ARTICLE TEMPLATE

Beta Regression for Modeling the ROC as a function of Continuous Covariates

Sarah Stanley[a] and Jack D. Tubbs[b]

Department of Statistical Science, Baylor University, Waco, Texas, USA

ARTICLE HISTORY
Compiled December 9, 2019

ABSTRACT
The receiver operating characteristic (ROC) curve is a well-accepted measure of accuracy for diagnostic tests. In many applications, a test's performance is affected by covariates. As a result, several regression methodologies have been developed to model the ROC as a function of covariate effects within the generalized linear model (GLM) framework. In this article, we present an alternative to two existing, a parametric and semi-parametric, methods for estimating a covariate adjusted ROC. The parametric and semi-parametric methods utilize generalized linear models for binary data who's expected value is the probability that the test result for a diseased subject exceeds that of a non-diseased subject with the same covariate values. This probability is referred to as the placement value. The new method directly models the placement values. The proposed method is compared with the existing models with simulation and two clinical studies.

KEYWORDS
Placement values

1. Introduction

A long standing problem in the testing literature is to determine and control how covariates affect a test's ability to distinguish between two, non-diseased and diseased, populations. The receiver operating characteristic (ROC) curve is commonly used as a measure of accuracy for diagnostic tests. Pepe(1998) provides a review of three major approaches to ROC regression. In this article, we focus on the approach that directly models the ROC as opposed to; modeling the underlying distributions of test response for the diseased and non-diseased populations. Advantages to this approach include; the accommodation of multiple test types, use of continuous covariates, and the ability to restrict the model to the portion of the ROC that is of interest. When originally proposed, Pepe's approach was difficult to implement, but simplifications have been made. In particular, Pepe(2000) proposed a generalized linear model framework for the ROC given by

$$ROC_X(t) = g(h_0(t)) + X'\beta), \tag{1}$$

CONTACT S. Stanley. Email: sarah_stanley@baylor.edu
I'm making suggests in this version - use as suggestions only - rewrite as you see best

for $t \in (0,1)$ where g is a monotone link function, X is a vector of covariates, $h_0(\cdot)$ is a monotonic increasing function and β is a vector of the model parameters.

Alonzo and Pepe(2002) further increased the utility of (1) by specifying a parametric form for $h_0(\cdot)$ and using a binary indicator as an outcome variable. Thus, rather than perform pairwise comparisons between each observation (as in the Mann-Whitney statistics), Alonzo and Pepe compared each diseased observation to a specified set of covariate-adjusted quantiles for the non-diseased population. The resultant binary value could then be modeled using a logistic regression approach.

I don't think placement value when considering Alonzo. Should we save this for the next section? In addition to the reduced number of comparisons, the advantages of Alonzo and Pepe's method include a simplified conceptual framework in which one can interpret the ROC in terms of placement values. A placement value is a right hand probability that can be measured by the survival function. Thus, for a diseased observation Y_D, we have that the placement value in terms of the reference survival function is defined as $PV_D = S_{\bar{D}}(Y_D)$. That is, given a diseased observation (Y_D), the placement value is found by mapping Y_D onto the reference population and finding the probability that a randomly selected reference individual will have a test response greater than Y_D. The ROC curve is thus equivalent to the cdf of the diseased placement values PV_D. We have

$$\begin{aligned} P[PV_D \leq t|X] &= P[S_{\bar{D}X}(Y_D) \leq t|X] \\ &= P[Y_D \geq [S^{-1}_{\bar{D}X}(t)|X] \\ &= ROC_X(t). \end{aligned}$$

Cai(2004) proposed a semi-parametric model by demonstrating that (1) is equivalent to $h_0(PV_D) = -X'\beta + \epsilon$, where $h_0(\cdot)$ is unknown. Implementation of this model is dependent upon pairwise comparisons of the placement value, given by $PV_D = \Pr[Y_D > Y_{\bar{D}}|X]$, to estimate the covariate effects β that are then included as an offset in the estimation of h_0.

