Automated Statistical Modeling for Data Mining

David Stephenson[1]

Abstract. We seek to bridge the gap between basic statistical data mining tools and advanced statistical analysis software that requires an expert operator. In this paper, we explore the automation of the process of statistical data analysis via model scoring functions and search algorithms through the space of statistical models. In particular, we focus on automated modeling using generalized linear statistical models and especially models for categorical data analysis. By automating the process of selecting, building and solving statistical models, a computer can compare hundreds or thousands of possible models for a data set and produce a highly accurate statistical predictor with essentially no intermediate input from the operator. One application of this process is in expanding the statistical components of data mining packages.

1.0. Introduction. Data sets that do not regress well to a linear function of the predictor variables may be better fit by regressing to a polynomial whose terms are products of the predictor variables. In such multi-level models, where the response variable is fit to the sum of products of the predictor variables, we are able to better model the interactions between variables and arrive at a more accurate predictive model. This comes at the price of losing model generality (measured in degrees of freedom) and adding work to the computation of the maximum likelihood estimators so that, instead of least squared methods, we must use general multivariate optimization methods, such as Newton-Raphson.

Our efforts focus on the family of **generalized linear models** (GLZs), which generalize the family of general linear models (GLMs), which, in turn, generalize linear models. Roughly speaking, the main idea behind GLZs is that there is a random response variable Y and a smooth, differentiable link function E such that $E(Y)$ regresses to a polynomial function, g, of the predictor variables. We will focus our attention on cases where g is a multi-level hierarchical function, where summands involve an arbitrary product of predictor variables (i.e., terms contain products of predictor variables with exponents of zero or one). The term **hierarchical** refers to the requirement that a model containing a higher-level interaction term must also include all corresponding lower-level interactions. For example, the existence of the 3-level interaction term XYZ in a model requires the existence of the terms X, Y, Z, XY, XZ, and YZ. The family of generalized linear models encompasses a great many of the data sets in which we at Wagner Associates were interested.

In particular, consider the case where the predictor variables are either naturally categorical (e.g., species, gender, professional occupation, nationality, etc.) or may be binned into categories (e.g., age or weight brackets). If the categorical response variable is multinomial, taking one of several discrete values, then we may fit our sample data using logistic (a.k.a. logit) regression or using probit regression (perhaps the less popular model choice). If the categorical response variable represents an integer count, where the response is assumed to take values drawn from a Poisson distribution, then we may fit the sample data using a loglinear model.

Because a great number of predictor variables are naturally categorical, categorical models have a wide range of applications, such as in the areas of credit scoring, marketing, behavioral studies, and epidemiology (see, for example, Agresti (2002) or Hosmer & Lemeshow).

---

[1]David Stephenson is with Daniel H. Wagner Associates, in Malvern, Pa. Email: dstephenson@pa.wagner.com.
This work was supported under the Office of Naval Research contract N0001499C0424. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research. The process described in this paper is patent pending.
©2004 David Stephenson. Unauthorized reproduction or distribution is prohibited.

1.1. Our Efforts. In this paper, we describe an automated process for determining a preferred multi-level regression model for a data set. This process involves an algorithmic search through the space of possible models and a scoring function used to establish a complete ordering on all available models. In particular, we focus on the application of our process to some generalized linear models associated with categorical data analysis. That is, we focus on logistic regression and loglinear models. Basing our approach somewhat on forward and backward stepwise methods, we develop an approach which, in addition to allowing a great deal of flexibility in model space search algorithms, does not depend on an application of the likelihood ratio test statistic to compare nested models.

In our *Data Mining Tool Set*, we implemented such a process for categorical data analysis, programmed in Java and interfaced with data agents that retrieve data from a Sybase database. The object-oriented nature of the Java programming language nicely handles the commonalities shared by generalized linear models. Using this tool set, we were able to take automated statistical data mining to a more advanced level and to make more powerful statistical analysis available for use by non-statisticians.

< Proprietary information removed here pending patent approval >

4.0. Testing. We ran several simulations to test this automated statistical modeling component of our Data Mining Tool Set. In particular, we wanted to see how accurate the tool would be at recovering the distribution behind a simulation and also to get some idea as to how many sample events would be required in order to build an accurate model from the data set (we measured this in terms of average number of sample events per categorical classification).

