Current Research in Biostatistics

Original Research Paper

Application of a Beta Regression Model for Covariate Adjusted ROC

Xing Meng and J.D. Tubbs

Department of Statistical Science, Baylor University, Waco, USA

Article history
Received: 13-02-2020
Revised: 11-04-2020
Accepted: 15-06-2020

Corresponding Author:
J.D. Tubbs
Department of Statistical Science, Baylor University, Waco, USA
Email: jack_tubbs@baylor.edu

Abstract: The Receiver Operating Characteristic (ROC) curve and the area under the ROC (AUC) are widely used in determining the diagnostic capability of a binary classification procedure. Since the test performance is affected by covariates, the ROC and AUC have been utilized in a Generalized Linear Regression (GLM) setting. In this study, we revisit a problem where the AUC regression model was used in a clinical study with discrete covariates by considering ROC regression models with both discrete and continuous covariates. The two ROC regression models are based upon a widely used parametric model and a recently published model based upon fitting the placement values with the beta distribution. The two methods are illustrated using data from a clinic study.

Keywords: Placement Values, Beta Regression, ROC Regression

Introduction

The Receiver Operating Characteristic (ROC) curve and the area under the ROC (AUC) are widely used measure of accuracy for diagnostic test to distinguish between two populations. An important application of ROC curve is to determine how a test's performance is affected by covariates. One approach is to model the AUC of the ROC curve by modifying the Mann-Whitney statistics (MW) as a GLM (Pepe, 2003). Another approach is to model the ROC directly. Dodd and Pepe (2003) proposed a Generalized Linear Model (GLM) framework to directly model the ROC with covariates as follows:

$$ROC_X(t) = g^{-1}\left(h_0(t) + X'\beta\right), \qquad (1)$$

for $t \in (0,1)$ where g^{-1} is a monotone link function, X is a vector of covariates, $h_0(\cdot)$ is an unknown monotonic increasing function and b is a vector of the model parameters. Assumptions concerning $h_0(\cdot)$ define whether (1) is a parametric (Alonzo and Pepe, 2002) or semi-parametric (Cai, 2004). Although the two models differ, they are both based upon the conditional expectation of Mann-Whitney *U*-statistic.

Stanley and Tubbs (2018) presented an alternative GLM model for the ROC as a function of the covariate-adjusted placement values. They compared their model with the parametric and semi-parametric model using simulated normal and extreme value data.

The objective of this paper is to investigate the parametric and beta ROC regression models when compared with the AUC regression model presented by (Zhang *et al*., 2011) using data from a clinical trial concerning the efficacy of an active drug to treat stress urinary incontinence in North American women.

The outline for this paper is as follows. Section 2 presents a brief overview of the two ROC regression methods. The results for the two methods using the incontinence trial data are reported in section 3. The paper concludes with a discussion in section 4.

Methods

Let Y be a continuous random variable used to distinguish between the two populations. Assume that the non-diseased or control population is indicted by $D = 0$. Let $D = 1$ denote the diseased or cases population of interest and assume that large values of Y are more likely to be associated with the disease indicator. The classifier assigns a subject to the diseased group if $Y \geq c$. In which case, the true positive rate of the test is TPR(c) = Pr[$Y \geq c|D = 1$] and the false positive rate of the test is FPR(c) = Pr[$Y \geq c|D = 0$]. The ROC curve, is defined as a collection of all TPR-FPR pairings.

The placement value of Y, denoted as ($PV_{D=0}$), is the proportion of the reference or control population with observations greater than Y (the survival value for Y in the reference or control population). This is just a transformation of Y given by $PV_{D=0} = S_0(Y)$. It has been shown that the CDF for the placement is the ROC. That is:

$$ROC(t) = S_1\left(S_0^{-1}(t)\right) = \Pr\left(PV_{D=0} \leq t\right),$$

where, $t \in (0,1)$ and S_1, S_0 are the survival function for the diseased and non-diseased populations.

Considering the covariates, denoted by X, the covariate-adjusted ROC can be written as:

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

for $t \in (0,1)$ where $S_{1,X,X_D}(c) = \Pr(Y \geq c)| X, X_D, D = 1)$, $S_{0,X}(c) = \Pr(Y \geq c)|X, D = 0)$ and c is any threshold. Thus, $ROC_{X,X_D}(t)$ is the probability that the test result Y of the diseased subject is greater than or equal to the tth quantile of the test result adjusted by the covariates of the unaffected subject.

