Beta Regression for Modeling the ROC

As a Function of Continuous Covariates

Sarah Stanley, Dr. Jack Tubbs · Department of Statistical Science · Baylor University

Motivation

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. Our goal is to investigate the effects of covariates on a test's ability to distinguish between a normal and an affected population by presenting two existing methods (parametric and semiparametric) and introducing a new approach using beta regression which avoids the additional correlation induced by the existing methods.

Background

The ROC curve quanitifies the relationship between the true positive and false postive rates of a diagnostic test. Given that Y_D denotes the response of a subject from the diseased group, and $Y_{\bar{D}}$ the response from the non-diseased group, we can define the ROC curve in terms of survival functions as

$$ROC(t) = S_D\left(S_{\bar{D}}^{-1}(t)\right), \qquad t \in (0,1).$$

The AUC is a summary measure of the ROC, given by

$$P(Y_D > Y_{\bar{D}}).$$

We define the placement value of a diseased observation as

$$PV_D = S_{\bar{D}}(Y_D).$$

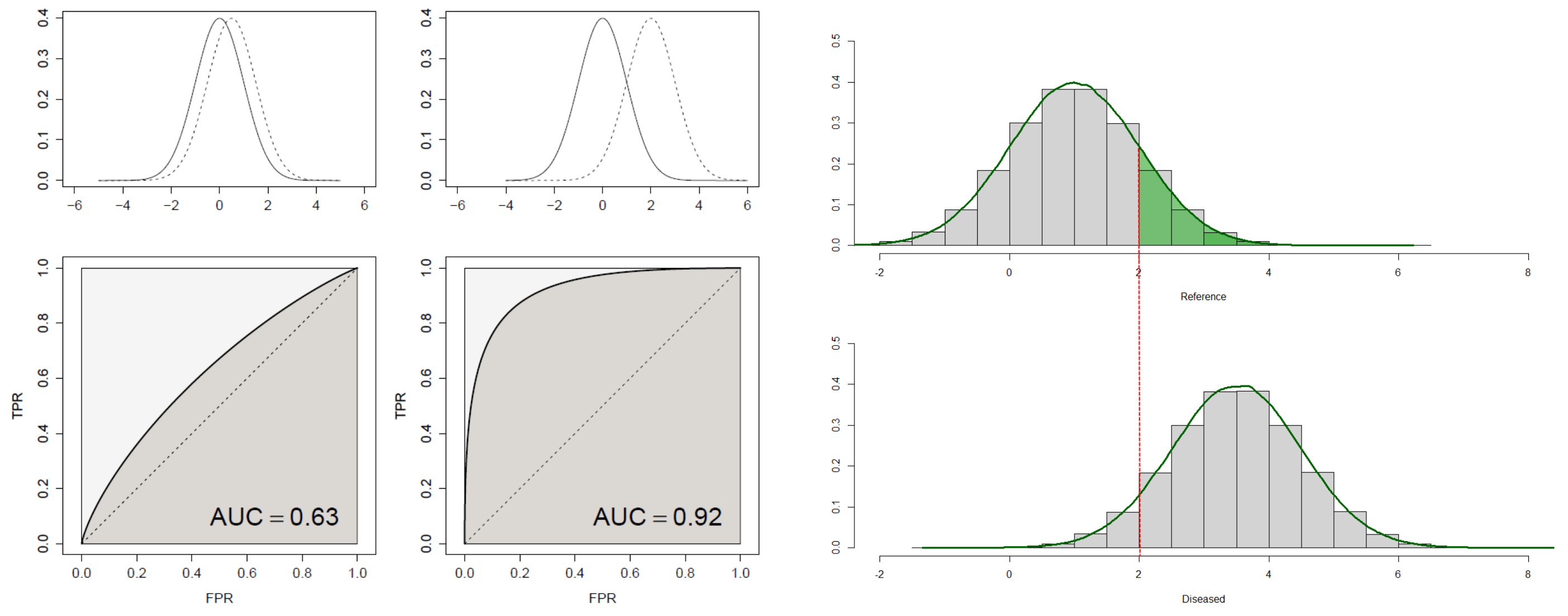

Figure 1: Illustration of the ROC, AUC, and PV_D

The ROC is equal to the cdf of PV_D.

$$P[PV_D \leq t|\boldsymbol{X}] = P[Y_D \geq [S_{\bar{D}\boldsymbol{X}}^{-1}(t)|\boldsymbol{X}] = ROC_{\boldsymbol{X}}(t).$$

A generalized linear model (GLM) framework to directly model the ROC with covariates is

$$ROC_{\mathbf{X}}(t) = g^{-1}(h_0(t) + \mathbf{X}'\boldsymbol{\beta}), \qquad t \in (0,1).$$

as defined by Pepe (2002). We extend this framework in the methods presented in the following section.

