SciencePG
Science Publishing Group

A Comparison of the Lehmann and GLM ROC Models

Melissa Innerst[1], Jack D. Tubbs[2], Musie Ghebremichael[3]

[1]Mathematics Department, Juniata College, Huntingdon, the United States

[2]Department of Statistical Science, Baylor University, Waco, the United States

[3]Biostatistics, Harvard University and Ragon Institute, Boston, the United States

Email address:

Melissa_Innerst1@baylor.edu (M. Innerst), jack_tubbs@baylor.edu (J. D. Tubbs),
Musie_Ghebremichael@DFCI.HARVARD.EDU (M. Ghebremichael)
*Corresponding author

Abstract: Recently, several regression methods have been developed to model the receiver operating characteristic curve (ROC), as a measure of accuracy for potential biomarker use in diagnostic testing and disease detection. In this paper, we investigate the Lehmann ROC regression model and compare it to more commonly used ROC regression methods that are found in the literature. The comparative performance of the methods are evaluated using simulated data from the normal, extreme value, and the Weibull distributions. Theory suggests that the Lehmann method should only work well when using the Weibull distribution. Our simulation results suggest that the performance of these methods is more complicated than the theory might suggest. The methods were demonstrated using data from a study concerning the clinical effectiveness of leukocyte elastase determination in the diagnosis of coronary artery disease (CAD).

Keywords: Cox's Model, Youden Index, Extreme Value, Beta and Weibull Distributions

1. Introduction

Receiver Operating Characteristic (ROC) curves and their associated area under the curves (AUC) are a widely-accepted measure of accuracy in diagnostic testing. Recently, the ROC has been used in a regression context so as to allow for the inclusion of covariates to improve the accuracy of a test. Several ROC regression methods utilizing a model proposed by ([16], [17]) have been proposed in recent years. Stanley and Tubbs investigated two of the more commonly used methods, the parametric model proposed by [1] and the semi-parametric model proposed by [6] in [20]. [20] proposed an alternative model that was based upon modeling the placement values as introduced in [5] with the GLLM Beta model. Their simulations using normal and extreme-value distributions indicated that the parametric model and the beta model performed well, whereas, the semi-parametric model was harder to use and required more specialized software. In addition, it did not preform as well as the other two models. It was for this reason that this approach was not considered in this paper.

There were two main purposes for this paper. One was to investigate the use of ROC regression model that is based on a method proposed by Lehmann, [13]. As far as we could tell, this model has not been used extensively when evaluating biomarkers for screening and classification purposes. Instead, its use has been primarily restricted to modeling data for which the proportional hazard function assumption holds as found in many survival studies.

Simulations for this approach mainly use exponential and weibull distributed data since they satisfy the Cox proportional hazard assumption, [12]. The second purpose is to compare these methods using data for which the proportional hazard assumption does not hold as with normal and extreme-value data. And to make this comparison with data for which the assumption does hold as with the exponential and the weibull distributed data. A third consideration is to determine how well the beta model using the placement values performs when compared with the other two approaches.

In this paper we have chosen to model data from three distributions which is hardly an exhaustive list of possible

choices. The three that we did choose were based upon whether or not they were used in modeling biomarker type data or if they satisfied the Cox proportional hazard assumption as commonly assumed when modeling survival data. The three methods that we chose to model are in some sense nonparametric methods in that we do not use the distribution of the underlying data within the modeling assumptions. One of the models used is referred to as a parametric ROC regression method given by Alonzo and Pepe in which they model the ROC with a normal assumption [1]. A second method uses the beta distribution to model the placement values which are found using a nonparametric method for estimating survival curves. Again, no distributional assumptions are made about the data Y.

The third method is a simple extension of the Cox proportional hazard model that is a nonparametric procedure for modeling survival or endpoint data.

This paper is organized as follows. The first section provides a brief overview of the ROC and AUC. The second section provides a brief overview of ROC regression as described by [20]. The third section introduces an alternative class of ROC regression models that are based on the Lehmann method.

In the fourth section, the ROC models are compared through a simulation using data simulated from several distributions.

In the fifth section, the ROC models are illustrated on leukocyte elastase data obtained from the Cardiology Department at Gala General Hospital.

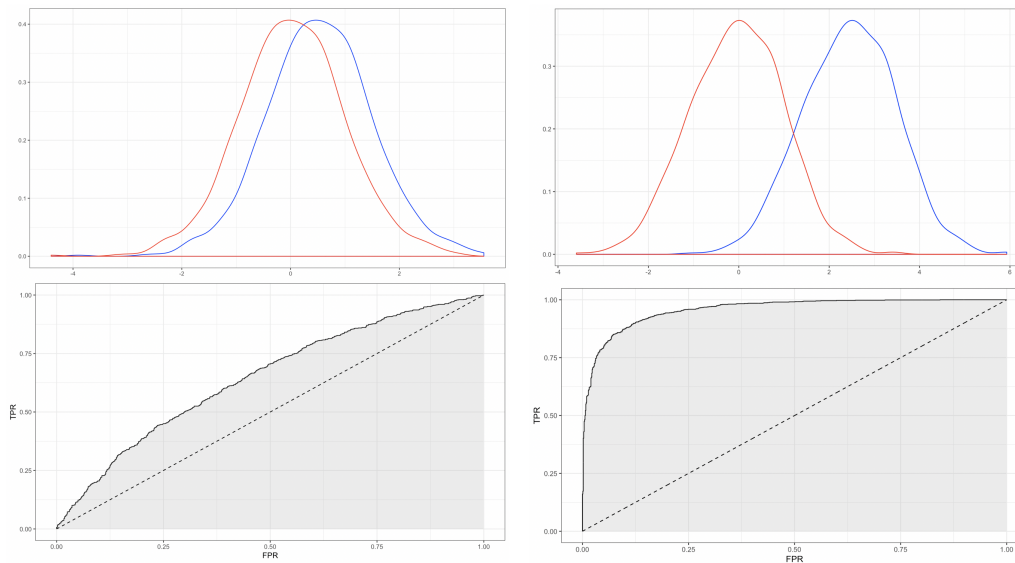The last section contains a summary and discussion of the results.

2. Background of the ROC and AUC

Before we begin the discussion of specific ROC regression techniques, a brief overview of the ROC methodology is provided.

Let Y denote measurements of a continuous biomarker that will be used to distinguish between the diseased population ($D = 1$) and reference population (D = 0). For a given threshold c, the convention is that a measurement $Y \geq c$ is indicative of disease.

In which case, the true positive rate (TPR) and false positive rate (FPR) of the diagnostic test are defined as $TPR(c) = \Pr[Y \geq c|D = 1]$ and $FPR(c) = \Pr[Y \geq c|D = 0]$.

The ROC curve is a plot of TPR versus FPR for various values of the threshold c, and is used to quantify the separation between the disease and reference populations. Figure 1 illustrates how the ROC curve behaves for populations with different levels of separation.

Figure 1. *The probability densities for the reference (red) and diseased (blue) populations. The two plots on the left show an example of populations with low separation and their corresponding ROC curve. The two plots on the right show an example of populations with high separation and their corresponding ROC curve.*

In the top left plot, we see the probability densities for reference and diseased populations which have low separation. The corresponding ROC curve, given in the bottom left plot, is close to the diagonal line. In general, the closer the ROC curve lies to this diagonal line, the lower the separation in the population densities. If there was no separation between the densities, the ROC curve would lie along the diagonal line. In the top right plot, we see the probability densities for reference and diseased populations which have high separation. The

corresponding ROC curve, given in the bottom right plot, is far from the diagonal line. In general, the farther the ROC curve lies from this diagonal line, the higher the separation in the population densities. If there was no overlap between the densities, the ROC curve would make a 90-degree angle in the top left-hand corner, away from the diagonal line.

A commonly used statistics for the ROC curve is the Area Under the ROC Curve (AUC). The AUC has a stochastic interpretation as the probability that a randomly selected

subject is classified into the correct population, $P(Y_1 > Y_0)$, where Y_0 and Y_1 are the biomarker values for the non-disease and diseased population, respectively. The AUC for the bottom left plot in Figure 1, is 0.6352 and is 0.9660 for the bottom right plot in Figure 1.

In general, populations with a low degree of separation will have an AUC close to 0.50, and populations with a high degree of separation will have an AUC close to 1.

In the literature, there are many notations used for the ROC. In this paper, we will use the survival and placement value notations as follows from [6] that

$$ROC(t) = S_1 \left(S_0^{-1}(t) \right) = \Pr \left[v_1 \leq t \right], \qquad (1)$$

for $t \in (0, 1)$ where S_1 is the survival function for the diseased population, S_0 is the survival function for the reference population, and v_1 is the placement values for the diseased population given by

$$v_1 = S_0(Y_1). \qquad (2)$$

The covariate adjusted ROC is a simple extension of (1) where the covariate adjusted ROC is given by

$$ROC_X(t) = S_{1_X} \left(S_{0_X}^{-1}(t) \right), \qquad (3)$$

for $t \in (0, 1)$ where S_{1_X} and S_{0_X} are the covariate adjusted survival functions for a specified covariate, X.

