SAS Macro for Estimating the Standard Error of Area Under the Curve with an Application to Receiver Operating Curves

Terry L. Mashtare Jr.[a,b], Alan D. Hutson[a,b]

[a]*Department of Biostatistics, University at Buffalo, 249 Farber Hall, 3435 Main Street, Buffalo, NY 14214-3000, U.S.A.*
[b]*Roswell Park Cancer Institute, Elm & Carlton Streets, Buffalo, NY 14263, U.S.A*

Abstract

We propose a novel SAS macro to utilize PROC NLMIXED in SAS to estimate the standard error for AUC when the AUC estimator is based on the trapezoidal rule. Our method accounts for the trapezoidal region-by-region correlation structure. An example is provided for the standard error of the AUC under the ROC curve.

Key words: Asymptotic Theory; Multivariate Normality; Numerical Integration

1. Introduction

There are many applications in statistics where the estimation of area under the curve (AUC) and its corresponding standard error is of interest. Examples include AUC of the concentration-time curve for a drug in pharmacokinetic studies, and for AUC of receiver operating characteristic (ROC) curves. In general, the common approach to estimating AUC is by the trapezoidal rule, which is a straightforward approach for estimating AUC [1]. Due to the complex covariance structure of the adjacent trapezoidal regions, statistical software packages generally do not have built-in method for estimating the variance. We provide an extention that utilizes SAS NLMIXED[2] within a SAS macro. We provide a detailed algorithm for estimating the AUC and its corresponding variance structure. Sample code for estimating AUC of an ROC binormal curve and illustration with a simulated data set is provided.

2. Computational Methods and Theory

Let $X_1, \ldots, X_n$ denote i.i.d. $f_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$, a $p \times 1$ vector of unknown parameters. Let $\hat{\boldsymbol{\theta}} \sim AN(\boldsymbol{\theta}, b_n^2 \boldsymbol{\Sigma})$, where $b_n \to 0$ as $n \to \infty$. For example, $\hat{\boldsymbol{\theta}}$ can be the maximum likelihood estimate or method of moments estimate of $\boldsymbol{\theta}$. Let $g(y; \boldsymbol{\theta})$ be a real valued function which is continuous over the interval $[a, b]$. Suppose we wish to estimate

$$A(\boldsymbol{\theta}) = \int_a^b g(y; \boldsymbol{\theta}) \mathrm{d}y \tag{1}$$

with

$$A(\hat{\boldsymbol{\theta}}) = \int_a^b g(y; \hat{\boldsymbol{\theta}}) \mathrm{d}y \tag{2}$$

Equation (2) is often in practice does not have a close form solution. Using the trapezoidal rule, we divide $[a, b]$ into $m+1$ subintervals of width $(b-a)/m$. Then the function

$$A_t(\hat{\boldsymbol{\theta}}) = \frac{b-a}{m} \left\{ \frac{g(a; \hat{\boldsymbol{\theta}}) + g(b; \hat{\boldsymbol{\theta}})}{2} + \sum_{i=1}^{m-1} g(a + \frac{i(b-a)}{m}; \hat{\boldsymbol{\theta}}) \right\} \tag{3}$$

approximates the integral $A(\hat{\boldsymbol{\theta}})$.

For simplicity of notation, write (3) as

$$A_t(\hat{\boldsymbol{\theta}}) = \frac{b-a}{m} \left[\sum_{i=0}^{m} h_i(\hat{\boldsymbol{\theta}}) \right], \tag{4}$$

where $h_0(\hat{\boldsymbol{\theta}}) = \frac{1}{2} g(a; \hat{\boldsymbol{\theta}})$, $h_m(\hat{\boldsymbol{\theta}}) = \frac{1}{2} g(b; \hat{\boldsymbol{\theta}})$, and $h_i(\hat{\boldsymbol{\theta}}) = g(a + \frac{i(b-a)}{n}; \hat{\boldsymbol{\theta}}), i = 1, \ldots, m-1$.

