Paper 226-27

An Explicit Functional Form Specification Approach to Estimate the Area under a Receiver Operating Characteristic (ROC) Curve

Paul Stober, GlaxoSmithKline, Collegeville PA.
Shi-Tao Yeh, GlaxoSmithKline, Collegeville PA.

ABSTRACT

The Receiver Operating Characteristic (ROC) curve is a curve presented in a probability scale graph and is used to judge the discrimination ability of various statistical methods for predictive purposes. The area under the ROC curve can be measured and converted to a single quantitative index for diagnostic accuracy.

An explicit functional form approach is proposed as an alternative estimation method to evaluate the area under a ROC curve. This paper provides an explicit functional form to represent the ROC curve through SAS code for parameter estimation and the area under the curve calculation. The empirical ROC curves produced from this approach are much smoother with convexity of the curves.

The SAS products used in this paper are base SAS® , SAS/STAT® and SAS/GRAPH®, with no limitation of operating systems.

INTRODUCTION

The clinicians tried to determine "….whether a patient would need hospitalization or would he/she benefit from just treatment-" in a clinical trial study. This question can be converted into the following statements: In a sample of *n* patients, suppose n_1 patients are observed to have a certain condition or event. Then *n* patients undergo a test for predicting the event and the test is based on the estimated probability of the event. Higher values of this estimated probability are assumed to be associated with the event.

Receiver Operating Characteristic (ROC) curves are popular as a tool for detection of clinical related events or various conditions such as asymptomatic dysfunction or disease. . The *sensitivity* is the probability of a positive test, given the subject's true response is positive. the *specificity* is the probability of a negative test, given the subject's true response is negative. The ROC curve, shown in Figure 1 by plotting of *sensitivity* versus 1 - *specificity*, is used to judge the discrimination ability of various statistical methods for predictive purposes

The area under a ROC curve, shown as the shaded area in Figure 1, is a summarized quantitative index. This index, varies between 0.5 (no discrimination power) to 1.0 (perfect accuracy) as the ROC travels towards the left and top boundaries of the graph. The meaning of the area under an ROC curve, namely the index, is a "probability of correctly ranking a (normal, abnormal) pair". In other words, the index is a probability of correct pairwise rankings.

There are several methods to estimate the area under a ROC curve; iterative maximum likelihood estimation, fitted curve method, the Wilcoxon or Mann-Whitney statistical method and trapezoidal rule. An explicit nonlinear functional form with curve fitted computation methods is proposed in this paper.

Different methods of estimating the area under a ROC curve are performed in this paper for comparison purposes.

The functional form approach used in this paper assumes that the underlying distributions for normal and abnormal groups are Gaussian. The functional form specified must satisfy the mathematical properties to assure that any estimated ROC can be correctly presented on the probability scale and graph.

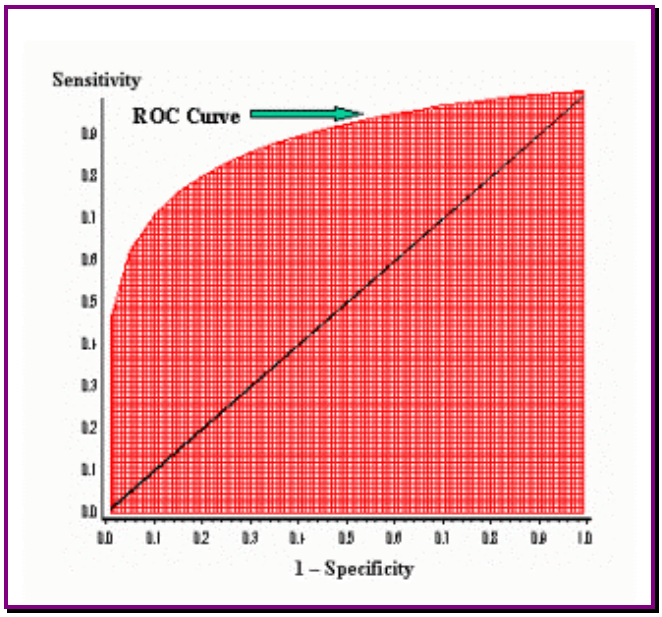

Figure 1. ROC Curve

This paper is comprised of five parts. Part 1 includes the abstract and the introduction. Part 2 describes the functional form specification and its mathematical properties. Part 3 is devoted to the application. Part 4 involves the comparison of estimation results from different estimation methods. Part 5 concludes the paper.