The outline for this article is as follows. In section 2, we describe in greater detail the three models considered in this article. Section 3 contains simulation results comparing the performances of the three methods. Section 4 includes a data example, and we conclude with a discussion in section 5.

2. Methods

In this section, two existing methods are briefly discussed before we present a new method. Each of the methods make use of a term defined by Pepe (???), called the placement value. For completeness, we provide a brief discussion of placement vales.

2.1. *Placement Values - PV_D*

A placement value is a right hand probability that can be measured by the survival function. Thus, for a diseased observation Y_D, we have that the placement value in

terms of the reference survival function is defined as $PV_D = S_{\bar{D}}(Y_D)$. That is, given a diseased observation (Y_D), the placement value is found by mapping Y_D onto the non-diseased population and finding the probability that a randomly selected non-diseased subject will have a test response greater than Y_D. The ROC curve is thus equivalent to the cdf of the placement values PV_D, where

$$\begin{aligned}\Pr[PV_D \leq t|X] &= \Pr[S_{\bar{D}X}(Y_D) \leq t|X] \\ &= \Pr[Y_D \geq S^{-1}_{\bar{D}X}(t)|X] \\ &= ROC_X(t).\end{aligned}$$

2.2. *Parametric Approach*

Our objective is to determine the effect a covariate X has on the accuracy of a diagnostic test Y, where larger values of Y indicate disease. Let Y_D denote the test result for an observation from the diseased population and $Y_{\bar{D}}$ denote the test result for an observation from the non-diseased (reference) population. Suppose that we classify a subject as being from the diseased population if $Y \geq c$. The test's true positive rate is, $\text{TPR}(c) = \Pr[Y \geq c|D]$. Similarly, the test's false positive rate is, $\text{FPR(c)} = \Pr[Y \geq c|\bar{D}]$. go ahead and use the equation that defines the ROC as a function of these two rates. The ROC curve defined as the set of all TPR-FPR pairs quantifies the separation between the diseased and healthy populations.

Let X denote covariates common to both populations, such as age and bmi. Let X_D denote covariates that are specific to the diseased group, such as, disease duration, severity or previous treatment. The ROC can be written as

$$\text{ROC}_{X,X_D}(t) = S_{D,X,X_D}(S^{-1}_{\bar{D},X}(t)), \tag{2}$$

for $t \in (0,1)$, $S_{D,X,X_D}(c) = P(Y_D \geq c|X, X_D)$ and $S_{\bar{D},X}(c) = P(Y_{\bar{D}} \geq c|X)$ are survival functions at threshold c. In which case, the $\text{ROC}_{X,X_D}(t)$ is the probability that a test result, Y_D, for a diseased subject is greater than or equal to the t^{th} quantile for the covariate adjusted test results of non-diseased subjects.

Alonzo and Pepe (2002) proposed a parametric extension of (1) as,

$$\text{ROC}_{X,X_D}(t) = g(\gamma_1 h_1(t) + \gamma_2 h_2(t) + \beta X + \beta_D X_D), \tag{3}$$

where $h_1(t) = 1, h_2(t) = \Phi^{-1}(t)$, and $g(\cdot) = \Phi(\cdot)$ where $\Phi(\cdot)$ is the cdf of the standard normal. Note, this approach is known as a parametric distribution free method because a parametric model is specified for the ROC, but no assumptions are made about the distributions of the test results Y_D and $Y_{\bar{D}}$.

The parametric model, (3), follows from Pepe(2000) where the ROC is written as the expectation of the binary indicator $U_{ij} = I[Y_{D_i} \geq Y_{\bar{D}_j}]$ for all pairs of observations $\{(Y_{D_i}, Y_{\bar{D}_j}), i = 1, \ldots, n_D; j = 1, \ldots, n_{\bar{D}}\}$, with n_D and $n_{\bar{D}}$ denoting the number of observations from the diseased and reference populations, respectively. Alonzo and Pepe(2002) proposed a modification by replacing $Y_{\bar{D}_j}$ with $S^{-1}_{\bar{D},X_i}(t)$, for $t \in T = \{n_T \text{ chosen values of FPRs } \in (0,1)\}$. In which case, the binary indicator becomes $U_{it} = I[Y_{D_i} \geq S^{-1}_{\bar{D},X_i}(t)]$. the benefit of this method is not the computation but rather the ability to adjust the reference population with covariates, one can't do this with