4.1. Test Case. An agency interested in sonar applications is in the process of collecting a database indicating situations in which their sonar detected or failed to detect a test object. Their data includes the following likely predictor variables:

1. the Closest Point of Approach (CPA) of the sonar to the object (binned as 0-10, 10-20, 20-50, 50-100, 100-150, and 150-200 yards),
2. the Bottom Type in the area (recorded as type A, B, C, or D),
3. the Clutter Density in the area (recorded as type 1, 2, or 3), and
4. the Sound Speed Profile (SSP) of the water (positive indicates that sound speed is increasing with depth and negative indicates that sound speed is decreasing with depth).

The agency performs numerous test runs, recording when their sonar detects or fails to detect the object. The number of tests may vary for each categorical classification (representing a different combination of the factors) and there may be scenarios for which no tests are run. They would like to use their limited test data to answer the following questions:

1. Which of the factors most heavily affects the probability of detection?
2. For any combination of Bottom Type, Clutter, and SSP, what is the probability of detection as a function of CPA?

4.2. Simulation Results. Because a comprehensive data set was not readily available to us, we ran a simulation to test the statistical tool. We started by constructing a sample probability of detection function that we expected would closely resemble real-world probability curves. The probability of detection was considered to be a non-separable function of the four categories in question and is illustrated in Figure 1 below. We then randomly determined the number of test runs for each of the 192 categorical classifications ((6 CPA bins) $\times$ (4 Bottom Types) $\times$ (4 Clutter Densities) $\times$ (2 SSP profiles)). The number of runs per bin was uniformly distributed between 0 and 9.

For each event in each categorical classification, we used our sample probability of detection function to randomly determine whether the run was successful. The measured data was then fed to our automated statistical modeling software in a simple ASCII format and the module provided a recommendation of the logistic regression model that best fit the data set. The end result was a modeled probability of detection curve for each categorical classification (i.e., each operational scenario).

Once our automated tool determined the recommended statistical model, so that we were equipped with statistical estimators for all categories of interest, we were interested in plotting the probability of detection as a function of closest point of approach (CPA) of the sonar to the submerged object. For purposes of illustration, we will focus on the effects of one particular category, the bottom type in the area of operation. Examples of actual bottom types would be *sand*, *pebbles*, *rocks*, *etc*. Figure 1 illustrates the true detection curves, which formed the basis for our simulation, plotted as a function of CPA and aggregated over the other two categories (clutter type and sound speed). Figure 2 illustrates the same curves when they are constructed using the raw simulation results. Figure 3 illustrates the modeled detection curves, which were produced by the automated statistical tool using only the simulated data set. The curve labeled "All" is the probability of detection curve averaged over all bottom types.

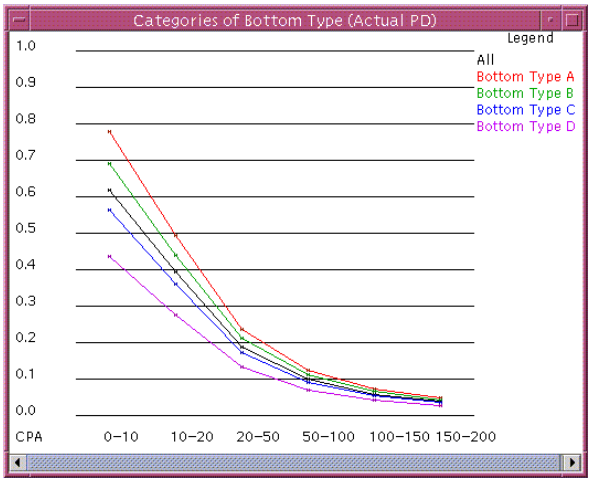


Figure 1. Actual Probability of Detection Curves for Individual Bottom Types (Y-axis indicates probability of detection)

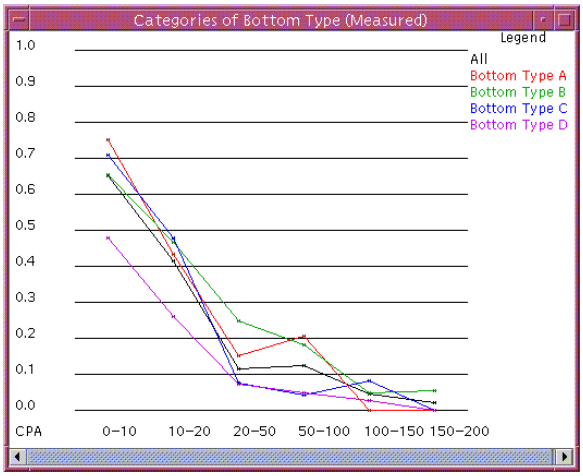


Figure 2. Simulated Sample Probability of Detection Curves for Individual Bottom Types

