Variable 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 15, 2000

- Methods that automate variable selection
  - What they are.
  - How they work.
  - What they achieve.
- Application in credit modeling
  - Over-sampling
  - Sparse data with rare events
  - Heteroscedasticity
- Further challenges and applications

Credit Models

Goal

For an individual credit-card holder, predict

- probability of bankruptcy
- credit risk, profitability

Data

For each account, have (ordered by "access" cost)

- Current account status (spend, balance, bureau)
- Demographic background (e.g., application)
- Historical series (some are incomplete)
- Transaction data
- Activity in other credit lines

Model construction

- Pick type of "regression" model:
  linear, logistic, neural nets, CART, MARS, ...
- Challenge: Identify predictors
  - Acquired expertise (Hoadley's talk last year)
  - Exploratory analysis
  - Automatic procedures
    - Iteractions
    - Nonlinearity
    - Subsets

Predicting Personal Bankruptcy

Goal

Identify customers at "high" risk of declaring bankruptcy.

"High" might mean Pr{Bankrupt next month} = 0.10.

Data

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

Linear model

Consider selecting from

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

including interactions (nonlinear) and group indicators.

Needle in the haystack

Bankruptcy is a rare event in our data:

2,500 events in 6,000,000 months of data

Traditional Variable Selection Criteria

Context

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

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

Akaike Information Criterion – AIC

- Unbiased estimate of the prediction MSE
- Identify the order of an autoregression (sequential)
- Selects X_j if (orthogonal regression)

$$|t_j| > \sqrt{2}$$

- Picks many predictors: 16% when no signal

C_p, Cross-validation

Equivalent to AIC for large sample sizes.

Bayesian Information Criterion – BIC

- Estimates the Bayes factor, ratio of posterior $\Pr\{H_0\}$
- Selects X_j if (orthogonal regression)

$$|t_j| > \sqrt{\log n}$$

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

Hard Thresholding and Bonferroni

Minimax variable selection

- Which predictors minimize maximum prediction MSE?

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

- Answer: (disappointing) Constant risk — pick them all!

Competitive analysis

- Which predictors minimize *ratio* of prediction 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)

$$\text{Pick predictors whose } |t_j| > \sqrt{2 \log p}$$

Heuristic for hard thresholding

- It's almost Bonferroni! ($\sqrt{2 \log p}$ is a bit more strict)
- Fisher's (1927) test for the max of periodogram.
- If have a sample of $X_1, \ldots, X_p \sim N(0,1)$ then

$$\Pr\left\{\max(|X_1|, \ldots, |X_p|) > \sqrt{2 \log p}\right\} \to 0.$$

Adaptive Variable Selection

Sources of prediction error

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

Weakness of "Bonferroni"

Includes too few predictors: prediction error dominated by omitting useful predictors.

Adaptive variable selection (a.k.a. multiple testing)

- 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 $\sqrt{2 \log p/q}$

$\Rightarrow$ Once you find one variable, easier to add more.

- Related to empirical Bayes and info theory

Prediction error of adaptive model is within a factor of the prediction error of "expert" model that knows the true β_j, but not the coordinates!

Variable Selection in Bankruptcy Model

Setting

- *Sample* $n = 16,000$ customer–months.
- *Select* from $p = 67,000$ variables.
- *Validate* with sample of 48,000 records.
- *Over-sample* BR cases since so few.

Key property of "good" selection criterion

- Validate prediction error using hold-out sample.

- Why not just use cross-validation all the time?
  $\Rightarrow$ Use all 2,500 bankrupt events for fitting.

Results for Thresholding

Results shown last year

Bonferroni picks too many, and adaptive picks even more!

How can this happen?

Explanations?

Adaptive thresholding

Pick X_j if

$$|t_j = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}| > \sqrt{2 \log p/q}$$

Distribution of the t_j's evidently is fatter than expected when the true coefficient is zero.

Assumptions

- Independence
- Constant variance
- Normality

Collinearity *reduces* probability of exceeding the bounds.

Standard error

- Sampling weights
- Intrinsic heteroscedasticity

Effects of Sample Weights

Over-sample bankrupt events

- Use *all* bankrupt events in analysis, but not all of the other data.
- Unlike logistic regression, linear regression slopes are *biased unless adjust* for sampling.

Weighted least squares estimator

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

Standard error

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

Last step holds only when

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

which is not likely since w_i are sampling weights.

Homoscedastic case

Assuming constant variance, 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.

Dare We Assume Constant Variance?

