Paper SP03-2009

Illustrative Logistic Regression Examples using PROC LOGISTIC: New Features in SAS/STAT® 9.2

Robert G. Downer, Grand Valley State University, Allendale, MI
Patrick J. Richardson, Van Andel Research Institute, Grand Rapids, MI

ABSTRACT

PROC LOGISTIC has many useful features for model selection and the understanding of fitted models. The standard generated output will give valuable insight into important information such as significant variables and odds ratio confidence intervals. However, proper utilization of output files, graphical displays and relevant options can further enhance justification of model choice and understanding of model fit. In this fairly general paper, a variety of logistic regression topics such as model building, model fitting and the ROC curve will be reviewed. The discussion will introduce the “PLOTS=” option, as well as the ROCCONTRAST statement as new features which are available in SAS/STAT® 9.2.

INTRODUCTION

Logistic regression is a common and popular technique for describing how a binary response variable is associated with a set of explanatory variables. The data can come in one of two forms. In one format, one will have the number of successes out of a sample of independent trials (i.e. a set of binomial counts). In the other and most common, one will have an observation as a 1 or 0 (success or failure) for an individual trial and the row of data corresponds to a single individual/subject.

THE MULTIPLE LOGISTIC REGRESSION MODEL

We consider the log odds of success versus failure p/(1-p) as a linear function of the predictor variables and the logistic regression model for predictors $X_1 \ldots . X_k$:

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 x_1 + \ldots \beta_k x_k$$

The multiple logistic regression model above is fit through maximum likelihood in PROC LOGISTIC. Standard output from PROC LOGISTIC includes these maximum likelihood parameters $\hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_k$ and their standard errors. The subsequent fitted probability $\hat{p}$ and its standard error can be obtained for each observation. Graphical display of the estimated probability function versus each predictor is a useful display particularly for continuous responses such as dosage or age. Odds ratios are also frequently an emphasis of a study or a study report. Estimated adjusted odds ratios for a given predictor are provided by PROC LOGISTIC as well as approximate confidence intervals.

THE RECEIVER OPERATING CHARACTERISTIC CURVE (ROC)

The practicality of a logistic regression is often evaluated in terms of its predictive ability. In a logistic regression, a two by two table classification table can be created for any cut-off value of the fitted probability and hence the sensitivity and specificity are then available for this particular table. The fraction calculated as count of predicted positives divided by the actual total of positives is the sensitivity and the fraction calculated as the count of predicted negatives divided by the total negatives will be the specificity.

A series of cut-offs from 0 to 1 and the resulting two by two tables will give plotted pairs (1-specificity, sensitivity). These pairs constitute the Receiver Operating Characteristic (ROC) curve. Points far above the 45 degree line are desirable and one hopes to have this curve rise as quickly as possible from the origin. The 45 degree line in the unit square would correspond to an area under the curve (AUC) of 0.5 and represents where the fraction true positives and false negatives are equal and hence the diagnostic would be no better than flipping a coin.

The concordance index, denoted “c”, as provided by PROC LOGISTIC gives the area under the curve (AUC) for a given model. A nonparametric approach to the comparisons of correlated ROC curves was proposed by Delong et al. and is utilized in the new syntax of PROC LOGISTIC. Contrasts are constructed and comparisons are made using the empirical ROC curves of specified models.

NEW FEATURES OF PROC LOGISTIC IN SAS/STAT® 9.2

SAS/STAT® 9.2 contains valuable additions to PROC LOGISTIC which enhance the visualization of model fit and comparisons between two or more models. The ROC and ROCCONTRAST statements provide this enhanced functionality. This paper will explore the application of these new statements, review basic model fitting strategies using PROC LOGISTIC and illustrate the utilization of receiver operating characteristic (ROC) curves. The new ODS graphics capabilities of SAS® 9.2 can provide production quality graphics of the estimated fitted probabilities and ROC curves.

EXAMPLE DATA SETS

The first utilized data set is originally from Gaylor and also shown by Rao. For full citations, see the reference section. In the SAS code to follow, the data set is referred to as “toxdat” and there are only six observations with three variables. For each toxin dosage, a count is given of the number of test animals with tumors and the total number of animals tested is also provided. The example illustrates logistic regression for grouped binomial counts and illustrates the usage of PROC LOGISTIC for a single continuous predictor (dose).