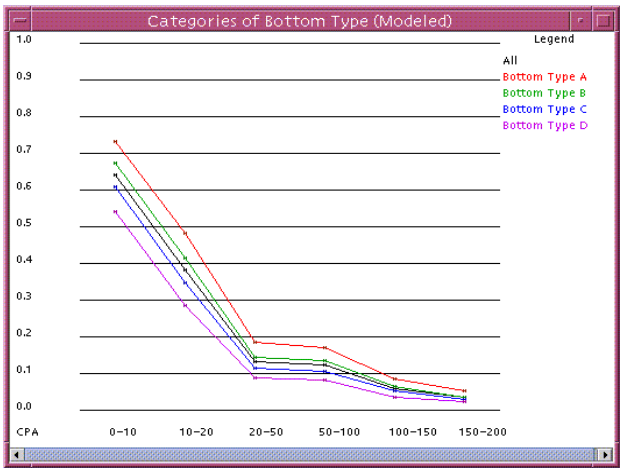
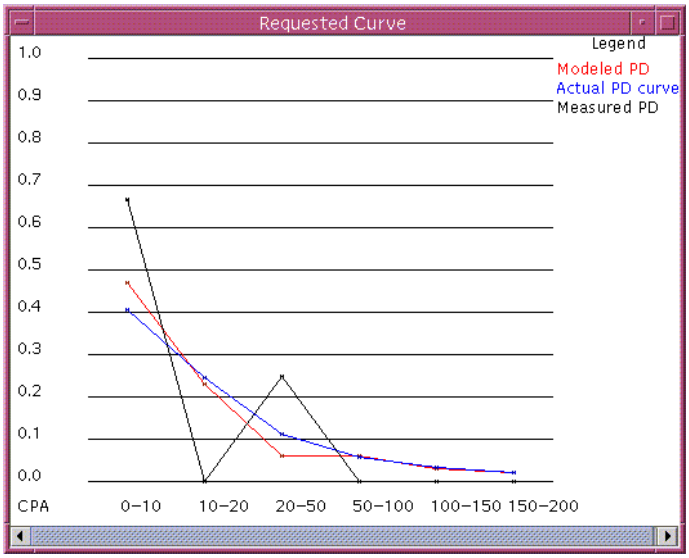


Figure 3. Curves from Computer Generated Logistic Regression Model

Figure 4 presents these three types of curves (actual, sample and modeled) for the specific categorical classification Bottom Type C, Clutter Density 3, and Negative SSP, along with the actual simulation results. Note that several CPA bins had no successful sample results, producing a very jagged sample curve. This sparse measured data gives a very poor indication of the true shape of the probability of detection curve, but the model built from the measured data using the automated statistical modeling tool is very similar to the actual probability of detection curve.



Measured Data for this Scenario		
CPA (yds)	Successes	Failures
0-10	4	2
10-20	0	1
20-50	1	3
50-100	0	4
100-150	0	9
150-200	0	7

Figure 4. Probability Curves for Bottom Type C, Clutter Density 3, Negative SSP

4.3. Analysis. Although this automated process lacks the careful statistical analysis that an experienced statistician could bring to bear, it has the advantage, as a data mining tool, of providing an automated functionality which requires only that the raw data be entered in a standard ASCII format. To this end we have incorporated a set of data agents capable of querying a Sybase database and assembling the required input files for the automated statistical tool, the result being a fully automated statistical data mining system.

Once the system has produced the model, it is a straightforward process to use the model estimates to construct additional information that would be of use to an analyst. In our application, for example, one thing that we do is to automatically check for instances of Simpson's Paradox. In addition, we analyze the modeled variance across categorical classifications and present the analyst with a ranking of which categories seem to most significantly affect the dependent variable. The analyst can use this ranking as a basis on which to view aggregated data (as in Figures 1-3).

4.4. Data Requirements. In order to get a better idea of the amount of measured data needed to produce a good statistical model, we ran a progressive series of simulations. The tests were conducted as follows:

1. We randomly distributed a number of simulated events among the possible categorical classifications, beginning with an average of 0.2 events per categorical classification.

2. For each event, we randomly determined success or failure based on the probability of success for the categorical classification in which it occurred.

3. We fed the resulting success/failure table into the automated statistical tool, which then generated a 'best' statistical model.

4. We computed the error in the modeled probabilities, measured against the true probabilities. We did the same for the probabilities based purely on the measured data. Note that the model was able to provide probabilities even for categorical classifications where no test cases occurred, but such categorical classifications had to be discarded in this comparison.