3. ROC Regression Models

The literature for adjusting for covariates in ROC curve regression and AUC estimation is well developed. [22] employed generalized linear models (GLMs) to categorical rating data to estimate and analyze ROC curves. This yields a flexible model that is not restricted to the usual assumption of binormality and can estimate covariate-adjusted ROC curves through two regression equations that correspond to location and scale. [9] presented a nonparametric approach to AUC analysis for correlated ROC curves that uses the theory on generalized U-statistics to estimate a covariance matrix. [21] proposed the partial AUC as a new measure of accuracy for ROC curves.

[3] and [11] extended the ideas presented in [22] by using random effects models for ordinal test results. [16] provided a review of two regression methods for covariate-adjusted ROC curves [[22] and [21]]. [16] proposed a new approach that directly models the ROC curve using parametric, distribution-free methods.

[17] proposed using GLM binary regression framework for estimating covariate-adjusted ROC curves where the model is given by

$$ROC_X(t) = g(h_0(t) + X'\beta), \qquad (4)$$

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

(4) by specifying a parametric form for $h_0(\cdot)$, where they created a binary outcome variable by comparing each diseased observation to a specified set of covariate-adjusted quantiles for the non-diseased population. These binary values are modeled using the logistic or probit regression model.

[18] extended Pepe's method [17] by proposing a nonparametric form for $h_0(\cdot)$ based on placement values, (2).

[5] proposed a semiparametric form for $h_0(\cdot)$ by demonstrating that (4) is equivalent to $h_0(v_1) = -X'\beta + \epsilon$, where $h_0(\cdot)$ is unknown and v_1 is the set of placement values for the diseased observations. The placement values are the survival probabilities of the disease values for the non-diseased population.

[20] propose an alternative to the parametric and semiparametric methods by modeling the placement values, v_1, with a beta regression.

The next two sections provide a brief description of two methods that are currently in the ROC regression literature. They are the widely used parametric model proposed by [1] and a more recently proposed model by [20].

3.1. Parametric Method

[17] proposed a generalized linear model framework for the ROC given by

$$ROC_{X, X_D}(t) = g \left(h_0(t) + \beta X + \beta_D X_D \right) \qquad (5)$$

is a function of covariates common to both diseased and non-diseased subjects, X, covariates exclusive to diseased subjects, X_D, and a function $h_0(t) = \sum_{k=1}^{K} \gamma_k h_k(t)$ for specified K and known γ_k. [1] considered a special parametric case of (5), called the binormal model, where $h_1(t) = 1$, $h_2(t) = \Phi^{-1}(t)$, and $g(\cdot) = \Phi$, the CDF of the standard normal distribution. It should be noted that their model is a parametric, distribution free approach since a parametric model is specified for the ROC, but no distributional assumptions are made concerning the biomarker Y.

[17] proposed a method for fitting (5) using binary indicators $U_{ij} = I[Y_{1_i} \geq Y_{0_j}]$ for each combination of paired of observations, Y_{1_i} and Y_{0_j} for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_0$, where n_1 and n_0 denote the number of observations in the diseased and reference groups, respectively. [1] modified the binary indicator by replacing Y_{0_j} with $S_{0, X_i}^{-1}(t)$, for $t \in (0, 1)$. In this case, the binary indicator becomes $U_{it} = I[Y_{1_i} \geq S_{0, X_i}^{-1}(t)]$. A key observation is that

$$
\begin{aligned}
E[U_{it}] &= E \left[I[Y_{1_i} \geq S_{0, X_i}^{-1}(t)] \right] \\
&= Pr \left[S_{0, X_i}(Y_{1_i}) \leq t \right] \\
&= Pr \left[v_{1_i} \leq t \right],
\end{aligned}
$$

where v_{1_i} is the placement value for the observation Y_{1_i} given covariate X. An algorithm [20] for the parametric method is as follows:

1. Specify a set $T = \{ t_\ell : \ell = 1, \ldots, n_T \} \in (0, 1)$ of false positive rates (FPRs).

2. Estimate the covariate specific survival functions for the reference population at each $t \in T, j = 1, \ldots, n_t$ using quantile regression.
3. Calculate the placement values $v_i = \widehat{S}_{0,X_i}(Y_1), i = 1, \ldots, n_1$.
4. Calculate the binary indicator $\widehat{U}_{it} = I[v_i \leq t], t \in T$.
5. Fit the model $E\left[\widehat{U}_{it}\right] = g^{-1}\left[\sum_{k=1}^{2} \gamma_k h_k(t) + X\beta\right]$.

3.2. Beta Regression Method

[17] and [1] models use binary random variables defined by the placement values of the diseased responses as referenced with the non-diseased population as the dependent variable. The beta method [20] models the ROC as the CDF of the placement values directly, and does not need the binary response variable. Instead, they modeled the placement values with the beta generalized linear model.

A brief introduction to the beta generalized linear model given in [10] is given below. Suppose that $Z \sim \text{Beta}(\alpha, \beta)$. Then, the mean and variance of Z are given by

$$E(Z) = \frac{\alpha}{\alpha + \beta}, \tag{6}$$

and

$$Var(Z) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{7}$$

Let $\mu = \alpha/(\alpha + \beta)$ and $\phi = \alpha + \beta$. Then, the beta distribution can be parameterized to have mean and variance

$$E(Z) = \mu, \tag{8}$$

and

$$Var(Z) = \frac{\mu(1 - \mu)}{(1 + \phi)}. \tag{9}$$

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 is given by

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\gamma_i = x_t'\gamma, \tag{10}$$

where γ is a vector of regression parameters, $x_{t1}, \ldots, x_{tk}$ are observations on k covariates, and g is a monotonic link function. When $g(\cdot)$ is the logit link, then

$$\mu_t = \frac{1}{1 + e^{-x_t'\gamma}}. \tag{11}$$

In which case, the parameters α and β in the original parameterization of the beta distribution are estimated by

$$\hat{\alpha} = \frac{\hat{\phi}}{1 + e^{-x_t'\hat{\gamma}}}, \tag{12}$$

and

$$\hat{\beta} = \hat{\phi}\left(1 - \frac{1}{1 + e^{-x_t'\hat{\gamma}}}\right). \tag{13}$$

An algorithm [20] for the beta regression method is as follows:
1. Specify a set $T = \{t_\ell : \ell = 1, \ldots, n_T\} \in (0, 1)$ of false positive rates (FPRs).
2. Estimate the covariate specific survival functions for the reference population at each $t \in T, j = 1, \ldots, n_t$ using quantile regression.
3. Calculate the placement values $v_i = \widehat{S}_{0,X_i}(Y_1), i = 1, \ldots, n_1$.
4. Perform a beta regression on the placement values to obtain estimates of γ and ϕ.
5. Transform to obtain $\alpha = \mu\phi$ and $\beta = (1 - \mu)(\phi)$.
6. Calculate the CDF of the placement values using the Beta(α, β) distribution found above to obtain the ROC.

The next section presents a brief discussion of the Lehmann ROC model as given in [13].

4. The Lehmann ROC Regression Model

The Lehmann family of ROC curves, **(author?)** [14], are motivated by the semiparametric relationship

$$S_1(t) = [S_0(t)]^\theta, \tag{14}$$

where S_0 and S_1 are the survival functions of the reference and diseased groups. (14) is known as the Lehmann assumption. The relationship between the survival functions is determined solely by θ. Suppose that the larger marker values are indicative of disease. In which case, $0 < \theta < 1$ will ensure the correct orientation of the survival functions in (14). The ROC curve can be formulated as

$$ROC(t) = S_1\left[S_0^{-1}(t)\right] \tag{15}$$

for $t \in (0, 1)$. Combining (14) and (15) yields the general form of the Lehmann family of ROC curves,

$$ROC(t) = t^\theta. \tag{16}$$

As seen in Figure 2, as θ increases, the ROC curve approaches the diagonal line.
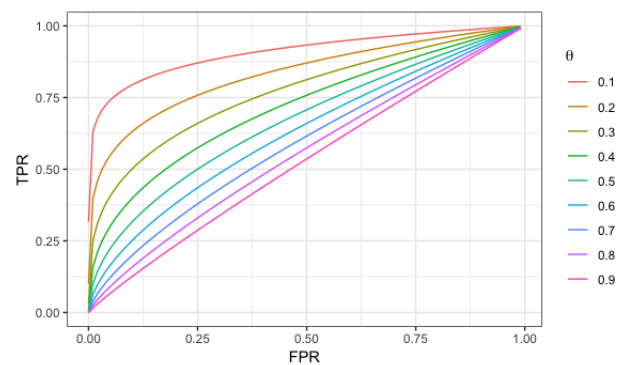

Figure 2. *The Lehmann family of ROC curves with parameter θ ranging from 0.10 to 0.90.*

Alternatively, (14) can be expressed in terms of the hazard functions for the two groups. In which case, the Lehmann model is the familiar Cox proportional hazards model and θ is the hazard ratio [7, 8]. That is,

$$\frac{\lambda_1(t)}{\lambda_0(t)} = \theta, \tag{17}$$

where the hazard function is defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