The second dataset used is from Pine et al. from an article published in, Archives of Surgery. This particular study looked at the incidence of organ malfunction and death for patients who had intra-abdominal sepsis found during a surgical procedure. Variables collected for analysis included age, as well as the binary variables, malnutrition, alcoholism, shock and bowel infarction, (where 0 indicated that the symptom was absent and 1, that it was present). In the SAS data set referred to as “pinedat”, the response variable, survival, was coded as 0=Alive and 1=Deceased. In this work, we have used the authors’ original data and applied PROC LOGISTIC using the ROC and ROCCONTRAST statements to assist in determining the best model fit for the available predictor variables.

EXAMPLE 1 (ILLUSTRATING PLOTS = EFFECT, PLOTS = ROC OPTION)

With ODS graphics invoked in SAS/STAT 9.2, consider the following run of PROC LOGISTIC:

```
proc logistic data=toxdat plots=EFFECT plots=ROC;
model count/n = dose / outroc = rocout;
output out=estimated predicted=estprob l=lower95 u=upper95;
run;
```

The plots option produces the following graphics.

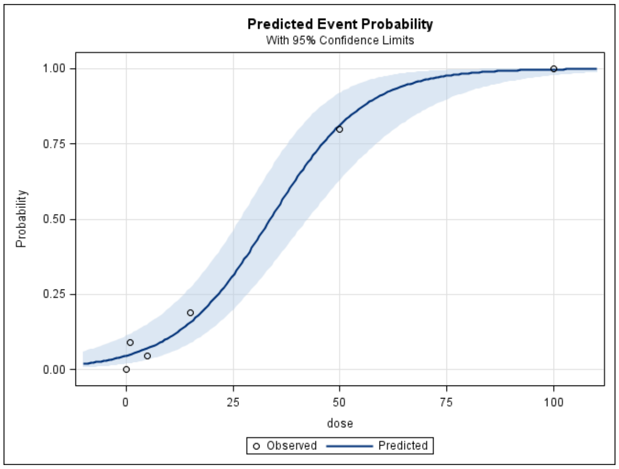


Figure 1 – Effect Plot for toxdat data

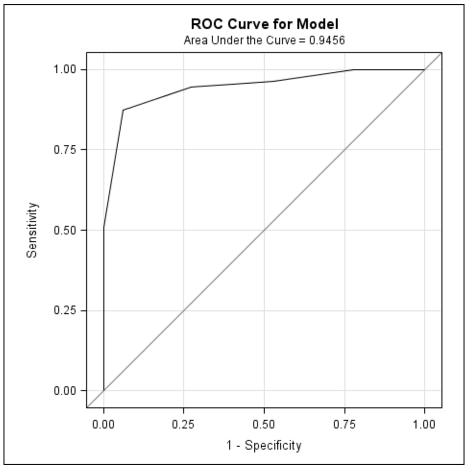


Figure 2 – ROC Curve for toxdat data

Figure 1 is the ODS graphics display from the PLOTS = EFFECT option on the PROC LOGISTIC line in SAS® 9.2. The logistic curve is displayed with prediction bands overlaying the curve. The ROC Curve, shown as Figure 2, is also now automated in SAS® 9.2 by using the PLOTS=ROC option on the PROC LOGISTIC line.

SAS® 9.2 eliminates the need for the output data set creation in order to obtain and plot the fitted logistic curve and ROC curve. The code given above also shows the traditional methodology of obtaining the estimated probabilities. These fitted probabilities and other possible quantities of the event of interest are placed in the output data set `estimated` via the OUTPUT statement in PROC LOGISTIC. The keywords PREDICTED =, LOWER=, UPPER= will name variables for the estimated probability as well as upper and lower confidence limits for these estimated probabilities.

Using the output data set, the following run of PROC GPLOT in SAS/GRAPH® could be used for display of the fitted S-shaped logistic curve.

```
symbol1 interpol=spline v=circle
c=red pointlabel=('#dose') ;
proc gplot data=estimated ; /*generates logistic curve*/
plot estprob*dose;
run;
```

