Science Journal of Applied Mathematics and Statistics
2018; 6(4): 110-118
http://www.sciencepublishinggroup.com/j/sjams
doi: 10.11648/j.sjams.20180604.11
ISSN: 2376-9491 (Print); ISSN: 2376-9513 (Online)

Beta Regression for Modeling a Covariate Adjusted ROC

Sarah Stanley, Jack Tubbs*

Department of Statistical Science, Baylor University, Waco, USA

Email address:
Jack_Tubbs@baylor.edu (J. Tubbs)

*Corresponding author

To cite this article:
Sarah Stanley, Jack Tubbs. Beta Regression for Modeling a Covariate Adjusted ROC. *Science Journal of Applied Mathematics and Statistics*. Vol. 6, No. 4, 2018, pp. 110-118. doi: 10.11648/j.sjams.20180604.11

Received: July 24, 2018; **Accepted**: August 9, 2018; **Published**: September 11, 2018

Abstract: *Background:* 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 parametric and semi-parametric methods for estimating a covariate adjusted ROC. The existing methods utilize GLMs for binary data when the expected value equals 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. *Objective:* The new method directly models the placement values through beta regression. *Methods:* We compare the proposed method to the existing models with simulation and a clinical study. *Conclusion:* The proposed method performs favorably with the commonly used parametric method and has better performance than the semi-parametric method when modeling the covariate adjusted ROC regression.

Keywords: Placement Values, Beta Regression, ROC Regression

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 populations. A widely used measure of accuracy for diagnostic tests is the receiver operating characteristic (ROC) curve. One approach was to modify the nonparametric Mann-Whitney statistics (MW) using GLM for the AUC regression [1]. Additional results followed this approach for modeling the MW in the presence of covariates [2-5]. Pepe chose to model the ROC directly where she provides a review of three major approaches to ROC regression that account for covariate effects [6]. In this article, we direct our attention to the approach that directly models the ROC as opposed to modeling the underlying distributions of test responses 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 proposed a generalized linear model framework for the ROC given by

$$ROC_X(t) = g(h_0(t) + X'\beta) \quad (1)$$

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 [7].

Alonzo and Pepe expanded the utility of the ROC-GLM in (1) by specifying a parametric form for $h_0(\cdot)$ and using a binary indicator as an outcome variable [8]. Thus, rather than perform pairwise comparisons between each observation from the diseased and non-diseased samples (as in the Mann-Whitney statistic), they compared each diseased observation to a specified set of covariate-adjusted quantiles for the non-diseased population [8]. The resultant binary values could then be modeled using a logistic regression approach.

Pepe and Cai extended the parametric ROC-GLM by allowing a non-parametric form for $h_0(\cdot)$ [9]. This semi-parametric approach hinged upon the idea that the ROC is the cumulative probability distribution of placement values, where a placement value is the probability that the test result for a diseased subject exceeds that of a non-diseased subject at the same covariate value. Cai further developed the semi-

parametric approach by demonstrating that (1) is equivalent to $h_0(PV_D) = -X'\beta + \varepsilon$, where $h_0(\cdot)$ is unknown and PV_D is the set of placement values for the diseased observations [10]. Implementation of the semi-parametric model is dependent upon pairwise comparisons of the placement values to estimate the covariate effects β that are then included as an offset in the estimation of $h_0(\cdot)$.

Pepe and Cai's use of placement values motivates the development of an alternative approach to modeling the covariate-adjusted ROC [9-10]. Given that the ROC is the cumulative distribution of the placement values for the diseased observations, we propose a new method that directly models the placement values using beta regression. In this article, we show that this third approach is not only easy to implement, but it also removes the need for pairwise comparisons, eliminating the dependency among the response variable induced by the preceding methods.

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

Recall that our objective is to determine the effect of covariates on the accuracy of a diagnostic test. Before detailing three methods to achieve this objective, we briefly introduce the ROC as a measure of test accuracy as well as the notation used for a covariate adjusted ROC.

Let Y denote the variable that will be used to distinguish between the reference or non-disease population (D = 0) and the diseased population (D = 1). Suppose that we classify a subject as being from the diseased population if $Y \geq c$. Then the test's true positive rate is $TPR(c) = Pr[Y \geq c \mid D = 1]$. Similarly, the test's false positive rate is $FPR(c) = Pr[Y \geq c \mid D = 0]$. The ROC curve, defined as the set of all TPR-FPR pairs, quantifies the separation between the diseased and non-diseased populations. The ROC has many forms in the literature. In this paper, we restrict our attention to the survival curve and the placement values given by,

$$ROC(t) = S_1(S_0^{-1}(t)) = Pr[PV_D \leq t],$$

for $t \in (0,1)$, where S_1, S_0 are survival functions for the diseased and non-diseased populations, respectively and PV_D represents the placement values for the diseased subjects.

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, disease severity, or previous treatment.