the observed $Y_{\bar{D}}$. Note, one can write

$$U_{it} = I[Y_{D_i} \geq S^{-1}_{\bar{D},X_i}(t)] = \Pr[S_{\bar{D},X_i}(Y_D) \leq t] = \Pr[PV_D \leq t],$$

where PV_D is the placement value for the observation Y_D given the covariate vector X. An algorithm for (3) can be written as

(1) Specify a set $T = \{t_\ell : \ell = 1, ..., n_T\} \in (0, 1)$ of FPRs.
(2) Estimate the covariate specific survival function $S_{\bar{D}X}$ for the reference population at each $t \in T$ using quantile regression.
(3) For each diseased observation y_{D_j}, calculate the placement values $PV_j = \hat{S}_{\bar{D}\mathcal{X}_{D_j}}(y_{D_j})$.
(4) Calculate the binary placement value indicator $\hat{B}_{jt} = I[PV_j \leq t], t \in T, j = 1, \ldots, n_D$.
(5) Fit the model $E[\hat{B}_{jt}] = g^{-1}\left(\sum_{k=1}^{K} \alpha_k h_k(t) + X'\beta\right)$.

In step (1), we specify a set of n_T FPRs, T, that are usually equally spaced. Recall, the ROC can be summarized by the AUC which is an extension of the Mann Whitney statistic formed by making n_d to $n_{\bar{d}}$ comparisons. In the parametric approach, Alonzo and Pepe make n_d to n_T comparisons. In step (2), we estimate the covariate adjusted survival curve for the reference group using quantile regression from which we obtain a non-diseased marker for each $t \in T$. This set of n_T markers will be used in step (3) to calculate the placement values. Estimation of the reference survival curve is illustrated in Figure 1. I didn't include graph since I didn't have it

To calculate the placement values of the diseased points, recall that for a diseased observation Y_D the placement value is found by evaluating the estimated reference survival curve at Y_D. Thus, in step (3) we determine where each of the n_d diseased points lies in relation to the markers found from the quantile regression and calculate the right hand probability. After obtaining the n_d placement values, we create a binary indicator $\hat{B}$ by performing n_d to t comparisons between the placement values and the set of FPRs t. Note, the ROC can be modeled as the conditional expectation of $B_{D_t} = I[PV_D \leq t]$. Thus, in step (5), we model $E[\hat{B}_{jt}]$ using a probit link to obtain a covariate adjusted estimate of the ROC.

2.3. *Semi-parametric Approach*

Cai and Pepe(2002) proposed a semi-parametric method by allowing an arbitrary non-parametric baseline function $h_0(\cdot)$ in (1). Their approach required the simultaneous estimation of $h_0(\cdot)$ and β. Cai(2004) introduced a new method of estimating parameters for the semi-parametric model by showing that (1) is equivalent to $h_0(PV_D) = -X'\beta + \epsilon$, where ϵ is a random variable with known distribution g and $h_0(\cdot)$ is an unspecified increasing function. Cai used pairwise comparison of placement values to estimate β, before estimating the baseline function $h_0(\cdot)$. An algorithm for implementing the semi-parametric approach is as follows.

(1) Specify a set $T = \{t_\ell : \ell = 1, ..., n_T\} \in (0, 1)$ of FPRs.
(2) Estimate the covariate specific survival function $S_{\bar{D}X}$ via quantile regression.
(3) Calculate the placement values $PV_j = \hat{S}_{\bar{D}\mathcal{X}_{D_j}}(y_{D_j})$.

(4) Calculate the binary placement value indicator

$$\hat{B}_{jt} = I[PV_j \leq t], t \in T, j = 1, \ldots, n_D.$$

(5) For each pair of observations in Y_D, calculate

$$\widehat{PV}_{j\ell} = I[PV_j \leq PV_\ell], \text{ and } x_{j\ell} = x_{D_j} - x_{D_\ell}$$

with $j, \ell = 1, \ldots, n_D, j \neq \ell$.

