The ROC Curve Model from Generalized-Exponential Distribution

Ehtesham Hussain
Department of Statistics
University of Karachi, Karachi, Pakistan
ehussain@uok.edu.pk

Abstract

In biomedical studies bio-markers are used to distinguish between two groups of subjects usually, diseased (high risk) and non-diseased (low risk) subjects. Many diagnostic biomarkers results are continuous measurements. Some example include, serum antigen or enzyme concentrations (Zweig and Campbell 1993) are continuous in nature. The Receiver Operating Characteristic (ROC) curves have been widely used for evaluating the accuracy and discriminating power of a biomarker or statistical model. In this regard Generalized-Exponential Distribution model is suggested for analyzing such data. The model can be applied in the situations when the other well-known parametric models (e.g. the bi-normal one) cannot be used.

In this paper the parametric equation of the Receiving Operating Characteristic (ROC) curve model is established under the assumptions of bi-distributional population based on pair of Generalized-Exponential Distributions. Also its maximum likelihood estimator MLE, sampling distribution, equivalence test statistic and exact confidence interval are derived.

Keywords: ROC curve, Generalized-Exponential Distribution, MLE, Equivalence test statistic, Exact confidence interval.

1. Introduction

The receiving operating characteristic (ROC) curve model is a useful graphical and statistical modeling tool for evaluation performance in any two population (group) classification task, e.g. in medicine two groups of subjects (Diseased, Healthy). In biomedical applications there has been an increased use of ROC curve models for assessing the effectiveness of continuous diagnostic marker values (e.g. tumor volume) in distinguishing between diseased and healthy individuals. A person is assessed as diseased (positive) or healthy (negative) depending on whether the corresponding marker value is greater than or less than equal to a given threshold value. Associated with any threshold value is the probability of true positive (sensitivity) and the probability of true negative (specificity).

Let X and Y denote the diagnostic variable conditional on healthy and diseased, respectively, and assume that for a generic subject, the disease is diagnosed, i.e. the test is positive, if the associated diagnostic variable is greater than a suitable threshold c, (Y>c). If the distribution of Y is F and the distribution function of X is G, then the ROC curve is a plot of 1-F(c), i.e. sensitivity, versus 1-G(c), i.e. the 1-specificity, across all possible threshold c. A typical ROC curve model is shown in Figure1.

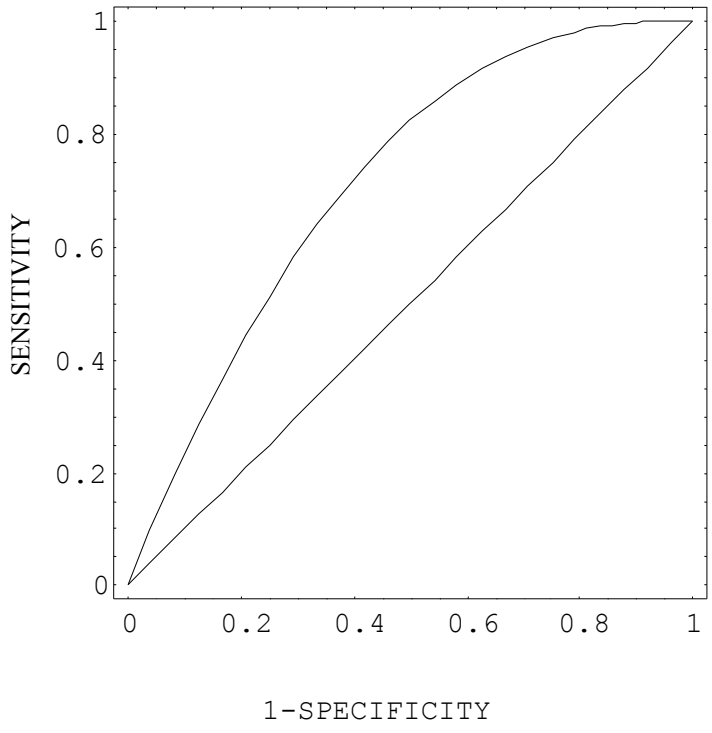

Figure 1. A TYPICAL ROC CURVE MODEL

Alternatively, the ROC curve model be defined as a function of (1-specificity)