(17) can be written as

$$\lambda_1(t, D) = \lambda_0(t) \exp\{\beta D\}$$

and $\theta = \exp\{\beta\}$. This expression allows for the use of proportional hazards regression modules commonly found in statistical software, such as R and SAS, to estimate β and subsequently estimate θ. Let $\hat{\beta}$ be the Cox proportional likelihood estimate for β. Then $\hat{\theta} = \exp\{\hat{\beta}\}$ and the estimated ROC is given by

$$\widehat{ROC} = t^{\hat{\theta}} \tag{18}$$

and the estimated area under the curve (AUC) is given by

$$\widehat{AUC} = \int_0^1 t^{\hat{\theta}} = (\hat{\theta} + 1)^{-1}.$$

The model can be extended to estimate a covariate-adjusted ROC curve. [22] showed that this could be done using a regression model with an interaction term that exists when the covariate Z provides a multiplicative effect with D. For the Lehmann family of ROC curves this is accomplished with the proportional hazards model

$$\lambda(t|D, Z) = \tilde{\lambda}(t) \exp\{\beta_1 D + \beta_2 Z + \beta_3 DZ\}, \tag{19}$$

where $\tilde{\lambda}$ is the unspecified baseline hazard function, D is a binary indicator where $D = 1$ represents the diseased group and $D = 0$ represents the reference group, and Z is a covariate common to both the diseased and reference groups. Then (17) becomes

$$\frac{\lambda(t|D = 1, Z)}{\lambda(t|D = 0, Z)} = \exp\{\beta_1 + \beta_3 Z\} \tag{20}$$
$$= \theta(z)$$

and (18) becomes

$$ROC(t, z) = t^{\theta(z)}. \tag{21}$$

The ROC model described above provides a semiparametric method that models the marker values directly. The advantages of using this approach are numerous and can be seen in the simplicity of model fitting, inference, and model specification diagnostics [13]. Although there are numerous advantages

in using the Lehmann model, its use has been limited in the ROC regression context when compared to the generalized linear models (GLM) framework since these data often violate the proportional hazard function assumption. A question of interest might be, can this model be used when the proportional assumption is not met?

In the next section, a simulation study comparing the existing ROC regression models based on the GLM framework are compared with the Lehmann model.

5. Simulations

The simulation given in this section compares the parametric, beta, and Lehmann ROC regression methods using data generated from the normal, extreme value, and Weibull distributions. [19] and [20] provide a comparison of ROC regression methods, including the parametric and beta models, using normal and extreme value data. [12] provide a discussion of the Lehmann model's performance for Weibull data. For comparative purposes the data models in this paper were intentionally chosen to be similar to those found in these papers. As far as we can tell, the performance of parametric and beta models have not been investigated using the weibull nor has the Lehmann model been evaluated using data for which the Cox proportional hazard assumption does not hold, as with the normal and extreme-value data. An additional benefit of using these types of data in the simulation is that we can specify the true ROC. In some cases we will know the close form value for the population AUC. In the extreme-value case the AUC can be computed using numerical integration of the ROC. These values will be used when evaluating the MSE for each method.

Two metrics of interest from the simulation are the area under the curve (AUC) and the Youden index Youden [23].

The Youden index, also known as Youden's J statistic, is the maximum vertical distance between the diagonal line (AUC = 0.50) and the ROC curve. The Youden index is the Kolmogorov-Smirnov test statistic when testing $F(x) \sim U(0, 1)$ ([15]). Our purpose for the simulation was to investigate the ROC regression methods in terms of their parameter estimates rather than their use as an inferential tool. If one was interested in hypothesis testing concerning the diseased versus the control groups one should use the Youden index as it compares the CDFs rather than medians or means as traditional associated with the AUC.

5.1. Simulated Normal Data

Assume that $Y_1 \sim \text{Normal}(\mu_1, \sigma_1^2)$ and $Y_0 \sim \text{Normal}(\mu_0, \sigma_0^2)$. Then the true ROC is given by

$$ROC(t) = \Phi\{a + b\Phi^{-1}(t)\} \tag{22}$$
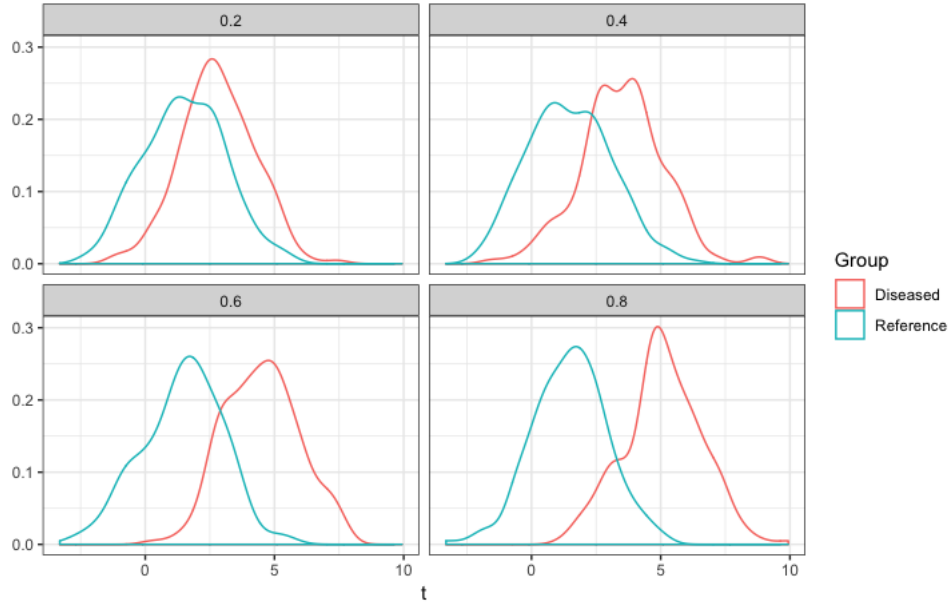
and the corresponding true AUC is given by

$$AUC = \Phi\left\{\frac{a}{\sqrt{1 + b^2}}\right\}, \tag{23}$$

where $a = (\mu_1 - \mu_0)/\sigma_1$ and $b = \sigma_1/\sigma_0$. The following models were used for the simulation
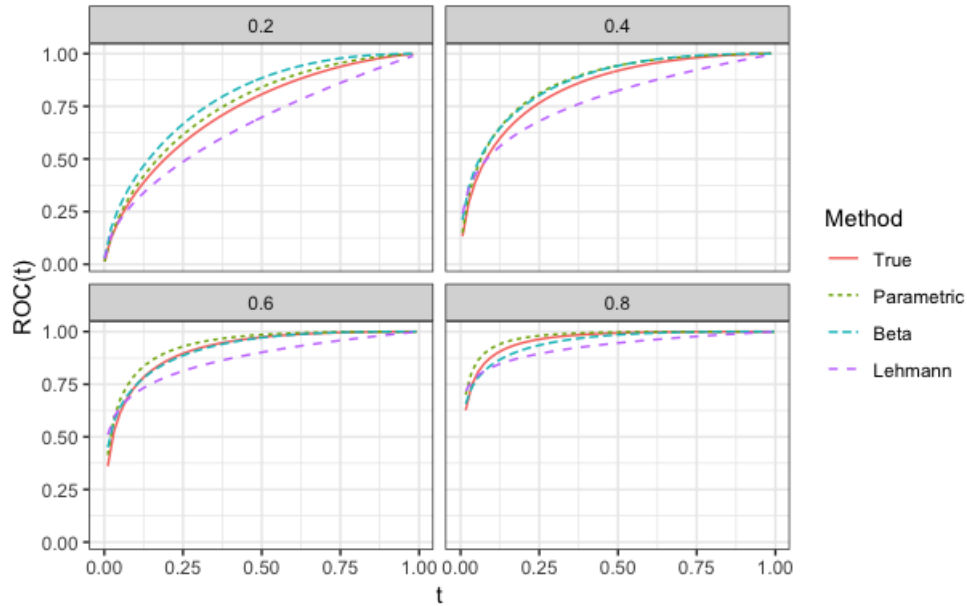
$$Y_0 = 1.5 + \epsilon_0$$
$$Y_1 = 2 + 4x + \epsilon_1$$

where $X \sim U(0,1)$ and $\epsilon_0, \epsilon_1 \sim N(0, 1.5^2)$. From these models, 1000 data sets of size $n_0, n_1 = 100$ are generated. An illustrative plot of the simulated probability densities for the reference and diseased populations at four covariate levels, $x = 0.2, 0.4, 0.6, 0.8$, are given in Figure 3. The separation between the two densities increases as the value of the covariate, x, increases.