The ROC is the CDF of the placement values PV_D (Pepe and Cai, 2004). The covariate-adjusted notation is given by:

$$\begin{aligned}
&\Pr\left(PV_D \leq t\right) = \Pr\left(S_{0,X}(Y) \leq t \mid X, D=0\right) \\
&= \Pr\left(Y \geq S_{0,X}^{-1}(t)\right) \mid X, D=0\right) \\
&= ROC_X(t)
\end{aligned}$$

Stanley and Tubbs (2018) provide a description of the algorithms used to model the ROC with the parametric presented by (Alonzo and Pepe, 2002) and the beta placement value model. A brief description of both methods are included for completeness.

Parametric Method

Alonzo and Pepe (2002) extended the use of ROC-GLM by considering the ROC curve as a parametric function of covariates and using the binary indicator as the dependent variable. The parametric function of covariates is reflected in parametric form of $h_0(\cdot)$. The binary indicators compare the test result for a diseased subject to a specified set of covariate-adjusted quantiles of the distribution of test results from non-diseased subjects. Then the binary values can be modeled using logistic regression methods. Their parametric form for $h_0(\cdot)$ is:

$$h_0(t) = \gamma_1 h_1(t) + \gamma_2 h_2(t),$$

where, $h_1(t) = 1$ and $h_2(t) = \Phi^{-1}(t)$. In which case, we have:

$$ROC_X(t) = g\left(h_0(t) + \beta' X\right),$$

for $t \in (0,1)$. The algorithm for parametric the method can be written as:

1. Specify a set of FPRs: $T = \{t_l: l = 1, \ldots, n_T\} \in (0,1)$
2. Estimate the covariate specific survival function S_{0,X_j} for the reference population at each $t \in T$, $j = 1,2, \ldots, n_0$ using quantile regression
3. For each diseased observation $y_{i|D=1}$, calculate $PV_i = \hat{S}_{0,x_i}(y_{i|D=1})$, $i = 1,2, \ldots, n_1$
4. Calculate the binary placement value indicator $\widehat{U_{it}} = I(PV_i \leq t)$, $t \in T$
5. Fit the model $E[\hat{U}_{it}] = g^{-1}(\widehat{h_0(t)} + X'\hat{\beta}))$ to obtain $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\beta}$

Beta Regression Method

Stanley and Tubbs (2018) proposed a method that models the placement values using beta regression. This method is easy to implement and it eliminates the dependency in models that use binary variables when using the logit or probit models.

The Beta regression model can be written as a GLM (Ferrari and Cribari-Neto, 2004) in terms of its mean $\mu = E(Y)$ and precision parameter $\phi = a + b$ where the mean and variance for $Y \sim Beta(a, b)$:

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

can be written as:

$$E(Y) = \mu$$

and:

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

The beta regression model can be written as:

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

Using the logit link, we have $\mu_t = \left(1 + e^{-x_t'\beta}\right)^{-1}$. From which we obtain the original parameters a and b as $\hat{a} = \frac{\hat{\phi}}{1 + e^{-x_t'\hat{\beta}}}$ and $\hat{b} = \hat{\phi}\left(1 - \frac{1}{1 + e^{-x_t'\hat{\beta}}}\right)$.

The algorithm for beta regression method can be written as:

1. Specify a set of FPRs: $T = \{t_l: l = 1, \ldots ,n_T\} \in (0,1)$
2. Estimate the covariate specific survival function S_{0,X_j} for the reference population at each $t \in T$ using quantile regression
3. For each diseased observation y_{1x_j}, calculate $PV_j = \hat{S}_0\left(y_{1x_j}\right)$
4. Perform a beta regression on the PVs to obtain estimates $\hat{\beta}$ and $\hat{\phi}$
5. Transform to obtain $\hat{a} = \hat{\mu}\hat{\phi}$ and $\hat{b} = (1-\hat{\mu})\hat{\phi}$
6. Calculate the CDF of the placement values using the Beta($\hat{a}, \hat{b}$) distribution to obtain $\widehat{ROC}$ and the $\widehat{AUC}$

Application

Zhang *et al*. (2011) presented results for AUC regression using data from a placebo-controlled study to determine the efficacy of an active drug to treat stress urinary incontinence in menopausal women. Their primary endpoint was the relative Percent reduction in Incontinence Episode Frequency (PIEF) from baseline to the final visit (12 weeks), where larger PIEF reduction indicates the desired treatment effect. They considered two discrete covariates; strata and horm50. The covariate strata indicate the severity of disease at baseline where 1 indicates the lowest level and 4 represents the highest number of episodes. The second covariate, horm50, is binary where 1 indicates that the subject had hormone replacement therapy prior to the start of the study. Zhang *et al*. (2011) indicated that they elected to reduce the computational complexity of their example by selecting a 10% random sample (n = 407) from the total available subjects.