5. We increased the average number of test cases per categorical classification and repeated steps one through four.

The results are shown in Figure 5 below. They suggest that an average of at least two to three test cases should be gathered for each categorical classification. Figure 5 also illustrates the decrease in error resulting from the use of automatically generated model probabilities rather than measured probabilities.

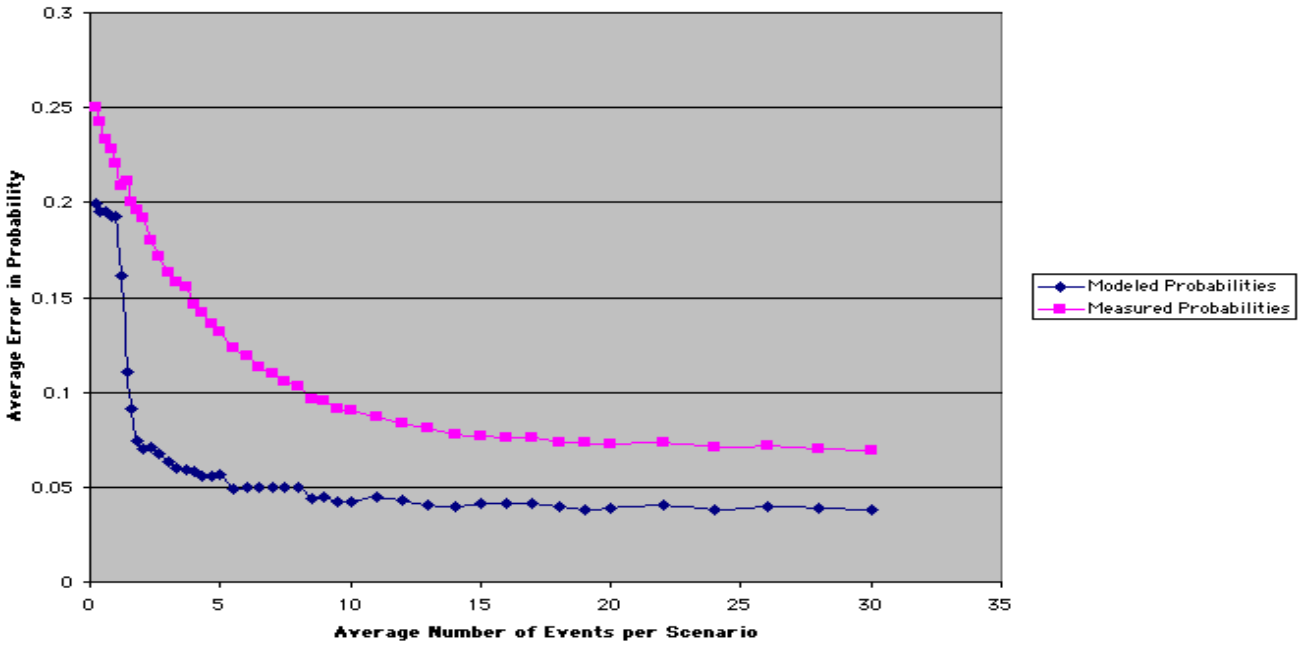


Figure 5. Error in probabilities as a function of sample size

As mentioned above, Figure 5 only reflects measurements from those categorical classifications that contained simulated events. Thus, this figure does not illustrate another significant advantage of modeled data, that it provides estimates for these untested categorical classifications. To give some idea of the sparsity involved in these simulation runs, we include Figure 6 below, which shows the average percentage of categorical classifications for which we had at least one simulated event.