FUNCTIONAL FORM SPECIFICATION

The ROC graph is presented on X, Y axis with (X, Y) values varied from 0 to 1. When the graph is rotated

on the equalitarian line (a 45 degree line from the origin) the ROC curve becomes the well known Lorenz curve in social sciences. The Lorenz curve is a rotated ROC curve during symmetry with respect to equalitarian line.

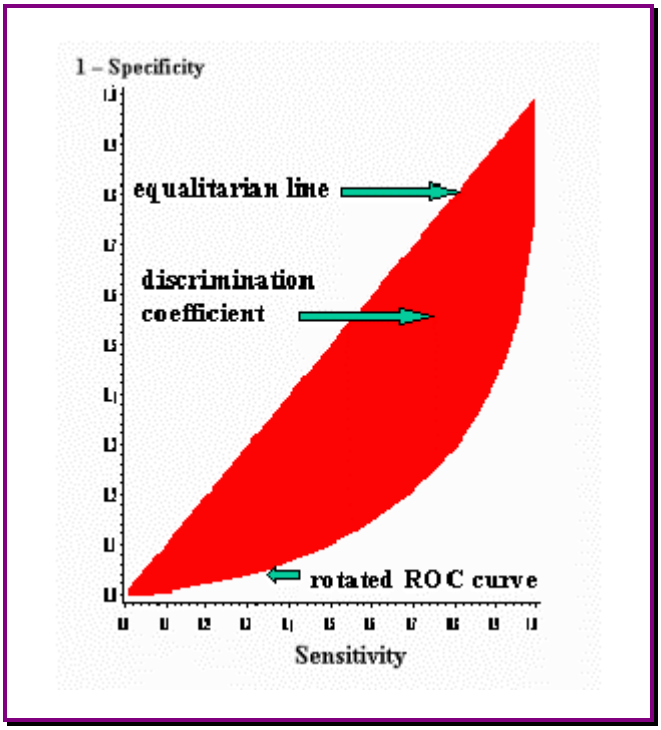

Figure 2. Rotated ROC Curve and Area of Discrimination Coefficient

The functional form

$$y = f(x) \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (1)$$

represents the rotated ROC curve when it satisfies the following properties:

(i) $f(0) = 0$;
(ii) $f(1) = 1$;
(iii) $f'(x) \geq 0$, for $0 \leq x < 1$;
(iv) $f''(x) \geq 0$, for $0 \leq x \leq 1$;
(v) $f(x) \leq x$, for $0 < x < 1$;
(vi) $0 \leq \int_0^1 f(x)\,dx \leq 1/2$.

The functional form that satisfies all of the properties, proposed by Rache, Gaffney, Koo, and Obst [1] , is used in this paper. The explicit functional form is as follows:

$$y = [\,1 - (\,1 - x\,)^{\alpha}\,]^{1/\beta}$$

where $0 < \alpha \leq 1,\ 0 < \beta \leq 1$;

SAS MACRO FOR PARAMETER ESTIMATION

The NLIN procedure in SAS/STAT is employed for the rotated ROC curve parameter estimation.

```
%macro nlreg(dsin,x,y,dsout);
 proc nlin data= &dsin
      maxiter = 30
      converge = .00001;
      parms a1 = 0.5
            b1 = 0.6;
      bounds 0 < a1 <= 1,
             0 < b1 <= 1;
      model &y = ( 1 - ( 1 - &x)**a1)**(1
/ b1 );
      output out=&dsout parms = a1 b1;
%mend nlreg;
```

The macro arguments for SAS macro ***nlreg*** are:

dsin: input dataset name
x: independent variable or values on equalitarian line
y: target variable for analysis
dsout: output dataset name

SAS MACRO FOR DISCRIMINATION COEFFICIENT ESTIMATION

The computation of the discrimination coefficient, shown in shaded area in Figure 2, is based on the functional form specified in equation (1). It is defined:

$$DC = 1.0 - 2.0 \int_0^1 [\,1 - (\,1 - x\,)^{\alpha}\,]^{1/\beta}\,dx,$$

Substituting variables

$$u = 1 - (\,1 - x)^{\alpha},$$

this is equal to:

$$DC = 1.0 - 2.0\,(1/\alpha) \int_0^1 (\,1 - u\,)^{1/\beta}\, u^{1/\alpha - 1} du$$

$$= 1.0 - 2.0\,(1/\alpha) * B\,(\,1/\alpha\,,\,1/\beta + 1)$$

where B represents the beta distribution.

The SAS function of GAMMA is then used for the computing of the beta distribution. The macro **%dc** is designed for computing the discrimination coefficient. The arguments of this function are:

dsin : dataset name which is the output file from the execution of macro %nlreg,

a1, b1: estimate of parameters

n: data selection cutpoint.

