Variable Selection in Data Mining:
Building a Predictive Model for Bankruptcy

Dean P. Foster and Robert A. Stine [*]

Department of Statistics

The Wharton School of the University of Pennsylvania

Philadelphia, PA 19104-6302

November, 2001

Abstract

We develop and illustrate a methodology for fitting models to large, complex data sets. The methodology uses standard regression techniques that make few assumptions about the structure of the data. We accomplish this with three small modifications to stepwise regression: (1) We add interactions to capture non-linearities and indicator functions to capture missing values; (2) We exploit modern decision theoretic variable selection criteria; and (3) We estimate standard error using a conservative approach that works for heteroscedastic data. Omitting any one of these modifications leads to poor performance.

We illustrate our methodology by predicting the onset of personal bankruptcy among users of credit cards. This application presents many challenges, ranging from the rare frequency of bankruptcy to the size of the available database. Only 2,244 bankruptcy events appear among some 3 million months of customer activity. To predict these, we begin with 255 features to which we add missing value indicators and pairwise interactions that expand to a set of over 67,000 potential predictors. From

[*] Research supported by the Wharton Financial Institutions Center

these, our method selects a model with 39 predictors chosen by sequentially comparing estimates of their significance to a sequence of thresholds. The resulting model not only avoids over-fitting the data, it also predicts well out of sample. To find half of the 1800 bankruptcies hidden in a validation sample of 2.3 million observations, one need only search the 8500 cases having the largest model predictions.

Key Phrases: AIC, C_p, risk inflation criterion *RIC*, hard thresholding, stepwise regression, Bonferroni, step-up testing.

1 Introduction

Large data sets and inexpensive computing are symbols of our time. The modelers of the data-mining community have enthusiastically embraced this combination. This pragmatic, energetic community eagerly adopts and customizes computational methods like neural nets to suit the problem at hand, aiming for predictive tools rather than asymptotic theorems (see Breiman's discussion of **?**). Statisticians have been more cautious, stressing the importance of domain knowledge and careful data analysis while warning of parameter bias and other consequences of over-fitting (e.g. **?**). Many interesting problems, particularly classification problems such as diagnosing a disease to identifying profitable customers in a mailing list or at a web site, have become the province of "data mining" rather than applications of statistics (e.g., see **?** and the November, 1999, issue of *Communications of the ACM* (Volume 42, Number 11) which offers several articles on knowledge discovery and machine learning).

Our purpose here is to show that a combination of statistical methods routinely handles "data-mining" problems quite well. Our key tool is familiar to all statisticians who model data: stepwise regression. What distinguishes our use of this familiar, often maligned tool is the expansive way that we use it. For our illustration, we fit a stepwise regression beginning with over 67,000 candidate predictors. To show that it works, we use the chosen model to predict a large validation sample. Like the data miners, our goal is out-of-sample predictive accuracy rather than the interpretation of specific model parameters.

Our use of stepwise regression succeeds because it finds real signal while protecting against over-fitting. Stepwise regression is well-known to have problems when fitting models with many parameters, claiming a better fit than it actually obtains (e.g. **??**). To address these real concerns, the last decade in decision theory produced several important techniques that avoid over-fitting. These cannot, however, be routinely used in data analysis. All evaluate the merit of a predictor through a measure of its statistical significance derived from, essentially, a t-statistic, the ratio of the estimate to its standard error. These theoretical results presume accurate standard errors which abound in the theoretical world where one can assume knowledge (or very good estimates) of key features like the noise variance σ^2. For example, simulations illustrating implementations of these results most often consider orthogonal regression, typically a wavelet regression whose "small" coefficients define an estimate of σ^2. In data analysis, however, a small degree of over-fitting produces biased standard errors. The biases may be small, such as a 10% change in the t-statistic from 2.0 to 2.2 or 4.0 to 4.4. Because the theory relies on the thin tail of the normal distribution, such bias with 67,000 estimates can lead to many false positives and produce yet more bias and a spiral of over-fitting.

To illustrate this methodology, we build a model to predict the onset of personal bankruptcy. The credit industry makes extensive use of statistical modeling (**?**), and decision automation systems incorporating statistical models have saved creditors millions of dollars (**?**). Though well-publicized, personal bankruptcy remains relatively rare in the US. Nonetheless, bankruptcy events are expensive to creditors and it would be valuable to anticipate them. The underlying data base for our study holds longitudinal data for 244,000 active credit-card accounts, as described further in Section 2. For the 12 month period considered in this analysis, only 2,244 bankruptcies are scattered among the almost 3,000,000 months of account activity. Though certain aspects of our application are specific to this problem of predicting a rare, discrete event, we believe that the approach has wider applicability. In seeking to demonstrate this generality, we structured the bankruptcy data in a way suited to regression modeling. The longitudinal nature of the data suggests a variety of relatively esoteric models, such as time-to-event or proportional hazards models. Though more specialized models may be more capable of predicting bankruptcy, the simplicity and familiarity of regression makes it more suited as a test case for describing our methodology. In the end, its predictions appear impressive as well.

While our modeling procedure would surely benefit from more expertise in the choice of predictors, the approach taken here is fully automatic. Beginning from a rectangular array of data, the algorithm expands the set predictors, searches these in a familiar stepwise manner, and chooses the predictors expected to generate significant predictive benefit. Put bluntly, we build a stepwise regression from a very large set of predictors expanded to include all pairwise interactions, here a set of 67,160 predictors. The presence of the interactions allows the linear model to capture local curvature and subset behavior, features often associated with more specialized methods (e.g., neural nets and CART). The novelty in our approach lies in how we avoid selection bias – choosing predictors that look good in-sample but predict new data poorly – without assuming the existence of a large hold-out sample. It is well-known that naive variable selection methods are typically overwhelmed by selection bias. For example, in an orthogonal regression, the AIC or C_p criterion chooses predictors whose absolute t-statistic exceeds $\sqrt{2}$. If we imagine a simplified problem in which none of the predictors is useful (all of "true" slopes are zero and the predictors are uncorrelated), then AIC would choose about 16% of them for the model even though none of them represents a useful predictor (e.g., see **?**).

In comparison to rules such as AIC that use a fixed threshold, our variable selection procedure employs a changing threshold. As such, the criterion adapts to problems such as our prediction of bankruptcy with few useful predictors (low "signal to noise" ratio) as well as other problems in which many predictors are valuable (high signal to noise ratio). One need not pretend to know *a priori* whether few or many predictors are useful. The criterion instead adapts to the problem at hand. Deferring the details to Section 3, the procedure yields a set of predictors whose out-of-sample mean squared error is about as good as might be achieved if one knew which variables were the "true" predictors. Our rule selects predictors whose t-statistic exceeds a threshold, much as described for AIC above. The difference from AIC is that the rule first compares the largest coefficients to a very large threshold and then gradually reduces the threshold to accommodate more predictors as significant effects become apparent.

Because of its central role, the standard error of the estimated coefficients must be accurately determined, lest a greedy stepwise procedure such as ours cascade into a spiral of worse and worse models: once the selection procedure incorrectly picks a useless predictor, it becomes easier for it to choose more and more. With an excess of possible predictors, our selection procedure uses conservative estimates of the standard error of the coefficient estimates. In addition to adjusting for sampling weights and natural sources of heteroscedasticity in this problem, a conservative variation on a standard econometric estimator is not only easier to compute in a stepwise setting but also avoids pitfalls introduced by the sparse nature of the response. (See equation (22) for the precise estimator.)

Adaptive variable selection consequently reduces the need to reserve data for validation. Validation or hold-back samples can serve one of two purposes: selecting variables or assessing the accuracy of the fitted model. In the first case, one adds predictors until the out-of-sample error estimated from the validation data set increases. Whereas the in-sample prediction error, for example, inevitably decreases as predictors are added during the optimization of the fitted model, given enough data, the out-of-sample squared error tends to increase once spurious predictors enter the model. One can thus use the validation sample to decide when to halt the selection process. When modeling a rare event such as bankruptcy, however, such a hold-back sample represents a major loss of information that could otherwise be used to find an improved model (see, e.g., Miller's discussion of **?**). Nonetheless, to convey the predictive ability of our model, we reserved a substantial validation sample, 80% of the available data.

