

PRODUCING ROC CURVES FOR DISCRIMINANT ANALYSIS USING PROC LOGISTIC

Keith Dunnigan
Trilogy Consulting Company

Introduction:

It often arises in various types of studies that one wishes to use explanatory variables to attempt to categorize a subject (or item) into one of two possible groups. One method of doing this is to run an experiment where subjects belonging to both groups can be measured on each of the explanatory variables. The data attained from the experiment can be used to build a classification rule using Logistic regression (and SAS/STAT® Proc Logistic). Once this is done then the SAS® System may also be used to assess the performance of the classification rule.

To illustrate the technique and develop the example SAS programs, an example from the Geriatrics literature is used. In this example (Carr et al, 1997) a sample of elderly persons were selected from the Alzheimer's Disease Research Center at Washington University in St. Louis and given a battery of 10 tests (variables x1 - x10). It was known whether they were demented (variable y=1) or nondemented (variable y=0). One of the methods of discrimination studied was the logistic regression of y on x1-x10. More details aren't relevant to this example, but are available to the reader in the literature.

Using Logistic Regression to Develop a Classification Rule:

First define the states of a dependent random variable as the two possible groups into which we wish to classify a subject. Here we arbitrarily define the states of a random variable Y as 1 and 0. We also define the probability that Y = 1 as p. The logit(y) is defined then as $\log(p/(1-p))$ and the logistic regression equation becomes:

$$\text{Logit}(Y) = \beta_1 + \beta_2 + \dots + \beta_n$$

Thus the probability that the subject has variable Y=1 may be estimated as follows:

$$P(Y = 1) = \frac{e^{\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n}}{1 + e^{\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n}}$$

A decision rule may be developed for classifying subjects based on this estimated probability equation. If a cutpoint is chosen for the estimated probability, then a decision rule may be developed from it for classifying subjects. For example, if a cutpoint of 51% is chosen, then any subject with an estimated probability of greater than 51% of having Y=1 could be so classified. Alternately cutpoints of 60%, 70%, etc. could be chosen. The estimated parameters from a standard proc logistic can be substituted into the above equation to develop a classification rule.

Evaluating the Performance of the Classification Rule:

Numerically how well the classification rule works is evaluated by comparing the classification of the rule with the true classification of the individual. There are many different statistics that can be used to quantify this, among the most common are sensitivity (percentage of events correctly classified), specificity (percentage of nonevents correctly classified), false positive rates (percentage of subjects classified as events that were not), and false negative rates (percentage of subjects classified as nonevents, that were not). There are a number of similar but different definitions of these quantities which should always be made clear to the reader in the write-up of a study. I am keeping with the definitions that the SAS system and literature uses (2), which are shown below.

First, a few preliminary definitions are in order:

Define:

```

tev = total number of events in the sample
tnev = total number of nonevents in the sample
cev = the number of events the rule correctly identifies
iev = the number of events the rule incorrectly identifies

cnev = tnev - iev = correctly identified non events
inev = tev - cev = incorrectly identified non events

```

Then the sensitivity, specificity, false positive and false negative rates can be defined as:

```

sensitivity= cev/tev
specificity= cnev/tnev
false positive rate (assuming sample proportions) = iev/(cnev + inev)
false negative rate (assuming sample proportions) = inev/(cnev + inev)

```

And the percent correct as:

```

%correct = (cev + cnev)/(tev + tnev)

```

As an example consider the Geriatrics Driving Signs study mentioned earlier. If Y=1 depicts the event that the elderly subject tested is found to be demented and the scores of the 10 questions on the test are denoted by X1 - X10 then the SAS code is as shown below:

```

proc logistic data=testdat descending;
model y = x1 x2 x3 x4 x5 x6 x7 x8 x9 x10/ ctable pprob=0.51;
run;

```

In Proc Logistic the probability cutpoint for constructing the classification rule is specified by the pprob option and the classification table used for evaluating the performance of the classification rule produced in the output by using the ctable option. When using a two value dependent variable, SAS Proc Logistic always models the probability that it equals the lower value (in this case 0). Since we wish to model the probability that Y=1, we use

the descending option to model this. The classification table produced by the above code for the geriatrics example is shown below:

Proc Logistic ctable Printed Output for the Geriatrics Driving Signs Study pprob = 0.51, pevent option not used										
The LOGISTIC Procedure										
Classification Table										
Prob Level	Correct		Incorrect		Percentages					
	Event	Non- Event	Event	Non- Event	Correct	Sensi- tivity	Speci- ficity	False POS	False NEG	
0.510	32	60	6	10	85.2	76.2	90.9	15.8	14.3	