Methods

Parametric

Alonzo and Pepe (2002) proposed a parametric ROC-GLM model by specifying $h_0 = \sum_{k=1}^{K} \alpha_k h_k(t)$ in the direct ROC-GLM model. A computational algorithm for the parametric approach is

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

Semiparametric

Cai (2004) developed a semiparametric approach based on the idea that the direct ROC-GLM model is equivalent to

$$h_0(PV_D) = -\mathbf{X}'\boldsymbol{\beta} + \epsilon.$$

The semiparametric algorithm is identical to the parametric approach until step 5 where pairwise comparisons of the diseased placement values are used to estimate $\boldsymbol{\beta}$, and the estimates for $\boldsymbol{\beta}$ are then used as an offset in the estimation of $h_0(\cdot)$.

Beta Regression

Note that in the above methods, correlation is introduced when making pairwise comparisons. In the following approach, we avoid correlation problems by modeling the placement values directly.
We reparameterize the Beta(a,b) distribution in terms of $\mu = E(Y)$ and a scale parameter $\phi = a + b$ so that the beta regression model can be written as

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

Choosing the logit link allows us to obtain the original parameters a and b from the beta distribution by calculating

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

The algorithm for the beta approach begins with the same three steps as in the parametric with the first difference appearing in step 4.

4. Perform a beta regression on PV_D to obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$;
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.

Simulation

Binormal ROC

When both populations are normally distributed, exact solutions for the ROC and AUC exist. 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,

$$Y_D \sim N(2 + 4X, 1.5^2) \text{ and } Y_{\bar{D}} \sim N(1.5 + 3X, 1.5^2).$$

Thus, the true AUC at covariate value $X = x_0$ is

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

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.

	$x_0 = 0.5$	$x_0 = 0.6$	$x_0 = 0.7$	$x_0 = 0.8$
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

Table1: Mean AUC estimates across 300 data sets for each method

We include a plot for the ROC estimates resulting from each method in Figure 2.

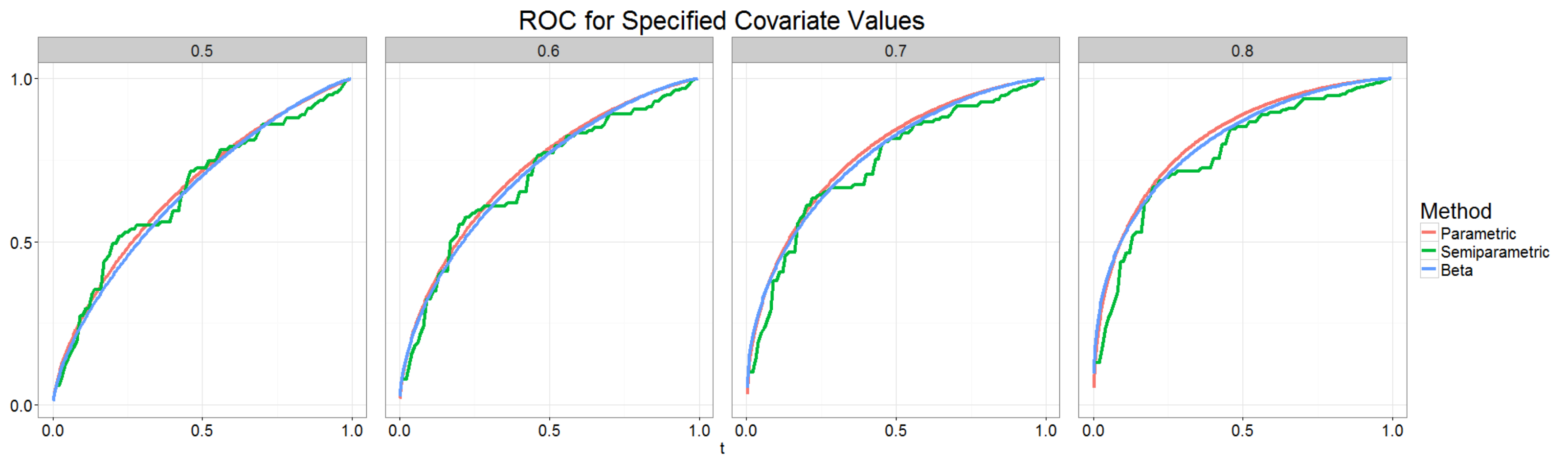

Figure 2: Comparison of ROC estimates for each method

References

Alonzo, T. and M. Pepe (2002), "Distribution-free ROC analysis using binary regression techniques," *Biostatistics*, 3, 421-432.

Cai, T. (2004), "Semi-parametric ROC regression analysis with placement values," *Biostatistics*, 5, 45-60.

Rodriguez-Alvarez, M.X. et. al. (2011) "Comparative Study of ROC regression techniques," *Computational Statistics and Data Analysis*, 55, 888-902.