Using the validation sample, the lift chart in Figure 1 offers a graphical summary of the predictive ability of the model found by our selection procedure. To motivate this chart, consider the problem of a creditor who wants to target customers at risk of bankruptcy. The creditor knows the bankruptcy status of the people in the estimation sample, and now must predict those in the validation sample. Suppose this creditor decides to contact those most a risk of bankruptcy, in hopes of changing their behavior. Because of budget constraints, the creditor can afford to call, say, only 1000 of the 2.3 million customers in the validation sample. If the creditor selects customers to call at random, a sample of 1000 customers has on average fewer than one bankruptcy. Alternatively, suppose that the creditor sorts the customers in the validation sample by their predicted scores from the fitted model, and then calls the first 1000 in this new list. Table 1 shows the number of observations and bankruptcies in the validation sample with predicted scores above several levels. When ordered by predicted scores, the creditor will find 351 bankruptcies among the first 999 called, almost 20% of the 1786 bankruptcies found in the validation sample. Continuing in this fashion, the creditor could reach about half of the bankruptcies hidden in the validation sample by contacting 8500 customers. Resembling an ROC curve, the lift chart in Figure 1 graphically displays the results in Table 1, varying the level of the predicted score continuously. The horizontal axis shows the proportion of the validation sample called, and the vertical axis shows the proportion of bankruptcies found. The diagonal line in the plot represents the expected performance of random selection. The concave curve

Figure 1: Lift chart for the regression model that uses 39 predictors to predict the onset of personal bankruptcy. The chart shows the percentage of bankrupt customers in the validation data found when the validation observations are sorted by predicted scores. For example, the largest 1% of the predictions holds 60% of the bankruptcies. The diagonal line is the expected performance under a random sorting.

shows the performance when the validation sample is sorted by our fitted model. Its lift chart rises swiftly; for example, calls to about 1% of the population as sorted by the model find 60% of the bankruptcies. For the sake of comparison, we note that the R^2 of the fitted model is only 9%. This traditional summary fails to convey the predictive value of the model.

We have organized the rest of this paper as follows. Section 2 that follows describes in more detail the bankruptcy data featured in our analysis. Section 3 describes our approach to variable selection and Section 4 provides the details to finding the estimates and standard errors needed for this type of modeling. Section 5 gives further details of the fitted model, and the concluding Section 6 offers some topics for discussion and suggests directions for enhancing this type of modeling.

2 Data Processing

This section describes how we constructed the "rectangular" data set used in our regression modeling. As often seems the case in practice, the task of preparing the data for analysis absorbed much of our time, more than the modeling itself. We obtained the data as part of a research project studying bankruptcy at the Wharton Financial Institutions Center, and we acknowledge their support.

The original data obtained from creditors present a longitudinal view of 280,000 credit-card accounts. The time frequency of the data vary; some measurements are monthly (such as spending and payment history), whereas others are quarterly (e.g., credit bureau reports) or annual (internal performance measures). Yet others are fixed demographic characteristics, such as place of residence, that are often gathered as part of the initial credit application. We merged the data from various files using a variety of Unix and lisp scripts into a single, unified data file. Because of changes in the

Table 1: Counts of bankruptcies when observations in the validation sample are sorted by predicted model score. In the validation sample, for example, 25 customers received scores of 0.60 or larger; of these 18 (72%) declared bankruptcy.

Predicted	Number above Level		
Level	Total	Bankrupt	% Bankrupt
1.00	4	3	75.00
0.90	4	3	75.00
0.80	4	3	75.00
0.70	10	7	70.00
0.60	25	18	72.00
0.50	45	30	66.67
0.40	99	58	58.59
0.30	205	99	48.29
0.20	545	222	40.73
0.15	999	351	35.14
0.10	2187	524	23.96
0.05	8432	855	10.14
0.00	852991	1678	0.20
-.05	2333087	1782	0.08

scope of data collection, we focus on a 12 month period during 1996-1997 that permits us to use a consistent set of predictors. For subsequent discussion, we number these months $t = 1, \ldots, 12$. To accommodate lagged predictors, the data set includes several prior months of activity. At this stage of processing, we removed 35,906 accounts that appeared inactive or closed and continued with the remaining 244,094 accounts.

We next align the data as though collected at a single point in time. Let Y_{it} denote the response for the i^{th} account in month t, and let $\mathbf{X}_{it}$ denote the corresponding collection of predictors. We model the data as though the conditional mean of the response has finite memory

$$\begin{aligned} E(Y_{it}|\mathbf{X}_{i,t-1}, \mathbf{X}_{i,t-2}, \ldots) &= E(Y_{it}|\mathbf{X}_{i,t-1}, \mathbf{X}_{i,t-2}, \ldots, \mathbf{X}_{i,t-\ell}) \\ &= \mu_{it}(\mathbf{X}_{i,t-1}, \mathbf{X}_{i,t-2}, \ldots, \mathbf{X}_{i,t-\ell}) \ . \end{aligned}$$

We limit our attention here to the prior $\ell = 4$ months of data because of a suspicion that bankruptcy events are sudden rather than long term. (We explored other time horizons and found similar results.) We further assume stationarity of the regression and model the conditional mean as invariant over accounts and months during the studied year,

$$\mu_{it}(\mathbf{X}) = \mu(\mathbf{X}), \qquad \forall i,\ 1 \le t \le 12 \ . \tag{1}$$

The assumption of stationarity allows us to pool the bankruptcy events to estimate the mean function. Rather than treat the responses for one normal account $Y_{i1}, \ldots, Y_{i12}$ as a dependent sequence to be modeled together, we treat these as 12 uncorrelated responses. Similarly, we align all bankruptcies to the same "point" in time. Thus each non-bankrupt account contributes 12 observations to the final data set, and each bankrupt account contributes at most 12. After alignment, the data set of 244,094 accounts expands to 2,917,288 monthly responses Y_{it} and the accompanying lagged predictors. The former association of these observations as part of an original longitudinal sequence is not used in the analysis. While this assumption simplifies the presentation of our methodology, it does make it harder to predict bankruptcy. For example, consider an account that declares bankruptcy in month 7. It is an error to predict bankruptcy for any of the previous six observations even though these may presage the event.

It may concern some readers that the data set contains multiple observations from most credit card accounts. We shared this concern and so performed an analysis of a smaller data set constructed in the following manner. Each account showing a bankruptcy was sampled, but we used only the month in which the bankruptcy occurred. For example, if the bankruptcy occurred in month 3, then the only response generated by this account is Y_{i3}. We then sampled 25% of the remaining accounts, picking one randomly chosen month as the single response for the account. We modeled these data in the same fashion as that presented here, finding essentially similar results. We were troubled, however, by the impact of subsampling and the required weighting on the validation error calculations and hence chose to use all of the data instead.

Each observation at this stage consists of the 0/1 indicator for the occurrence of bankruptcy and 255 predictors. To capture trends or seasonal variation in the rate of bankruptcy, we retain the month index t as a possible predictor. We treat the time trend as both a continuous predictor and as a collection of 12 seasonal dummy variables. Other categorical predictors are handled in the usual fashion, converting a k-level categorical variable into k dummy variables. We also merge some of the indicators, such as the state of residence, to reduce the number of predictors. Since our search procedure does not treat these indicators as a set, to keep or ignore as a group, the categorical variables are converted into the over-determined set rather than leaving out one category. Missing data in such categorical variables simply defines another category. Handling missing data in continuous predictors is done in a different manner described next.

The prevalence of missing data led us to adopt a simple approach to incomplete cases. We treat any continuous variable with missing data as the interaction of an unobserved, complete variable with a "missingness" indicator. This procedure is easy to implement. A scan through the time-aligned data set flags any continuous predictors with missing data; of the 255 raw predictors, 110 have missing data. In each case, we fill the missing values with the mean $\overline{X}_j$ of the observed cases and add an indicator, say B_j, to the data set. This augmentation of the set of predictors thus adds 110 more dichotomous predictors, giving a total of $110 + 255 = 365$ predictors. An important by-product of filling in missing data in this way is the possible introduction of heteroscedasticity. If indeed X_j is an important predictor of bankruptcy, then filling missing values with the observed mean $\overline{X}_j$ introduces heteroscedasticity.

Since we anticipated none of the predictors to have a large effect, this source of heteroscedasticity was not expected to be large. In any event, our estimation allows for such heteroscedasticity.

The penultimate stage of preparing the data adds *all* of the second-order interactions to the set of predictors. Because the resulting data set would be huge, this step is implicit and our code computes interactions between predictors as needed in subsequent calculations. With the set of predictors expanded in this way, the search procedure is able to identify local quadratic nonlinearity and subset differences, at a cost of a dramatic expansion in the number of candidate predictors. The addition of the interactions expands the data set to $365 + 365(366)/2 = 67,160$ predictors. In our analysis of this data, we treat interactions just like any other predictor, violating the so-called principle of marginality. This principle requires, for example, that a model containing the interaction $X_j * X_k$ must also include both X_j and X_k. Our reasoning is simple: if the model benefits from having the base linear terms, then these should be found by the selection procedure. We also allow "overlapping" interactions of the form $X_j{*}X_{k_1}$ and $X_j{*}X_{k_2}$, unlike **?**. In fact, our search for predictors of bankruptcy discovers a number of such overlapping interactions, so many as to suggest a multiplicative model (see Section 5 and remarks in the concluding discussion).