$$R(t) = 1 - F\left(G^{-1}(1-t)\right) \text{ , for } 0 \le t \le 1 \qquad (1)$$

See, e.g. Lloyd [1]. The parametric ROC curve model can be obtained for symmetric and skewed markers values under bi-distributional models assumptions [2]. The frequently used Bi-Normal model assumes independent normal distributions with different population means and variances [3]. Other models considered in the literature include Bi-Logistic [4]. Bi-Exponential [5] and Bi-Gamma [6], etc. Here we propose Bi-Generalized-Exponential-Distribution model [7] for strongly positively skewed scores.

In the following sections certain results related to use one parameter Generalized-Exponential-Distribution GED(θ) under the bi-distributional assumptions for the ROC curve model is derived. In subsequent sections results regarding maximum likelihood estimator (MLE), its distribution, test of equivalence statistic and exact confidence interval are derived [8].

2. The Generalized Exponential Distribution

The one parameter generalized exponential distribution GED(θ) was introduced by Kundu and Gupta [7], as an alternative to the gamma and Weibull distributions for analyzing failure time data. The GED(θ) is a uni-modal density function. The advantage of employing the GED(θ) is that the cumulative distribution function (cdf) can be written in closed form. It is reported by Kundu and Gupta [9] that GED(θ) is quite flexible and can be used very effectively in analyzing positive time data in place of well known gamma and Weibull models. In particular, the probability density function (pdf) and the cumulative (cdf) of GED(θ) are

$$f(x) = \theta e^{-x}(1-e^{-x})^{\theta-1} \qquad x>0,\ \theta>0 \qquad (2.1)$$

$$F(x) = (1-e^{-x})^{\theta} \qquad x>0,\ \theta>0 \qquad (2.2)$$

where θ is the shape parameter. GED(θ) becomes more and more symmetric if θ increases.

For ready reference the graphs of (2.1) for certain parameter values are shown in Figure 2.

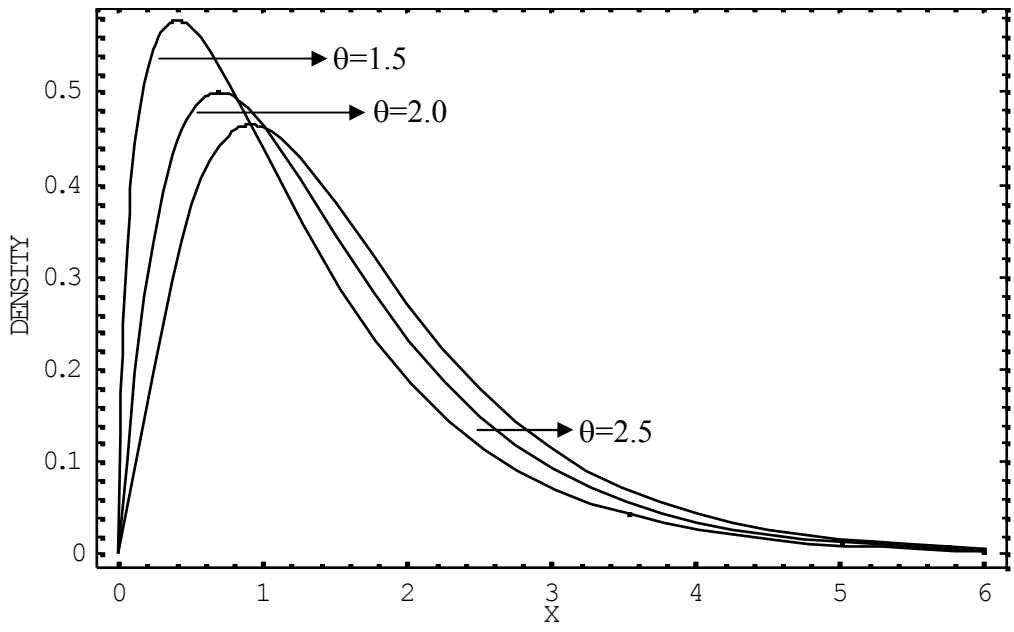

Figure 2. Generalized-Exponential Density Plots for θ=1.5, 2, 2.5.