(6) Fit the following GLM without an intercept to estimate $\boldsymbol{\beta}$

$$g(\widehat{PV}) = -\boldsymbol{X}'\boldsymbol{\beta}.$$

(7) Estimate $h_0(\cdot)$ using $\hat{\boldsymbol{\beta}}$ and $\hat{B}_{jt}$ as follows

$$g(E[\hat{B}_{jt}]) = intercept + \text{offset}(\boldsymbol{X}'\hat{\boldsymbol{\beta}}).$$

Note that steps (1) - (4) are identical to those of the parametric method. The first difference between the two approaches appears in step (5), where we create a second binary indicator describing the relationship between each pair of placement values. In this step, we also calculate the pairwise differences for each covariate value. We then fit a GLM without an intercept to the binary indicator created in step 5, adjusting for covariates using the pairwise differences. From this model, we obtain an estimate for β. In step (7), we then estimate $h_0(\cdot)$ by modeling the binary indicator $\hat{B}$ as a function of the intercept and an offset term that accounts for $\hat{\beta}$ from step (6).

2.4. *Beta Approach*

The parametric and semi-parametric approaches to estimating the covariate adjusted ROC given in equation (1) made use of the binary random variable defined by the placement values of the diseased response as referenced with the non-diseased population. In this section, we present a method for modeling the covariate adjusted ROC with the placement values as opposed to a binary indicator based upon the placement values. Our method makes use of the beta regression models that are readily available with GLM software.

Before describing an algorithm for the proposed method, we briefly introduce the beta regression model as referenced in Ferrari 2004. Recall that the mean and variance of $Y \sim \text{Beta}(a, b)$ are

$$E(Y) = \frac{a}{a+b} \text{ and } Var(Y) = \frac{ab}{(a+b)^2(a+b+1)}.$$

We will define the beta regression model in terms of $\mu = E(Y)$ and a precision parameter $\phi = a + b$ so that the reparameterized beta distribution mean and variance are

$$E(Y) = \mu \text{ and } Var(Y) = \frac{\mu(1-\mu)}{1+\phi}.$$

Let $y_1, \ldots, y_n$ be independent random variables from a beta density with mean μ_t, t

= 1,. . . ,n and scale parameter ϕ. Then the beta regression model can be written as

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t,$$

where $\boldsymbol{\beta}$ is a vector of regression parameters, $x_{t1}, \ldots, x_{tk}$ are observations on k covariates, and g is a monotonic link function. Using the logit link, we have $\mu_t = \frac{1}{1+e^{-x_t'\beta}}$. We can thus obtain the original parameters a and b from the beta distribution by calculating

$$\hat{a} = \frac{\hat{\phi}}{1 + e^{-x_t'\hat{\beta}}} \text{ and } \hat{b} = \hat{\phi}\left(1 - \frac{1}{1 + e^{-x_t'\hat{\beta}}}\right).$$

An algorithm for the beta method can be written as follows.

(1) Specify a set $T = \{t_\ell : \ell = 1, \ldots, n_T\} \in (0, 1)$ of FPRs.
(2) Estimate the covariate specific survival function $S_{\bar{D}X}$ via quantile regression.
(3) Calculate the placement values $PV_j = \hat{S}_{\bar{D}\mathcal{X}_{D_j}}(y_{D_j})$.
(4) Perform a beta regression on the placement values to obtain estimates of $\boldsymbol{\beta}$ and ϕ.
(5) Transform to obtain $a = \mu\phi$ and $b = (1 - \mu)\phi$.
(6) Calculate the cdf of the placement values using the Beta(a,b) distribution found above to obtain the ROC and the AUC.

Steps (1) - (3) are identical to the parametric and semi-parametric cases. In step (4), we model the placement values directly using beta regression to obtain estimates of β and ϕ, instead of calculating a binary indicator. We then apply a transformation to return to the original beta parameters a and b and calculate the cdf of the placement values using the resulting Beta(a, b) distribution which yields an estimate for the ROC.