As the final stage of preparations, we randomly divide the time-aligned data set into an estimation sample and a validation sample. This random split essentially defines the population as the time-aligned data set and scrambles the dependence that may be introduced by time alignment. We randomly sampled 20% of the account months from the time-aligned data for estimation, obtaining a sample of 583,116 account months with 458 bankruptcies. The remaining 80% of the data (2,334,172 account months with 1,786 bankruptcies) constitute the validation sample. Our choice to reserve 80% is perhaps arbitrary, but reflects our bias for an accurate assessment of the predictive ability of the selected model. While expanding the proportion used for estimation might lead to a better model, decreasing the size of the validation sample makes it hard to recognize the benefit. Heuristically, our split implies that validation sums of squared errors possess half the standard error of those computed from the estimation sample. We did not shrink the estimation sample further since preliminary calculations indicated that the model fitting would deteriorate with smaller estimation samples; the number of bankruptcy events becomes quite small relative to the number of potential predictors.

3 Variable Selection

We use variable selection to identify features that predicts well when applied to new observations. The object is to add those predictors whose improvement in the accuracy of prediction overcomes the additional variation introduced by estimation, a classic bias/variance trade-off. Adding more predictors reduces bias at a cost of more variance in the predictions. Our criterion judges the accuracy of the predictions by their mean squared error (*MSE*), combining bias and variance. Other metrics (such as classification error or likelihoods) are sometimes chosen. **?** offers further discussion and examples of variable selection in regression.

Procedures for variable selection are simplest to describe in the context of an orthonormal least squares regression. For this section, let Y denote a vector with n elements and let X_j, $j = 1, \ldots, p$, denote $p \le n$ orthogonal predictors, normalized so that $X_j' X_j = 1$. We suppose that the data follow the familiar normal linear regression model

$$Y = \mu + \sigma\epsilon, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0,1) \ , \tag{2}$$

where the mean vector μ has the form

$$\mu = \beta_1 X_1 + \cdots + \beta_p X_p \ . \tag{3}$$

Some of the β_j may be zero and σ^2 is the *known* error variance. Let $\gamma = \{j_1, \ldots, j_q\}$ denote a subset of $q = |\gamma|$ integers in the range 1 to p, and define the associated fitted values

$$\hat{Y}(\gamma) = \sum_{j \in \gamma} \hat{\beta}_j X_j \ ,$$

with $\hat{\beta}_j$ estimated by least squares,

$$\hat{\beta}_j = X_j' Y \ .$$

The challenge for the variable selection criterion is to identify the set of predictors γ that minimizes the mean squared error,

$$MSE(\gamma) = E \, \|\mu - \hat{Y}(\gamma)\|^2 \ , \tag{4}$$

where for vectors x, $\|x\|^2 = \sum x_i^2$. Notice that minimizing $MSE(\gamma)$ is equivalent to minimizing the expected squared error when predicting an independent copy of the response, $Y^* = \mu + \epsilon^*$. Models that estimate the mean μ well also predict well out-of-sample.

Given the goal of minimizing the mean squared error, a logical place to begin is with an unbiased estimate of $MSE(\hat{\gamma})$. This is the path taken by the Akaike information criterion *AIC* (**?**) and Mallow's C_p (**?**). These criteria choose the set of predictors that minimizes an unbiased estimate of $MSE(\gamma)$. For an orthonormal regression, the unbiased estimator is

$$mse(\gamma) = RSS(\gamma) + 2\,q\,\sigma^2\;, \quad q = |\gamma|, \tag{5}$$

where *RSS* is the residual sum of squares,

$$RSS(\gamma) = \|Y - \hat{Y}(\gamma)\|^2\;. \tag{6}$$

The second summand $2q\sigma^2$ on the right-hand side of (5) acts as a penalty term, increasing with the dimension of the model. To pick the best model, then, one computes $mse(\gamma)$ for various collections of predictors (search is another matter) and selects the set, say $\hat{\gamma}$, that obtains the smallest such estimate. The form of (5) also suggests how *AIC* generalizes to other models and distributions. In the normal regression model, the residual sum of squares is essentially the minimum of twice the negative of the log of the likelihood,

$$RSS(\gamma) = \min_{\beta_\gamma} -2\log L(Y_1, \ldots, Y_n; \beta_\gamma) + c_n\;,$$

where c_n is an additive constant for all models. Consequently, *AIC* not only picks the model that minimizes a sum of squares, it also selects the model that maximizes a penalized likelihood.

The act of choosing the model with the smallest estimated mean squared error leads to selection bias. The minimum of a collection of unbiased estimates is not unbiased. This effect is small when *AIC* is used, for example, to select the order of an autoregression. The problem becomes magnified, however, when one compares many models of equal dimension, as in regression (see **?**, for a discussion of this issue). This selection bias produces a criterion that chooses too many variables when none are in fact useful. This effect is easily described for an orthonormal regression. Order the predictors so that $\hat\beta_j^2 \geq \hat\beta_{j+1}^2$ and observe that the residual sum of squares drops by $\hat\beta_j^2$ when X_j is added to the model. Now, *AIC* implies that one should add X_{q+1} to the model with predictors $X_1, \ldots, X_q$ if the residual sum of squares drops by enough to compensate for increasing the penalty. In this setting this condition reduces to

$$\text{add } X_{q+1} \quad \Longleftrightarrow \quad \hat\beta_{q+1}^2 > 2\sigma^2 \ , \tag{7}$$

or, equivalently, if the absolute z score for X_{q+1}, $|z_q = \hat\beta_{q+1}/\sigma| > \sqrt{2}$. In the null case with $\beta_j = 0$ for all j, $z_q \sim N(0,1)$, *AIC* selects about 16% of the predictors even though none actually reduces the mean squared error. In fact, each added superfluous predictor increases the *MSE*. In a problem such as ours with 67,000 predictors, most with little or no effect on the response, a procedure that selects 16% of the predictors would lead to a rather poor model. The variation introduced by estimating so many coefficients would outweigh any gains in prediction accuracy.

A simple way to reduce the number of predictors in the model is to use a larger threshold – but how much larger? The literature contains a variety of alternatives to the penalty $2q\sigma^2$ in (5) (as reviewed, for example, in **?**). At the extreme, the Bonferroni criterion selects only those predictors whose two-sided p-value is smaller than α/p, where p is the number of possible predictors under consideration and α is the type I error rate, generally $\alpha = 0.05$. In contrast to *AIC*, the Bonferroni criterion selects on average only a fraction of one predictor under the null model. Because the p-values implied by the Bonferroni criterion can be so small (on the order of $.05/67000 \approx 0.0000007$ in our application), many view this method as hopelessly conservative.

Despite such reactions, the small p-values associated with Bonferroni in large problems such as ours are appropriate. In fact, recent results in statistical decision theory show that variable selection by the Bonferroni criterion is optimal in a certain minimax sense. These optimality properties are typically associated with a method called hard thresholding (**?**) or the risk inflation criterion (**?**). These procedures, which we will refer to as *RIC*, select predictors whose z-score is larger than the threshold $\sqrt{2 \log p}$. *RIC* is optimal in the sense of the following minimax problem. A statistician is competing against nature, and nature knows which predictors have non-zero coefficients. The statistician chooses a data-driven variable selection rule, and nature then chooses the regression coefficients. The objective of the statistician is to minimize the ratio of the mean squared error of his model to that obtained by nature who includes all predictors for which $\beta_j \neq 0$. In a regression with Gaussian errors and q non-zero β_j scattered among the p elements of β, **?** show that the best possible ratio of mean squared errors is about $2 \log p$,

$$\min_{\hat\beta} \max_{\beta} \frac{E \, \|Y - X\hat\beta\|^2}{q\sigma^2} = 2 \log p - o_p(\log p) \ . \tag{8}$$

The minimum here is over *all* estimators of β and is asymptotic in the size of the model p, holding q fixed. The model identified by *RIC* attains this competitive mean squared error. That is, if one selects predictors by choosing only those whose absolute z-scores exceed $\sqrt{2 \log p}$, then the *MSE* of the resulting model is within a factor of $2 \log p$ of that obtained by estimating the true model, and this is the best asymptotic performance.

These same claims of optimality apply to the Bonferroni criterion because it implies essentially the same threshold. To get a better sense of the similarity of the Bonferroni criterion and *RIC*, consider again the simplifying situation of an orthonormal regression with known error variance σ^2. The Bonferroni criterion implies that one should select those predictors whose absolute z-scores exceed a threshold τ_α defined by

$$\frac{\alpha}{2p} = 1 - \Phi(\tau_\alpha) \ . \tag{9}$$

In this expression $\Phi(x)$ is the cumulative standard normal distribution,

$$\Phi(x) = \int_{-\infty}^{x} \phi(t) dt, \ \text{ with } \phi(x) = e^{-x^2/2}/\sqrt{2\pi} \ .$$

To show that τ_α is close to the *RIC* threshold $\sqrt{2 \log p}$ for large p, we use the well-known bounds for the cumulative normal (e.g. **?**, page 175):

$$\phi(x) \left(\frac{1}{x} - \frac{1}{x^3} \right) < 1 - \Phi(x) < \frac{\phi(x)}{x}, \quad x > 0 \ . \tag{10}$$

Since our interest lies in large x, we can simplify these bounds to

$$\frac{3}{4} \frac{\phi(x)}{x} < 1 - \Phi(x) < e^{-x^2/2}, \quad x > 2 \ . \tag{11}$$

At the Bonferroni threshold τ_α, equation (9) holds and the upper bound in (11) implies

$$\tau_\alpha < \sqrt{2 \log p + 2 \log \tfrac{2}{\alpha}} \ . \tag{12}$$

Figure 2: Thresholds implied by the risk inflation criterion $\sqrt{2\log p}$ (short dashes, - - -), the optimal threshold τ^* (long dashes, — —), and Bonferroni with $\alpha = 0.05, 0.20$ (solid curves in black and gray, respectively). The *RIC* threshold is a compromise between the Bonferroni threshold $\tau_{0.05}$ and the optimal threshold τ^*.