2.1 Estimation of Parameter (θ) Maximum Likelihood Estimation

Let $x_1, x_2, \ldots, x_n$ be the random sample taken from GED with parameter θ. Then, likelihood and log-likelihood functions based on the above sample are as follows:

$$L(\theta : X) = \theta^n e^{-\Sigma x_i} \prod_{i=1}^{n} (1 - e^{-x_i})^{\theta - 1}$$

and

$$\ell n(\theta : X) = n \ell n \theta - \sum x_i + (\theta - 1) \sum \ell n(1 - e^{-x_i})$$

The associated gradient is found as follows:

$$\frac{\partial \ell n(\theta : X)}{\partial \theta} = \frac{n}{\theta} + \sum \ell n(1 - e^{-x_i}) = 0$$

Hence MLE of the parameter θ is obtained as

$$\hat{\theta} = \frac{-n}{\sum \ell n(1 - e^{X_i})} \qquad (2.1.1)$$

It is observed in Kudu and Gupta and [9], that

$$Y = -2\theta \sum \ell n(1 - e^{X_i}) \sim \chi^2_{(2n)}$$

$$\hat{\theta} = \frac{2n\theta}{Y} \qquad (2.1.2)$$

The distribution of $\hat{\theta}$ is same as the distribution of $\frac{n\theta}{Y}$, where Y follows chi-squared distribution with 2n degrees of freedom.

3. Bi-Generalized-Exponential-Distribution Model

A Bi-Generalized-Exponential-Distribution model assumes that, decision values follow two independent Generalized -Exponential-Distributions: one for diseased (negative) examples, one for healthy (positive). Let X and Y are from healthy and diseased populations with underlying GED functions *g(x), f(y)* respectively.

Hence for healthy population GED(θ_0)

$$g(x / H) = \theta_0 e^{-x} \left(1 - e^{-x}\right)^{\theta_0 - 1} \qquad x>0,\ \theta_0>0 \qquad (3.1)$$

and for diseased population GED(θ_1)

$$f(y / D) = \theta_1 e^{-y} \left(1 - e^{-y}\right)^{\theta_1 - 1} \qquad y>0,\ \theta_1>0 \qquad (3.2)$$

Also

Specificity:

$$\overline{G}(t) = \int_{t}^{\infty} g(x/H)dx \tag{3.3}$$

and Sensitivity:

$$\overline{F}(t) = \int_{t}^{\infty} f(y/D)dx \tag{3.4}$$

These define the theoretical ROC curve model for GED. The explicit equation of ROC curve for sensitivity as a function of (1-specificity) as defined in equation 1 is obtained as

$$R(t) = 1 - \left\{(1-t)^{\frac{\theta_1}{\theta_0}}\right\} \qquad 0 \le t \le 1$$

$$= 1 - \left\{(1-t)^{\beta}\right\} \tag{3.5}$$

where $\beta = \dfrac{\theta_1}{\theta_0}$ $\beta > 0$

Note in particular the ROC model equation (3.5) is function of two parameters θ_1 and θ_2 . The functional form depends on θ's through their ratio. The ratio β can be utilized as a measure of the size difference between the two distributions.