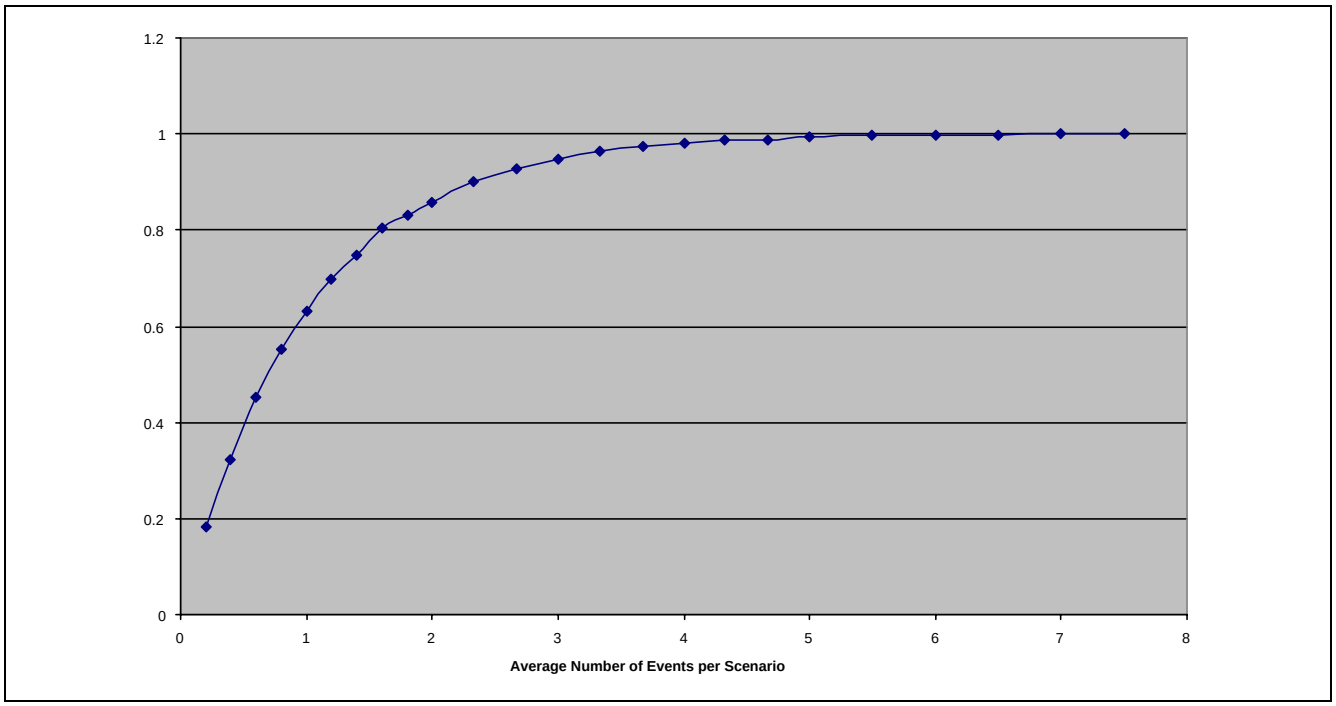


Figure 6. Percent of classifications with sample events as a function of average number of events

5. Conclusion. Initial tests of our statistical data mining process appear promising. We do not have reason to believe that these tests were atypical or that they provided an unfair advantage to the automated process, nevertheless, we would very much like to test the system in a broader range of applications. We would also look to verify our initial indications that modeling is most accurate when there are, on average, at least twice as many events as categorical classification. This minimum level will be necessary in determining when ordinal bins can be subdivided and when they must be consolidated and also in determining the number of distinct categories that can be analyzed without compromising requirements on the minimum average number of events per categorical classification.

There are additional areas in which we would like to refine the tool. Maximum likelihood (ML) estimators for loglinear model parameters can be obtained using numerical optimization techniques or using the Iterative Proportional Fitting method. The latter method proved to be faster and more stable than Newton-Raphson, but was not an option for logistic regression models. Because we are solving for such a large number of MLEs, we wish to implement the fastest, most stable methods for multivariate optimization, and so we hope to improve performance by incorporating additional optimization methods. In addition, there is an interesting application of combinatorial optimization that appears to be useful in circumventing difficulties inherent with zero-event bins, and we would like to explore in more detail the possibility of incorporating this application into the automated statistical modeling process.

References

Agresti, A. 1990. Categorical Data Analysis, first edition. Wiley-Interscience Publication.

Agresti, A. 2002. Categorical Data Analysis, second edition. Wiley.

Billingsley, P. 1985. Probability and Measure, Second Edition, Wiley, New York.

Deming, W. E., and Stephan, F. F. 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Stat.* **11**: 427-444. Sited in Agresti (2002).

Hosmer and Lemeshow. 1989. Applied Logistic Regression, Wiley-Interscience Publication.

Kendall, Stuart & Ord. 1987. Kendall's Advanced Theory of Statistics, Volume 1, fifth edition. Oxford University Press.

Kendall, Stuart & Ord. 1987. Kendall's Advanced Theory of Statistics, Volume 2, fifth edition. Oxford University Press.

Monach, W.R. 1999. Environmental Data Fusion for MCM: Phase I Final Report. Daniel H. Wagner Associates.

StatSoft, Inc. (2004). Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/stathome.html.

Wright, S. and Nocedal, J. 1999. Numerical Optimization. Springer-Verlag.