3. Simulation

To compare the three presented methods we perform two simulations, one with binormal data and one using the extreme value distribution. When both populations are normally distributed, exact solutions for the ROC and AUC exist. We have $ROC(t) = \Phi(a + b\Phi^{-1}(t))$, and $AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$, where $a = \frac{\mu_D - \mu_{\bar{D}}}{\sigma_D}, b = \frac{\sigma_{\bar{D}}}{\sigma_D}$. We are thus able to compare AUC estimates from each of the three presented methods with the truth. For this example we simulate data from

$$Y_D = 2 + 4X + \epsilon_D \text{ and } Y_{\bar{D}} = 1.5 + 3X + \epsilon_{\bar{D}},$$

where $X \sim U(0, 1)$ and $\epsilon_D, \epsilon_{\bar{D}} \sim N(0, 1.5^2)$. That is, $Y_D \sim N(2 + 4X, 1.5^2)$ and $Y_{\bar{D}} \sim N(1.5 + 3X, 1.5^2)$. Thus, the true AUC at covariate value $X = x_0$ is

$$AUC(x_0) = \Phi\left(\frac{\mu_D - \mu_{\bar{D}}}{(\sigma_D^2 + \sigma_{\bar{D}}^2)^{1/2}}\right) = \Phi\left(\frac{0.5 + x_0}{\sqrt{4.5}}\right).$$

We simulate 300 data sets from the binormal scenario presented above, and summarize the mean of the AUC estimates for each method at specified covariate values in Table 1. Beyond showing that the means across the three methods are comparable and addressing standard deviations, what should I be emphasizing here? Does coverage make sense?

	$x_0 = 0.5$	$x_0 = 0.6$	$x_0 = 0.7$	$x_0 = 0.8$
Parametric	0.6791	0.7426	0.7983	0.8456
Semiparametric	0.6642	0.7124	0.7567	0.7964
Beta	0.6699	0.7305	0.7831	0.8274
Truth	0.6813	0.6980	0.7142	0.7300

Table 1. Mean AUC estimates across 300 data sets for each method

Extreme Value Results

4. Example

Alonzo example?
DME Protocol I???

5. Discussion

Given the broad acceptance of the ROC curve as a measure of accuracy for diagnostic tests, our intent was to investigate the effect of covariates on a test's performance through ROC regression. As noted, several regression methodologies have been developed to model the ROC as a function of covariate effects within the generalized linear model (GLM) framework. In particular, the parametric and semi-parametric approaches estimate the ROC using binary indicators. The use of such indicators, however, leads to additional correlation in the model that must be accounted for through methods such as bootstrapping. In this paper, we proposed a new approach that implements beta regression to model the placement values directly, thereby eliminating the additional correlation induced by the pre-existing methods. We compared our beta methodology with the parametric and semi-parametric approaches via simulation, showing that the new method yields comparable ROC estimates without inducing additional correlation.
This is currently a rewording of the abstract... more work to be done here.

6. Bibliography

Need to create a bibliography for your dissertation that can be used for all your work

7. References

{Transfer to Bibtex OK}

- Alonzo, T. and M. Pepe (2002), "Distribution-free ROC analysis using binary regression techniques," *Biostatistics*, 3, 421-432.
- Bamber, D. (1975), "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology*, 12, 387-415.
- Cai, T. (2004), "Semi-parametric ROC regression analysis with placement values," *Biostatistics*, 5, 45-60.
- Ferrari, S. and Cribari-Neto, F. (2004), "Beta Regression for Modelling Rates and Proportions," *Journal of Applied Statistics*, 31, 799-815.
- Pepe, M. and T. Cai (2002), "The analysis of placement values for evaluating discriminatory measures," *UW Biostatistics Working Paper Series.* Working Paper 189.
- Rodriguez-Alvarez, M.X. et. al. (2011) "Comparative Study of ROC regression techniques," *Computational Statistics and Data Analysis,* 55, 888-902.