The covariate-adjusted ROC can then be written as

$$ROC_{X,X_D}(t) = S_{1,X,X_D}\left(S_{0,X}^{-1}(t)\right), for\ t \in (0,1), \qquad (2)$$

where $S_{1,X,X_D}(c) = Pr[Y \geq c \mid X, X_D, D = 1]$ and $S_{0,X}(c) = Pr[Y \geq c \mid X, D = 0]$ are survival functions at threshold c. Thus, $ROC_{X,X_D}(t)$ is the probability that a test result, Y, for a diseased subject is greater than or equal to the t^{th} quantile for the covariate adjusted test results of non-diseased subjects. For completeness, we again note that the ROC is the cumulative distribution function of the placement values PV_D as seen in the following covariate-adjusted notation,

$$Pr[PV_D \leq t|X] = Pr[S_{0,X}(Y) \leq t|X, D = 0] = Pr\left[Y \geq S_{0,X}^{-1}(t)\middle|X, D = 0\right] = ROC_X(t).$$

Having established notation, we now introduce three ROC-GLM methods.

2.1. Parametric Approach

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

$$ROC_{X,X_D}(t) = g(\gamma_1 h_1(t) + \gamma_2 h_2(t) + \beta X + \beta_D X_D), \quad (3)$$

with γ_1, γ_2, β, and β_D as model parameters, $h_1(t) = 1$, $h_2(t) = \Phi^{-1}(t)$, and $g(\cdot) = \Phi(\cdot)$ where $\Phi(\cdot)$ is the cdf of the standard normal [8]. Their approach is known as a parametric distribution free method because a parametric model is specified for the ROC. It should be noted that no assumptions are made about the distributions for Y for D = 0 or D = 1 [7].

The parametric model, (3), follows from [1] where the ROC is written as the expectation of the binary indicator $U_{ij} = I[Y_{i|D=1} \geq Y_{j|D=0}]$ for all pairs of observations $\{(Y_{i|D=1}, Y_{j|D=0}), i = 1,\ldots, n_1; j = 1,\ldots, n_0\}$, with n_1 and n_0 denoting the number of observations from the diseased and non-diseased populations, respectively, and I denoting the indicator function.

Alonzo and Pepe proposed a modification by replacing $Y_{j|D=0}$ with $S_{0,X_i}^{-1}(t)$, for $t \in T = \{n_T$ chosen values of FPRs $\in (0, 1)\}$ [8]. In this case, the binary indicator becomes $U_{it} = I[Y_{i|D=1} \geq S_{0,X_i}^{-1}(t)]$. Note, the expected value of U_{it} satisfies $E(U_{it}) = E(I[Y_{i|D=1} \geq S_{0,X_i}^{-1}(t)]) = Pr[S_{0,X_i}(Y_{i|D=1}) \leq t] = Pr[PV_D \leq t]$, where PV_D is the placement value for the observation $Y_{i|D=1}$ given the covariate vector X. An algorithm for (3) can be written as

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_{0,X_j} for the reference population at each $t \in T$, $j = 1,\ldots, n_0$ using quantile regression.
3. For each diseased observation $Y_{i|D=1}$, calculate the placement values $PV_i = \hat{S}_{0,X_i}(Y_{i|D=1})$, $i = 1, \ldots, n_1$.
4. Calculate the binary indicator $\widehat{U}_{it} = I[PV_i \leq t], t \in T$.
5. Fit the model $E[\widehat{U}_{it}] = g^{-1}[\sum_{k=1}^{K} \gamma_k h_k(t) + X'\beta]$.

In step (1), we specify a set of n_T false positive rates (FPRs), where in practice the FPRs are equally spaced. In step (2), we estimate the covariate-adjusted reference survival curve using quantile regression on the set of FPRs. The quantile regression yields n_T covariate adjusted estimates of the reference survival curve for each $Y_{j|D=0}$. In step (3), we calculate the placement values for each diseased observation $Y_{i|D=1}$. The placement values are calculated by evaluating the covariate-adjusted reference survival curve at each $Y_{i|D=1}$,

resulting in n_1 probabilities. We next create a binary indicator $\widehat{U}_{it}$ in step (4) by performing n_1 to n_T comparisons between the placement values and the set of FPRs. Note that step (4) is similar to the Mann Whitney statistic formed by making n_1 to n_0 comparisons from which we can derive the area under the curve (AUC), [11]. In step (5), the covariate adjusted ROC is obtained by modeling the expectation of $\widehat{U}_{it}$ using a probit link.

2.2. Semi-parametric Approach

Pepe and Cai extended the parametric approach by proposing a semi-parametric method allowing for an arbitrary non-parametric baseline function $h_0(\cdot)$ in (1) [9]. Their approach required the simultaneous estimation of $h_0(\cdot)$ and β. Cai introduced a method of estimating parameters for the semi-parametric model þy demonstrating that (1) is equivalent to $h_0(PV_D) = -X\beta + \varepsilon$, where ε is a random variable with known distribution g, $h_0(\cdot)$ is an unspecified increasing function, and PV_D represents placement values for the diseased observations [10]. 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 $T = \{t_\ell: \ell = 1, \ldots, n_T\} \in (0,1)$ of FPRs.
2. Estimate the covariate specific survival function S_{0,X_j} for the reference population at each $t \in T$, $j = 1,\ldots, n_0$ using quantile regression.
3. Calculate the placement values $PV_i = \hat{S}_{0,X_i}(Y_{i|D=1})$, $i = 1, \ldots, n_1$.
4. Calculate the binary placement value indicator $\widehat{U}_{it} = I[PV_i \leq t], t \in T$.
5. For each pair of observations in Y_D, calculate $V_{ij} = I[PV_i \leq PV_j]$ and $x_{ij} = x_{D_i} - x_{D_j}$ with $i, j = 1, \ldots, n_1, i \neq j$.
6. Fit the following GLM without an intercept to estimate $g(V) = -X'\beta$.
7. Estimate $h_0(\cdot)$ using $g(E[\widehat{U}_{it}] = \text{intercept} + \text{offset}(X'\hat{\beta})$.

