General Discrimination Index, the Area under the ROC Curve (*overall C*) for Survival Time Model

Byung-Ho Nam
Boston University
Statistics and Consulting Unit
Department of Mathematics and Statistics
111 Cummington Street
Boston, Massachusetts 02215
E-Mail Address: byungho@bu.edu

Ralph B. D' Agostino
Boston University
Statistics and Consulting Unit
Department of Mathematics and Statistics
111 Cummington Street
Boston, Massachusetts 02215
E-Mail Address: ralph@bu.edu

The accuracy of fit of a mathematical predictive model is the degree to which the predicted values coincide with the observed outcome. When the outcome variable is dichotomous and predictions are stated as probabilities that an event will occur, models can be checked for good discrimination and calibration. In case of multiple logistic regression model for binary outcomes, the area under the ROC(Receiver Operating Characteristic) curve is the most used measures of model discrimination. The area under the ROC curve is identical to Mann-Whitney statistic. In this talk, we present a development of general description of overall discrimination index *C*(*overall C*) which we can extend to survival time model such as Cox regression model. The general theory of rank correlation is applied in developing the *overall C*. The *overall C* is a linear combination of three independent components: event vs. non-event, event vs. event and event vs. censored. By showing that these three components are asymptotically normally distributed, the *overall C* is shown to be asymptotically normally distributed:

Suppose we have n individuals, among which n_1 developed events in time t (*event*), n_2 did not develop events in time t (*non-event*) and n_3 were censored in time t (*censored*).
($n = n_1 + n_2 + n_3$).
Define
T_i = survival time for i^{th} individual, i=1,2,..,n
Y_i = predicted probability for developing an event in time t for i^{th} individual, i=1,2,..,n
Then, we have n pairs of (T_1,Y_1), (T_2,Y_2), ..,(T_n,Y_n)

Define

$$\text{Overall } C = \frac{1}{Q}\sum_{i=1}^{n-1}\sum_{j=1}^{n} a_{ij}b_{ij}$$

Where,
Q = the total number of comparisons made
$a_{ij} = 1$ if $T_i < T_j$, and at least one of pair(T_i,T_j) is for *event*, i,j=1,2, ..,n

= 0 otherwise

b_{ij} = 1 if $Y_i > Y_j$, and at least one of pair(Y_i,Y_j) is for *event*, i,j=1,2, . .,. n

= 0 otherwise

Hence we have

T_{1i} = survival time for *event*, i=1,2,. .,.n_1

Y_{1i} = predicted probability for *event*, i=1,2,. .,.n_1

T_{2j} = survival time for *non-event*, j=1,2,. .,.n_2

Y_{2j} = predicted probability for *non-event*, j=1,2,. .,.n_2

T_{3j} = survival time for *censored*, j=1,2,. .,.n_3

Y_{3j} = predicted probability for *censored*, j=1,2,. .,.n_3

We can have three sets of comparisons:

1. *event* vs. *non-event*: comparing those who developed events against those did not
2. *event* vs. *event*: comparing those who developed events against those who also developed events
3. *event* vs. *censored*: comparing those who developed events against those who were censored

Note that these three comparisons are *independent* one another.

The C for the first set (C_1) is identical to the Mann-Whitney statistic, which has asymptotic normal distribution. Hence C_1 is also asymptotically normally distributed. The C for the second set (C_2) is shown to have a linear relationship with the rank correlation coefficient $\hat{t}$ and $\hat{t}$ is asymptotically normally distributed. Hence C_2 is also asymptotically normally distributed. The C for the third set (C_3) is a conditional Mann-Whitney statistic, which is asymptotically normally distributed.

REFERENCE

1. Bamber, D. ' The area above the ordinal dominance graph and the area below the receiver operating graph' , *Journal of Mathematical Psychology*, **12**, 387–415 (1975).
2. D' Agostino, R.B. ' Development, Use and Modification of Health Risk Appraisal Functions: The Framingham Study' , *Proceedings of Statistics Canada Symposium, From Data to Information – Methods and Systems*, 57–62 (1995).
3. Hanley, J.A. and McNeil, B.J. ' The meaning and use of the area under a receiver operating characteristic (ROC) curve' , *Radiology*, **143**, 29–36 (1982).
4. Harrell, F.E., Lee, K.L. and Mark, D.B. ' Tutorial in Biostatistics, Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors' , *Statistics in Medicine*, **15**, 361–387 (1996).
5. Kendall, M.G. *Rank Correlation Methods*. Griffin, London, 1970.
6. Lehmann, E.L. ' Consistency and unbiasedness of certain nonparametric tests' , *Annals of Mathematical Statistics*, **22**, 165–179 (1951).
7. Lehmann, E.L. and D' Abrera, H.J.M. *Nonparametrics(Statistical Methods based on Ranks)*. Holden-day, California, 1975.