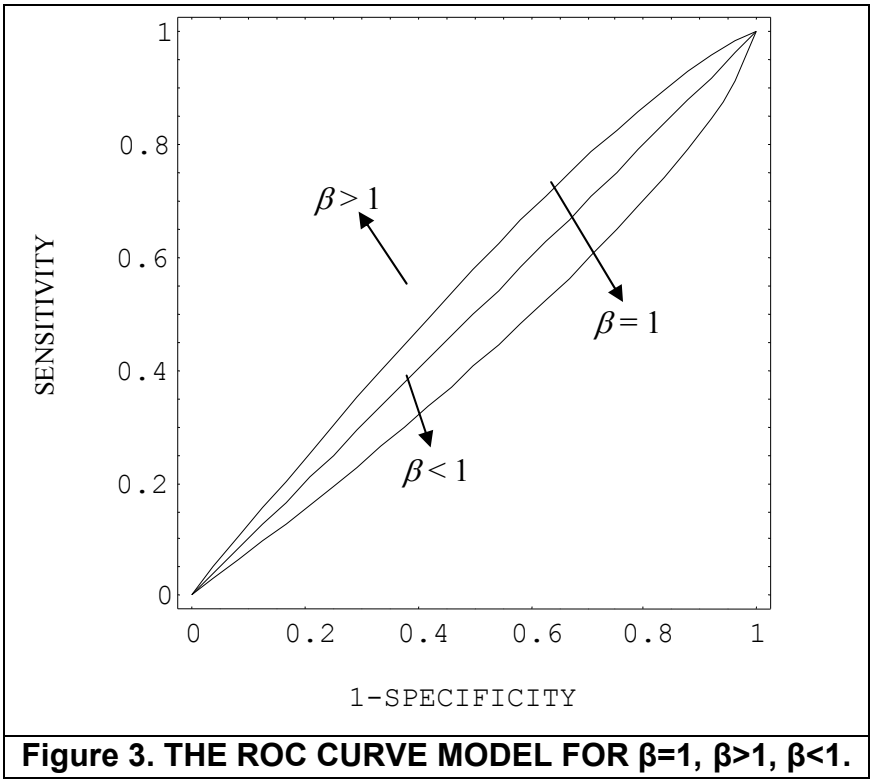

Figure 3. THE ROC CURVE MODEL FOR β=1, β>1, β<1.

If the ratio β = 1, i.e. $\theta_0 = \theta_1$, the curve is the random ROC curve model i.e. X and Y are identically distributed.

If the (β > 1, i.e. $\theta_0 < \theta_1$,) the curve is concave, lying above the main diagonal, and the corresponding marker is better than the random one i.e. the distribution of X lies entirely below the distribution of Y. The last option β < 1, $\theta_0 > \theta_1$, leads to a convex ROC curve indicating that the labels of groups Healthy and Diseased are possible swapped. All the above mentioned situations are shown in Figure. 3.

3.1 Estimation of the Parameter β

Now our parameter of interest is β. Since it the ratio of two unknown parameters θ_0 and θ_1. $\hat{\beta}$ the MLE of β is obtained using the plug in property of MLE i.e. $\hat{\beta} = \dfrac{\hat{\theta}_0}{\hat{\theta}_1}$.

Let $\{x_1, x_2, ..., x_{n1}\}$ and $\{y_1, y_2, ..., y_{n2}\}$ be the two independent random samples taken from the Generalized-Exponential-Distribution with parameters θ_0, and θ_1 respectively. The MLE, $\hat{\theta}_0$ and $\hat{\theta}_1$ are obtained as using equation (2.1.1). The MLE of β is of the form

$$\hat{\beta} = \frac{\dfrac{-n_2}{\sum \ln\left(1 - e^{-Y_i}\right)}}{\dfrac{-n_1}{\sum \ln\left(1 - e^{-X_i}\right)}} \qquad (3.1.1)$$

and note that $t_1 = -2\theta_0 \sum \ln\left(1 - e^{-X_i}\right)$ and $t_2 = -2\theta_1 \sum \ln\left(1 - e^{-Y_i}\right)$ have chi-squared distributions with ($2n_1$) and ($2n_2$) degrees of freedoms (d.f.) respectively.

3.2 The Sampling Distribution of $\hat{\beta}$

The sampling distribution of $\hat{\beta}$ can be obtained as follows:
Since

$$-2\theta_0 \sum \ln\left(1 - e^{X_i}\right) \sim \chi^2(2n_1)$$

and $$-2\theta_1 \sum \ln\left(1 - e^{-Y_i}\right) \sim \chi^2(2n_2)$$

where $\chi^2(m)$ is the pdf of the chi-squared distribution with m degrees of freedoms.

Hence the ratio of

$$\frac{\chi^2_{(2n_1)}}{\chi^2_{(2n_2)}} \sim F(2n_1, 2n_2)$$

where *F*(*a*, *b*) denotes the Snedelor's F-distribution with *a* and *b* degrees of freedoms. Analogously $\frac{n_2}{n_1}\frac{\beta}{\hat{\beta}} \sim F(2n_2, 2n_1)$

The knowledge of the distribution of $\hat{\beta}$ can be used for sample inference.