Note that steps (1) - (4) are identical to those of the parametric method. The 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. 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 $\widehat{U}$ as a function of the intercept and an offset term that accounts for $\hat{\beta}$ ([12]).

It should be noted that the parametric and semi-parametric models enable one to use readily available GLM software with either the logit or probit link functions. However, in both cases the binary data are no longer independent, hence the resultant standard errors produced by the software are not correct. The standard fix is to estimate the standard errors for the regression coefficients using bootstrap estimates [8]. In the next section, a procedure for modeling the covariate adjusted ROC is presented where the above constraint is no longer a consideration in the modeling problem. The approach is to model the placement values directly using the beta regression model.

2.3. Beta Approach

The parametric and semi-parametric approaches to estimating the covariate adjusted ROC given in equation (1) were dependent upon a binary random variable defined by the placement values of the diseased response as referenced with the non-diseased population. In this section, we present an alternative method that models the covariate-adjusted ROC as the cdf of the placement values directly and bypasses the need for a binary random variable. The beta regression model is used in this approach.

A brief introduction to the beta generalized linear model given in [12] is presented here. Suppose that Z ~ Beta(a, b), in which case,

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

By letting $\mu = \frac{a}{a+b}$ and $\phi = a + b$ we obtain a beta distribution with mean and variance $E(Z) = \mu$ and $Var(Z) = \frac{\mu(1-\mu)}{1+\phi}$.

Let $z_1,\ldots, z_n$ be independent random variables from a beta density with mean μ_t, $t = 1,\ldots, n$ 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 β 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 $\mu_t = \frac{1}{1+e^{-x_t'\beta}}$. Estimates of the original parameters a and b are

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

An algorithm for the proposed method using the beta distribution for the placement values 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_{0,X_j} for the reference population at each $t \in T$, $j = 1,\ldots, n_t$ using quantile regression.
3. Calculate the placement values $PV_i = \hat{S}_{0,X_i}(Y_{i|D=1})$, $i = 1, \ldots, n_1$.
4. Perform a beta regression on the placement values to obtain estimates of β and φ.
5. Transform to obtain $a = \mu\varphi$ and $b = (1-\mu)\varphi$.
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. We then apply equation (4) to obtain beta parameters a and b and calculate the cdf of the placement values using the resulting Beta(a, b) distribution that yields an estimate for the ROC. The AUC is obtained by integrating the Beta(a, b) cdf, which results in b/(a + b) by Fubini's theorem [14].

3. Simulation Studies

We compare the parametric, semi-parametric and beta ROC regression methods through two simulations, one using normally distributed data and the other using data from an extreme value distribution. Rodriquez-Alvarez et. al. provide a comparison of several indirect and direct ROC regression methods including the parametric and semi-parametric for binormal and extreme value data [12]. The data models in this section are similar to those in [12]. For simplicity, we consider one continuous covariate from a uniform distribution. The models and results follow.

3.1. Binormal Data

Suppose that $Y_1 \sim N(\mu_1,\sigma_1)$ and $Y_0 \sim N(\mu_0,\sigma_0)$. Then using $ROC(t) = S_1(S_0^{-1}(t)), for\ t \in (0,1)$, we derive the binormal ROC and AUC,

$$ROC(t) = \Phi[a + b\Phi^{-1}(t)]\ and\ AUC = \Phi[\tfrac{a}{\sqrt{1+b^2}}],$$

where $a = (\mu_1 - \mu_0)/\sigma_1$ and $b = \sigma_0/\sigma_1$.

The following models were used for the binormal simulation $Y_1 = 2 + 4X + \varepsilon_1$, and $Y_0 = 1.5 + 3X + \varepsilon_0$, where $X \sim U(0, 1)$ and $\varepsilon_1, \varepsilon_0 \sim N(0, 1.5^2)$. Given the model, the true ROC and AUC at covariate $X = x_0$, $t \in (0,1)$ are

$$ROC(t) = \Phi\left[\frac{0.5 + x_0}{1.5} + \Phi^{-1}(t)\right] and\ AUC(x_0) = \Phi\left[\frac{0.5 + x_0}{\sqrt{4.5}}\right].$$

We generate 1000 data sets of size $n_1, n_0 = 200$ from which we calculate the ROC and AUC for each of the three methods. We also compute the mean squared error (MSE) for the AUC of each method. Boxplots of the MSE values for each method are given in Figure 1. The summary statistics for the MSE of the AUC are given in Table 1. We note that the mean MSE and standard deviation for the parametric method are smaller than the corresponding results for the beta and semi-parametric methods. The beta mean MSE is, however, within one standard deviation of the parametric mean MSE. Plots of the simulated and true ROC curves are included in Figure 2 for covariate values $x_0 = \{0.2, 0.5, 0.8\}$. The dotted lines represent plus and minus two standard deviations from the simulated mean ROC. The cdf of a Uniform(0,1) distribution representing the ROC for identical populations is included for reference. Observe that the AUC increases with an increase in the covariate value.