Figure 3. *The simulated probability densities for the reference (blue) and diseased (red) populations for normally distributed data.*

Plots of the true and representative simulated ROC curves for the three methods at four covariate levels are given in Figure 4. Note these curves are consistent with the separation of the two density functions for the normal data as given in Figure 3. The curve for the Lehmann method is much flatter than the curves for the other two methods. This is to be expected since the normal data does not, in general, satisfy the Lehmann assumption.
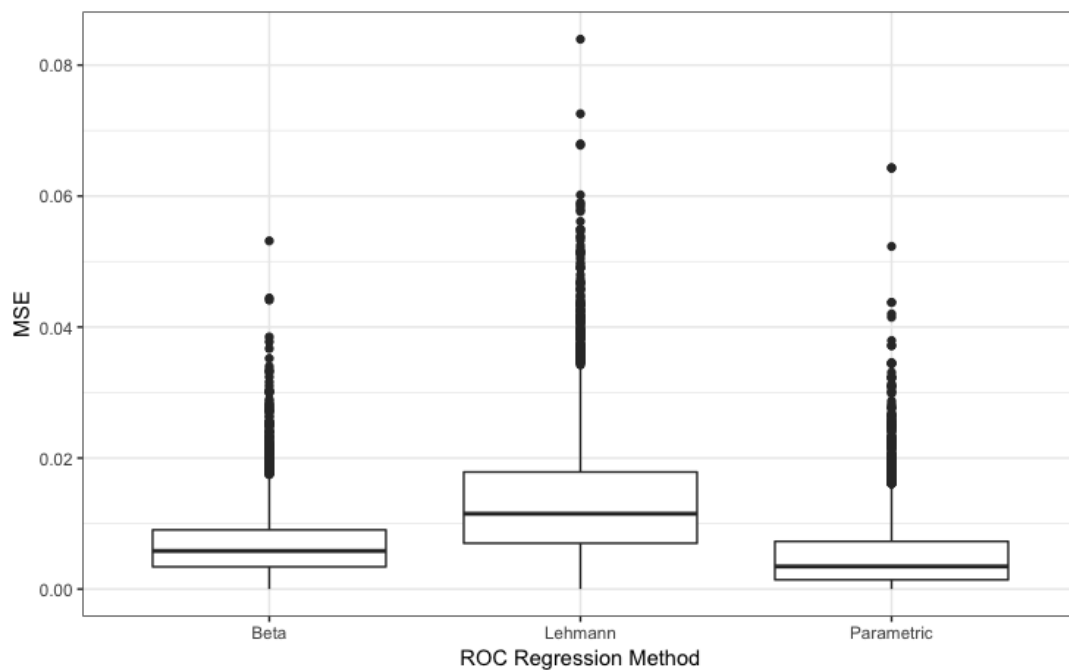

Figure 4. *A representative case of the simulated covariate-adjusted ROC curves for normally distributed data.*

A summary of the MSE for the AUC for simulated data from a normal distribution is given in Table 1.

Table 1. The average MSE for the AUC for the three methods at each of the four covariate levels in the normal data case.

Method	x = 0.2	x = 0.4	x = 0.6	x = 0.8
Parametric	0.0080	0.0073	0.0041	0.0016
Beta	0.0097	0.0089	0.0084	0.0053
Lehmann	0.0142	0.0175	0.0137	0.0081

Box plots of the adjusted MSEs for the four covariate levels are given in Figure 5. The parametric method has the lowest median MSE and the Lehmann method has the highest median MSE. Additionally, the MSEs resulting from the parametric and beta methods have less variability than those resulting from the Lehmann method.

Figure 5. Box plots of the MSE for the three methods based on 1000 estimates with $n_{\bar{D}}, n_D = 100$ in the normal data case.

A summary of the MSE for the Youden index is given in Table 2.

Table 2. The average MSE for the Youden index for the three methods at each of the four covariate levels in the normal data case.

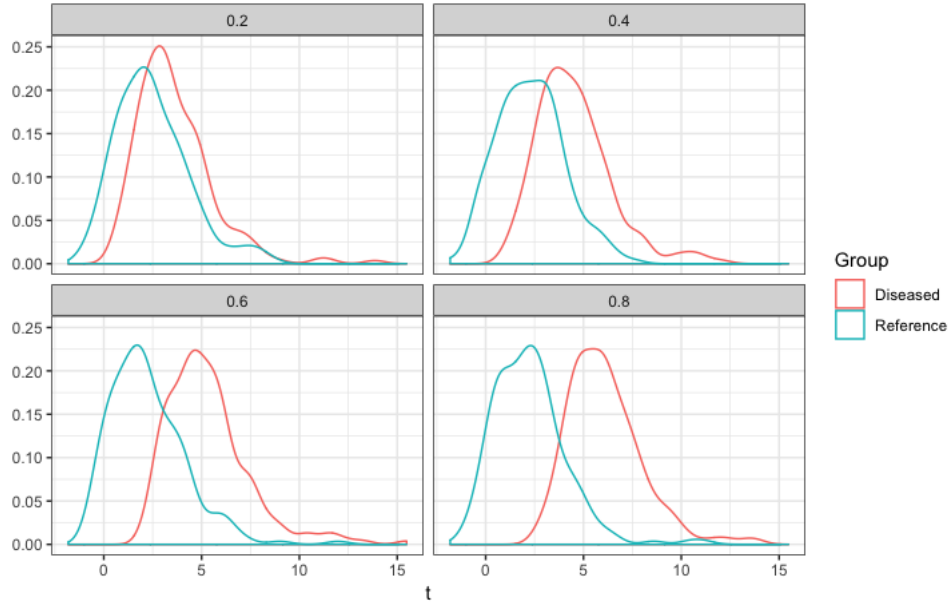| Method	x = 0.2	x = 0.4	x = 0.6	x = 0.8
Parametric	0.0048	0.0027	0.0024	0.0024
Beta	0.0073	0.0029	0.0056	0.0101
Lehmann	0.0081	0.0067	0.0086	0.0108

5.2. Simulated Extreme Value Data

The extreme value distribution, also known as the Gumbel or double exponential distribution, has many forms. In this paper, an extreme value distribution that has a CDF given by $F(x) = \exp\{-\exp[-(x - \mu)/\beta]\}$ where $\mu \in \mathbb{R}$, $\beta > 0$, and $x \in (-\infty, \infty)$ is used. Specifically, the following models are used for this simulation

$$Y_0 = 1.5 + \epsilon_0$$
$$Y_1 = 2 + 4x + \epsilon_1$$

where $X \sim U(0, 1)$ and ϵ_0, ϵ_1 have an extreme value distribution with $\mu = 0$ and $\beta = 1.5$. Again, 1000 data sets of size $n_0, n_1 = 100$ are generated from these models. Plots of the simulated probability densities for the reference and diseased populations at four covariate levels, $x = 0.2, 0.4, 0.6, 0.8$, are given in Figure 6. Again, as the value covariate increases, the separation between the two densities increases.

Figure 6. *The simulated probability densities for the reference (blue) and diseased (red) populations for extreme value data.*
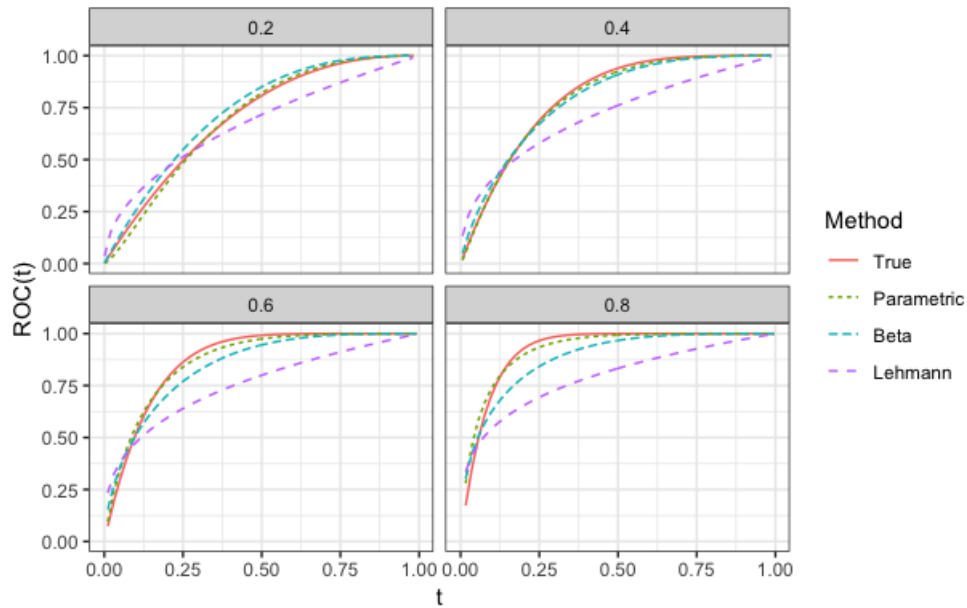
The true value of the ROC for the specified model is given by

$$ROC_X(t) = 1 - \exp\left\{-\exp\left[\ln\left(-\ln(1-t)\right) - \frac{0.5 + 4x}{1.5}\right]\right\}. \tag{24}$$

There is not a closed-form expression for AUC of this extreme value distribution. Instead, the AUC is calculated using numerical integration.

Plots of the true and representative simulated ROC curves for the three methods at four covariate levels are given in Figure 7. Note these curves are consistent with the separation of the two density functions for the extreme value data as given in Figure 6. The curve for the Lehmann method is much flatter than the curves for the other two methods. This is to be expected since the extreme value data does not, in general, satisfy the Lehmann assumption.