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 $n_0 \gg n_1$)

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

Correct test for mean shift One-sample t

$$t_1 = \frac{\sqrt{n_1}\,\overline{Y_1}}{s_1}$$

Two-sample t test *assuming* homoscedastic

$$t_2 = \frac{\overline{Y_1} - \overline{Y_2}}{s\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}} = t_1 \times \frac{\sqrt{n_0}}{\sqrt{n_1}}$$

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

'Robust' Variance Estimate

White's estimator

$$\mathrm{Var}(\hat{\beta}_w) = (X'WX)^{-1}\,(X'W\ \underbrace{\mathrm{Var}(Y)}\ WX)\,(X'WX)^{-1}$$

by the squared residuals from fitted model

$$E = \mathrm{diag}(e_i), \quad e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \cdots - \hat{\beta}_q X_q$$

Obtain SE from

$$\mathrm{var}(\hat{\beta}_w) = (X'WX)^{-1}\,(X'W\ E^2\ WX)\,(X'WX)^{-1}$$

(Or, use the binomial variance form in this special case)

Further complication Rows of intermediate output

Term	Estimate	Homo SE	t	Hetero SE	t
$X_{20} * X_{317}$	0.9992	0.01970	51	0.00003	29857
$X_{25} * X_{348}$	0.9992	0.01970	51	0.00003	29857

What happened?

$Y = 1$	300	2
$Y = 1$	15000	0
	$X = 0$	$X = 1$

Rate 1/30

Search method? How to do stepwise now?

Stylized Testing Problem

Stylized problem

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

Standard two-sample test

Inflated t-statistic since data lacks constant variance and observe many more in group at $X = 0$.

One-sample test

Fails for *sparse data* when only one value in group at $X = 1$.

Conservative one-sample test

Estimate assuming no signal using conservative estimate of variance

$$t_1' = \frac{\sqrt{n_1}\,\overline{Y_1}}{s_1'}, \qquad s_1'^2 = \frac{\sum_i Y_i^2}{n_1}$$

Current Procedure

Estimator

Use sampling weights, but not heteroscedastic weights:

- WLS down-weights observations with large variance. which tends to occur for "large" $\hat{Y} > 0.25$, say
- WLS down-weights observations of most interest.

Standard error

Recognize the heteroscedasticity when estimating SE.

Search procedure

Stepwise search, with adjustment for survey weight to get correct forward selection...

- Sort omitted predictors by change in residual SS
- Add variable with most explanatory power **if** "look ahead" SE is significant using adaptive threshold.
- Estimate using current, not updated residuals. i.e., to evaluate at step k, estimate SE using residuals from step $k-1$.

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

- Bonus: easier to compute look-ahead selection step.

Results with Bankruptcy Model

Setting

- Select a sample of $n = 16,000$ customer–months.
- Stepwise regression from $p = 67,000$ possible X's.
- Validate with independent sample of 48,000.

Validation error

Plot shows the out-of-sample prediction error, and the vertical line locates the adaptive rule cut-off.

Interpreting Bankruptcy Model

Coefficients from fitted model

Arranged by order of variable indices reveals same "base terms" appearing in selected terms:

Term	Estimate	Homo SE	t	Hetero SE	t
$X_{59} * X_{263}$	0.002063	0.00022	9.20	0.00036	5.66
$X_{59} * X_{292}$	0.000460	0.00004	10.99	0.00012	3.89
$X_{59} * X_{284}$	0.000275	0.00003	10.07	0.00006	4.80
$X_{215} * X_{292}$	0.000021	0.00000	13.41	0.00000	6.15
$X_{284} * X_{292}$	0.002664	0.00011	25.30	0.00058	4.62
$X_{292} * X_{298}$	-0.007760	0.00066	-11.77	0.00291	-2.67

"Interpretation" of model

- Combination of terms suggests *multiplicative* model.
- Logistic regression requires only base terms – the interactions are no longer needed!

Discussion

Adaptive variable selection

- Powerful technique with strong claims
- Crucial role of standard error estimates
- Adaptive cut-off coincides with turning point of CVSS

Key aspects of modeling

- Selection of model structure (linear vs logit)
- Identification of predictors
- Validation of model accuracy
- Search methods and computing

Implications for practice

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

Next steps Better searching ...

- Improved backward elimination
- Searching for other interactions (detecting multiplicative)
- Logistic regression as base model.

Effects of Biased Variance Estimate

Assume small bias

Suppose that estimated error variance is too small

$$\hat{\sigma}^2 = (1 - \delta)\sigma^2$$

Hard thresholding picks too many

- Use threshold

$$|t| > \sqrt{2(1 - \delta)\log p}$$

- Number chosen grows geometrically

$$\sqrt{2(1 - \delta)\log p} = \sqrt{2\log \underbrace{p^{1-\delta}}}$$

- Pick fraction

$$\frac{1}{p^{1-\delta}} = p^{\delta}\ \frac{1}{p}$$

- **PLOT** of the number chosen as function of δ

Adaptive picks even more

Cascading effect as picks more and more predictors.

Sources of bias and solutions

- Selection bias:

$$\min\{\text{unbiased estimates}\} = \text{biased}$$

- Honest s^2 methods.

Simulation Results

"Stepwise" in practice

Allow extra terms into the model (low p-to-enter) to find interesting structure, then remove with backward elimination.

Simulation

Simulation results for hard/adaptive thresholding in situations that do/do not lead to biased variance estimates.