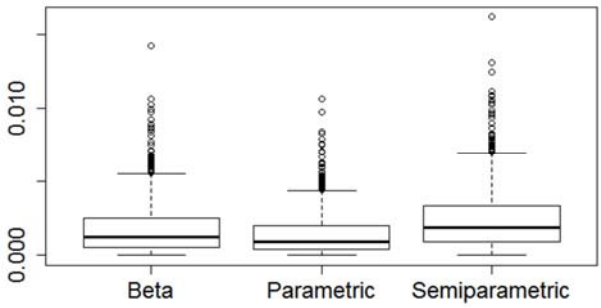

Figure 1. *Boxplots of the estimated MSE for the AUC of each method based on 1000 estimates ($n_1 = n_0 = 200$).*

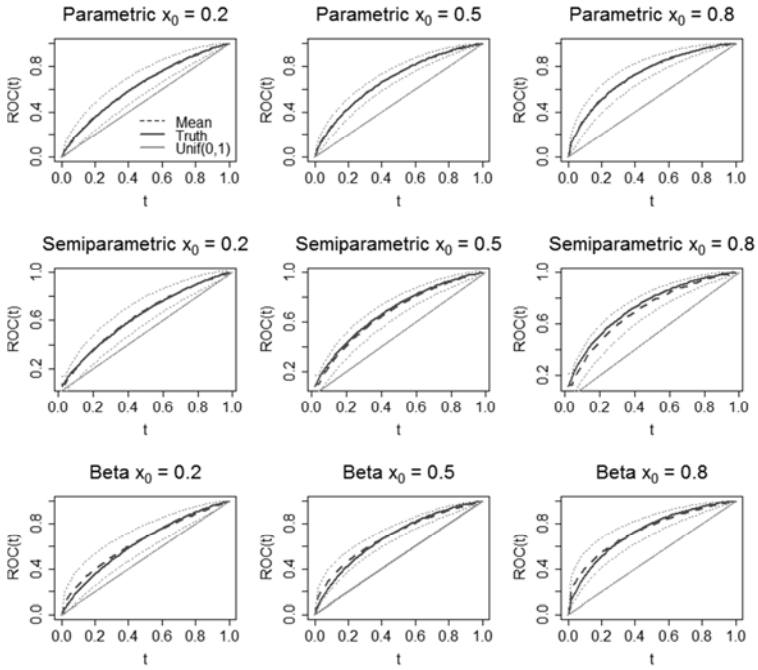

Figure 2. *Comparison of simulated ROC and true ROC for binormal data.*

Table 1. *Summary of MSEs for binormal.*

Method	Q1	Q2	Q3	Mean	SD
Beta	0.000521	0.001251	0.002544	0.001819	0.001831
Parametric	0.000383	0.000936	0.001996	0.001398	0.001446
Semi-parametric	0.000958	0.001912	0.003369	0.002459	0.002088

3.2. Extreme Value Data

The extreme value distribution used in the following models has a cdf of the form $F(x) = \exp\{-\exp[-(x-\mu)/\beta]\}$, where $\mu \in R$, $\beta > 0$, $x \in (-\infty,\infty)$. This extreme value distribution is also known as the Gumbel or double exponential distribution [15]. In choosing a model for simulation, we note that the extreme value distribution exhibits more sensitivity than the normal distribution to differences in location and scale for the two populations. Highly separated populations will yield an AUC of one regardless of covariate value. We thus consider scenarios such as the following in which the covariate effect can be assessed, $Y_1 = 2+2.5X + \varepsilon_1$, and $Y_0 = 1+2X + \varepsilon_0$, where $X \sim U(0,1)$ and ε_1, ε_0 have an extreme value distribution with $\mu = 0$ and $\beta = 1.5$. The true value of the ROC when $X = x_0$ is

$$ROC_X(t) = 1 - exp[-exp\{ln[-ln(1-t)] - \frac{1+0.5x_0}{1.5}\}].$$

We approximate the AUC using numerical integration. A plot of the densities for Y_1 and Y_0 appears in Figure 3 as well as the true ROC at covariate values 0, 0.5, and 1.