Since we do not have access to the same sample used by (Zhang *et al*., 2011), we present the results for the data set that we have access (n = 2200) and for four (4) random subsets of size 420 with a 1:1 split for the treatment and control, in hopes of understanding data variability and dependency on the performance of the ROC regression methods. In addition, we will consider two additional covariates, a binary indicator of high level of BMI (BMI > 30) and a continuous covariate, BMI.

Discussion

Table 1 reproduces the results of a table given in (Zhang *et al*., 2011) where we have highlighted the potential significant terms in red. Tables 2 and 3 contain the results obtained when using the ROC methods with the strata and horm50 as covariates. The interaction terms were included. Our objective in this study was to use both ROC regression models to obtain estimates for the regression coefficients without being overly concerned about the significance of the terms as was done in (Zhang *et al*., 2011). When comparing our results with those given in Table 1 it is doubtful that any terms are significant when using the results of the beta model whereas the parametric method may have found some significant AUCs. It appears that the parametric ROC model produces estimates that are closer to those given in Table 1 than the beta method. This shouldn't be that surprising since the parametric ROC model modifies the use of the Mann-Whitney statistic used in (Zhang *et al*., 2011). Although we have elected not to be overly concerned about the standard errors of our estimates in this study, it should be mentioned that the standard errors for the beta method are obtained directly from the beta regression model whereas the parametric method makes use of bootstrapped estimates.

Table 1: Zhang *et al*. (2011) $\widehat{AUC}$ estimates for strata with horm50

	Model			
Strata	Horm50	$\widehat{AUC}$	Bootstrap	DeLong
1	No	0.458	(0.068, 0.848)	(0.138, 0.778)
	Yes	0.467	(0.019, 0.914)	(0.140, 0.793)
2	No	0.576	(0.420, 0.732)	(0.412, 0.739)
	Yes	0.437	(0.289, 0.585)	(0.300, 0.575)
3	No	0.596	(0.435, 0.757)	(0.428, 0.764)
	Yes	0.878	(0.763, 0.993)	(0.751, 1.000)
4	No	0.654	(0.497, 0.811)	(0.514, 0.794)
	Yes	0.625	(0.484, 0.766)	(0.492, 0.759)

Table 2: Beta $\widehat{AUC}$ estimates for strata with horm50

	Model	Sample				
Strata	Horm50	all	1	2	3	4
1	No	0.569	0.535	0.345	0.661	0.699
	Yes	0.638	0.728	0.656	0.683	0.634
2	No	0.622	0.646	0.621	0.578	0.649
	Yes	0.611	0.657	0.620	0.688	0.617
3	No	0.555	0.595	0.541	0.574	0.608
	Yes	0.593	0.524	0.602	0.637	0.682
4	No	0.567	0.578	0.572	0.565	0.590
	Yes	0.589	0.624	0.621	0.535	0.561

Table 3: Parametric $\widehat{AUC}$ estimates for strata with horm50

	Model	Sample				
Strata	Horm50	all	1	2	3	4
1	No	0.549	0.566	0.415	0.685	0.724
	Yes	0.547	0.543	0.653	0.701	0.516
2	No	0.566	0.558	0.581	0.454	0.598
	Yes	0.587	0.592	0.630	0.609	0.576
3	No	0.597	0.633	0.583	0.611	0.622
	Yes	0.633	0.589	0.637	0.637	0.680
4	No	0.580	0.615	0.547	0.622	0.586
	Yes	0.628	0.635	0.670	0.611	0.602

Table 4: Beta $\widehat{AUC}$ estimates for strata with high BMI

		Sample				
Strata	Model BMI	all	1	2	3	4
1	Low	0.631	0.721	0.724	0.661	0.710
	High	0.607	0.604	0.467	0.658	0.696
2	Low	0.610	0.697	0.651	0.649	0.652
	High	0.642	0.677	0.733	0.721	0.619
3	Low	0.592	0.624	0.558	0.629	0.608
	High	0.632	0.706	0.669	0.655	0.604
4	Low	0.589	0.610	0.603	0.601	0.611
	High	0.577	0.587	0.530	0.600	0.601

Table 5: Parametric $\widehat{AUC}$ estimates for strata with high BMI

		Sample				
Strata	Model BMI	all	1	2	3	4
1	Low	0.575	0.571	0.579	0.536	0.595
	High	0.578	0.589	0.449	0.652	0.594
2	Low	0.606	0.654	0.595	0.663	0.621
	High	0.636	0.636	0.670	0.709	0.607
3	Low	0.602	0.604	0.583	0.617	0.624
	High	0.621	0.655	0.667	0.660	0.570
4	Low	0.604	0.622	0.604	0.608	0.612
	High	0.604	0.625	0.548	0.613	0.626