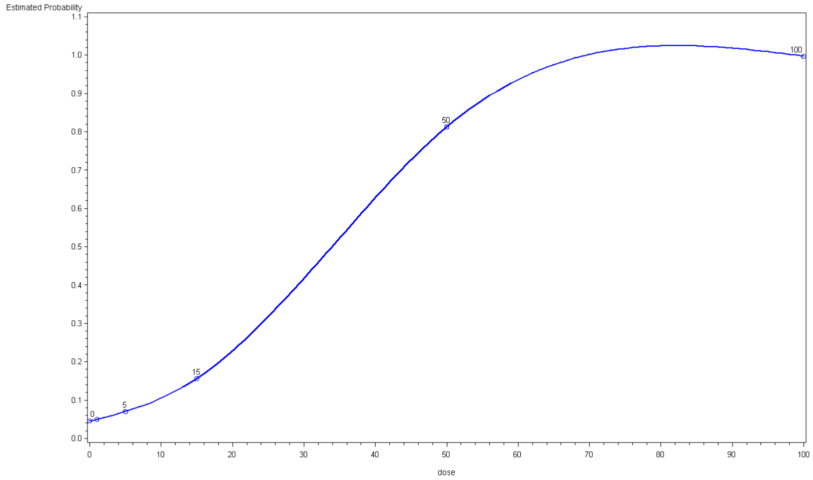


Figure 3 – GPLOT creation of logistic curve

Figure 3 suggests that PROC GPLOT using a spline interpolation for the estimated probabilities has some potential difficulties. This fitted logistic curve exceeds 1 here as there are only a few sparse observations. This impossibility degrades the ability of the curve to always show reasonable fitted probabilities through a quick visual reference. The new PLOTS= option in SAS® 9.2 eliminates this difficulty in the high quality display above in Figure 1.

Additionally, although the ROC curve given below (from PROC GPLOT) provides appropriate information it is still a less appealing graphic and it took additional effort to create it. The new ODS graphics capabilities coupled with the PLOTS= option in LOGISTIC quickly creates aesthetically pleasing graphics, suitable for publication or presentation.

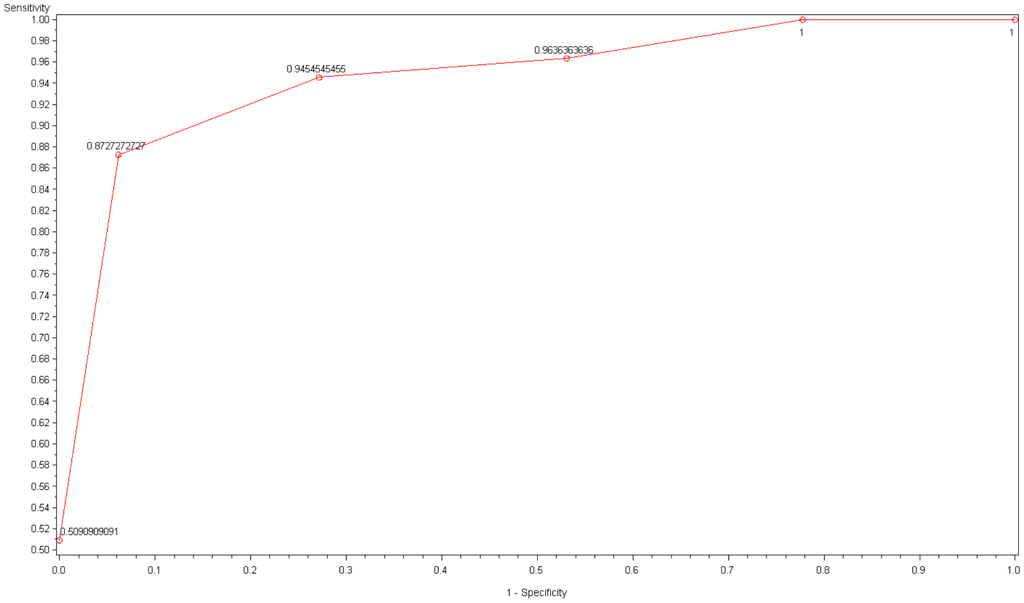


Figure 4 – GPLOT display of ROC curve

A variety of plots can be obtained using the EFFECTS option, as well as setting the PLOTS= option to EFFECTS. In the case of continuous predictors, a plot of the estimated probability versus the first continuous predictor (fixing all other continuous covariates at their means) is displayed. In this first example, there is only a single continuous predictor dose. For factor variables specified in a CLASS statement, the estimated probability versus the first CLASS variable at each level of the second CLASS variable, if any, (holding all other CLASS predictors at their reference levels) is displayed.