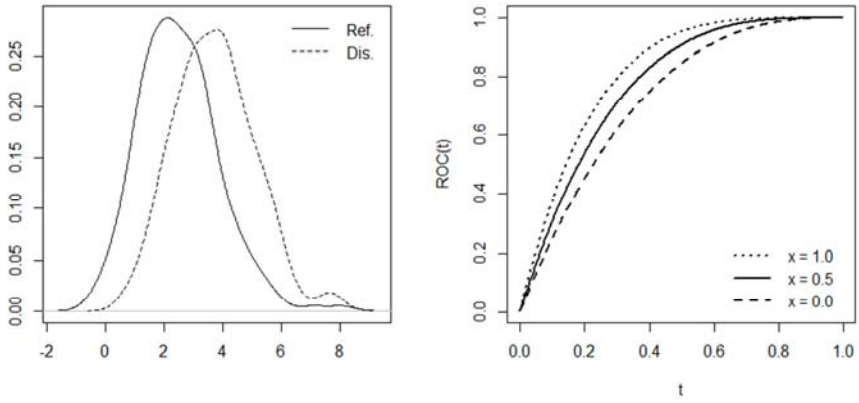

Figure 3. *Density plots and ROCs for extreme value data.*

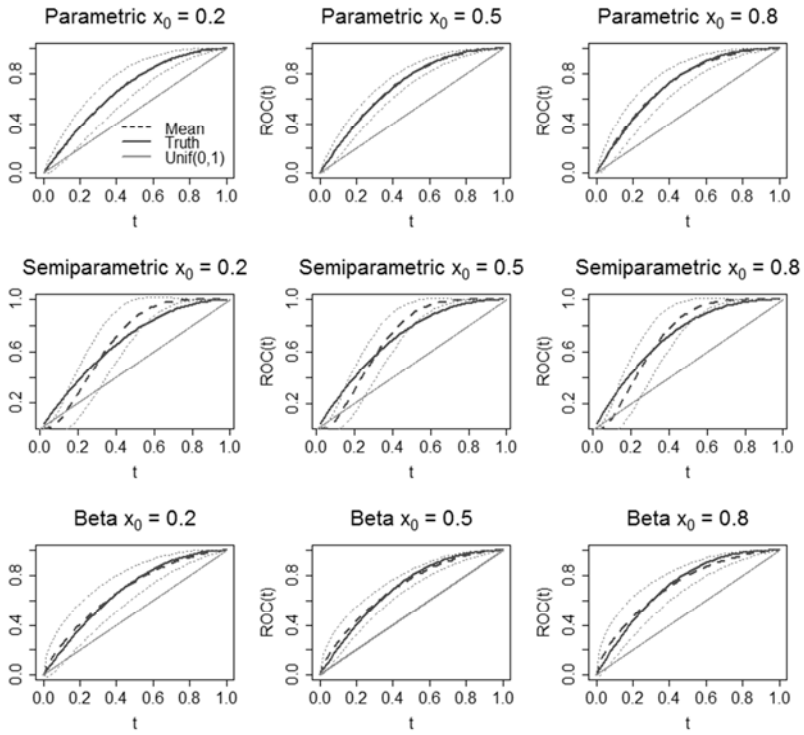

Figure 4. *Comparison of simulated ROC and true ROC for extreme value data.*

Table 2. *Summary of MSEs for extreme value.*

Method	Q1	Q2	Q3	Mean	SD
Beta	0.000528	0.001248	0.002578	0.001878	0.002567
Parametric	0.000406	0.000969	0.001932	0.001449	0.002046
Semi-parametric	0.000432	0.001038	0.002278	0.001712	0.002422

As in the binormal simulation, we generate 1000 data sets of size $n_1, n_0 = 200$, calculating the resulting ROC and AUC estimates from each of the parametric, semi-parametric, and beta methods and comparing to the truth using the MSE. The summary statistics for the MSE of the AUC are given in Table 2. We observe that the beta mean MSE and standard deviation are slightly larger than the corresponding results for the parametric method although the means are within one standard deviation of each other as in the binormal simulation. The semi-parametric mean MSE is smaller than that of the beta, but examination of Figure 3.2 shows that the beta method provides a better estimation of the true ROC. Plots of the simulated and true ROC curves are included in Figure 4 for covariate values $x_0 = \{0.2, 0.5, 0.8\}$. The dotted lines represent plus and minus two standard deviations from the simulated mean ROC.

3.3. Discussion

The binormal and extreme value simulations provide a comparison of the three ROC regression methods. When preforming the simulation with the normal data, we would have expected the parametric method to produce the best results, in terms of accuracy and minimum mean-square error. Yet, we were pleased that the proposed method performed very comparably and performed better than the semi-parametric method. In examining the ROC curves, one notes that the proposed method has somewhat higher values for small t. The placement value produces a large number of small values when the distributions being compared are somewhat widely separated, hence the resultant estimated beta parameters reflect this characteristic when generating the ROC. One should exercise care when using this method in these cases. However, one usually isn't bothered by widely separated density functions when defining a classifier or determining if the diseased population is widely different from the reference population. Rather, the cases of interest are when the two density functions are much closer together. In these cases, the proposed method performed very well. Similar results were found when using the extreme value distribution. We had hope that the proposed method would perform better than the parametric method with these data but that was not the case. The ROC produced by the proposed method performed better than the semi-parametric method. It is worth mentioning again that the advantage of the beta regression model over the pre-existing methods is the ability to directly model the placement values without the use of a binary indicator. Furthermore, these placement values are independent when the data are random. In which case, the resultant standard errors for the regression parameters are correct, thus avoiding the need for bootstrapping the standard error for the regression parameters. We have shown through simulation that the beta method is a viable alternative to the parametric and semi-parametric models and merits additional exploration. We further illustrate the performance of the parametric and beta approaches by considering clinical study as our motivating example.