From the lower bound in (11), we have

$$e^{-\tau_\alpha^2/2} < c\,\frac{\tau_\alpha}{p}$$

with the constant $c = 2\alpha\sqrt{2\pi}/3$. If we replace τ_α on the right-hand side of this expression by the just-found upper bound (12), we eventually arrive at the lower bound

$$\tau_\alpha > \sqrt{2\log p - \log\log p - \frac{\log\frac{2}{\alpha}}{\log p} - c'}\ , \tag{13}$$

where the constant $c' = 2\log c + \log 2$. Combining these bounds, we see that the Bonferroni threshold is asymptotically sandwiched for large p between $\sqrt{2\log p - \log\log p}$ and $\sqrt{2\log p}$. To the accuracy of the theorems in **?** or those in **?**, these thresholds are equivalent and both obtain the optimal minimax competitive ratio (8).

This asymptotic equivalence ignores, however, constants and terms that go to zero for large p. A plot of these thresholds for varying values of p clarifies matters. Figure 2 plots several thresholds as functions of the size p of the problem. One is the *RIC* threshold $\sqrt{2\log p}$. Two of the others are Bonferroni thresholds, $\tau_{0.05}$ and $\tau_{0.20}$. The fourth is a bit different and requires some explanation. The underlying minimax problem admits an optimal solution; that is, for any p we can find the threshold τ^* that minimizes the maximum value of the competitive ratio $E\|Y - \hat{Y}(\hat{\gamma})\|/q\sigma^2$. This threshold is found by finding first that β that nature would choose to maximize the ratio, and then finding the threshold best suited for this choice. **?** describe the necessary calculations. The figure shows that for $100 \le p \le 100,000$, the *RIC* threshold $\sqrt{2\log p}$ roughly corresponds to a Bonferroni threshold with $\alpha \approx 0.20$. The optimal threshold is smaller still. For a model with $p = 67,000$ predictors, *RIC* is equivalent to Bonferroni with $\alpha \approx 0.16$ whereas the optimal threshold for this p corresponds to $\alpha = 0.475$, as seen in Figure 3. The *RIC* threshold is seen to be a compromise, lying between the "traditional" Bonferroni threshold at $\tau_{0.05}$ and the optimal threshold τ^*.

Figure 3: α-level associated with the threshold τ_p^* that minimizes the maximum risk inflation.

Before leaving Figure 2, we make two remarks. First, Bonferroni thresholds are not so large as one might expect, thanks to the thin tails of the normal distribution. A common reaction to Bonferroni-type methods is to think that these make it virtually impossible to find important predictors. Even with $p = 67,000$ and $\alpha = 0.05$, the Bonferroni threshold is 4.95. Important predictors, those with z-scores larger than 10, say, are still quite easy to detect. The second point is that *RIC* eventually becomes much more conservative. The "small" values of p in Figure 2 obscure this property. Eventually, for fixed α and growing p, the *RIC* threshold is larger than any τ_α. Rather than admit a fixed fraction of the predictors in the null case, *RIC* admits on average fewer and fewer as p grows. This property of *RIC* follows from the upper bound in (10) which shows that the expected number of coefficients larger than the *RIC* threshold goes to zero as p increases,

$$p(1-\Phi(\sqrt{2\log p})) < \frac{p\,\phi(\sqrt{2\log p})}{2\log p} < \frac{1}{\sqrt{2\log p}} \ .$$

Though *RIC* is optimal in the sense of (8), the size of the threshold makes it impossible to find smaller, more subtle effects. Another look at the asymptotics of **?** suggests that one can do better, at least in models with more effects. When the number of non-zero coefficients q is near zero, the minimax result (8) implies one can do little better than Bonferroni. Situations with more non-zero terms, however, offer room for some improvement. In particular, a small change in the proof of (8) leads to the following revised claim:

$$\min_{\hat\beta}\max_{\beta} \frac{E\,\|Y - X\hat\beta\|^2}{q\sigma^2} = 2\log p/q - o_p(\log p) \ , \tag{14}$$

so long as the proportion of non-zero coefficients diminishes asymptotically, $q/p \to 0$ as $p \to \infty$. When q is a substantial fraction of p, say 10%, the bound is considerably smaller than what Bonferroni obtains. Modifications of the arguments in **?** show that a variation on the following approach, however, provably obtains the better bounds. The idea is simple and intuitively appealing. Rather than compare *all* of the estimated coefficients $\hat\beta_j$ to the Bonferroni threshold or $\sqrt{2\log p}$, reduce the threshold as more significant features are found. The resulting procedure can also be motivated by ideas in empirical Bayes estimation (**?**), multiple hypothesis testing or step-up testing (**?**), and information theory (**?**).

Adaptive variable selection automatically adjusts the threshold to accommodate problems in which more predictors appear useful. Instead of using the same threshold for all of the coefficients, only the largest z-score is compared to the Bonferroni threshold. Order the z-scores as

$$z^2_{(1)} \geq z^2_{(2)} \geq \cdots \geq z^2_{(p)} \ .$$

To identify the model, first compare $z^2_{(1)}$ to the *RIC* threshold $2 \log p$ (or, alternatively, compare its p-value to $\frac{\alpha}{2p}$). If $z_{(1)}$ exceeds its threshold, then add the associated predictor, X_{j_1} say, to the model. Otherwise no predictors are utilized and we are left with the null model. Assuming the largest z-score passes this test, then consider the second largest. Rather than compare $z^2_{(2)}$ to $2 \log p$, however, compare it to the reduced threshold $2 \log p/2$. The selection process stops if $z_{(2)}$ fails this test; otherwise, add X_{j_2} to the model and continue on to examine the third predictor. In general, the process adds the qth most significant predictor if

$$z^2_{(q)} > 2 \log \tfrac{p}{q} \ .$$

In terms of p-values (as done in step-up testing) the approximately equivalent procedure compares the p-value of $z_{(q)}$ to $1 - \Phi(\frac{q}{p}\,\frac{\alpha}{2})$. The covariance inflation criterion (**?**) works similarly in orthogonal problems, though with a larger leading constant 4 rather than 2.

Adaptive thresholds let more predictors enter the model, ideally without overfitting. Figure 4 plots the adaptive threshold $\sqrt{2 \log p/q}$ along with the corresponding thresholds implied by step-up testing, again with $\alpha = 0.05, 0.20$ and $p = 67,000$. All three thresholds drop off quickly. The adaptive threshold $\sqrt{2 \log p/q}$ closely corresponds to a step-up procedure with $\alpha = 0.20$ over the shown range of q; indeed, the approximation holds for $q < 1000$. The adaptive threshold starts at 4.71 and drops to 4.56 if the first predictor passes the test. For $q = 10$, the adaptive threshold is 4.20, and once $q = 68$, the adaptive threshold is a full standard error below the threshold applied to the first predictor. The discovery of only 68 significant effects among 67,000, about 1 in 1000, reduces the threshold from 4.71 down to 3.71. The data analysis in

Figure 4: Adaptive thresholds given by the risk inflation criterion $\sqrt{2\log p/q}$ (dashed, - - -) and step-up comparisons with $\alpha = 0.05, 0.20$ (solid black and gray, respectively). In each case the number of potential predictors $p = 67,000$.

the next section uses both *RIC* and the adaptive threshold. The model found with *RIC* is more parsimonious, but does not attain as low an out-of-sample mean squared error in the validation data. The additional predictors found by the adaptive method do indeed improve the predictions.