A summary of the MSE for the AUC is given in Table 3. The parametric and beta methods have a lower mean MSE than the Lehmann method for all levels of the covariate.
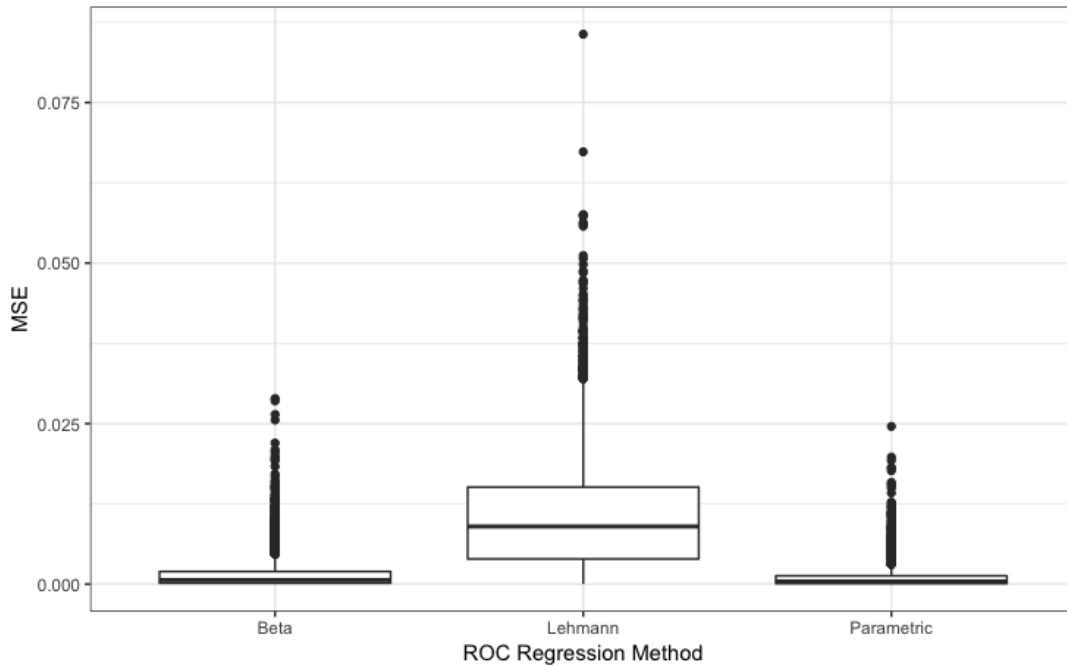

Figure 7. *A representative case of the simulated covariate-adjusted ROC curves for extreme value data.*

Table 3. The average MSE for the AUC for the three methods at each of the four covariate levels in the extreme value data case.

Method	x = 0.2	x = 0.4	x = 0.6	x = 0.8
Parametric	0.0019	0.0011	0.0007	0.0005
Beta	0.0028	0.0012	0.0011	0.0012
Lehmann	0.0073	0.0105	0.0130	0.0134

Box plots of the adjusted MSEs for the three ROC regression methods overall across all four covariate levels are given in Figure 8. Note that the parametric method has the lowest median MSE and the Lehmann method has the highest median MSE. Additionally, the MSEs resulting from the parametric and beta methods have significantly less variability than those resulting from the Lehmann method.

Figure 8. Box plots of the estimated MSE for each of the three ROC regression methods based on 1000 estimates with $n_{\bar{D}}, n_D = 100$ for the extreme value data case.

A summary of the MSE for the Youden index is given in Table 4. Again, a similar pattern can be seen in these results.

Table 4. The average MSE for the Youden index for the three methods at each of the four covariate levels in the extreme value data case.

Method	x = 0.2	x = 0.4	x = 0.6	x = 0.8
Parametric	0.0058	0.0036	0.0103	0.0214
Beta	0.0050	0.0032	0.0032	0.0042
Lehmann	0.0322	0.0212	0.0142	0.0110

5.3. Simulated Weibull Data

In this section a simulation is used to determine how the methods perform when using data from a Weibull distribution. [4] presented a method for generating Weibull data with a specified hazard function for the Cox models. It is given by

$$T = \left(-\frac{\ln(U)}{\lambda \exp(\beta' x)}\right)^{1/\nu} \qquad (25)$$

where the scale parameter is λ, the shape parameter is ν for the hazard function given by $h(t \mid x) = \lambda \exp(\beta' x)\nu t^{\nu-1}$ and $U \sim U(0, 1)$. Let the data from the reference population, Y_0, have a Weibull distribution with scale and shape parameters equal to 1.
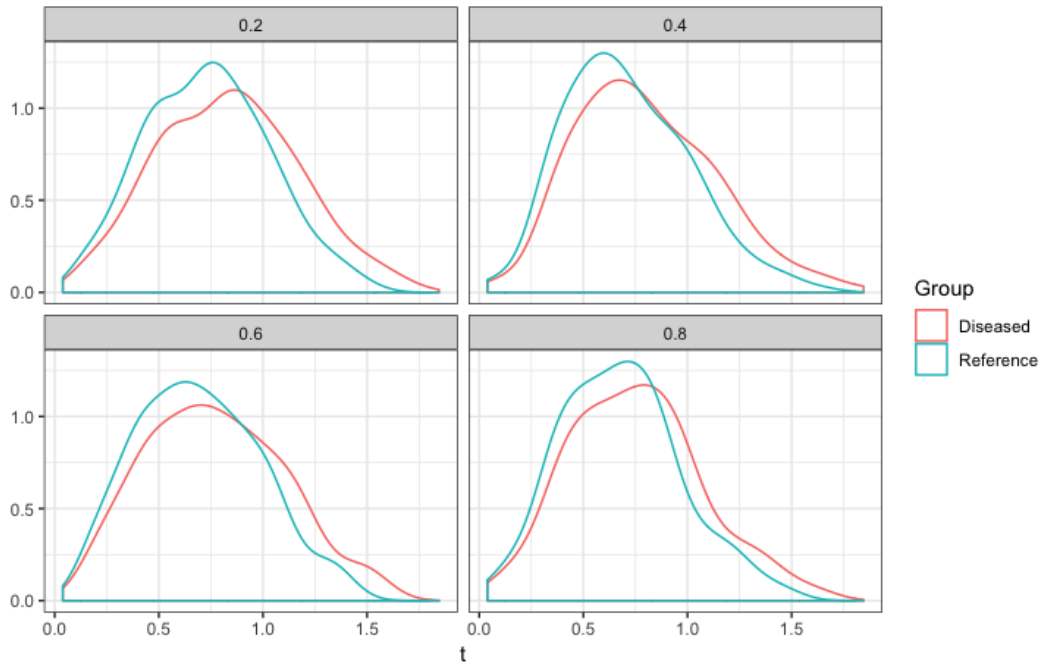
The data from the diseased population, Y_1, have a Weibull distribution with scale parameter $\theta^{-1/s}$, for $\theta = 3/7$ and $s = 1$. The shape parameter is still equal to 1.

As before, 1000 data sets of size $n_0, n_1 = 100$ are generated.
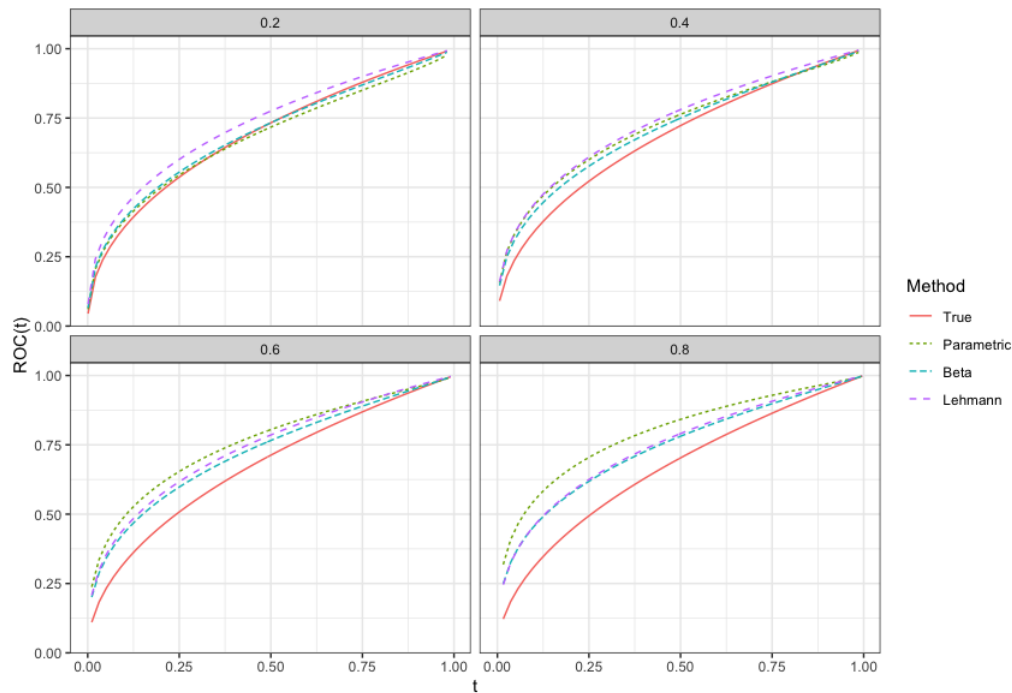
Plots of the simulated probability densities for the reference and diseased population at four covariate levels, $x = 0.2, 0.4, 0.6, 0.8$ are given in Figure 9. From an intuitive viewpoint, the Weibull is much more difficult to work with than the normal or extreme value distributions [4]. From Figure 9, one notes that the separation between the reference and disease groups lessens as x increases. In fact, the density for the reference group (baseline hazard) is the same while the hazard for the disease group is reduced as x increases.

