Beta regression and modeling the ROC as a function of continuous covariates

Sarah Stanley

Joint Statistical Meetings

August 3, 2017

BAYLOR
UNIVERSITY

Motivation and Objective

- The ROC curve is a well-accepted measure of accuracy for diagnostic tests.
- In many applications, a test's performance is affected by covariates.
- Ignoring covariate effects can lead to faulty conclusions.
- Our goal is to investigate the effects of covariates on a test's ability to distinguish between a normal and an affected population.
- We present two existing methods (parametric and semiparametric) and introduce a new approach.

Outline

ROC and AUC

- Suppose
  Y_D = response of a subject from the diseased group
  $Y_{\bar{D}}$ = response of a subject from the non-diseased group.
- In terms of the survival function, we have

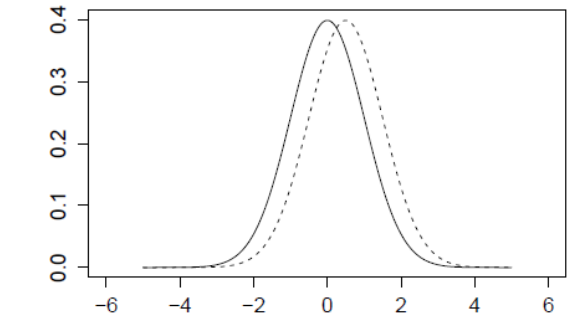
$$ROC(t) = S_D\left(S_{\bar{D}}^{-1}(t)\right), \qquad t \in (0,1)$$

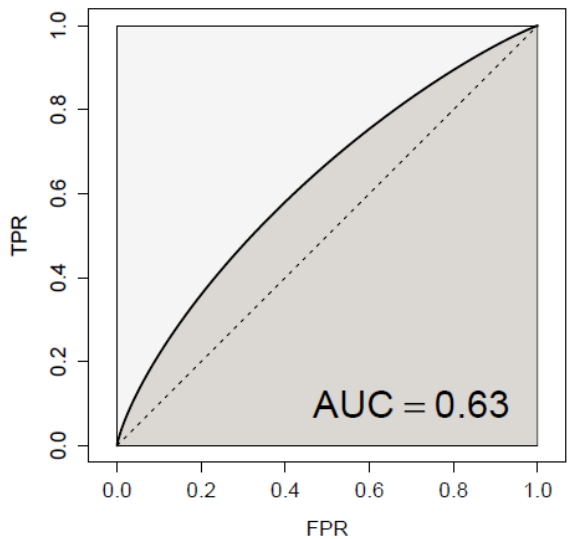
- The AUC, a summary measure of the ROC, given by

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

  is the probability that a randomly selected subject is classified into the correct group.