Before turning to the application, we draw attention to an important assumption that underlies thresholding methods: namely that one can convert $\hat{\beta}_j$ into a z-score. The error variance σ^2 is crucial in this calculation. For tractability, most theoretical analyses assume σ^2 is known or can be estimated to high precision. Our application falls in the latter case. While we cannot claim σ^2 is known, we have more than enough data to compute an accurate estimate of the error variance. This luxury need not be the case, particularly when the number of coefficients q in the fitted model is large relative to the sample size n . In such cases, the usual estimator of σ^2 can become biased during the fitting process. This bias is easy to describe in an orthonormal regression. Suppose all of the coefficients $\beta_j = 0$, the null model, and we have just as many predictors as observations, $p = n$. Under these conditions, $\hat{\beta}_j \sim N(0, \sigma^2)$ and $\sum_i Y_i^2 = \sum_j \hat{\beta}_j^2$. Notice that

$$s_0^2 = \sum_i Y_i^2/n$$

is an unbiased estimate of σ^2. Now suppose that the predictors are ordered so that $\hat{\beta}_1^2 \geq \hat{\beta}_2^2 \geq \cdots \geq \hat{\beta}_n^2$. If the most significant predictor X_1 is added to the model, then the estimator obtained from the residual sum of squares,

$$s_1^2 = \frac{\sum_i (Y_i - \hat{Y}_i(1))^2}{n-1} = \frac{\sum_{j=2}^n \hat{\beta}_j^2}{n-1} ,$$

is biased. For large n, the size of the bias is about

$$E s_1^2 - \sigma^2 = \frac{\sigma^2 - E\beta_1^2}{n-1} \approx \frac{-2\sigma^2 \log n}{n} .$$

Initially small, such bias rapidly accumulates as more predictors enter the model. Since the bias reduces the estimate of σ^2, once a predictor has been added to the model, it becomes easier to add the next and to produce more bias. Asymptotically, this bias has a dramatic effect. Suppose for some small $\delta > 0$ that we use the biased value $\nu^2 = \sigma^2 - \delta/2$ for the variance in the standard error of the slope estimates in an orthonormal regression. The lower bound in (10) shows that the expected number of test statistics $\hat{\beta}_j/\nu$ that exceed the threshold $\sqrt{2 \log p}$ grows without bound,

$$\lim_{p \to \infty} p \left(1 - \Phi(\sqrt{(2-\delta) \log p}) \right) = \infty .$$

Small bias in σ^2 can overcome even the *RIC* threshold and similarly affects the performance of the adaptive threshold. Though not an issue in the following bankruptcy regression, we have encountered smaller data sets where such bias overwhelms Bonferroni procedures (e.g., see the case "Using Stepwise Regression for Prediction" in ?).

4 Estimators and Standard Error

The theory of the previous section is clearly oriented toward variable selection in models that have orthogonal predictors, such as a wavelet regression or classical orthogonal analysis of variance. Given a sequence of any predictors, however, one can use Gram-Schmidt to construct a sequence of orthogonal subspaces, converting the original predictors into orthogonal predictors. That is the approach that we will take, with the ordering defined by a greedy search. We treat the predictors as unstructured, using of none of the features of their creation, such as the fact that one is the missing value indicator of another or that one is a component of an interaction. This "anonymous" treatment of the predictors is more in the spirit of data-mining tools like neural nets that shun external side-information.

The task for variable selection then is two-fold. First, we must choose the sequence of predictors. For this, we use stepwise regression, modified to accommodate the size of the problem and heteroscedasticity. Consideration of 67,160 predictors breaks the most implementations and 583,115 observations slows iterations. Second, we have to decide when to stop adding predictors. For this, we use adaptive thresholding as defined in Section 3, comparing the t-statistic of the next slope to the adaptive threshold $\sqrt{2 \log p/q}$. Our algorithm features a conservative estimator of the standard error of the slopes, sacrificing power to avoid over-fitting.

Our division of the data into estimation and validation subsets leaves 583,115 account months for estimation. Combined with 67,160 predictors, calculations of covariances are quite slow, needlessly so. Since the estimation data contains only 458 bankruptcies, we subsample the data, selecting all of the bankruptcies and 2.5% (on average) of the rest. Doing so randomly generates a data file with 15,272 account months. In the sense of a case-control study, we have about 40 controls for every bankruptcy, much larger than the rules of thumb suggested in that literature (e.g. **?**). We denote the associated sampling weights as

$$w_i = \begin{cases} 1\ , & \text{if } Y_i = 1 \quad \text{(bankrupt case)}\ , \\ 40\ , & \text{if } Y_i = 0\ . \end{cases} \tag{15}$$

Each bankrupt observation represents one case, whereas each non-bankrupt observation represents 40.

Although the response variable indicating bankruptcy is dichotomous, we nonetheless fit a linear model via least squares. The choice is not merely one of convenience, but instead is motivated by our measure of success (prediction) and nature of the data (bankruptcy is rare). Obviously, stepwise linear regression is easier to program and faster to run than stepwise logistic regression. Although this concern is non-trivial in a problem with so many predictors, our preference for a linear regression runs deeper. The logistic estimate for the probability of bankruptcy is bounded between 0 and 1 and its intrinsic nonlinearity provides a smooth transition between these extremes. Because bankruptcy is so rare in our data, however, it is hard to expect a statistical model to offer probability predictions much larger than 0.25, with the preponderance essentially at 0. A linear model that allows quadratic components is capable of approximating the local curvature of the logistic fit in the tails. That said, our procedure generates a set of predictors that do in fact suggest a multiplicative model (see Section 5). Since logistic regression is multiplicative (for the probability), our automatic modeling leads one to discover this structure rather than impose it from the start.

A second argument in favor of logistic regression lies in its efficiency of estimation. Least squares is not efficient since it ignores the heteroscedasticity induced by the 0/1 response, so one is tempted to use weighted least squares (*WLS*). Let $\hat{Y}_i$ denote the fit for the ith response. A *WLS* search for efficient estimates of β minimizes the loss function

$$L_w(\hat\beta) = \sum_i \frac{(Y_i - \hat Y_i)^2}{\hat Y_i(1-\hat Y_i)} \ .$$

Concern for efficiency of estimation, however, leads one to forget which observations are important in the analysis. The weighted loss function L_w *down-weights* the most interesting observations, those that have some probability of bankruptcy. By giving high weight to the overwhelming preponderance of non-bankrupt data with $\hat Y_i \approx 0$, weighting for efficiency of estimation conceals the observations that would seem to have most information about the factors that predict bankruptcy. In contrast, the least squares loss function

$$L(\hat\beta) = \sum_i (Y_i - \hat Y_i)^2$$

weights all of the data equally. With so much data, efficiency in the estimation of $\hat\beta$ for a given set of predictors is much less important than finding useful predictors. Thus, we have chosen to minimize $L(\hat\beta)$ rather than the weighted loss. Part of the appeal of using a well-understood technique like regression lies in the flexibility of choosing the loss function. While we eschew weighting in the loss function, we do allow for heteroscedasticity when estimating the standard error of $\hat\beta$, though we do not impose the binomial form $p(1-p)$. This is a special structure, and our goal is an approach that can be applied routinely without needing a model for the variance.

To compute the regression estimates, we implemented a variation of the standard sweep operator. In order to describe our algorithm, it is helpful to review the sweep algorithm and how it is used in a stepwise regression. Suppose that we use the $n \times q$ matrix X of predictors in a regression model for the response Y. The standard sweep algorithm (**??**) provides the least squares estimator

$$\hat\beta = (X'X)^{-1}X'Y$$

and its estimated covariance matrix

$$\text{var}(\hat\beta) = s^2(X'X)^{-1} \ , \quad s^2 = \frac{e'e}{n-q},$$

where the residual vector is $e = Y - X\hat\beta$. The sweep algorithm transforms the partitioned cross-product matrix in place, mapping

$$\left[\begin{array}{c|c} Y'Y & Y'X \\ \hline X'Y & X'X \end{array}\right] \quad \Rightarrow \quad \left[\begin{array}{c|c} e'e & -\hat\beta' \\ \hline \hat\beta' & (X'X)^{-1} \end{array}\right] \ .$$

After "sweeping out" the q predictors in X, the resulting array holds the residual sum of squares in its upper left corner and $-\hat{\beta}'$ in the rest of the first row.

The value of the sweep operator in stepwise regression lies in extending the regression to more predictors. Suppose that $p-q$ other predictors form the array Z and denote the projection matrix $H = X(X'X)^{-1}X'$. Sweeping out X from the expanded cross-product matrix yields the following transformation,

$$\left[\begin{array}{c|c|c} Y'Y & Y'X & Y'Z \\ \hline X'Y & X'X & X'Z \\ \hline Z'Y & Z'X & Z'Z \end{array}\right] \quad \Rightarrow \quad \left[\begin{array}{c|c|c} e'e & -\hat{\beta}' & e'(I-H)Z \\ \hline \hat{\beta}' & (X'X)^{-1} & X'(I-H)Z \\ \hline Z'(I-H)e & Z'(I-H)X & Z'(I-H)Z \end{array}\right] .$$

This new array has all of the information needed to select the next predictor which gives most improvement in the current fit. The quadratic form $e'(I-H)Z$ in the first row is $n-q$ times the estimated partial covariance between Y and Z_j given X, and the diagonal of $Z'(I-H)Z$ is the corresponding sum needed for the partial variances of Z. Combining these as the ratio

$$\frac{e'(I-H)Z_j}{\sqrt{(e'e)(Z'(I-H)Z)_{jj}}}$$

gives the estimated partial correlation. The predictor Z_j with the largest partial correlation offers the largest improvement to the current fit and is the choice of standard stepwise selection.