Theorem 1. *Let $A_t(\hat{\boldsymbol{\theta}})$ be defined as in (4) with $\hat{\boldsymbol{\theta}} \sim AN(\boldsymbol{\theta}, b_n^2 \boldsymbol{\Sigma})$. Then $A_t(\hat{\boldsymbol{\theta}})$ is $AN(A_t(\boldsymbol{\theta}), Var(A_t(\hat{\boldsymbol{\theta}})))$ where*

$$Var(A_t(\hat{\boldsymbol{\theta}})) = \frac{(b-a)^2}{m^2} \{ \sum_{i=0}^{m} Var[h_i(\hat{\boldsymbol{\theta}})] + 2 \sum_{j=i+1}^{m} \sum_{i=0}^{m} Cov[h_i(\hat{\boldsymbol{\theta}}), h_j(\hat{\boldsymbol{\theta}})] \} \tag{5}$$

Proof. By Theorem A, section 3.3 in Serfling[3] we have

$$h_i(\hat{\boldsymbol{\theta}}) \sim AN(h_i(\boldsymbol{\theta}), Var[h_i(\hat{\boldsymbol{\theta}})])$$

for $i = 1, \ldots, n$. The formula for $Var(A_t)$ follows directly from the generalization of Theorem 4.5.6 in Casella and Berger[4]. □

3. ROC Example

We illustrate the trapezoid method for estimating the AUC and standard error for a binormal ROC curve. Estimating the parameters is straightforward using methods discussed by Gönen[5]. For the purpose of comparing the two methods, we choose a binormal ROC model, in which the standard error can be directly estimated using PROC NLMIXED in SAS. Note however, that the approach used in the example easily extends to models where closed form solutions for AUC and/or standard errors for AUC do not exist.

Let $Y|D = 0 \sim N(\mu_0, \sigma_0^2)$ and $Y|D = 1 \sim N(\mu_1, \sigma_1^2)$. Then the equation of the binormal ROC curve is given by $g(x; \boldsymbol{\theta}) = \Phi[a + b\Phi^{-1}(x)]$, where $\boldsymbol{\theta} = (\mu_0, \sigma_0, \mu_1, \sigma_1)'$, $a = (\mu_1 - \mu_0)/\sigma_1$, and $b = \sigma_0/\sigma_1$. For the binormal ROC curve, there is a closed form solution.

$$\int_0^1 \Phi[a + b\Phi^{-1}(x)]\mathrm{d}x = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right), \tag{6}$$

where $-\infty < a < \infty$ and $b > 0$. We estimate (6) by

$$A_t(\hat{\boldsymbol{\theta}}) = \frac{1}{m}\left\{\frac{1}{2} + \sum_{i=1}^{m-1} \Phi[a + b\Phi^{-1}(i/m)]\right\}. \tag{7}$$

So, $h_0(\hat{\boldsymbol{\theta}}) = 0$, $h_m(\hat{\boldsymbol{\theta}}) = 1$, and $h_i(\hat{\boldsymbol{\theta}}) = \Phi[a + b\Phi^{-1}(i/m)]$ for $i = 1, \ldots, m-1$. Equation (8) reduces to

$$Var(A_t(\hat{\boldsymbol{\theta}})) = \frac{1}{m^2}\{\sum_{i=1}^{m-1} Var[h_i(\hat{\boldsymbol{\theta}})] + 2\sum_{j=i+1}^{m-1}\sum_{i=1}^{m-1} Cov[h_i(\hat{\boldsymbol{\theta}}), h_j(\hat{\boldsymbol{\theta}})]\} \tag{8}$$

We simulated a data set so that $Y|D = 0 \sim N(0, 1)$ and $Y|D = 1 \sim N(2, 1.5^2)$ and $n = 50$ in each group. Under these assumptions, the true AUC is 0.866. Using maximum likelihood estimation for this data set we get AUC = 0.880 with standard error estimate of 0.0332. Using the trapezoid rule, we get estimated AUC = 0.879 and standard error estimate of 0.0329.

References

[1] Yeh, Shi-Tao. (2002) Using Trapezoidal Rule for the Area Under a Curve Calculation. SUGI 27 Proceedings.

[2] SAS (v9.1) SAS Institute Inc., Cary, NC, USA.

[3] Serfling, R.J. (1980) Approximation Theorems of Mathematical Statistics. John Wiley & Sons, Inc.

[4] Casella, G. and Berger R.L. (2002) Statistical Inference 2nd ed. Duxbury.

[5] Gönen, M. (2007) Analyzing Receiver Operating Characteristic Curves with SAS. Cary, NC: SAS Institute Inc.