4. Application to a DME Study

In this example, the parametric and beta methods are used for subject-specific data from the Protocol I study in the Diabetic Retinopathy Clinical Research Network (NCT 00444600) [16]. The study was designed to determine the efficacy of ranibizumab alone and ranibizumab in combination with laser therapy as compared to the efficacy of laser therapy alone in the treatment of diabetic macular edema (DME). In the study, each patient had been previously diagnosed with either type 1 or type 2 diabetes as well as diabetic macular edema affecting the center of the macula. The patients were randomized to one of four treatment groups. For the purpose of our example, we will consider two groups: A – a sham injection with laser treatment and B – a 0.5 mg injection of intravitreal ranibizumab along with laser treatment given three to ten days after injection. The primary outcome was visual acuity at one year adjusted for baseline acuity. Visual acuity was measured with Optical Coherence Tomography (OCT) which detects changes in retinal thickness, and the ETDRS test which records the number of letters that a patient can correctly identify. In this context, a favorable result is a decrease in retinal thickness which corresponds to vision improvement.

We define treatment A (laser therapy alone) to be the reference population and treatment B the comparator population. To investigate the performance of the ROC regression models, we define the response of interest to be the amount of decrease in retinal thickness from baseline at one year. If treatment B is effective, the amount of decrease in retinal thickness should be higher for patients in the comparator population (treatment B) than for those in the reference population (treatment A). Density plots of the response (decrease in OCT from baseline) and one year OCT values for each treatment group appear in Figure 5. We note a high degree of overlap in the responses for the two groups, implying that the resulting ROC will be close to the diagonal line.

Table 3. *Summary statistics for OCT and age by gender and duration.*

Subset	N	OCT Baseline	OCT One Year	Age
Trt A		Mean (SD)	Mean (SD)	Mean (SD)
Females	100	323.4(116.2)	303.0(112.2)	62.8(10.7)
Dur(0)	56	336.4(109.3)	296.3(85.8)	62.1(10.7)
Dur(1)	44	306.9(123.7)	311.6(139.3)	63.7(9.9)
Males	141	346.7(134.8)	305.4(111.7)	62.1(9.9)
Dur(0)	81	352.5(133.5)	313.2(119.1)	62.1(10.0)
Dur(1)	60	338.9(137.3)	294.8(100.9)	63.3(9.9)
Trt B		Mean (SD)	Mean (SD)	Mean (SD)
Females	55	297.5(103.7)	256.0(87.7)	63.0(10.8)
Dur(0)	23	300.3(124.6)	239.0(87.0)	61.0(11.6)
Dur(1)	32	295.5(87.7)	269.7(64.4)	64.4(10.1)
Males	64	306.2(93.4)	261.5(67.0)	60.7(9.8)
Dur(0)	29	324.4(106.3)	260.7(64.6)	58.1(8.8)
Dur(1)	35	291.1(79.6)	262.1(69.9)	62.8(10.3)

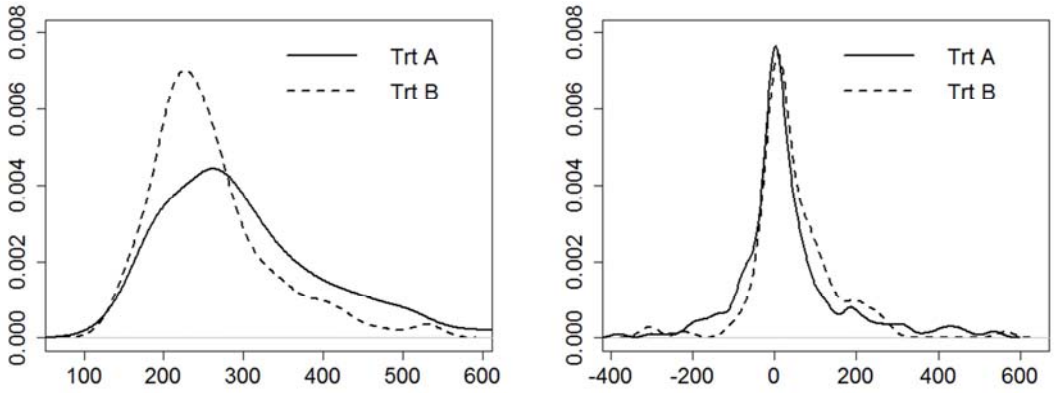

Figure 5. *Density plots of one year OCT measurements on the left and one year decrease in OCT from baseline on the right.*

We are interested in the effect of covariates on the separation between the populations. Covariates common to both populations are gender and age at enrollment, and for illustrative purposes, we assume that duration of diabetes is a covariate associated with the comparator population. In this example, duration is a binary variable with a value of 1 if the duration is greater than or equal to the median of 17 years, and zero otherwise. A summary for each population and covariates of interest is included in Table 3, and boxplots of age appear in Figure 6.