Prior to SAS® 9.2, the ROC curve for a single model would have been typically constructed by first obtaining the sensitivity and specificity from an output data set as generated by the OUTROC= option on the model statement (output data set roc out above in Example 1). Subsequently, one might again use SAS/GRAPH® to create the ROC curve. As seen in Figure 2 above, the plots=ROC option on the PROC LOGISTIC LINE creates a high quality display with very little extra effort. Additional display features are now possible such as using the ID= option to label certain points on the ROC curve. In the modeling process, if automated selection is utilized via the option SELECTION= , then an overlaid plot of all the ROC curves for each step will be displayed. If ROC statements are specified then ROC curves for ROC statement models will be displayed.

EXAMPLE 2 (ILLUSTRATING ROC AND ROCCONTRAST STATEMENTS)

For the Pine et al. (1993) data in which the response y is binary, let's first consider the fit of a logistic regression with only a single predictor, age.

```
ods graphics on;
proc logistic descending data = pinedat plots =EFFECT plots=ROC(id=prob);
model y = age  ;
roc 'Age' age;
run;
ods graphics off;
```

The DESCENDING option on the PROC LOGISTIC line models the probability of a 1 (death) as it is the higher value of the numerical response variable (a ref= option is also commonly used on the model statement). Age is significant in the fitted model (p=.0006) and the (unadjusted) odds of death is estimated to be between 1.023 and 1.085 times for each 1 year increase in age. In Figure 5, the estimated logistic curve (from the plots = EFFECT option) shows that the estimated probability of death for a patient of age 65 is approximately 0.25 (with confidence limits as shown).

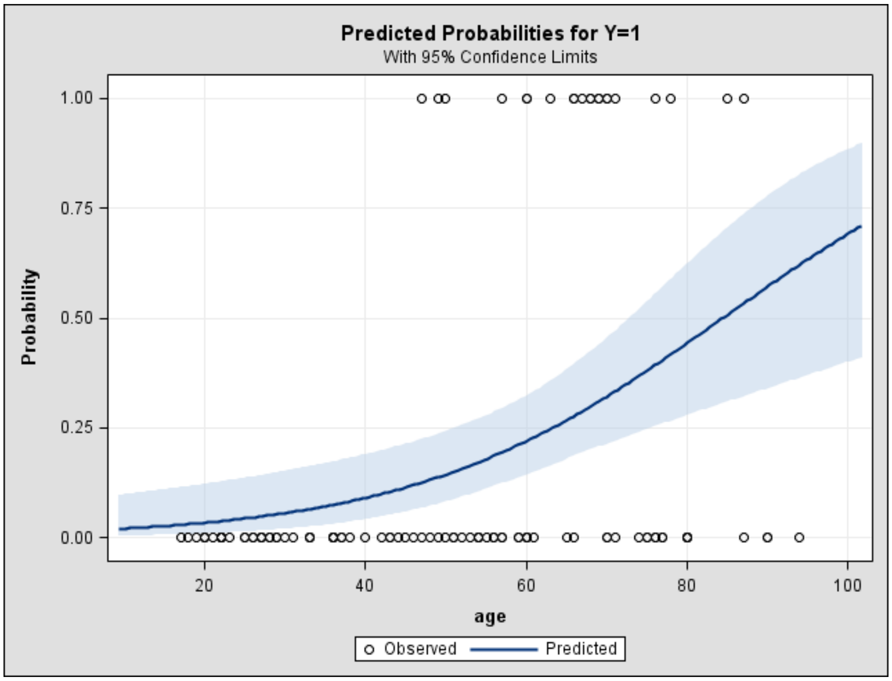


Figure 5 -Estimated Logistic Curve for Age Only Model

In Figure 6, we see the ROC curve with the concordance index c of .7678 (area under the curve). It shows that for the (labeled) cut-off of estimated probability 0.25, we can quickly see that we have a sensitivity of around 0.73 and 1-specificity near 0.23. Since there is only one model (age as a single predictor), the 'ROC curves for comparison' plot (generated by having the ROC statement) only contains this same ROC curve.