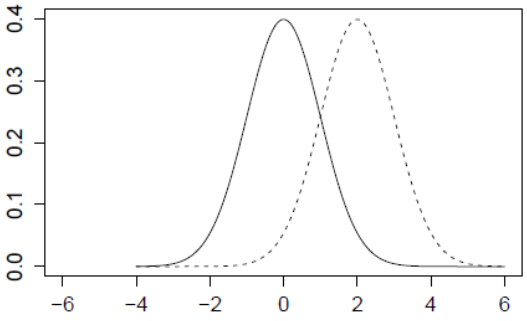
Illustrating the AUC

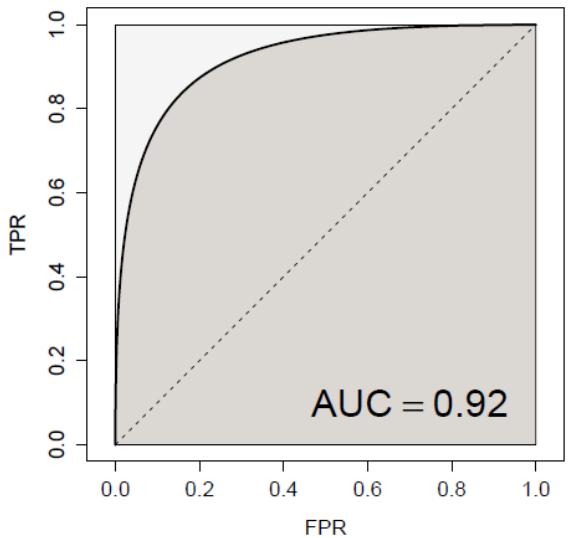




- Low separation
- $ROC(t) = S_D\left(S_{\bar{D}}^{-1}(t)\right)$
  - Survival curves are nearly identical
  - ROC is close to the diagonal
- $AUC = P(Y_D > Y_{\bar{D}})$
  - Close to 0.5

Illustrating the AUC

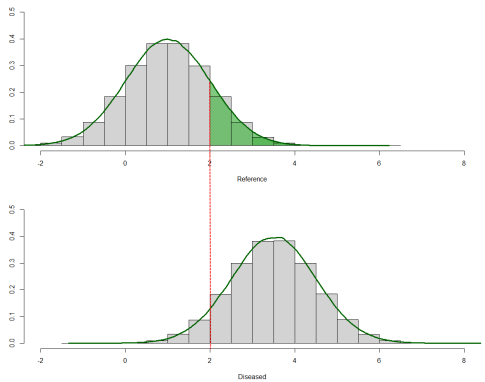




- High separation
- $ROC(t) = S_D\left(S_{\bar{D}}^{-1}(t)\right)$
  - Survival curves are different
  - ROC rises more steeply
- $AUC = P(Y_D > Y_{\bar{D}})$
  - Close to 1

Placement Values

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



- The ROC is equivalent to the cdf of PV_D.

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

- Note also that the ROC curve can be thought of as the conditional expectation of $B_{Dt} = I[PV_D \leq t]$.

Relationship between the Mann Whitney Statistic and the AUC

- The Mann-Whitney (MW) U-statistic for two independent random samples, $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} I(x_i > y_j).$$

- The MW statistic can be used as a nonparametric unbiased estimate of the AUC [Bamber(1975)].

Methodology

Direct ROC Regression Methodology

- Pepe (2002) proposed a generalized linear model (GLM) framework to directly model the ROC with covariates as follows

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

  where g is a monotone link function, $\mathbf{X}$ is a vector of covariates, β is a vector of the model parameters, and h_0 is an unknown monotonic increasing function.
- Note that the dependent variable is not directly observable, we thus estimate $ROC_{\mathbf{X}}(t)$ with either the cdf of the placement values or the conditional expectation of B_{Dt}.