Our first modification to this algorithm handles the large number of predictors. We cannot explicitly compute the entire cross-product matrix with 67,000 predictors. Rather, we defer some of the calculations and only form those portions of the cross-product matrix as needed for identifying the next predictor. In particular, when considering the omitted predictors that form Z, our implementation computes the full sweep for the relatively small augmented matrix $[Y \| X]$, providing standard errors, slopes, and residuals e. For evaluating the omitted predictors, the algorithm computes the vector $e'(I-H)Z_j$ and *only the diagonal* of the large matrix $Z'(I-H)Z$. Combined with the base sweep of X, we have all of the information needed to find the next predictor.

Our use of sampling weights leads to a second, more fundamental modification to the basic calculations. In a linear regression, over-sampling biases $\hat{\beta}$ and necessitates a weighted estimator. Let W denote an $n \times n$ diagonal matrix with the sampling weights w_i from (15) along its diagonal, $W_{ii} = w_i > 0$. For a model with the q predictors X, the weighted estimator is

$$\hat{\beta}_W = (X'WX)^{-1}X'WY \ ,$$

with variance

$$\mathrm{Var}\,(\hat{\beta}_W) = (X'WX)^{-1}X'W(\,\mathrm{Var}\,Y)WX(X'WX)^{-1} \ . \tag{16}$$

One normally encounters weighted least squares when the weights are inversely proportional to the variances of the observations,

$$\mathrm{Var}\,(Y) = \sigma^2 \ W^{-1} \ . \tag{17}$$

In this case, the variance of $\hat{\beta}_W$ reduces to

$$\mathrm{Var}\,(\hat{\beta}_W) = \sigma^2(X'WX)^{-1} \ .$$

One can evaluate all of these expressions by applying the standard sweep algorithm to the weighted cross-product matrix. With sampling weights, however, one must augment the standard sweep in order to find appropriate standard errors. (We refrain from calling this the "WLS estimator" since that name is most often used when using variance weights, not sampling weights.)

Calculation of the standard errors of the estimated coefficients is critical for adaptive thresholding. Without an accurate standard error, we cannot use the t-statistic to judge which predictor to add to the model. The fact that both the response as well as many predictors are dichotomous and sparse complicates this calculation. Under homoscedasticity, the variance of $\hat{\beta}_W$ simplifies to

$$\mathrm{Var}\,(\hat{\beta}_W) = \sigma^2(X'WX)^{-1}(X'W^2X)(X'WX)^{-1} \ . \tag{18}$$

Not only is this hard to compute (it's not a convenient extension of the extant weighted sweep calculations), but it is also wrong. If there is some signal in the data, then those cases with higher risk of bankruptcy (say $EY_i \approx 0.10$) have higher variance since $\mathrm{Var}(Y_i) = (EY_i)(1 - EY_i)$. It is tempting to ignore this effect, arguing that the differences among the $E\,Y_i$ are likely small. While that may be the case, the paucity of bankruptcy events exacerbates the situation.

To explain the issues that arise when calculating standard errors for the slope estimates, we turn to the case of a simple regression of a dichotomous variable Y on a single dichotomous predictor X. Many of our predictors are, like the response, dichotomous. For example, recall that 110 of the 365 base predictors are indicators of missing values and other predictors are indicators for geographic location, time period, and credit history features. Further, many of the interactions are themselves the product of dichotomous predictors, and some of these products are quite sparse. The small, hypothetical table below gives the number of cases with each of the four combinations of X and Y.

$Y = 1$	500	k
$Y = 0$	14,500	$n_1 - k$
	X=0	X=1

The counts used in this table are meant to resemble some that we encountered in a preliminary analysis of this data, and so we assume that $n_1 \ll 15,000 = n_0$ (with $n_j = \#\{X_i = j\}$). For this illustration, we treat all of the observations as equally weighted; weighting does not materially affect the implications though it would obfuscate some of the expressions. For the data in this table, the least squares slope is the mean difference in the sample proportions for $X = 0$ and $X = 1$. We denote these sample proportions $\hat{p}_0$ and $\hat{p}_1$, the observed difference $\hat{d} = \hat{p}_1 - \hat{p}_0$, and the underlying parameters p_0 and p_1. Thus, declaring a significant slope is equivalent to rejecting the null hypothesis $H_0 : p_0 = p_1$. A counting argument gives a simple test of H_0. With so much data at $X = 0$, estimate $\hat{p}_0 = 1/30$ and use this estimate to compute the probability of k successes in the small number of trials where $X = 1$. With $n_1 = k = 2$, (i.e., two successes at $X = 1$) we find a "p-value" $= 1/900$ which we can associate with a z-statistic $z = 3.26$. This test statistic would not exceed the *RIC* threshold $\sqrt{2 \log p} \approx 4.71$ in the bankruptcy regression. (We focus on the case $n_1 = k = 2$ for illustrations because this is the specific situation that we encountered in our analysis.)

If we treat the data as homoscedastic, the standard error is inflated when the data have the pattern suggested in our table. Modeling the data as having constant variance in the bankruptcy regression via (18) is equivalent to forming a pooled estimate of the standard error when testing $H_0 : p_0 = p_1$,

$$se_{pool}(\hat{d}) = se(\hat{p}_1 - \hat{p}_0) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_0} + \frac{1}{n_1}\right)}\ , \quad \hat{p} = \frac{n_0\hat{p}_0 + n_1\hat{p}_1}{n_0 + n_1}\ . \tag{19}$$

In the special case with $n_1 = k = 2$, $\hat{d} = 29/30$ and the pooled standard error $se_{pool} \approx 0.127$ gives an inflated $z = 7.60$. Were we to use this estimate of standard error in the bankruptcy regression, this predictor would now appear significant. This test statistic is over twice the size of that obtained by the previous counting argument and implies that we cannot use this approach to estimate the standard error of $\hat{\beta}_W$.

The literature contains a number of alternative variance estimates that are robust to heteroscedasticity. While the Bernoulli expression $p(1-p)$ seems an obvious choice, we use a procedure that does not presume such knowledge of the variance. The expression (16) for $\text{Var}(\hat{\beta}_W)$ suggests a very simple sandwich estimator. Rather than plug in an assumed formula for $\text{Var}(Y)$, estimate it directly from the variance of the residuals. Since the size of the fitted model varies in this discussion, we add an argument to our notation that gives the size of the current model. For example, $e(k) = Y - X(k)\hat{\beta}_W(k)$ is the residual vector for a model fit to the k predictors in $X(k) = [X_1, \ldots, X_k]$. With this notation, a sandwich estimator of the variance is

$$\text{var}\,(\hat{\beta}_W(k)) = ((X(k)'WX(k))^{-1}X(k)'W\ \text{diag}(e(k)^2)\ WX(k)(X(k)'WX(k))^{-1}\ . \tag{20}$$

Under mild conditions like those established by **?**, this expression provides a consistent estimator of the variance of $\hat{\beta}_W(k)$. The estimator is biased, particularly for small samples and with high-leverage points. **?** review the motivation for (20) and offer some recommendations for small-sample adjustments. **?** show that the bias is $O(1/n^2)$ and give an iterative scheme for reducing the bias. Such adjustments for differences in leverage have little effect in our application.

The standard implementation of the sandwich estimator fails in situations such as ours with sparse, dichotomous data. In a preliminary analysis, we added a predictor for which the homoscedastic t-statistic using (18) was computed to be 51. This value seemed inflated to us, as suggested by the previous stylized binomial example. When we computed a heteroscedastic t-statistic by replacing the standard error from (18) by one from (20), however, the resulting t-statistic soared to 29,857. Clearly, the heteroscedastic formula is even farther off the mark. We can explain what is happening in the regression for bankruptcy by returning to the comparison of proportions. Rather than use a pooled estimate like (19) to form the standard error, the sandwich formula (20) implies that we test $H_0 : p_0 = p_1$ using a different standard error. Parallel to

$$\mathrm{Var}(\hat{p}_1 - \hat{p}_0) = \frac{p_0(1-p_0)}{n_0} + \frac{p_1(1-p_1)}{n_1} \; ,$$

the sandwich estimator of the standard error reduces to

$$se_{sand}(\hat{d}) = \frac{\hat{p}_0(1-\hat{p}_0)}{n_0} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} \tag{21}$$

for the binomial comparison. With $n_1 = k = 2$, $\hat{p}_1 = k/n_1 = 1$ and this expression leads to an even smaller standard error and further inflated test statistic because the term $\hat{p}_1(1-\hat{p}_1)/n_1$ drops out of (21). Consequently the sandwich standard error essentially ignores the substantial variation at $X = 1$, giving the estimate $se_{sand} \approx 0.00147 \approx \sqrt{\hat{p}_0(1-\hat{p}_0/n_0}$. The claimed z-statistic rises to $\hat{d}/se_{sand}(\hat{d}) \approx 660$.