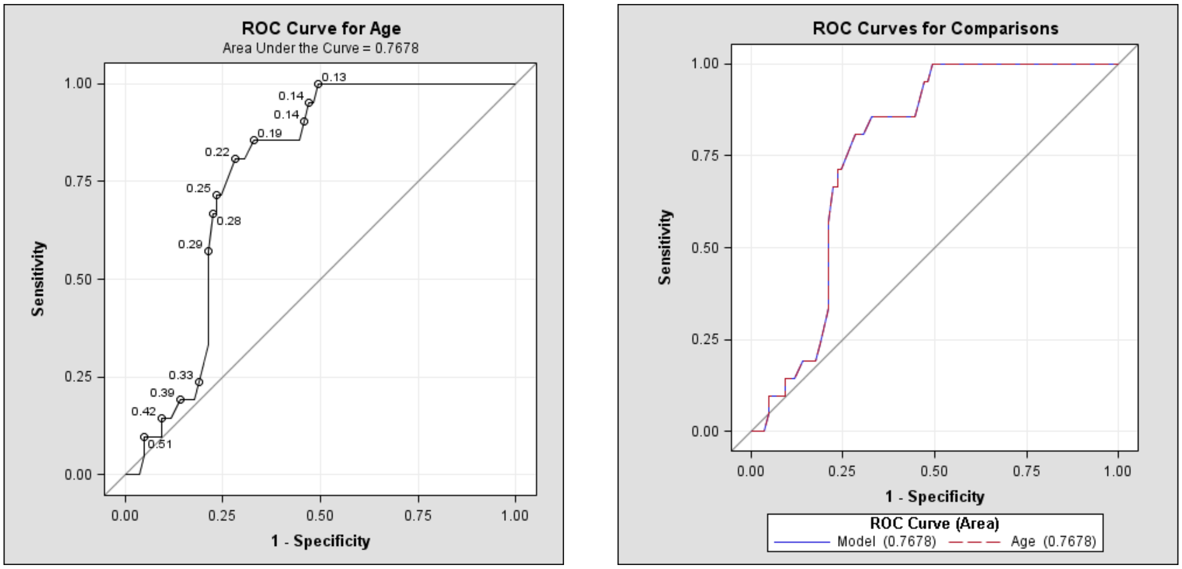


Figure 6 – ROC displays for model with age as single predictor.

Now let's see the impact on the fit and the resulting displays as we add the categorical predictor alcoholism to the model (coded 1 for presence):

```
ods graphics on;
title1 'Age, alcohol model(s)';
proc logistic descending data = pinedat plots=effect plots=ROC(id=prob);
class alc /descending param = glm;
model y  =   age alc ;
roc 'Age' age;
roc 'Alcohol' alc;
roc 'Age & Alcohol' age alc ;
roccontrast reference('Age & Alcohol') /estimate ;
run;
ods graphics off;
```

There are now 3 possible models (one model for each of the single predictors plus the model with both predictors) and hence there are three fits available for comparison. In the fit with alcoholism as the sole predictor, the plots = EFFECT options gives the overlaid fitted logistic curves for each level of the alcohol variable. The ROC statements coupled with the plots = ROC option results in display of each of the ROC curves and the 'ROC curves for Comparisons' contains each of them in an overlaid display (shown in Figure 7 along with the ROC curve for the model with both predictors).

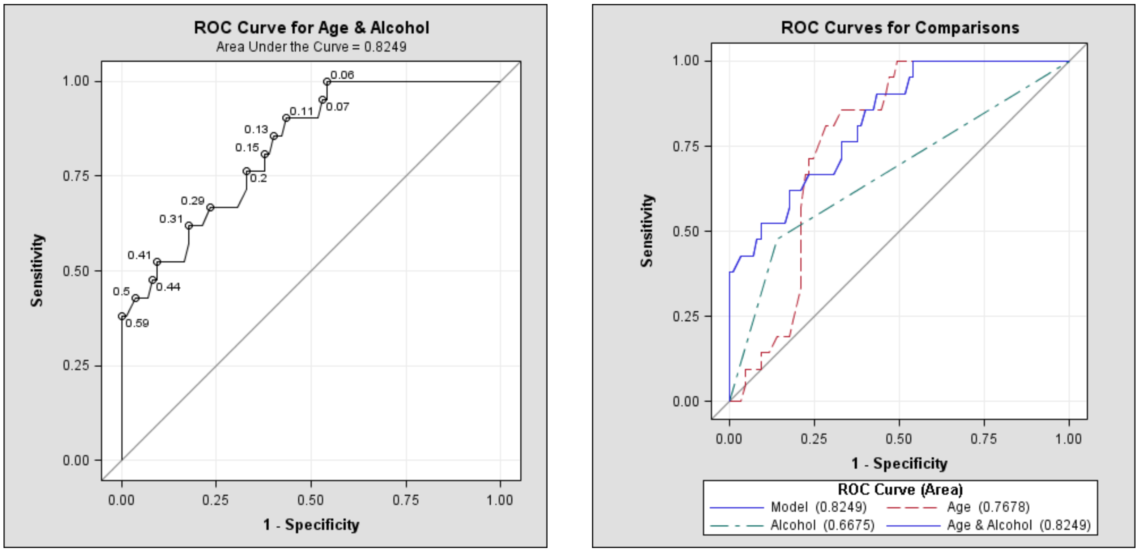


Figure 7 - ROC curve for age & alcohol model, display comparison of all 3 ROC curves

In the model with alcoholism alone, alcoholism is significant (p=.0014) and in the model with both predictors present, each predictor is significant (p=.0009 and p=.0018 respectively). The overlaid ROC comparison display suggests that from a practicality standpoint, the model with only alcoholism is likely to be less effective in prediction (as the ROC curve doesn't remain high and the AUC is only 0.6675).