Plots of the true and representative simulated ROC curves for the three methods and at the four covariate levels are given in Figure 10.

Figure 9. *The simulated probability densities for the reference (blue) and diseased (red) populations for Weibull data.*

Figure 10. *A representative case of the simulated covariate-adjusted ROC curves for Weibull data.*

The true value of the ROC curve is given by

$$ROC(t) = t^{\theta(x)}, \tag{26}$$
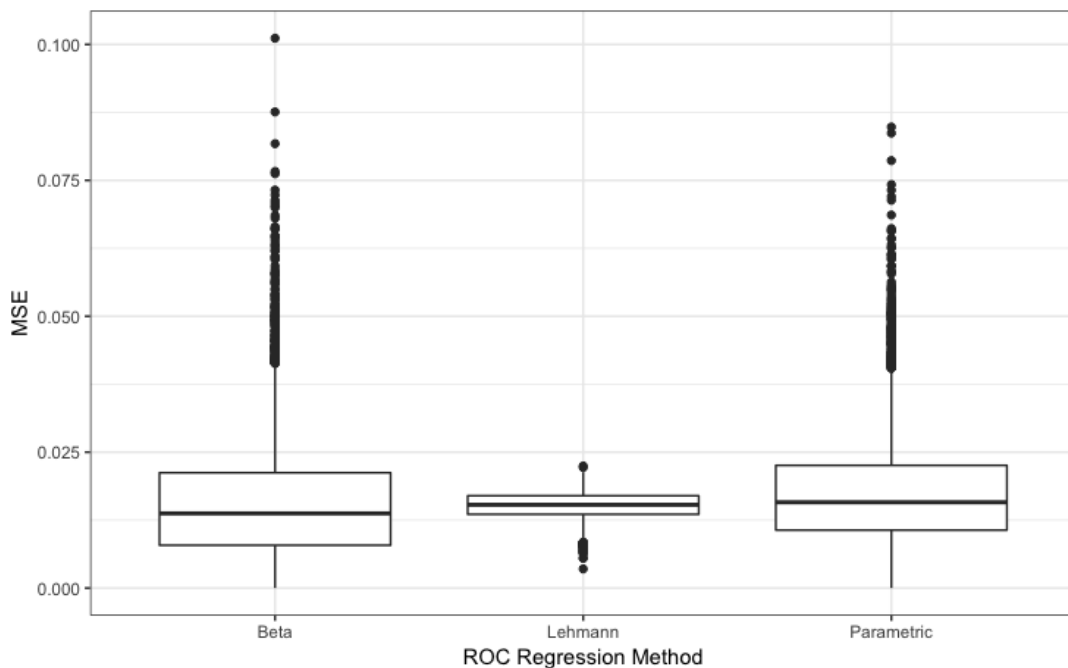
and the true value of the AUC is given by

$$AUC = \frac{1}{1 + \theta(x)}, \tag{27}$$

where $\theta(x) = \theta - (\beta_1' x - \beta_0' x)$. The MSE for the AUC and the Youden index are calculated for each of the three methods. A summary of the MSE for the AUC is given in Table 5. Note that the performance of the Lehmann method is now comparable with the performance of the parametric and beta methods. This is to be expected because the proportional hazards assumption is met in the Weibull data case.

Table 5. *The average MSE for the AUC for the three methods at each of the four covariate levels in the Weibull data case.*

Method	x = 0.2	x = 0.4	x = 0.6	x = 0.8
Parametric	0.0031	0.0010	0.0005	0.0012
Beta	0.0023	0.0007	0.0003	0.0009
Lehmann	0.0003	0.0004	0.0005	0.0008

Box plots of the adjusted MSEs for the three ROC regression methods overall across all four covariate levels are given in Figure 11. Note that the Lehmann method has a median MSE similar to that of the other two ROC regression methods. Additionally, the MSEs resulting from the parametric and beta methods have significantly more variability than those resulting from the Lehmann method.

Figure 11. *Box plots of the estimated MSE for each of the three ROC regression methods based on 1000 estimates with $n_{\bar{D}}, n_D = 100$ for the Weibull data case.*

A summary of the MSE for the Youden index is given in Table 6. Again, a similar pattern can be seen in these results.

Table 6. *The average MSE for the Youden index for the three methods at each of the four covariate levels in the Weibull data case.*

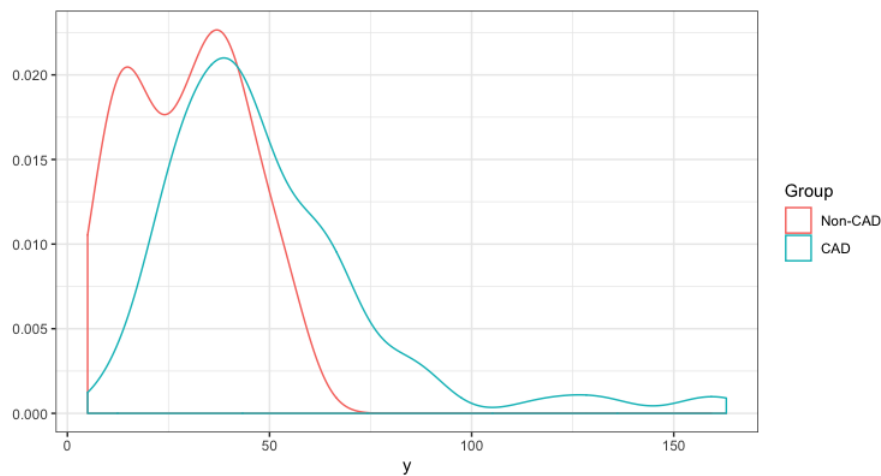| Method	x = 0.2	x = 0.4	x = 0.6	x = 0.8
Parametric	0.0056	0.0021	0.0012	0.0029
Beta	0.0038	0.0014	0.0013	0.0033
Lehmann	0.0008	0.0009	0.0014	0.0021

6. Real Data Application

In this section, we have applied the three ROC methods to a real data application. One of the reviewers questioned our choice of this dated data example for good reason. However, we have chosen not to replace it for two reasons. The first reason is the data set is very easy to obtain, thus allowing others to reproduce similar results using the methods presented in this paper. The second reason is we didn't want the underlying science to be seen as the dominate reason for writing this paper. Instead, we wanted to demonstrate how data of this type can be model using a variety of ROC methods while not introducing highly disputable model choices that one would likely have to make in the analysis of their own specific data problem or application. The results that we find with this data set are nor surprising and are somewhat predictable.

Elastase, a proteolytic enzyme released by neutrophils, plays a potentially critical role in the pathogenesis of coronary artery disease (CAD). [2] conducted a study to evaluate the clinical usefulness of leukocyte elastase determination in the diagnosis of CAD. A subset of the data from this study is given in the elas data set in the OptimalCutpoints R package.
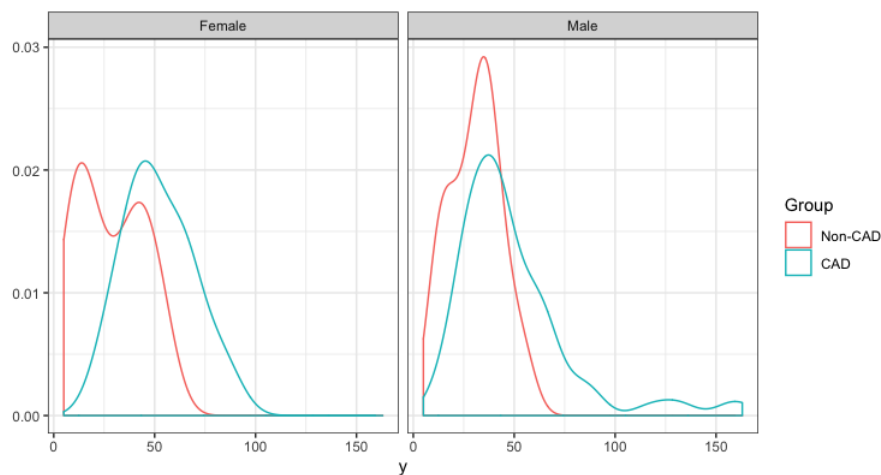
The patients in the study were characterized as belonging to either a non-CAD group or a CAD group based upon the presence (or absence) of angina, a family history of CAD, smoking status, diabetes mellitus, and hypertension. The 141 patients in the study were categorized into the CAD group (96) or the non-CAD group (45). The leukocyte count and plasma lipid and elastase concentrations were measured for each of the subjects in the study.