Parametric ROC-GLM

- Alonzo and Pepe (2002) proposed a parametric form for $h_0(\cdot)$ such that

$$h_0(t) = \sum_{k=1}^{K} \alpha_k h_k(t),$$

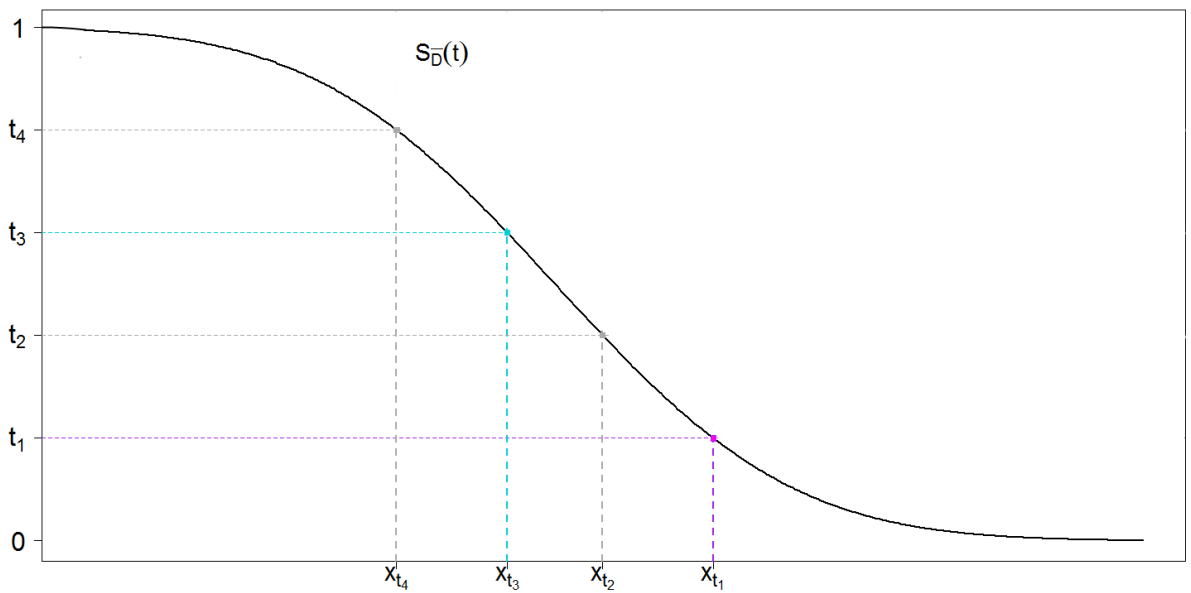
where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)$ is a vector of unknown parameters and $h(\cdot) = (h_1(\cdot), ..., h_K(\cdot))$ are known functions.

- Thus, a parametric ROC-GLM model is

$$ROC_{\mathbf{X}}(t) = g^{-1}\left(\sum_{k=1}^{K} \alpha_k h_k(t) + \mathbf{X}'\boldsymbol{\beta}\right), \qquad t \in (0,1).$$

Algorithm

1. Specify a set $T = \{t_\ell : \ell = 1, ..., n_T\} \in (0, 1)$ of FPRs;
2. Estimate the covariate specific survival function $S_{\bar{D}\boldsymbol{X}}$ for the reference population at each $t \in T$ using quantile regression.



Algorithm

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

Semiparametric ROC-GLM

- Developed by Cai(2004)
- Based on the idea that the ROC-GLM model

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

  is equivalent to

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

  where ϵ has known distribution g and $h_0(\cdot)$ is an unspecified increasing function.
- Essentially, 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 $h_0(\cdot)$.

Algorithm

1. Specify a set $T = \{t_\ell : \ell = 1, ..., n_T\} \in (0, 1)$ of FPRs;
2. Estimate the covariate specific survival function $S_{\bar{D}\boldsymbol{X}}$ via quantile regression;
3. Calculate the placement values $PV_j = \hat{S}_{\bar{D}\mathcal{X}_{Dj}}(y_{Dj})$;
4. Calculate the binary placement value indicator
$$\hat{B}_{jt} = I[PV_j \leq t], t \in T, j = 1, ..., n_D;$$
5. For each pair of observations in Y_D, calculate
$$\widehat{PV}_{j\ell} = I[PV_j \leq PV_\ell], \text{ and } x_{j\ell} = x_{Dj} - x_{D\ell}$$
with $j, \ell = 1, ..., n_D, j \neq \ell$;
6. Fit the following GLM without an intercept to estimate $\boldsymbol{\beta}$
$$g(\widehat{PV}) = -\mathbf{X}'\boldsymbol{\beta}.$$
7. Estimate $h_0(\cdot)$ using $\hat{\boldsymbol{\beta}}$ and $\hat{B}_{jt}$ as follows
$$g(E[\hat{B}_{jt}]) = \textit{intercept} + \text{offset}(\mathbf{X}'\hat{\boldsymbol{\beta}}).$$

Consequences of parametric and semiparametric procedures

- Correlation is introduced when making pairwise comparisons.
- The resulting standard errors are thus incorrect.
- Recall, however, that the cdf of the placement values from the diseased population is equivalent to the ROC.
- A method that models the placement values directly avoids the above correlation problems.
- We implement a direct model of the placement values through beta regression.

