

## **Logistic Regression Examples Using the SAS System**

By Robert G. Clauss of Fleet Bank

Author: Staff of the SAS Institute

Publisher: SAS Institute Inc.

ISBN: 1-55544-674-4

Order #: P55201

Price: \$24.95

I had the pleasure of using this book while on a two week vacation this past July. The bank had an important regulatory review coming up in August and I was the senior statistician working on the project. We were using logistic regressions to predict the credit worthiness of applicants for small business loans. I am a seasoned statistician but didn't have a lot of experience in using this specific technique. I needed something to illustrate the finer points of analysis and evaluation. Also I was just coming back to SAS after a five-year hiatus. I had worked with SAS for the last 5 months extracting the data, manipulating it, creating analysis tables, and modeling. Being a past master it was coming back quickly but I didn't remember all the tricks. So I had Logistic Regression Examples FED-EXed to the small coastal island off of New Hampshire on which I was vacationing. With my trusty 266Mhz, 32MB RAM, 5GB hard drive laptop I began working with the book in my spare time not devoted to communing with nature.

The organization, level of detail, and focus of the book is superb. There are 19 examples, each contained in its own chapter. The data is provided so that from a 'hands on' point of view you can work along with the examples. I had plenty of my own data on which I wanted to put the final touches of modeling and gain better insight, and this book allowed me to do that. The output is annotated and provides just enough theory, equations, and, especially, answers. This is not just a 'how' book, but also a 'why' book of the requirement or need of a specific technique. I believe this book will be most beneficial for an experienced statistician who needs to analyze dichotomous response variables or a practitioner who is working in the field but needs to know more. Although a beginner might derive some benefit, I'd recommend supplementing the beginners education with some of the excellent references.

One main difference in logistic regressions is how you measure goodness of fit. In ordinary least squares it is pretty straightforward. You look at the  $R^2$ , plots of the residuals by variable, eyeball the predicted versus actual, or evaluate the predictions of a hold out sample. With a logistic model you are not necessarily predicting one the analysis variables but the probability of an event. The analyst defines the event. In my case, I use historical underwriting data for my independent variables and look at actual loan performance over a 12-24 month period. I define an event of the performance being 'bad', such as charging off or being delinquent with a payment by 90 days or more. This leads to a model that predicts the performance of an applicant's loan being 'bad' based on his/her specific underwriting attributes. So getting back to goodness of fit we now need to look closely at misclassification (good vs. bad) issues. We must evaluate and trade off between misclassification of an event (bad), misclassification of a non event (good), and an overall rate of correctness. The relationship of these classifications is not simple or linear. The book does a good job of making and clarifying these distinctions.

In the medical field where the event may be the presence of a disease or response to a treatment the cost of misclassification may be more clear, but in my business case we must balance fairness to the applicant and solvency/profitability which requires more judgment and information. In the end you

must select a probability cutoff with which you can live. The tools and techniques presented here make it much easier.

Example 10 on the receiver operating curves (ROC) was very useful as it plots the relationship between sensitivity, correct prediction of an event, and specificity, the correct prediction of a non event, by probability cutoff. This graph lets you hone in on the cutoff quickly and allows for easy visual comparison between models. The ROC helps you pick the cutoff. But another important question is how well does the data fit the model specified? Example 7 introduces a very clear classification table using the Hosmer and Lemeshow goodness of fit test. This uses the novel null hypothesis that the model fits the data well (as opposed to the typical null hypothesis of no affect). In this case if you reject the null hypothesis, then you discard the model. As it turns out, the table and concept is very easy to explain to subject matter experts and lay people. I was so impressed with the novelty and clarity of the Hosmer and Lemeshow approach that I picked up their text book, Applied Logistic Regression, which is found in the references on page 18.

I strongly recommend Logistic Regression Examples. In a couple of weeks you can gain a year's insight into using SAS and the statistical techniques. Although all the examples are taken from the medical sciences, the techniques are easily applicable to other business problems. I was much more knowledgeable and understood my data and model much better through the use of this book. I was fully prepared to deal with the regulatory review.