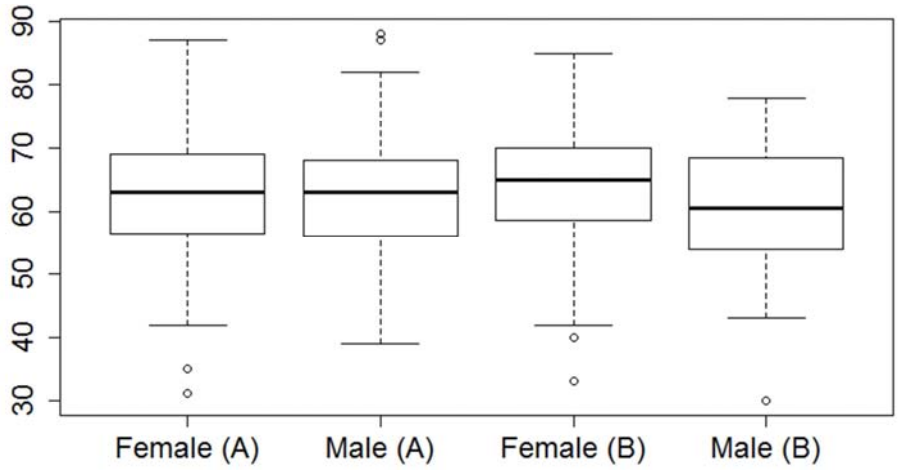

Figure 6. *Boxplots for age by treatment and gender.*

For the comparator treatment B, we note a slight difference in median age between males and females. The one year OCT measurements for treatment B are lower than those for treatment A. The amount of decrease in OCT measurements from baseline is slightly higher for those in treatment B and there appears to be very little gender effect. Each of the methods is performed for the following ROC-GLM

$$ROC_X(t) = g(h_0(t) + \beta_1 * age + \beta_2 * gender + \beta_3 * duration).$$

Plots of the resulting ROC curves for different covariate values appear in Figures 7 and 8 for the parametric and beta approaches, respectively. The dotted line represents a Uniform(0,1) cdf to illustrate the case for identical populations. Note that for both the parametric and beta methods, the AUC increases with age which indicates that the amount of decrease in OCT measurements from baseline was higher for older patients receiving treatment B.

As anticipated given the overlap in response densities (Figure 5), the ROC is nearly diagonal when accounting for covariates.

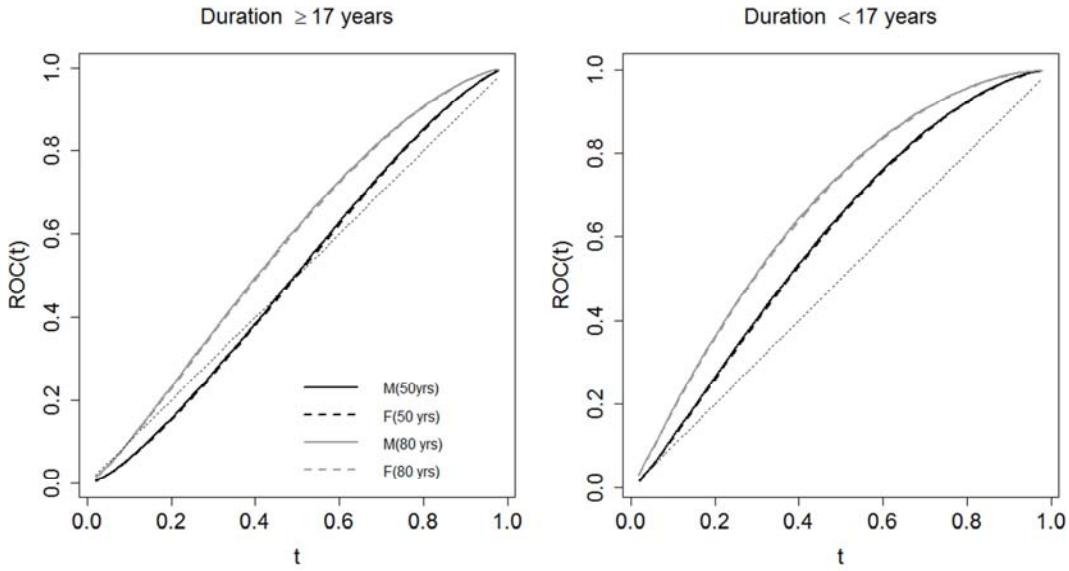

Figure 7. *Covariate-adjusted ROC curves from the parametric method for males and females at ages 50 and 80.*