The ROCCONTRAST statement and its ESTIMATE option provide the following results in Table 1:

ROC Contrast Test Results			
Contrast	**DF**	**Chi-Square**	**Pr > ChiSq**
Reference = Age & Alcohol	2	47.2029	<.0001

ROC Contrast Rows Estimation and Testing Results						
Contrast	**Estimate**	**Standard Error**	**95% Wald Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Model - Age & Alcohol	0	.	.	.	.	.
Age - Age & Alcohol	-0.0571	0.0438	-0.1430	0.0287	1.7019	0.1920
Alcohol - Age & Alcohol	-0.1574	0.0400	-0.2358	-0.0791	15.5103	<.0001

Table 1 – ROC contrast results for 3 models involving Age and Alcohol

The overall comparison test suggests that at least one difference exists among the 3 models (with the Age & Alcohol model as the reference). The ESTIMATE option shows that this difference exists between the reference model and the model with alcohol as the single predictor. This result was fairly evident from the initial comparison plot with the alcohol model only standing out as being lower (away from 1.0) and it indeed had the lowest AUC.

In the following code, variable names have been suppressed for brevity and space. The new variables (with full names in parentheses) are as follows: x1 (shock), x2 (malnutrition), x3 (alcoholism), x4 (age) and x5 (bowel infarction).

INVESTIGATION OF THE FINAL MODEL

A formal selection of a final model often involves incorporates several ideas and criterion. Among the priorities impacting the decision-making may be previous research involving the predictors and model simplicity. In this data set with 5 possible predictors, there are many possible sub-models involving interactions. In theory, one could consider a 5 way interaction term, 5 possible 4 way interactions terms, and 10 possible 3 way interactions. There will be further subsets of these combinations possible for inclusion (e.g. 4 models involving 4 of the 5 four-way interactions). For this data set, exploratory investigation of a model with all main effects and interactions of 3 or more factors suggested issues with maximum likelihood estimation of many coefficients for the fitted models (i.e. at least partial separation of points). There are a limited number of combinations of the binary categorical predictor variables X1, X2, X3 and X5 , only 106 total observations and only 21 observations of the event modeled (death). Hence there are very few cases or observations associated with the many combinations of factors.

A model which included all main effects plus all two way interactions was a logical starting point for model selection and ROC curve comparison for this data set. Since our focus here was on ROC curve comparison, a liberal inclusion significance level of 0.1 was used in the following run with backward automatic selection in order to extract possible models for ROC comparison. All main effects and the two way interactions of x2*x3 and x3*x4 remained using this criterion. The x3*x4 interaction remained only due to the liberal significance level. ROCCONTRAST was then used to compare a main effects model (‘main’) and the three models involving combinations of these two interactions:

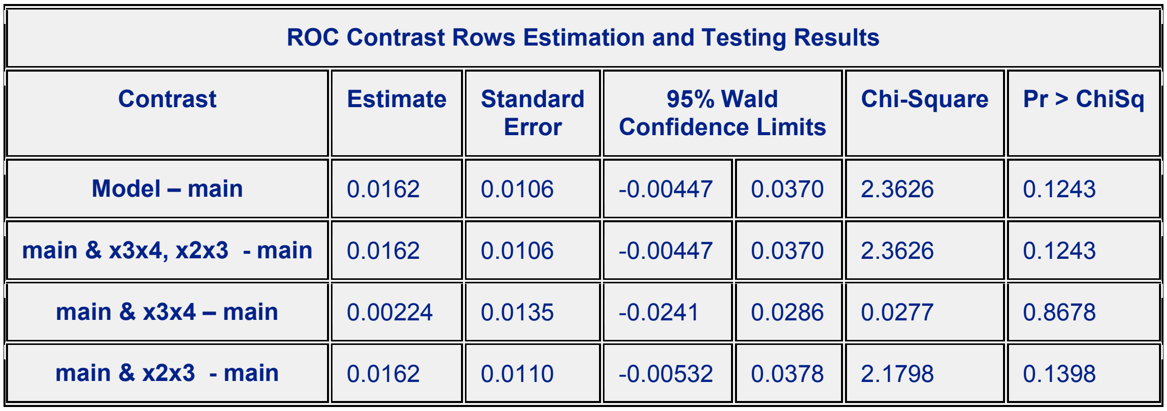
```
ods graphics on;
proc logistic descending data = pinedat;
class x1 (ref = 'A') x2 (ref = 'A') x3 (ref = 'A') x5 (ref = 'A') / param = ref;
       model y (event='Dead') =   x1 x2 x3 x4 x5 x2*x3 x3*x4 ;
  roc 'main & x3x4, x2x3'  x1 x2 x3 x4 x5 x2*x3 x3*x4 ;
  roc 'main & x3x4' x1 x2 x3 x4 x5 x3*x4;
  roc 'main & x2x3' x1 x2 x3 x4 x5 x2*x3;
  roc 'main' x1 x2 x3 x4 x5 ;
  roccontrast reference('main') / estimate e;
run;
ods graphics off;
```

As part of the resulting ODS output, one can note the area under the curve (AUCs) as given in Table 2:

ROC Model	Area
main & x3x4, x2x3	0.9457
main & x3x4	0.9317
main & x2x3	0.9457
Main	0.9294

Table 2 Area under ROC curve for four candidate final models

None of the ROC curve comparisons are significant as given by Table 3 below.



ROC Contrast Rows Estimation and Testing Results						
Contrast	**Estimate**	**Standard Error**	**95% Wald Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Model – main	0.0162	0.0106	-0.00447	0.0370	2.3626	0.1243
main & x3x4, x2x3 - main	0.0162	0.0106	-0.00447	0.0370	2.3626	0.1243
main & x3x4 – main	0.00224	0.0135	-0.0241	0.0286	0.0277	0.8678
main & x2x3 - main	0.0162	0.0110	-0.00532	0.0378	2.1798	0.1398

Table 3 – ROCCONTRAST results for four final model candidates

The best predictive power at an AUC of 0.9457 (see Table 2) and was obtained using the model of all main effects plus the interaction x2*x3 (alcoholism*malnutrition). If one goes to a separate investigation strictly based on backward elimination and significance level .05, this is also the resulting model selected.

Bridges et al. point out there is an association between alcoholism and malnutrition in a 1999 paper. This established link and maximum AUC gives us ample justification for using the following as our final logistic regression model:

$$\begin{aligned}\hat{\mathrm{P}}(\mathrm{Y{=}1|x}) &= -10.6648 + 3.6930(\mathrm{shock}) + 2.7663(\mathrm{malnutrition}) + 5.1725(\mathrm{alcoholism})\\ &+0.0891(\mathrm{age}) + 3.3840(\mathrm{infarction}) - 3.2816(\mathrm{malnutrition} * \mathrm{alcoholism})\end{aligned}$$

The resulting ODS graphic ROC curve comparison of the four possible final models gives a display of ROC curves that are very close together and is not presented here.

A re-run of PROC LOGISTIC was performed for this paper in order to show varying levels of predictive ability and a visual comparison of ROC curves relative to the final model. Age (x4) is the best single predictor among all the main effects so it has been included. The main effects model (x1-x5) are shown as well as “Model” which is all main effects plus the interaction term of x2*x3 (malnutrition*alcoholism).

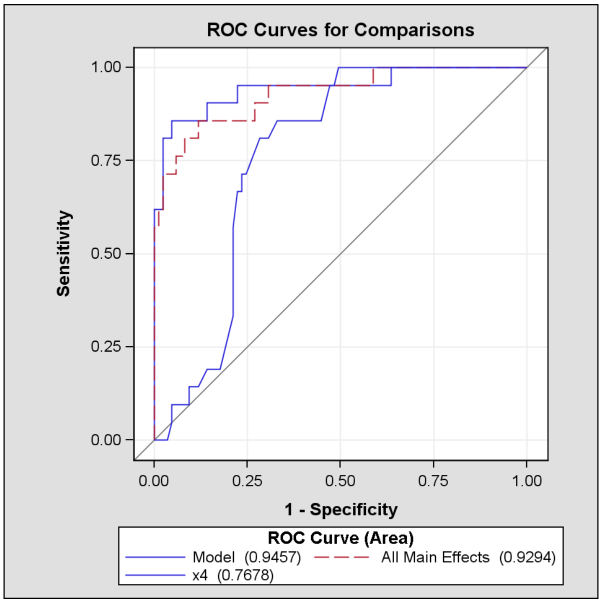


Figure 9 – ROC Curve comparison for final model, age model and main effects model

One caveat of note is that although ROCCONTRAST is an advantageous resource to utilize, care should be used in specifying curves for inclusion into the comparison chart. Specification of too many ROC curves can generate an unwieldy chart that can be potentially difficult to decipher and impractical to display (see Figure 10).