The earlier formulas for false positive and negative rates assume that the proportion of subjects with variable $Y=1$ in the sample is the same as that in the general population. In most cases this is not plausible, since the sample is usually drawn in such a way as to assure a sizable occurrence of both values of the dependent variable. If the true percentage of subjects with variable $Y=1$ in the population is denoted by p_0 , then the false positive and negative rates become:

$$\begin{aligned}\text{false positive rate:} &= [(1-\text{spec})*(1-p_0)]/[1-\text{spec} + p_0 * (\text{sens} - 1 + \text{spec})] \\ \text{false negative rate:} &= [(1-\text{sens})*p_0]/[\text{spec} - p_0 * (\text{sens} + \text{spec} - 1)]\end{aligned}$$

The SAS System assumes that the sample proportion is the population proportion, unless the pevent option is used. Adjusting the previous code to assume that 5%, 25%, or 40% of the population has dependent variable $Y=1$ produces the following code:

```
proc logistic data=testdat descending;
model y = x1 x2 x3 x4 x5 x6 x7 x8 x9 x10/
ctable pprob=0.51 pevent=0.05 0.25 0.40;
run;
```

Remember that the pevent option only affects the false positive and negative rates, and does not affect the other quantities calculated. The above code produces the following classification table for the Geriatrics Study:

Proc Logistic ctable Printed Output for the Geriatrics Driving Signs Study
pprob = 0.51, pevent= .05, 0.25, 0.4

The LOGISTIC Procedure

Classification Table

Prob Event	Prob Level	Correct		Incorrect		Percentages				
		Event	Non- Event	Event	Non- Event	Correct	Sensi- tivity	Speci- ficity	False POS	False NEG
0.050	0.510	32	60	6	10	90.2	76.2	90.9	69.4	1.4
0.250	0.510	32	60	6	10	87.2	76.2	90.9	26.4	8.0
0.400	0.510	32	60	6	10	85.0	76.2	90.9	15.2	14.9

Evaluating the Operating Point of the Classification Rule:

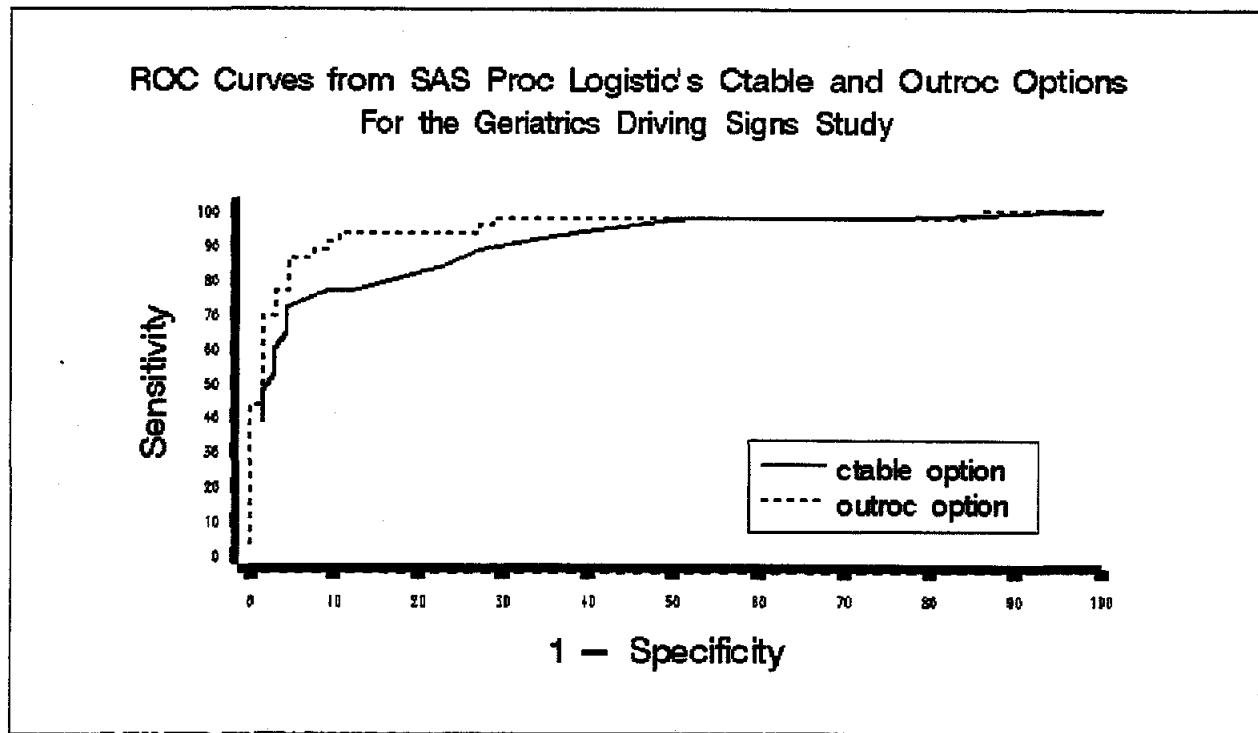
As noted earlier, a different rule (or operating point) can be obtained from the logistic regression model for each probability cutpoint used to classify a subject. For each of these operating points, the sensitivity and specificity can be calculated. Graphs of sensitivity versus (1 - specificity) are known as "Receiver Operating Characteristics" and can be used to determine a desired operating point (cutoff probability) to use in practice.