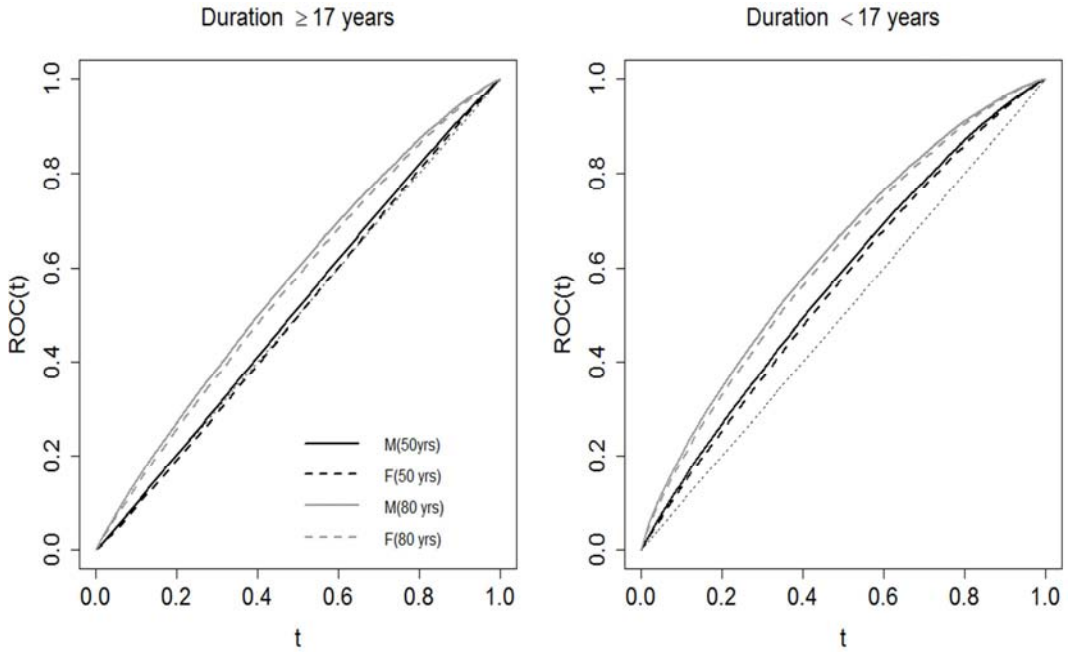

Figure 8. *Covariate-adjusted ROC curves from the beta method for males and females at ages 50 and 80.*

It should be noted that, in this study all the subjects are diabetic with macular edema. This disease is degenerative and patients seldom show long-term improvement with any of the treatment methods. We chose this study since we were able to have unidentified subject-wise responses with discrete and continuous covariates. Although the results of this study are not spectacular, one does note that the treatment (laser plus ranibizumab) is more effective in those subjects who have had diabetes for less than 17 years and are younger. There does not seem to be any differences in gender. Furthermore, the treatment is not effective when compared to the laser treatment alone when the duration of the disease exceeds 17 years, regardless of the subjects' age or gender.

5. Conclusion

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. In this paper, we proposed a new approach that implements beta regression to model the placement values directly, thereby eliminating the additional need for bootstrapping the standard errors for the regression parameters. We compared our beta methodology with the parametric and semi-parametric approaches via simulation, showing that the new method yields comparable ROC estimates using the parametric method. The simulation also confirmed that care needs to be taken when using the extreme value examples. If the two density functions are widely separated, then the placement values will be very close to zero and the resultant beta parameters could be compromised as in any zero-inflated model. This problem can be avoided by a simple graphical inspection of the data.

References

[1] Dodd, L. and Pepe, M. (2003). Semiparametric regression for the area under the receiver operating characteristic curve. Journal of the American Statistical Association, 98:409–417.

[2] Zhang, L., Zhao, Y. D., and Tubbs, J. D. (2011). Inference for semiparametric AUC regression models with discrete covariates. Journal of Data Science, 9(4):625–637.

[3] Buros, A., Tubbs, J., van Zyl, J. S. (2017). AUC Regression for Multiple Comparisons with the Jonckheere Trend Test. Statistics in Biopharmaceutical Research, 9(3), 279-285.

[4] Buros, A., Tubbs, J., van Zyl, J. S. (2017). Application of AUC Regression for the Jonckheere Trend Test. Statistics in Biopharmaceutical Research, 9(2), 147-152.

[5] van Zyl, J. S., Tubbs, J. (2018). Multiple Comparison Methods in Zero-dose Control Trials. Journal of Data Science, 16(2), 299-326.

[6] [6] Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. Biometrics, pages 124–135.

[7] Pepe, M. S. (2000). An interpretation for the ROC curve and inference using GLM procedures. Biometrics, 56(2):352–359.

[8] Alonzo, T. A. and Pepe, M. S. (2002). Distribution-free ROC analysis using binary regression techniques. Biostatistics, 3(3):421–432.

[9] Pepe, M. and Cai, T. (2004). The analysis of placement values for evaluating discriminatory measures. Biometrics, 60(2):528–535.

[10] Cai, T. (2004). Semi-parametric ROC regression analysis with placement values. Biostatistics, 5(1):45–60.

[11] Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of mathematical psychology, 12(4):387–415.

[12] Rodriguez-Alvarez, M. X., Tahoces, P. G., Cadarso-Suarez, C., and Lado, M. J. (2011). Comparative study of roc regression techniques – applications for the computer-aided diagnostic system in breast cancer detection. Computational Statistics and Data Analysis, 55(1):888–902.

[13] Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. Journal of Applied Statistics, 31(7):799–815.

[14] Fubini, G. (1907). Sugli integrali multipli. Rend. Acc. Naz. Lincei, 16:608–614.

[15] Balakrishnan, N. and Nevzorov, V. (2003). A Primer on Statistical Distributions. Wiley, New Jersey.

[16] Elman, M. J., Ayala, A., Bressler, N. M., Browning, D., Flaxel, C. J., Glassman, A. R., Jampol, L. M., and Stone, T. W. (2015). Intravitreal ranibizumab for diabetic macular edema with prompt versus deferred laser treatment: 5-year randomized trial results. Ophthalmology, 122(2):375–381.