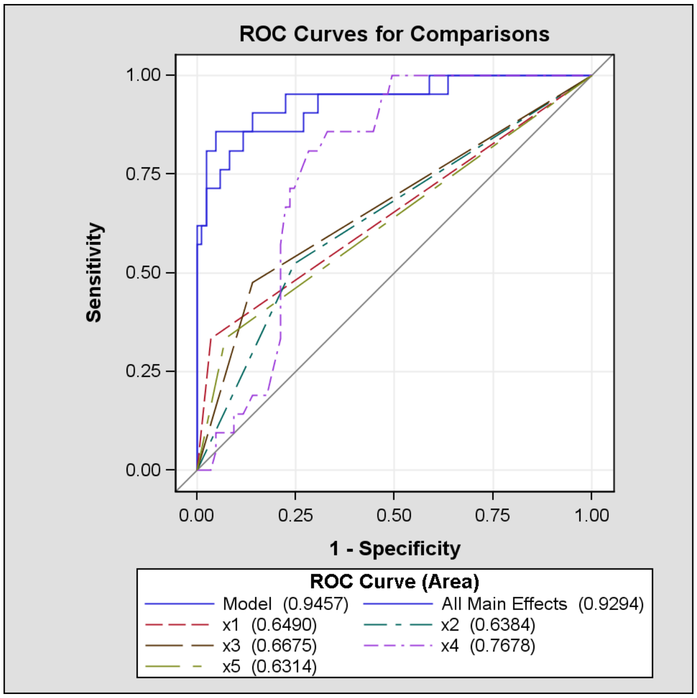


Figure 10 – Improper utilization of ROCCONTRAST

CONCLUSION

Within SAS/STAT® 9.2, the PLOTS =option and ROCCONTRAST statements are welcome additions to the LOGISTIC procedure. Utilized in conjunction with ODS GRAPHICS, the fitted logistic function and ROC curve generated by the PLOTS = option are high quality displays that can be quickly produced or referenced. In the tricky process of practical model selection, the ROCCONTRAST statement gives us an additional powerful technique which emphasizes predictive ability. The resulting output and graphical displays are comprehensive and easily produced. Researchers should seriously consider these valuable tools when fitting and choosing a logistic regression model.

REFERENCES

Bridges, K.J., Trujillo, E.B., and Jacobs, D.O., (1999) "Alcohol-related thiamine deficiency and malnutrition", *Critical Care Nurse*, **6**, 80-85.

DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988) "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Non-Parametric Approach", *Biometrics,* **44**, 837-845.

Pine, R. W., Wertz, M. J., Lennard, E. S., Dellinger, E. P., Carrico, C. J., & Minshew, B. H. (1983). "Determinants of organ malfunction or death in patients with intra-abdominal sepsis. A discriminant analysis." *Archives of Surgery 118:*242-249.

Gaylor, D.W. (1987) "Linear-nonparametric upper limits for low dose extrapolation." *American Statistical Association: Proceedings of the Biopharmaceutical Section.* 63-66.

Rao, P.V. (1998) *Statistical Research Methods in the Life Sciences.* Thompson Wadsworth

SAS Institute, Inc. SAS/STAT®. User's Guide, Version 9. Cary, NC: SAS Institute Inc. (2004).

van Belle, G., Fisher, L. D., Heagerty, P. J., & Lumley, T. (2004). *Biostatistics: A Methodology for the Health Sciences* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Robert G. Downer, PhD.
Biostatistics Director & Associate Professor
Department of Statistics
Grand Valley State University
1 Campus Drive
Allendale, MI 49401
downerr@gvsu.edu

Patrick Richardson, M.S.
Biostatistician - Program of Translational Medicine
Van Andel Research Institute
333 Bostwick Avenue NE
Grand Rapids, MI 49503
Patrick.Richardson@vai.org