The complete macro code is shown in the following block.

```
%macro dc(dsin, a1, b1, n);
 data t&n;
  set t&n;
   aa1 = 1 / &a1;
   bb1 = (1/ &b1) + 1;
   c1 = aa1 + bb1;
   cr = 1.0 - (2.0/ &a1) * (gamma(aa1) *
        gamma(bb1) / gamma(c1));
   roc = cr / 2 + 0.5;
   crt = &n;
 drop  aa1 bb1 c1;
%mend dc;
```

SAS MACRO USING TRAPEZOIDAL RULE FOR AREA CALCULATION

The macro is an alternative method to estimate the area under a curve.

The trapezoidal rule is a numerical method to be used to approximate the integral or the area under a curve. Using trapezoidal rule to approximate the area under a curve involves slicing up the area to be found into a number of strips of equal width approximating the area of each strip by the area of the trapezium formed when the upper end is replaced by a chord; the sum of these approximations then gives the final numerical result of the area under the curve. The trapezoidal rule cab be presented as follows:

Function $\int_a^b$ **f(x) dx** is a definite integral. The points of subdivision of the domain of the integration **[a, b]** are labelled $x_0, x_1, \ldots x_n$;

where $x_0 = a$, $x_n = b$, $x_r = x_0 + r (b – a) / n$.

Function **T(a, b, n)** can be defined as the procedure of trapezoidal rule that

T(a, b, n)
= ((b – a) / n) * (((f(a) + f(b)) / 2) + ∑ f (a+ i (b-a)/n))

The summation of the above equation is i = 1 to n – 1. **T(a, b, n)** approximates the definite integral

$\int_a^b$ **f(x) dx.**

Using trapezoidal rule with n number of intervals, provided **f(x)** is defined and that it is continuous in the domain **[a, b].** The following SAS macro performs trapezoidal rule for area under a curve calculation.

```
%macro trap(a,b,n, function);
data f1;
 do i = 0 to &n ;
  if i = 0 then do;
  x = &a ;
  y = (( &b - &a ) / &n ) * ( &function /
2 );
  output;
  end;
  else if i = &n then do;
  x = &b ;
  y = (( &b - &a ) / &n ) * ( &function /
2 );
  output;
  end;
  else do;
  x = ( &a + i * (( &b - &a) / &n ) );
  y = ((&b - &a) / &n) * &function;
  output;
  end;
 end;
run;
   proc summary data=f1;
     var y;
     output out=p1 sum=areau;
run;
   proc print data=p1;run;
%mend;
```

The arguments of this macro are:

a: lower limit of the integration,

b: upper limit of the integration,

n: number of intervals,

function: explicit functional form expression.

The invocation example for this macro is as follows:

```
%trap(a=0, b=1, n=40, function= 1/x)
```

APPLICATION AND COMPARISON

A hypothetical clinical trial protocol is designed to examine the value of Brain Natriuretic Peptides (BNP) in the assessment of asymptomatic Left Ventricular Dysfunction (LVD) in a population at high risk of developing Congestive Heart Failure (CHF). The measurement of the test characteristics of BNP can be used in conjunction with a two-dimensional echocardiography for the detection of LVD.

The patients in the study are required for the Left Ventricular Ejection Fraction (LVEF) data. Several different levels of LVEF are selected, which is then used for the definitions of LVD. Levels of BNP are selected for LVD prediction purpose. The cutpoints for LVEF are selected as less than 40%, less than 45%, less than 50%, less than 55% and less than 60%. Calculation of

test characteristics of chosen values of BNP for LVD is performed. The ROC curves can be constructed by varying the cutpoint that determines which estimated BNP probabilities are considered to predict the LVD. The plots of these test characteristics data are shown in Figure 3.