Appendix A

```
/*************************************************************/
/* Macro seAUC was created to estimate the AUC and standard */
/* error of the AUC for a binormal ROC curve using the      */
/* trapezoidal rule.  This macro can be easily modified to  */
/* include other distributions besides the binormal.        */
/* MACRO VARIABLE DEFINITIONS                               */
/*   INDATA = name of dataset to be analyzed                */
/*        Y = response variable name                        */
/*        D = disease status variable name(0=no 1=yes)      */
/*        N = number of intervals for trapezoid rule        */
/* OTHER VARIABLE DEFINITIONS                               */
/*    m1 = mean of Y|D = 1; m0 = mean of Y|D = 0            */
/*    s1 =  std of Y|D = 1; s0 =  std of Y|D = 0            */
/*     a = (m1 - m0)/s1;     b =  s0/s1                     */
/*   u_i = i/N;            u_j = j/N                        */
/*************************************************************/

%MACRO seAUC(INDATA,Y,D,N);
/*************************************************************/
/* 1. Estimate the parameters                                */
/*************************************************************/

PROC NLMIXED DATA=&INDATA;
    PARMS m1=0 m0=0 s1=1 s0=1;
    BOUNDS s1>0, s0>0;
    IF &D=1 THEN m=m1; ELSE IF &D=0 THEN m=m0;
    IF &D=1 THEN s=s1**2; ELSE IF &D=0 THEN s=s0**2;
    MODEL &Y ~ normal(m,s);
    ODS OUTPUT ParameterEstimates=parms;
RUN;

/*************************************************************/
/* 2. Create macro variables of the parameter estimates     */
/*************************************************************/

PROC TRANSPOSE DATA=parms OUT=parms1;
```

```
    VAR estimate;
    ID parameter;
RUN;

DATA parms1; SET parms1;
    CALL SYMPUT ('m0',m0); CALL SYMPUT ('m1',m1);
    CALL SYMPUT ('s0',s0); CALL SYMPUT ('s1',s1);
RUN;

/*************************************************************/
/* 3. Output dataset (n-1)(n-2)/2 times to estimate          */
/*    covariance components                                  */
/*************************************************************/

DATA CopyData; SET &INDATA;
    DO i=1 TO &N-1;
        DO j=i+1 TO &N-1;
            u_i=i/&N;
            u_j=j/&N;
            OUTPUT;
        END;
    END;
RUN;

PROC SORT DATA=CopyData;
    BY i j;
RUN;

/*************************************************************/
/* 4. Use PROC NLMIXED to estimate the standard errors of    */
/* each of the AUC components                                */
/*************************************************************/

ODS LISTING CLOSE;

PROC NLMIXED DATA=CopyData MAXITER=3; BY i j;
    PARMS m1=&m1 m0=&m0 s1=&s1 s0=&s0;
    IF &D=1 THEN m=m1; ELSE IF &D=0 THEN m=m0;
```

```
    IF &D=1 THEN s=s1**2; ELSE IF &D=0 THEN s=s0**2;
    a=(m1 - m0)/s1; b=s0/s1;
    h_i=PROBNORM(a+b*PROBIT(u_i));
    h_j=PROBNORM(a+b*PROBIT(u_j));
    sum_ij=h_i+h_j;
    MODEL &Y ~ normal(m,s);
    ESTIMATE 'VAR_i' h_i;
    ESTIMATE 'VAR_j' h_j;
    ESTIMATE 'SUM_ij' sum_ij;
    ODS OUTPUT AdditionalEstimates=out;
RUN;

ODS LISTING;
/*************************************************************/
/* 5. Get the var and cov of the AUC components              */
/*************************************************************/

DATA out1; SET out;
    KEEP i j label standarderror variance;
    variance=standarderror**2;
RUN;

DATA out2; SET out1; BY i;
    IF first.i then output;
    IF i=&N-2 and j=&N-1 and label='VAR_j' then output;
RUN;

DATA out2; SET out2;
    IF label='VAR_j' then i=&N-1;
RUN;

PROC TRANSPOSE DATA=out1 OUT=out3;
    VAR variance;
    ID label;
    BY i j;
RUN;

DATA out4; SET out3;
```

```
    _2cov=sum_ij-var_i-var_j;
RUN;

/*************************************************************/
/* 6. Sum the variance and covariance AUC components         */
/*    to get the final standard error estimate               */
/*************************************************************/
PROC MEANS DATA=out2 SUM;
    VAR variance;
    OUTPUT OUT=out5 SUM=sumvar;
RUN;

PROC MEANS DATA=out4 SUM;
    VAR _2cov;
    OUTPUT OUT=out6 SUM=sum2cov;
RUN;

DATA AUCse; MERGE out5 out6; by _TYPE_;
    se2=sqrt(sumvar+sum2cov)/&N;
    KEEP _TYPE_ se2;
RUN;

PROC PRINT DATA=AUCse;
RUN;
%MEND seAUC;
```