Beta Regression Model

We now introduce a beta regression model (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

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

Beta Regression Model

- Let $y_1, ..., y_n$ be independent random variables from a beta density with mean μ_t, t = 1,...,n and precision 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, $x_{t1}, ..., x_{tk}$ 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

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

Beta Algorithm

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

Example

Binormal ROC

Let

$$Y_D \sim N(\mu_D, \sigma_D^2), Y_{\bar{D}} \sim N(\mu_{\bar{D}}, \sigma_{\bar{D}}^2).$$

Then

$$ROC(t) = \mathbf{\Phi}(a + b\mathbf{\Phi}^{-1}(t)),$$

and

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right),$$

where

$$a = \frac{\mu_D - \mu_{\bar{D}}}{\sigma_D}, b = \frac{\sigma_{\bar{D}}}{\sigma_D}.$$

Binormal Example

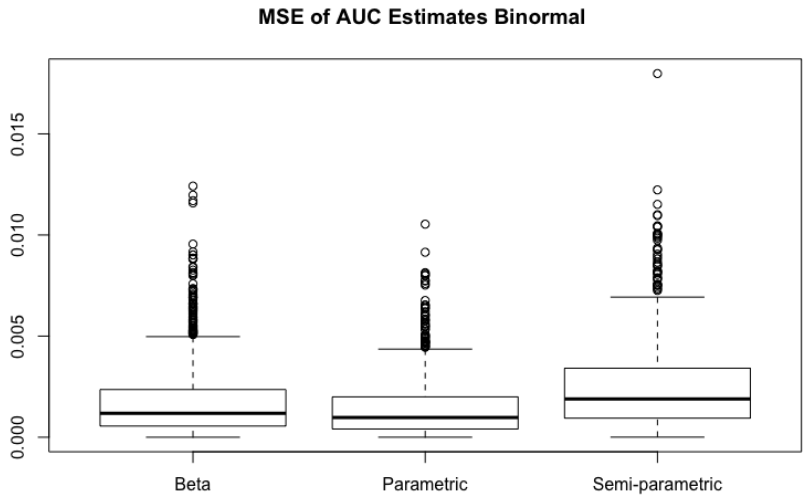
- Data simulated from

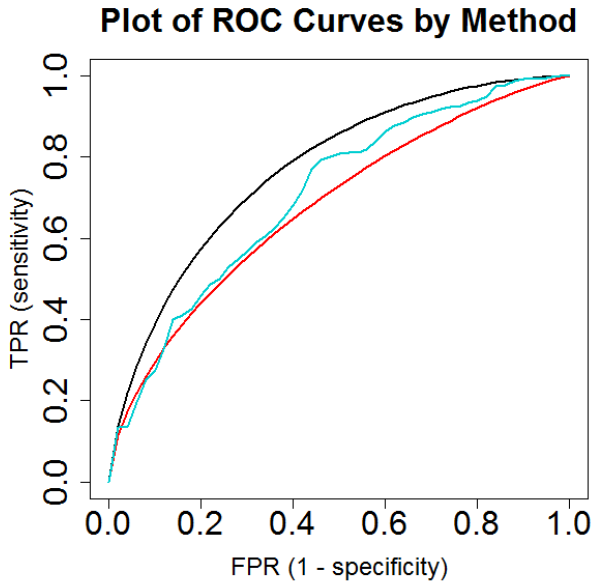
$$Y_D = 2 + 4X + \epsilon_D \text{ and } Y_{\bar{D}} = 1.5 + 3X + \epsilon_{\bar{D}},$$

  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

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

Binormal Results





	Median	Mean	St. Dev.
Parametric	0.001437	0.000975	0.001480
Beta	0.001183	0.001817	0.001877
Semi-parametric	0.001893	0.002450	0.002136

Table : *Summary of MSEs for binormal*

Conclusion

- Beta regression on the placement values yields comparable AUC estimates to those obtained via parametric and semiparametric approaches without inducing correlation.

Future Work

- Use of Historical Controls
- Meta-Analysis
- Bayesian Methods

References

- 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.