted least squares estimator

$$\hat{\beta}_w = (X'WX)^{-1}X'WY, \quad W = \text{diag}(w_i)$$

## Standard error

$$\begin{aligned} \text{Var}(\hat{\beta}_w) &= (X'WX)^{-1} (X'W \text{Var}(Y)WX) (X'WX)^{-1} \\ &= \sigma^2(X'WX)^{-1} \end{aligned}$$

**IF**

$$\text{Var}(Y) = \sigma^2W^{-1}$$

which is not likely since  $w_i$  are sampling weights.

## Homoscedastic case

Assuming constant variance  $\sigma^2$ , left with

$$\text{Var}(\hat{\beta}_w) = \sigma^2(X'WX)^{-1} (X'W^2X) (X'WX)^{-1}$$

which greatly complicates the *search* for predictors.

# Appendix: Sparse Response

## Discrete data

- Response  $Y$  is 0/1 indicator with most  $Y = 0$ .
- Many predictors are also 0/1:  
Indicators, missing data, interactions

## Stylized testing problem (assume $n_0 \gg n_1$ )

$$n_0 : Y_{0i} = 0 \text{ at } X = 0 \quad n_1 : Y_{1i} \sim N(0, 1) \text{ at } X = 1$$

## Correct test for mean shift One-sample t

$$t_1 = \frac{\sqrt{n_1} \bar{Y}_1}{s_1}$$

## Two-sample t test *assuming* homoscedastic

$$t_2 = \frac{\bar{Y}_1 - \bar{Y}_0}{s \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}} = t_1 \times \frac{\sqrt{n_0}}{\sqrt{n_1}}$$

and test statistic is inflated since  $n_0 \gg n_1$