A simple modification of the sandwich estimator, however, works nicely. In the bankruptcy regression, we need a way to estimate $\mathrm{Var}(\hat{\beta}_W)$ that handles heteroscedasticity without presuming that the predictor affects the response. The sandwich formula (21) achieves the first, but through its use of the residuals resulting from adding X_k to the model, fails at the second. One can attribute the behavior of the sandwich estimator in our example to the perfect fit at $X = 1$ that conceals the major component of the variance of the slope estimate. In a sense, such calculation of the standard error ignores the need to test $H_0 : p_0 = p_1$ and proceeds as though H_0 is false and $p_0 \neq p_1$. With this bit of motivation, we offer the following conservative estimate of standard error,

$$\mathrm{var}\,(\hat{\beta}_W(k)) = (X(k)'WX(k))^{-1}X(k)'W\ \mathrm{diag}(e(k-1)^2)\ WX(k)(X(k)'WX(k))^{-1} \; . \tag{22}$$

That is, we simply replace the residuals computed with X_k added to the model by the residuals from the *previous* iteration, one that assumes X_k has *no* effect. For the stylized binomial test, the corresponding standard error in effect only uses the n_1 values at $X = 1$ and the pooled proportion $\hat{p}$, $\hat{p}(1-\hat{p})/n_1$. With $n_1 = k = 2$, the resulting standard error is about $1/\sqrt{2}$, half the size of the simple counting test statistic. Although conservative in this fashion, the estimator (22) allows us to identify useful predictors without introducing spurious precision that distorts the selection procedure.

To recap our modeling process, our search algorithm proceeds as a forward stepwise regression. At each step, we compare the t-statistic based on the standard error from (22) for each excluded predictor to the adaptive threshold $\sqrt{2 \log p/q}$, where q denotes the current size of the model, starting from $q = 1$ for the model with the initial constant term. The search continues as long as a predictor is found that exceeds the threshold. As there may be several that exceed the threshold, we employ the following sorting strategy. For those predictors whose absolute t-statistic exceeds $\sqrt{2 \log p/q}$, we sort them based on their impact to the residual sum of squares. Of those judged significant, we choose the predictor offering the most improvement in the fit, add it to the model, and update the sweep calculations. This process continues until no further predictor attains the adaptive threshold. At that point, the algorithm formally stops. To see what would happen if the search continues past this cut off, we allowed the algorithm to go further. When the search moves beyond the adaptive cut-off, the following version of constraint relaxation obtains a "soft landing." Rather than sort the predictors that do not exceed the threshold by the change in the residual sum of squares, we sort them by the attained level of significance, i.e. by t-statistics based on the conservative standard error (22). This less-greedy approach avoids the steep rise in out-of-sample error often associated with over-fitting, at least in our application. Indeed, this ordering of the predictors produces out-of-sample mean squared errors that resemble those offered to show that boosting does not over-fit (see, e.g., Figure 1 in **?**).

5 Results

Figure 5 summarizes the step-by-step in-sample performance of our methodology. This plot shows the residual sum of squares (*RSS*) as the modeling proceeds. Without a validation sample, this plot, along with the accompanying parameter estimates, is all that one has to determine the choice of model. To put the *RSS* in perspective, recall that the estimation data set has 458 bankruptcies. The null model that predicts all as 0 (no bankruptcies) thus has a total residual sum of squares of 458. Each drop by one in the *RSS*, in a loose sense, represents "finding" one more bankruptcy. Initially, the *RSS* drops relatively quickly down to about 420 and then commences a steady, slow decline as the model expands. Using the adaptive criterion applied the *RSS* and fitted

Figure 5: Sums of squared residuals decline up to the selected model order at $q = 39$ (solid) and continue to decline slowly (dashed) beyond this cutoff. To interpret the scales, note that the estimation sample holds 458 bankrupt events.

Figure 6: Sums of squared validation errors decline up to the selected model order at $q = 39$ (solid). Beyond this limit, the validation error either grows slowly using the described "soft-landing" procedure to choose additional predictors (dashed) or abruptly when a greedy approach is used (triangles). To interpret the scales, note that the validation sample holds 1756 bankrupt events.

slopes, our procedure stops adding predictors at $q = 39$, represented by the vertical line in the figure. We allowed the selection process to continue to generate this figure, using the soft-landing procedure described at the end of the previous section to select predictors that did not meet the selection criterion.

Figure 6 shows how well the selected models predict the held-back observations in the validation sample. This plot is perhaps the best evidence that our procedure works, at least in the sense of finding predictive effects without over-fitting. As with the residual sum of squares, the validation sum of squares (*VSS*) drops rapidly as the initial predictors join the model. Again, to interpret the scales, note that the validation sample holds 1756 bankruptcies. The *VSS* drops down to a minimum at 1652, and then, in contrast to the *RSS*, begins to increase. The vertical line in the figure again highlights the model selected by the adaptive criterion. All of the predictors added by this criterion either improve the predictive accuracy of the model or at least do not cause appreciable damage. At a minimum, each predictor adds enough to the model to "pay" for the cost of estimating an additional parameter. The selected model lies in the midst of a collection of models that share comparable predictive accuracy. The *VSS* is essentially constant for the models with $30 \leq q \leq 44$. At both $q = 30$ and $q = 44$, $VSS = 1651.7$, with the smallest value (1651.50) occurring for $q = 31$. As the model expands by adding predictors that do not meet the adaptive criterion, the *VSS* gradually rises.

The gradual increase in the *VSS* obtained through the soft-landing expansion beyond the selected model stands in sharp contrast to what happens if we take a more greedy approach to adding further predictors. Figure 6 includes the *VSS* when the model is extended past the adaptive threshold by adding the predictor that minimizes the observed residual sum of squares. When insignificant predictors are sorted by their change in the *RSS* (rather than by their measured significance), the *VSS* rises rapidly when the model is grown past the adaptive cutoff. The *VSS* jumps from 1652 to 1748 if we add the 40th predictor to minimize the *RSS* rather than to maximize the conservative t-statistic. The size of the jump is so large that it requires some explanation, and we can again use the stylized simple regression used to motivate the conservative standard error in Section 4. Basically, by choosing predictors based on the observed change in residual sum of squares, the model over-fits, with the sparse nature of the data compounding the effects of the spurious estimate. The predictor X_{7291} added at step 40 that most improves the in-sample residual sum of squares is, not surprisingly, another interaction. In this case, it is an interaction of an indicator with a continuous predictor. In the sample of the estimation data used to fit the model, X_{7291} differs from zero for only six observations, and all six of these happen to be bankruptcies. This predictor is zero for all of the observations in the 2.5% sample of the non-bankrupt cases. This is exactly the situation described in the stylized examples of Section 4, only now X_{7291} is not dichotomous. While seeming a good predictor in-sample by this criterion, it fares poorly out of sample because the pattern of bankruptcies for this predictor differs markedly in the validation sample. Among the validation data, X_{7291} is nonzero for 299 cases, of which only 17 are bankruptcies. When these are predicted using the spurious slope computed from the estimation sample, the validation error increases dramatically.

The predictions of this model take on greater economic significance because they are well calibrated. By calibrated, we mean that the predictions satisfy $E(Y_i|\hat{Y}_i = p) = p$. Calibrated predictions imply that among those observations assigned a predicted score of 0.10, for example, 10% declare bankruptcy. Calibrated predictions allow a decision maker to optimize costs. Suppose the value of early detection of bankruptcy is, say, \$5000 and the cost of treating a good customer as though he might declare bankruptcy is \$50 (from loss of good will and annoyance perhaps). Given calibrated predictions of bankruptcy, these dollar values imply that the creditor maximizes profits by contacting all customers assigned a probability of $50/5050 \approx 0.01$ or larger. The calibration plot in

Figure 7: Calibration chart for the out-of-sample predictions of the bankruptcy model applied to the validation sample. Each point shows the proportion of bankrupt observations among those having predicted scores in the ranges 0–0.05, 0.05–0.10,... The vertical error bars indicate 95% pointwise confidence intervals.

Figure 7 shows the rate of bankruptcies among validation observations whose predicted values fall into equal-width bins, with the bins located at [0, 0.05), [0.05, 0.10), ..., [0.95, 1.0]. The diagonal line in the figure is the goal. The vertical lines at each point indicate a range of plus or minus two standard errors. The standard error bars have different lengths predominantly because of the imbalance of the sample sizes and become quite long for higher predicted scores where data is scarce. In general, the proportion of actual bankruptcies in the validation data rises with the predicted scores. The model does miss some curvature, in spite of the quadratic interactions and unweighted loss function, mostly for predictions in the range 0.2 to 0.35. Here, the model underestimates the proportion that declare bankruptcy.

For an application with, relatively speaking, so few useful predictors, one may suspect that hard thresholding would do well. To investigate how well hard thresholding performs, we also used this rule (i.e., comparing all predictors to $\sqrt{2\log p}$ rather than to $\sqrt{2\log p/q}$) to choose models for the bankruptcy data. The stepwise search using hard thresholding ends after choosing 12 predictors. Up to that point, the two procedures have found models with essentially the same predictive ability though different predictors. The 12-predictor model identified by hard thresholding obtains a validation sum of squares of 1664 whereas the 39-predictor model identified by adaptive thresholding obtains a sum of 1652. Interpreting these sums as before, the adaptive procedure has found about 11 more bankruptcies.