3.3 Testing Hypothesis of Equivalent

In this section we shall use the results of the pervious section, and testing approach of Lawless(1982,sec.3.3) for testing of hypotheses in (3.3.1).

Suppose $\{x_1, x_2, ..., x_{n1}\}$ is random sample of size n_1 from a GED(θ_0) and $\{y_1, y_2, ..., y_{n2}\}$ is independent random sample of size n_2 from a GED(θ_1). Where θ_0 and θ_1 are positive parameters their ratio, $\beta = \frac{\theta_1}{\theta_0}$ can be used to discriminate between the two distributions. To test distribution of diseased and healthy are identically distributed the hypotheses of interest and, a sized α test of is to reject H_0 if

$$\mathrm{H}_0 : \beta_0 = 1 \quad vs. \quad \mathrm{H}_1 : \beta_0 > 1 \qquad (3.3.1)$$

or

$$H_0 : \frac{\theta_1}{\theta_0} \le 1 \quad vs. \quad H_1 : \frac{\theta_1}{\theta_0} > 1$$

Using results of section (2.1), it is not difficult to show that the p value for testing (3.3.1) is given by

$$P\left(F_{(2n_2, 2n_1)} < \left(\frac{n_1}{n_2}\right)\frac{t_2}{t_1} \right)$$

where t_1 and t_2 defined as in section 3.1.

The null hypothesis in (3.3.1) will be rejected whenever this p-value is less than α.

3.4 Confidence Interval of β

The confidence limits for the ratio, $\beta = \frac{\theta_1}{\theta_0}$ can easily be obtained using the distributional result $\hat{\beta} = \frac{\hat{\theta}_0}{\hat{\theta}_1} \sim F(2n_2, 2n_1)$. Let $F_{\gamma 1}$ and $F_{1-\gamma 2}$ be the $(1-\gamma_1)$ and γ_2

quantiles of the $F(2n_2,2n_1)$ distribution respectively. The $(1-\gamma_1)$ – confidence interval for β is of the form

$$P\left[\hat{\beta}\frac{1}{F_{\gamma_2}} \le \beta \le \hat{\beta}\frac{1}{F_{\gamma_1}}\right] = (1-\alpha)$$

Acknowledgements

I am grateful to two referees for their valuable and painstaking comments. One of the referee has suggested to extend this work by producing simulated results. The work is in progress in direction and will be presented in next paper.

References

1. Dorfman, D.D, Berbaum, K.S and Metz, C.E. (1997). Proper receiver operating characteristic analysis: The bigamma model. Academic Radiology. 4: 138-49.
2. Dorfman, D.D., Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals-rating method data. Journal of Mathematical Psychology. 6: 487-496.
3. England, W. L. (1999). An exponential model used for optimal threshold selection on ROC curves. Medical Decision Making. 8: 120-131.
4. Goddard, M. J, Hinberg I. (1990). Receiver Operating Characteristic ROC curves and non-normal data: an empirical study. Statistics in Medicine. 9: 325-337.
5. Gupta, R.D.and Kundu, D. (1999). Generalized exponential distribution. The Australian and New Zealand Journal of Statistics. 41: 173-188.
6. Kotz, S., Lumelskii, Y., Penskly, M. (2003). The Stress-Strength Model and its Generalizations .World Scientific Publishing: New Jersey.
7. Kundu, D., Gupta, R.D. (2007). Generalized exponential distribution: Existing results and some recent developments. Journal of Statistical Planning and Inference. 136: 3130-3144.
8. Lawless, J.F. (1982). Statistical Models and Methods for Lifetime Data. John Wiley: New York.
9. Lloyd, C. J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. Journal of the American Statistical Association, 93, 1356-1364.
10. Walsh, S. J. (1999). Goodness-of-fit issues in ROC curve estimation. Medical Decisions Making. 19: 193-201.
11. Zweig, M. H. & Campbell, G. (1993). Receiver operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clinical Chemistry. 561-577.