Table 6: Beta $\widehat{AUC}$ estimates for strata with continuous BMI

		Sample				
Strata	Model BMI	all	1	2	3	4
1	25	0.624	0.684	0.670	0.653	0.708
	30	0.630	0.685	0.662	0.666	0.704
	35	0.636	0.685	0.655	0.678	0.699
2	25	0.616	0.690	0.670	0.665	0.647
	30	0.622	0.690	0.662	0.677	0.642
	35	0.628	0.691	0.655	0.689	0.636
3	25	0.599	0.651	0.591	0.627	0.610
	30	0.605	0.651	0.583	0.640	0.604
	35	0.611	0.652	0.576	0.652	0.599
4	25	0.579	0.600	0.581	0.588	0.612
	30	0.585	0.600	0.573	0.602	0.607
	35	0.591	0.601	0.565	0.615	0.601

Table 7: Parametric $\widehat{AUC}$ estimates for strata with continuous BMI

		Sample				
Strata	Model BMI	all	1	2	3	4
1	25	0.573	0.576	0.549	0.566	0.597
	30	0.581	0.577	0.544	0.582	0.591
	35	0.588	0.578	0.540	0.598	0.584
2	25	0.610	0.648	0.611	0.669	0.621
	30	0.618	0.649	0.607	0.683	0.615
	35	0.626	0.650	0.603	0.697	0.609
3	25	0.603	0.619	0.606	0.620	0.612
	30	0.610	0.620	0.601	0.635	0.607
	35	0.618	0.621	0.597	0.650	0.601
4	25	0.598	0.623	0.586	0.596	0.622
	30	0.605	0.624	0.582	0.612	0.617
	35	0.613	0.625	0.577	0.627	0.611

Table 4-7 summarize the results when using BMI as a covariate with the entire data set and 4 subset data sets. Table 4 and 5 summarize the results when using the discrete BMI covariate. There is a lot going on in these tables, but both methods indicate that the separation between the treatment and the control groups decreases as the BMI increases. We see similar results when using BMI as a continuous covariate with both ROC methods (Tables 6 and 7).

Conclusion

Our objective was to demonstrate how two ROC regression methods could be used instead of the more commonly used AUC regression models based upon the Mann-Whitney statistic when one has both discrete and continuous covariates. The ROC model provided believable results when used with data from a clinical study where the results from the AUC model were published. The parametric ROC model given by Alonzo and Pepe (2002) is widely used and is commercially available for use in R packages and Stata. The beta model based upon modeling the placement scores given by Stanley and Tubbs (2018) is not as widely used. Yet, it performed as well as the parametric model.

Funding Information

No external funding was used for this research and manuscript.

Author' Contributions

Xing Meng: Computation and computer programming support. Created tables and figures.

J.D. Tubbs: Problem formulation, writing and editing the manuscript.

Ethics

No ethical issues were encountered in connection with this manuscript.

References

Alonzo, T.A. and M.S. Pepe, 2002. Distribution-free ROC analysis using binary regression techniques. Biostatistics,3:421-432. DOI:10.1093/biostatistics/3.3.421

Cai, T., 2004. Semi-parametric ROC regression analysis with placement values. Biostatistics, 5: 45-60. DOI: 10.1093/biostatistics/5.1.45

Dodd, L. and M. Pepe, 2003. Semiparametric regression for the area under the receiver operating characteristic curve. J. Am. Stat. Assoc., 98: 409-417. DOI: 10.1198/016214503000198

Ferrari, S. and F. Cribari-Neto, 2004. Beta regression for modelling rates and proportions. J. Applied Stat., 31: 799-815. DOI: 10.1080/0266476042000214501

Pepe, M.S., 2003. The Statistical Evaluation of Medical Tests for Classification and Prediction. 1st Edn., Oxford University Press, Oxford, ISBN-10: 0198509847, pp: 302.

Pepe, M.S. and T. Cai, 2004. The analysis of placement values for evaluating discriminatory measures. Biometrics, 60: 528-535. DOI: 10.1111/j.0006-341X.2004.00200.x

Stanley, S. and J. Tubbs, 2018. Beta regression for modeling a covariate adjusted roc. J. Applied Math. Stat., 6: 110-118. DOI: 10.11648/j.sjams.20180604.11

Zhang, L., Y.D. Zhao and J.D. Tubbs, 2011. Inference for semiparametric AUC regression models with discrete covariates. J. Data Sci., 9: 625-637.