All 39 predictors in the adaptively chosen regression model are interactions. These interactions overlap and involve only 37 underlying predictors. Given the effects of selection bias on the estimated coefficients and the presence of obvious collinearity, we do not offer an interpretation of the regression coefficients attached to these predictors. Furthermore, a condition for using this data is that we would not describe the variables with identifiable precision. Nonetheless it is interesting to note the types of predictors that appear in multiple interactions. One predictor, a count of the number of credit lines, appears in 6 of the found 39 interactions. A missing data indicator appears in 5 interactions, and interest rates and a history of past problems each appear in 4 interactions.

One explanation for the presence of interactions as the chosen predictors argues that they are selected not because of their real value, but rather because of their number. After all, we have 365 linear predictors compared to almost 67,000 interactions. To get a sense of what is possible with just linear terms, we fit an alternative model as follows. We first forced all 37 linear terms that make up the interactions in the selected model into a regression. This model obtains a validation sum of squares of 1741, finding 45 bankruptcies in our heuristic sense. We then added interactions. Stepwise search restores 14 of the interactions from the original model and adds 7 other interactions not found in that model. The resulting model with 37+21=58 predictors obtains a validation sum of squares of 1654 (finds 132 bankruptcies), essentially matching – but not exceeding – the performance of the selected model with only interactions.

The presence of so many overlapping interactions suggests a multiplicative model. One can view our collection of pairwise interactions as a low-order approximation to a multiplicative regression function. We explored this notion by fitting a logistic regression using the 37 base predictors that constitute the interactions found by the stepwise search. When fit to the estimation sample, only 23 of these predictors obtain a p-value smaller than 0.20. (We identified this subset sequentially, removing the least significant predictor and then refitting the model.) With an eye toward prediction, we removed the others. When used to predict the validation sample, the sum of squared errors obtained by this trimmed-down logistic regression (using the estimated probability of bankruptcy as the predictor) is 1685, compared to 1652 obtained by the original stepwise model. The multiplicative structure of logistic regression does not eliminate the need for interactions. When combined with the results of expanding the linear predictors, these results suggest that interactions are inherently useful and cannot be removed without a noticeable loss of predictive accuracy.

6 Discussion and Next Steps

Several enhancements may improve the calibration of the model. While our use of over 67,000 predictors may be considered outrageous, this number is in fact artificially small, held down by a number of assumptions. As evident in Figure 7, the 39-predictor model underpredicts the chance of bankruptcy for those assigned predicted scores in the range 0.2–0.35. The use of higher-order interactions, say third-order interactions, offers a potential solution by providing a fit with higher curvature – closer to a multiplicative model. Similarly, we could expand the set of fitted models to include piecewise linear fits by partitioning continuous predictors into discrete bins as done in MARS. Both of these changes introduce so many new predictors that to make use of them requires us to modify the both the method of search as well as the thresholding rule described in Section 3. For example, a modified search procedure could recognize that with X_{jk_1} and X_{jk_2} chosen for the model that it should consider adding $X_{jk_1k_2}$ as well. The appropriate threshold is not so easily described as that used here, and is perhaps most well described as a coding procedure in the spirit of those in **?**.

We hardly expect automated algorithms such as ours to replace the substantively motivated modeling that now characterizes credit modeling. Rather, we expect such automated searches to become a diagnostic tool, used to learn whether data contain more predictors than careful search and science have missed. Given the expanse of available information and the relative sparsity of bankruptcy cases, the *a priori* choice of factors for the analysis is the ultimate determinant of the success of the fitted models. While we have automated this choice via stepwise selection among a host of predictors, such automation will inevitably miss certain predictors, such as say an exponential smooth of some aspect of past payment history. It remains the role for well-informed judgment to add such features to the domain of possible predictors. The analyst often has a keen sense of which transformations and combinations of predictors will be effective, and these should be added to the search space.

An evident weakness in our application is the absence of a standard for comparison. How well can one predict this type of bankruptcy data? Though the validation exercise demonstrates that we have indeed built a model that is predictive, this may not be as good as is possible. Such standards are hard to find, however. Credit rating firms like Experian or Fair-Issacs have invested huge resources in the development of models for credit risk. These models are proprietary, closely-held secrets and are typically specialized to the problem at hand. Without being able to apply their methodology to this data, it is hard to say that we have done well or done poorly. One further suspects that in a real credit scoring model, one might use a time-to-event analysis and leverage the longitudinal structure of this data. Such an analysis, while perhaps more suited to commercial application, would distract from our emphasis on the power of adaptive thresholding with stepwise regression.

There are domains that feature head-to-head comparisons of automatic tools for modeling. These competitions remove the benefit of specialized domain knowledge and afford a benchmark for comparison. We have tailor-made our procedure to be able to compete in this domain. The methodology does not assume any structure on the errors except independence. The methodology does not require complete data. The methodology can run on massive data sets. One such competition has been set up with a collection of data sets stored at UC Irvine (www.ics.uci.edu/∼mlearn). This collection allows experts in the use of each tool to run their favored technology on the same data. Since many of the systems we want to compete against require some degree of hand tuning, the best way to run a competition is to have knowledgeable users for each technique. We plan to test our methodology on all the data sets stored in this collection and compare our out-of-sample performance to that obtained by existing systems.

Acknowledgement

The authors thank Nick Souleles of the Wharton Finance Department and The Wharton Financial Institutions Center for providing the raw data used in this analysis. The Financial Institutions Center also funded the authors during the initial development of this methodology.

References

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," in *2nd International Symposium on Information Theory*, eds. Petrov, B. N. and Csàki, F., Akad. Kiàdo, Budapest, pp. 261–281.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. of the Royal Statist. Soc., Ser. B*, 57, 289–300.

Breslow, N. E. and Day, N. E. (1987), *Statistical Methods in Cancer Research*, vol. II, Lyon: International Agency for Research on Cancer.

Chatfield, C. (1995), "Model uncertainty, data mining and statistical inference (with discussion)," *J. of the Royal Statist. Soc., Ser. A*, 158, 419–466.

Cheng, B. and Titterington, D. M. (1994), "Neural networks: A review from a statistical perspective (with discussion)," *Statistical Science*, 9, 2–54.

Cribari-Neto, F., Ferrari, S. L. P., and Cordeiro, G. M. (2000), "Improved heteroscedasticity-consistent covariance matrix estimators," *Biometrika*, 87, 907–918.

Curnow, G., Kochman, G., Meester, S., Sarkar, D., and Wilton, K. (1997), "Automating credit and collections decisions at AT&T Capital Corporation," *Interfaces*, 27, 29–52.

Fayyad, U. and Uthurusamy, R. (1996), "Data mining and knowledge discovery in databases," *Communications of the ACM*, 39, 24–34.

Feller, W. (1968), *An Introduction to Probability Theory and Its Applications*, New York: Wiley.

Foster, D. P. and Stine, R. A. (1996), "Variable selection via information theory," Tech. rep., Northwestern University, Chicago.

Foster, D. P., Stine, R. A., and Waterman, R. P. (1998), *Business Analysis Using Regression: A Casebook*, Springer.

Freedman, D. A. (1983), "A note on screening regression equations," *American Statistician*, 37, 152–155.

Friedman, J., Hastie, T., and Tibshirani, R. (2000), "Additive logistic regression: A statistical view of boosting (with discussion)," *Annals of Statistics*, 28, 337–407.

Goodnight, J. H. (1979), "A tutorial on the SWEEP operator," *American Statistician*, 33, 149–158.

Gustafson, P. (2000), "Bayesian regression modeling with interactions and smooth effects," *Journal of the Amer. Statist. Assoc.*, 95, 795–806.

Hand, D. J., Blunt, G., Kelly, M. G., and Adams, N. M. (2000), "Data mining for fun and profit," *Statistical Science*, 15, 111–131.

Long, J. S. and Ervin, L. H. (2000), "Using heteroscedastic consistent standard errors in the linear regression model," *American Statistician*, 54, 217–224.

Mallows, C. L. (1973), "Some comments on C_p," *Technometrics*, 15, 661–675.

McQuarrie, A. D. and Tsai, C.-L. (1998), *Regression and Time Series Model Selection*, Singapore: World Scientific.

Miller, A. J. (1990), *Subset Selection in Regression*, London: Chapman& Hall.

Rencher, A. C. and Pun, F. C. (1980), "Inflation of R^2 in best subset regression," *Technometrics*, 22, 49–53.

Thisted, R. A. (1988), *Elements of Statistical Computing: Numerical Computation*, New York: Chapman and Hall.

Tibshirani, R. and Knight, K. (1999), "The covariance inflation criterion for adaptive model selection," *J. of the Royal Statist. Soc., Ser. B*, 61, 529–546.

White, H. (1980), "A heteroscedastic-consistent covariance matrix estimator and a direct test of heteroskedasticity," *Econometrica*, 48, 817–838.