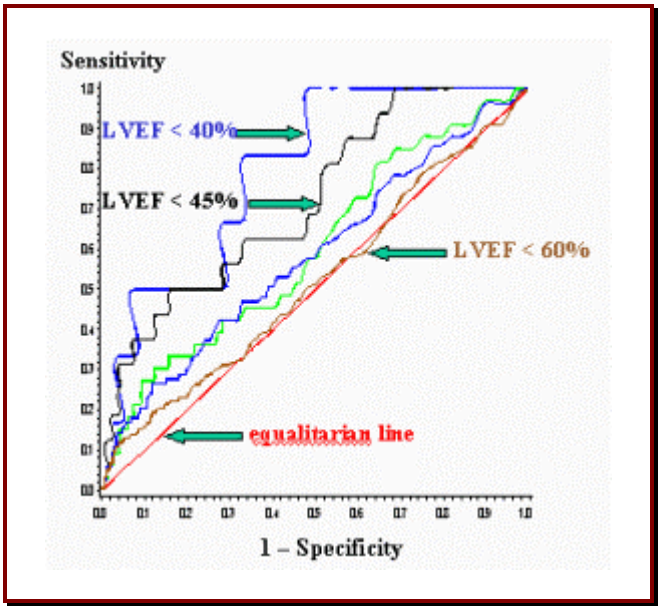

Figure 3. ROC Curves from Test Characteristic Data

Estimation by Functional Form Method

The same test characteristic data for Figure 3 plotting are used for parameter estimation by functional form method. The estimation results and ROC curves presentation are as follows:

LVEF CUTPOINT	ESTIMATED α	ESTIMATED β	DISCRIMINATION COEFFICIENT	ESTIMATED AREA UNDER ROC CURVE
< 40%	0.32887	0.91848	0.55003	0.77502
< 45%	0.57710	0.78732	0.39625	0.69812
< 50%	0.79896	0.86426	0.18826	0.59413
< 55%	0.99739	0.72524	o.16067	0.58033
< 60%	1.00000	0.92969	0.03644	0.51822

Table 1. Estimation Results from Functional Form Method

The estimation results show that when the LVEF cutpoint increases the area under the ROC curve decreases.

Estimation by Trapezoidal Rule with Functional Form Specified

The same functional form and estimated parameters are used for the area calculation by trapezoidal rule.

LVEF CUTPOINT	ESTIMATED α	ESTIMATED β	AREA UNDER ROTATED CURVE	ESTIMATED AREA UNDER ROC CURVE
< 40%	0.32887	0.91848	0.2256	0.7744
< 45%	0.57710	0.78732	0.3020	0.6980
< 50%	0.79896	0.86426	0.4059	0.5941
< 55%	0.99739	0.72524	o.4197	0.5803
< 60%	1.00000	0.92969	0.4818	0.5182

Table 2. Estimation Results from Trapezoidal Rule with Fuctional Form Specified

The area calculation results from both methods are shown in Tables 1 and 2. The calculation results are very close.

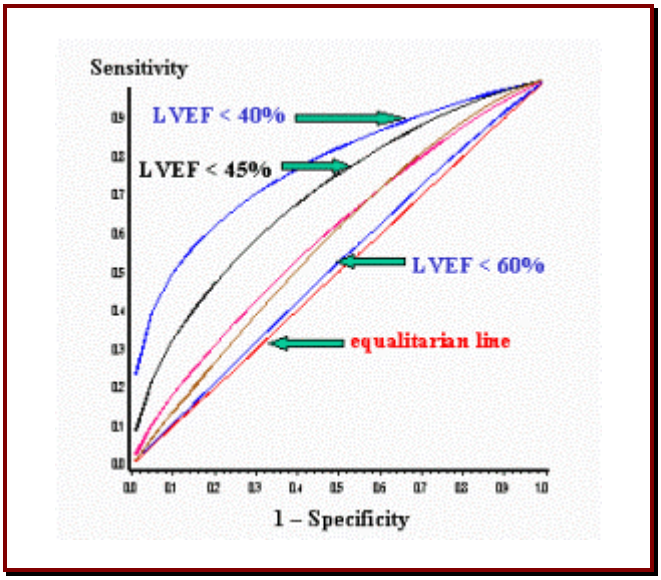

Figure 4. Estimated ROC Curves from Functional Form Method

Estimation by SAS Procedure LOGISTIC

The procedure LOGISTIC of SAS software allows you to compute the ROC curve. The OUTROC option in PROC LOGISTIC stores the needed measures, the variables' sensitivity and specificity to a data set. This data set can be used to construct the ROC curve by using the PLOT or GPLOT procedure and plotting for sensitivity (_SENSIT_) against 1 – specificity (_1MSPEC_). The area under the ROC curve, as determined by trapezoidal rule, is given by the statistic c in the "Association of Predicted Probabilities and Observed Responses" table.