The non-CAD group will be the reference population and the CAD group will be the diseased population. The response variable of interest is plasma leukocyte elastase concentration. Higher values suggest the presence of complex atheromatous plaques, an indicator of CAD [2]. Density plots of elastase concentration by group are shown in Figure 12.
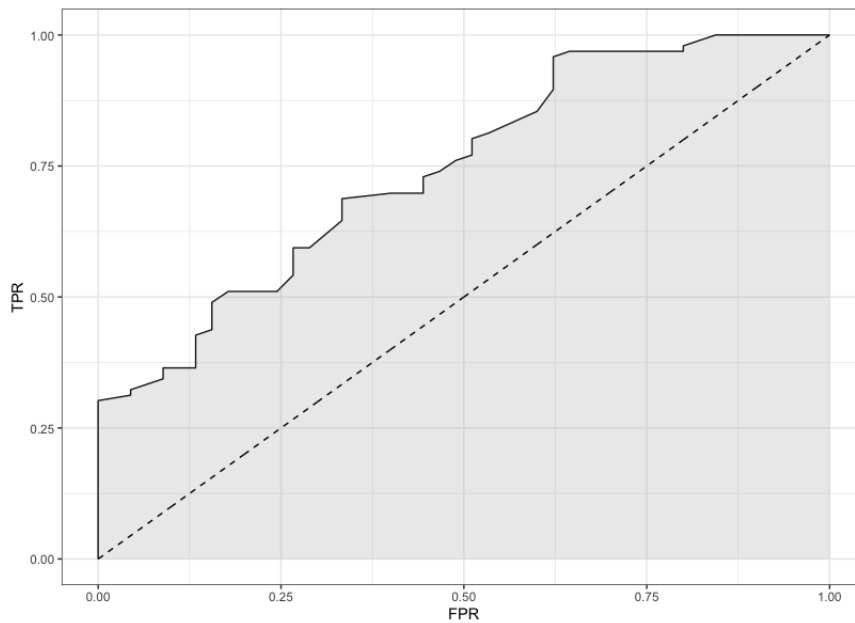

***Figure 12.** The density plots for the plasma leukocyte elastase concentration for the non-CAD (red) and CAD (blue) groups.*

The effect of covariates on the separation between the non-CAD and CAD groups is of interest. GENDER, is the only covariate that is included in the **elas** data set. The density plots of elastase concentration for each group by gender are shown in Figure 13.
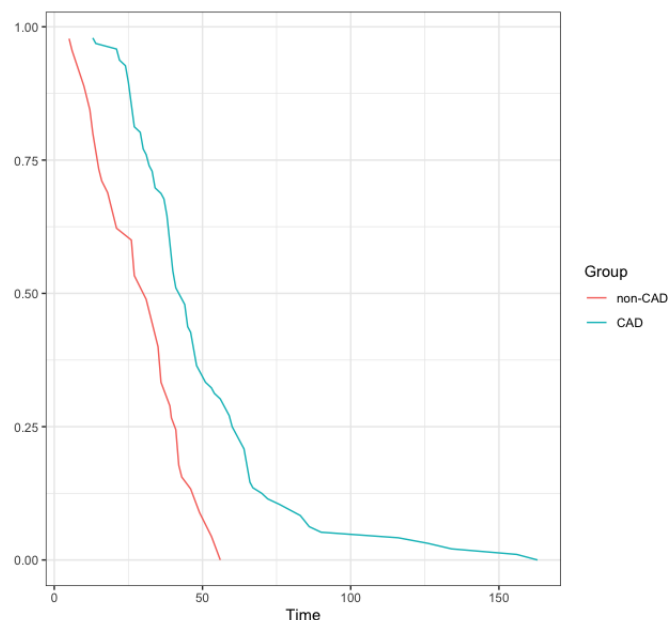

***Figure 13.** The density plots for the plasma leukocyte elastase concentration for the non-CAD (red) and CAD (blue) groups by gender.*

The empirical ROC curve, and the corresponding AUC based on the Mann-Whitney procedure were computed and will act as a comparative baseline for the ROC regression methods. The unadjusted empirical ROC curve is shown in Figure 14 and the corresponding empirical AUC is 0.7436.

Figure 14. The empirical ROC curve based on the Mann-Whitney procedure.
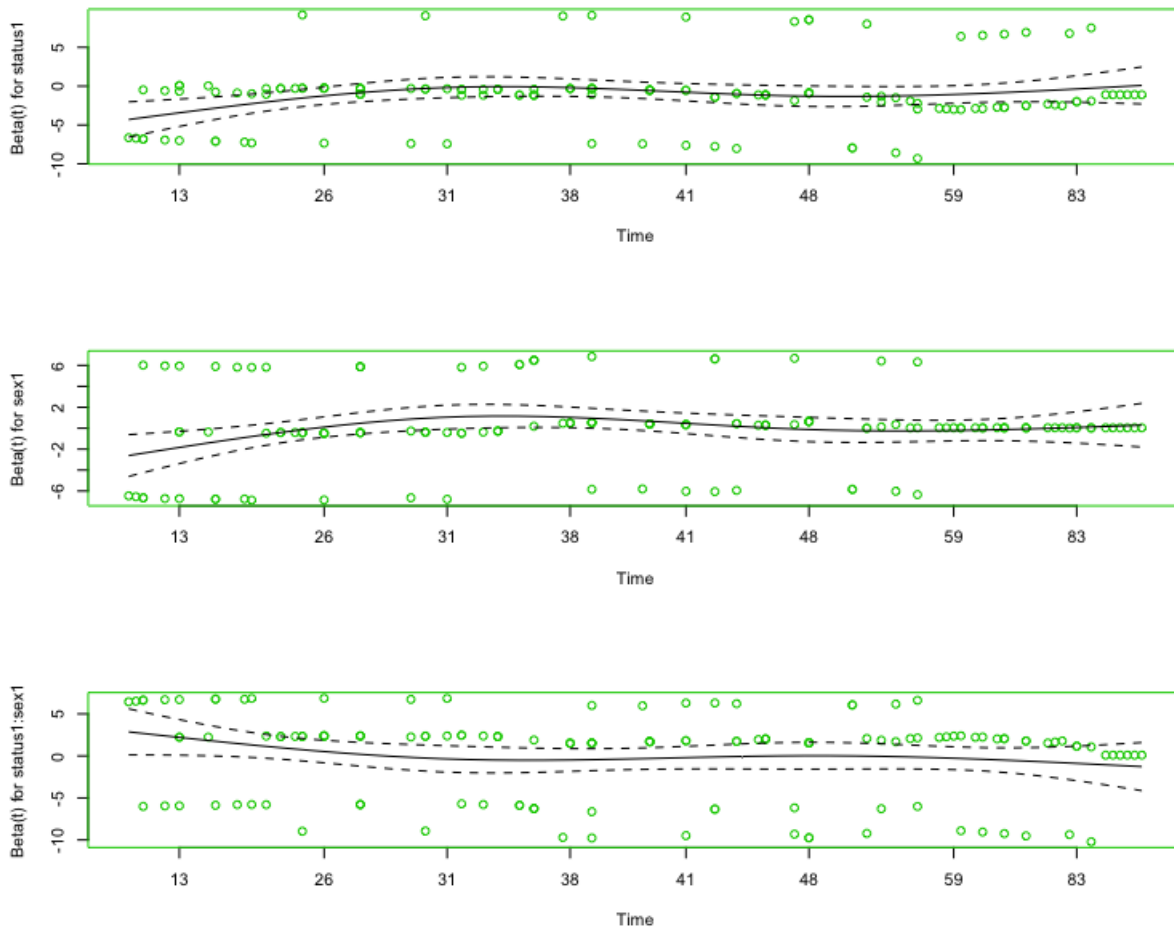
The results for the Lehmann method will be presented first. As seen from the simulations in the previous sections, it is imperative to determine whether or not the proportional hazards assumption holds before applying this method. A simple, visual check can be done by looking at the survival, hazard, or log-hazard curves for the groups. If these curves cross, then the assumption has been violated. The survival curves for the elastase data, shown in Figure 15.

Figure 15. The survival curves for the plasma leukocyte elastase concentration for the non-CAD (red) and CAD (blue) groups.

A secondary check can be done by inspecting the Schoenfeld residuals as shown in Figured 16. If the proportional hazards assumption is satisfied then these plots should be relatively constant over time. Visually, this corresponds to an absence of a clear pattern in the residuals where the plotted line has an approximate zero slope. Each plot has an associated p-value, which were

greater than 0.05. Thus, we are able to assume that the proportional hazards assumption has been satisfied.

Figure 16. *Plots of the Schoenfeld residuals over time for the status variable, the sex variable, and the interaction between the two variables (from top to bottom).*

The covariate-adjusted ROC curves for males and females estimated by the three methods are shown together in Figure 17. The results for the three methods indicate that plasma leukocyte elastase concentration is a sensitive diagnostic marker of CAD. Additionally, the performance of plasma leukocyte elastase as a diagnostic marker of CAD has a significant gender covariate where the separation between non-CAD and CAD is greater for females. This separation is the least when using the Lehmann method. This is likely attributable to the fact that the hazard for CAD is similar for males and females.

The Lehmann AUC for the female and male groups are 0.7494 and 0.7319, respectively. The Youden index values for the female and male groups are 0.3858 and 0.3563, respectively. This indicates that there is slightly more separation between the non-CAD and CAD groups for females. These values are consistent with what is to be expected from the empirical ROC and AUC values. It should be noted that the Lehmann method produced the least significant ROC when compared (with the diagonal line) and the other models. This result was quite common in the analysis

we have done with this model (not shown in this paper).