At present there are two ways of producing ROC curves using Proc Logistic. The first way is to proceed as we have thus far and run a separate Proc Logistic with the ctable option for each of a number of different probability cutpoints, taking the resulting sensitivities and specificities down by hand, and then plotting them. The other way is to use the outroc option in Proc Logistic which produces an output dataset containing the sensitivities, specificities, and other statistics for a number of probability cutpoints and can be plotted to produce an ROC curve. The code to implement this second method this for the geriatrics example is shown below along with a portion of the output dataset produced:

```
proc logistic data=testdat descending;
  model y = x1 x2 x3 x4 x5 x6 x7 x8 x9 x10/
    ctable pprob=0.51 pevent=0.40 outroc=roc;
run;
```

Printout of Dataset Produced by the Proc Logistic outroc Option (Geriatrics Driving Sign Study)							
OBS	_PROB_	_POS_	_NEG_	_FALPOS_	_FALNEG_	_SENSIT_	_1MSPEC_
1	0.99940	1	66	0	41	0.02381	0.00000
2	0.99928	8	66	0	34	0.19048	0.00000
3	0.99874	9	66	0	33	0.21429	0.00000
20	0.85712	25	65	1	17	0.59524	0.01515
21	0.84610	26	65	1	16	0.61905	0.01515
22	0.81242	27	65	1	15	0.64286	0.01515
23	0.77416	28	65	1	14	0.66667	0.01515
51	0.17649	39	48	18	3	0.92857	0.27273
52	0.15671	40	48	18	2	0.95238	0.27273
53	0.14629	40	47	19	2	0.95238	0.28788
85	0.012706	41	10	56	1	0.97619	0.84848
86	0.012447	42	9	57	0	1.00000	0.86364
87	0.011236	42	8	58	0	1.00000	0.87879
88	0.010205	42	7	59	0	1.00000	0.89394
89	0.009757	42	6	60	0	1.00000	0.90909
90	0.006982	42	1	65	0	1.00000	0.98485

While the second method is certainly easier than the first, the results produced are *not* the same. This is because the algorithm used to calculate the sensitivities and specificities is different for those produced by ctable than those produced by the outroc option. If the estimated parameters of a logistic regression are used to estimate a probability which is in turn used to classify a subject (as we are doing), then the sensitivities and specificities will be somewhat biased since the subject was used in estimating the regression parameters. If for a given subject the regression parameters are estimated *excluding* the subject being classified, then the resulting classifications and calculated quantities will be less biased. When the ctable option is specified SAS calculates the sensitivities and specificities based on a jackknifing technique similar in spirit to this. However the outroc option output does not use this technique. Thus the outroc option tends to produce a more optimistic (biased) ROC than does the ctable output. If both methods are used for the Geriatrics driving signs study, then the ROC curves calculated are shown below:



In using ROC curves to aid in choosing a classification rule, we usually look to balance the tradeoff between sensitivity and specificity. Using the outroc option in the geriatrics example would probably result in choosing an operating point in the above figure with sensitivity around 88% and 1-specificity around 9% (sensitivity around 91%). However a similar operating point using the ctable option reveals a sensitivity of about 79% and 1-specificity of around 17% (specificity of about 83%). Thus for this study the differences seen are not trivial with the outroc option producing statistics about 8 or 9 percent too optimistic. Unfortunately as of SAS 6.12 there is no way of getting the jackknifed statistics directly out of Proc Logistic and we must resort to using method 1.

In summary, SAS Proc Logistic provides very useful software for performing discriminant analysis for studies where a rule needs to be developed for classifying subjects into one of two classes. It also is useful for evaluating the performance of the classification rule through the calculation of statistics and ROC curves. One must be careful however in choosing which way to apply it for constructing ROC curves and to assure that the study write-up documents the method used to construct them.

References:

SAS Institute Inc., SAS/STAT Software: Changes and Enhancements through Release 6.11, Cary NC: SAS Institute Inc., 1996. pg. 420-422, 426-427.

Carr D., LaBarge E., Dunnigan K., and Storandt M. "Differentiation of Drivers with Dementia of the Alzheimer's Type by the Traffic Sign Naming Test", *Journal of Gerontology: Medical Sciences*, 1998, Vol. 53A, M135-139

SAS, SAS/AF, SAS/ASSIST, SAS/CONNECT, and SAS/FSP are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

For further information contact:

Keith Dunnigan
e-mail: Harry.Dunnigan@mkg.com