The output from PROC LOGISTIC is shown in the following table:

LVEF CUTPOINT	STATISTIC C
< 40 %	0.716
< 45 %	0.710
< 50 %	0.588
< 55 %	0.586
< 60 %	0.518

Table 3. Estimation Results from LOGISTIC Method

The ROC presentation from estimation by LOGISTIC is shown as follows:

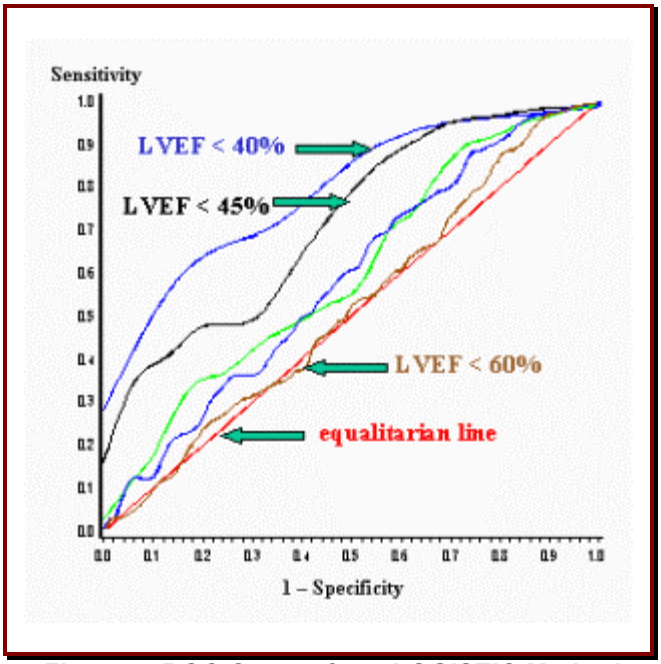

Figure 5. ROC Curves from LOGISTIC Method

Comparison of the computed area under the ROC curves by functional form specification method and the LOGISTIC procedure are as follows:

LVEF CUTPOINT	STATISTIC C	ESTIMATED AREA FROM FUNCTIONAL FORM	DIFFERENCE
< 40 %	0.716	0.775	-0.059
< 45 %	0.710	0.698	0.012
< 50 %	0.588	0.594	-0.006
< 55 %	0.586	0.580	0.006
< 60 %	0.518	0.518	0

Table 4. Comparison Estimation Results from Functional Form Method and LOGISTIC Method

Table 4 shows that the differences between alternative methods are ranging from –0.059 to 0.012.

CONCLUSION

This paper takes a comprehensive approach in the selection of functional form to represent a ROC curve. The SAS NLIN procedure is used to estimate the parameters. This functional form specification approach with non-linear estimation from NLINJ procedure makes best use of the SAS system capability of providing curve fitting and efficient parameter estimation.

In summary, this approach

* Provides a better graphic presentation by satisfying the mathematical properties
* Provides a smoother empirical ROC curve with convexity of the curve.
* Provides a compact, efficient, and yet simple code for parameter estimation and area under a curve calculation.
* Provides an estimation result that is close to other estimation methods.

ACKNOWLEGMENTS

The author would like to thank Nancy Asbel, Senior Statiscian, Biostatistics & Data Sciences, GlaxoSmithKline , and Ju Zhang, EDP Contract Services, for their discussion of this topic.

REFERENCES

[1] Hanley, James A., McNeil Barbara J.: *The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve*; Radiology 143, pp 29-36, April 1982

[2]. Lloyd, Chris J.: *Regression Models for Convex ROC Curves* Biometrics 56, pp 562-567, Sep. 2000

[3]. Pepe, Margaret, S.: *An Interpretation for the ROC Curve and Inference Using GLM Procedure* , Biometrics 56, pp 352-359 June 2000

[4]. _____: *Three Approaches to Regression Analysis of Receiver Operating Characteric Curves for Continuous Test Results* , Biometrics 54, pp 124-135 March 1998

[5] Rasche, R. H., J. Gaffney, A.Y. C. Koo, and N. Obst, *Functional Forms for Estimating the Lorenz Curve*, Econometrica, 48, pp. 1061-1062, 1980

[6] Reiser, B., J. and D. Faraggi, *Confidence Intervals for the Generalized ROC Criterion*, Biometrics 53, pp. 644-652, 1997

[7] Yeh, Shi-Tao ; *"Estimation of the Lorenz Curve and Concentration Ratio*", SUGI 18 Proceedings, pp. 873-877, May 1993

[8] _____ ; "*The Techniques to Improve Nonlinear Regression Curve Fit*", SUGI 16 Proceedings, pp. 1242-1245, February 1991

Authors
Paul W. Stober
(610) 917-6541
E-mail: paul_w_stober@gsk.com

Shi-Tao Yeh, Ph. D.
(610)917-5883(W)
E-mail: shi-tao_yeh-1@gsk.com