The second method to be illustrated is the placement value beta method. The AUC for the female and male groups are 0.8357 and 0.7858, respectively. The Youden index values for the female and male groups are 0.5319 and 0.4441, respectively. Again, the results indicate that there is more separation between the non-CAD and CAD groups for females than for males. As in the simulation study, this method tends to produce ROC curves with higher values for the AUC when compared with the Lehmann method but not as high when compared with the parametric binormal model.

The binormal AUC for the female and male groups are 0.9151 and 0.8426, respectively. The Youden index values for the female and male groups are 0.6708 and 0.5246, respectively. Again, the results indicate that there is more separation between the non-CAD and CAD groups for females than for males. This method estimates the greatest separation between the non-CAD and CAD groups. This method tends to produce the most significant ROC curves when compared to the other two methods.
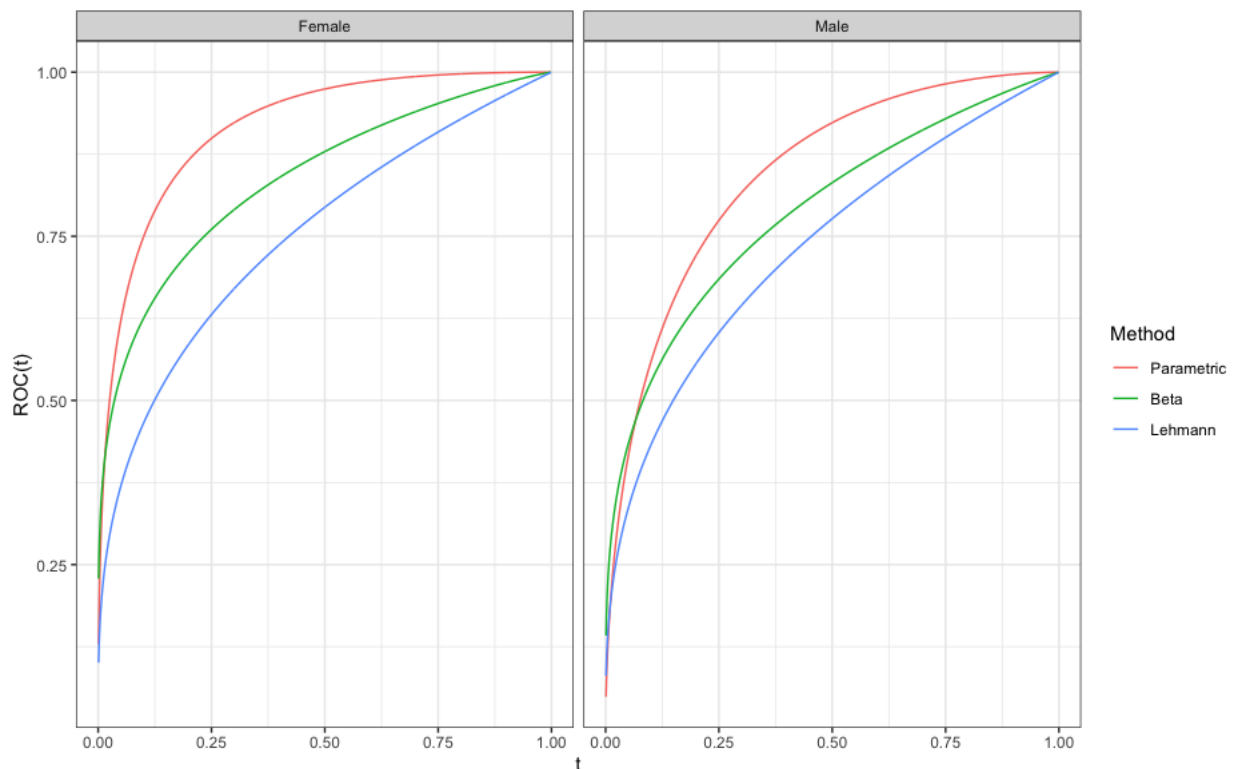
Figure 17. *The estimated covariate-adjusted ROC curves for the plasma leukocyte elastase data.*

7. Discussion

The normal, extreme value, and Weibull simulations provide a comparison of the three ROC regression methods. Recall, the advantage of the Lehmann method is the simplicity of the model. However, it was unknown how it would perform for data generated from the normal and extreme value distributions because these distributions may violate the proportional hazards assumption.

The parametric and beta methods have a lower mean MSE than the Lehmann method for all levels of the covariate for simulated data from a normal distribution. This is to be expected because the normal data does not satisfy the proportional hazards assumption necessary for good performance by the Lehmann method. Similarly, the parametric and beta methods have a lower mean MSE than the Lehmann method for all levels of the covariate for simulated data from an extreme value distribution. However, for the Weibull distribution, the performance of the Lehmann method is superior to that of the other two methods when the covariate x is close to zero. However, when x becomes large the advantage that the Lehmann method has over the other two disappears.

8. Conclusion

The Cox proportional hazard assumption is extremely important when using the Lehmann ROC method. Since the Weibull is the only distribution we used that satisfied this assumption, the Lehmann method preformed best when using Weibull data. The parametric method performed very well with the normal and extreme value data but had some difficulty with the Weibull data. The beta method surprisingly was comparable to the parametric method.

The Lehmann model is very easy to implement and it accommodates a number of modelling issues that are more difficult with the parametric method. It should be considered as a supplementary method when the proportional hazards assumption holds, as demonstrated with the real data application. The results of this application also reveals that the Lehmann method will likely find fewer differences between the reference and treated (diseased) groups.

Funding Information

Authors' Contributions

Melissa Innerst – computation and computer programming support. Created tables and figures.

Jack Tubbs – problem formulation, writing and editing the manuscript.

Musie Ghebremichael – problem formulation and manuscript review.

Ethics

No ethical issues were encountered in connection with this manuscript.

References

[1] Todd A Alonzo and Margaret Sullivan Pepe. Distribution-free roc analysis using binary regression techniques. *Biostatistics*, 3 (3): 421–432, 2002.

[2] A Amaro, F Gude, R Gonzalez-Juanatey, C Iglesias, F Fernandez-Vazquez, J Garcia-Acuna, and M Gil. Plasma leukocyte elastase concentration in angiographically diagnosed coronary artery disease. *European Heart Journal*, 16 (5): 615–622, 1995.

[3] CA Beam. Random-effects models in the receiver operating characteristic curve-based assessment of the effectiveness of diagnostic imaging technology: concepts, approaches, and issues. *Academic Radiology*, 2: S4, 1995.

[4] Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24 (11): 1713–1723, 2005.

[5] Tianxi Cai. Semi-parametric roc regression analysis with placement values. *Biostatistics*, 5 (1): 45–60, 2004.

[6] Tianxi Cai and Margaret Sullivan Pepe. Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association*, 97 (460): 1099–1107, 2002.

[7] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34 (2): 187–202, 1972.

[8] DR Cox. Partial likelihood. biometrika, 62, 269-276. *Mathematical Reviews (MathSciNet): MR400509 Zentralblatt MATH*, 312, 1975.

[9] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44 (3): 837–845, 1988.

[10] Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31 (7): 799–815, 2004.

[11] CA Gatsonis. Random-effects models for diagnostic accuracy data. *Academic Radiology*, 2: S14–21, 1995.

[12] Musie Ghebremichael, Haben Michael, Elijah Paintsil, and Jack Tubbs. Comparing the diagnostics accuracy of cd4+ t-lymphocyte count and percent as a surrogate markers of pediatric hiv disease. *Journal of Mathematics and Statistics*, 15: 55–64, 2019.

[13] Mithat Gönen and Glenn Heller. Lehmann family of roc curves. *Medical Decision Making*, 30 (4): 509–517, 2010.

[14] E. L. Lehmann. The power of rank tests. *Ann. Math. Statist.*, 24 (1): 23–43, 03 1953.

[15] Christos T Nakas, Todd A Alonzo, and Constantin T Yiannoutsos. Accuracy and cut-off point selection in three-class classification problems using a generalization of the youden index. *Statistics in Medicine*, 29 (28): 2946–2955, 2010.

[16] Margaret Sullivan Pepe. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, pages 124–135, 1998.

[17] Margaret Sullivan Pepe. An interpretation for the roc curve and inference using glm procedures. *Biometrics*, 56 (2): 352–359, 2000.

[18] Margaret Sullivan Pepe and Tianxi Cai. The analysis of placement values for evaluating discriminatory measures. *Biometrics*, 60 (2): 528–535, 2004.

[19] María Xosé Rodríguez-Álvarez, Pablo G Tahoces, Carmen Cadarso-Suárez, and María José Lado. Comparative study of roc regression techniquesłapplications for the computer-aided diagnostic system in breast cancer detection. *Computational Statistics & Data Analysis*, 55 (1): 888–902, 2011.

[20] Sarah Stanley and Jack Tubbs. Beta regression for modeling a covariate adjusted roc. *Science Journal of Applied Mathematics and Statistics*, 6 (4): 110, 2018.

[21] Mary Lou Thompson and Walter Zucchini. On the statistical analysis of roc curves. *Statistics in Medicine*, 8 (10): 1277–1290, 1989.

[22] Anna N Angelos Tosteson and Colin B Begg. A general regression methodology for roc curve estimation. *Medical Decision Making*, 8 (3): 204–215, 1988.

[23] William J Youden. Index for rating diagnostic tests. *Cancer*, 3 (1): 32–35, 1950.