Journal of Data Science 13(2014), 328-

General Semiparametric Area Under the Curve Regression Model with Discrete Covariates

Som B. Bohora[a], Yan D. Zhao[b†] and Tatiana N. Balachova[a]

[a] *Department of Pediatrics, The University of Oklahoma Health Sciences, Oklahoma City, Oklahoma 73104, USA*
b Department of Biostatistics and Epidemiology, College of Public Health, The University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA

Abstract: In this article, we considered the analysis of data with a non-normally distributed response variable. In particular, we extended an existing Area Under the Curve (AUC) regression model that handles only two discrete covariates to a general AUC regression model that can be used to analyze data with unrestricted number of discrete covariates. Comparing with other similar methods which require iterative algorithms and bootstrap procedure, our method involved only closed-form formulae for parameter estimation. Additionally, we also discussed the issue of model identifiability. Our model has broad applicability in clinical trials due to the ease of interpretation on model parameters. We applied our model to analyze a clinical trial evaluating the effects of educational brochures for preventing Fetal Alcohol Spectrum Disorders (FASD). Finally, for a variety of simulation scenarios, our method produced parameter estimates with small biases and confidence intervals with nominal coverage probabilities.

Key words: AUC; clinical trial; discrete covariates; nonparametric; semiparametric.

1. Introduction

Linear models are commonly used to assess the association between a continuous response variable and a set of covariates. Such models are flexible in that model terms can be continuous or discrete covariates as well as interactions among these covariates. Despite many advantages of using linear models, there are data sets where linear models may not be appropriate to apply. For example, data sets exist where the response variables are not normally distributed. Even with transformations, the transformed response variables may still be non-normally distributed. For these data sets, statistical methods other than linear models are desired.

Nonparametric statistical methods are routinely used to analyze continuous response variables when linear models are inappropriate to use. Many nonparametric methods are rank-based and therefore are applicable not only to continuous but also to ordinal response variables. The Wilcoxon-Mann-Whitney test can be performed for comparing non-normal responses

† Corresponding author. E-Mail: daniel-zhao@ouhsc.edu

between two treatment groups (Mann & Whitney, 1947; Wilcoxon, 1945). The test does not make normality assumption and loses only a small amount of power as compared with the two-sample t-test even when the normality assumption holds. However, this test cannot adjust for any other covariates, which limits its use in applications where additional covariates are available for analysis. In situations where a stratum effect exists, where the stratum is defined by one discrete covariate or combinations of several discrete covariates, the van Elteren test can be conducted to compare the response variable between two groups while adjusting for the stratum effect (Van Elteren, 1960). However, in more complex scenarios, the van Elteren test is no longer applicable. For example, this test is unable to handle the interaction between the treatment and the stratum effect.

In many applications, comparing two groups while adjusting for multiple covariates is desired for the statistical analysis. For instance, in clinical trials, adjusting for covariates is a necessary aspect of the statistical analysis in order to improve the precision of the treatment comparison and to assess effect modification. For this type of data and in the context of nonparametric methods, Dodd and Pepe (2003) and Brumback, Pepe, and Alonzo (2006) used the area under the curve (AUC) regression model to compare two treatment groups while adjusting for continuous and/or discrete covariates. Their method applied the logistic regression model on a derived binary outcome variable with correlate values for parameter estimation and then the bootstrap method for standard error estimation. However, the bootstrap method involved heavy computation and did not produce an explicit analytical solution for the standard error estimation. In addition, although their method can be used to adjust for continuous covariates, they did not provide a method for checking the linearity assumption between the logits and the continuous covariates. Therefore, their method could lead to bias when the linearity between the logits and the continuous covariates does not hold.

Recently, new methods have also been developed to perform the AUC regression that can adjust for covariate effects. A non-parametric AUC regression method adjusting for a continuous covariate was proposed by (Rodríguez-Álvarez, Roca-Pardiñas, & Cadarso-Suárez, 2011). However, a bootstrap method must be used for testing the effect of the covariate on the Receiver Operating Characteristic (ROC) curves with the proposed method. In addition, this methodology can adjust for only one continuous covariate and not for a discrete covariate or for a combination of both. (Rajan & Zhou, 2012) proposed a semiparametric AUC regression method for ordinal data that can account for both discrete and continuous covariates. They used a log link function and estimating equations for model estimation. With this method, the empirical variance-covariance matrix for parameter estimates was based on additional distributional assumption and thus, was difficult to calculate and needed to be calculated using a bootstrap method. Similar to Dodd and Pepe (2003) and Brumback et al. (2006), no method for checking the linearity assumption between the logits and the continuous covariates was provided. Moreover, additional contributions to the development of AUC models are recently added to the literature (Branscum, Johnson, Hanson, & Baron, 2015; Inacio de Carvalho, Jara, E. Hanson, & de Carvalho, 2013; Rodríguez & Martínez, 2014). Although these studies focus on the semiparametric AUC models, they use either Bayesian approach or do not directly

address the issue mentioned above. In addition, a recent study attempted to extend the AUC regression model to a case for ordered multiple treatment levels as an alternative hypothesis using the Jonckheere-Terpstra statistic (Buros, Tubbs, & Van Zyl, 2016). There are other methods developed with ROC curves for correlated measures for example, Gao, Xiong, Yan, Yu, and Zhang (2008) and Liu, Li, Cumberland, and Wu (2005), but these methods are either computationally extensive or impose additional distributional assumptions.

For data sets with discrete covariates, but no continuous covariates, a semiparametric AUC regression method that used a set of AUC estimates for obtaining parameter estimates and combined Delong's method and the delta method for computing parameter standard errors was proposed by (Zhang, Zhao, & Tubbs, 2011). Their method was computationally simple based on closed-form parameter and standard error estimation. However, (Zhang et al., 2011) applied their new algorithm for estimating parameters and standard errors for the AUC regression model with only two discrete covariates, which led to the initiation of our research. Therefore, in this article, we intend to extend the semiparametric AUC regression model developed by (Zhang et al., 2011) to a general framework that can adjust for any number of categorical covariates and their interactions.

The remainder of the paper is organized as follows. In Section 2, we describe our model and derive the estimation method. In Section 3, we conduct simulation studies to assess the performance our method under various simulation scenarios. We illustrate the application of our method and interpretation of our model results using a clinical trial. We provide a discussion of issues on using our method in Section 5.

2. Methods

2.1 Models

We consider applications that compare a response variable y between two groups (A and B) while adjusting for k categorical covariates $X_1, X_2, \dots, X_k$. The response variable y is a continuous or ordinal variable that is not necessarily normally distributed. Without loss of generality, we assume each covariate is coded such that $X_i = 1, \dots, n_i,$ for $i = 1, \dots, k$. For each combination of the levels of the covariates, we define the Area Under the ROC curve (AUC) in the following way:

$$\pi_{x_1 x_2 \dots x_k} = P(Y^A > Y^B | X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) + \frac{1}{2} P(Y^A = Y^B | X_1 = x_1, X_2 = x_2, \dots, X_k = x_k)$$

where $x_1 = 1, \dots, n_1, \dots, x_k = 1, \dots, n_k,$ and Y^A and Y^B are two randomly chosen observations from Group A and B, respectively. The second term in the above equation is for the purpose of accounting for ties.

For each covariate X_i, without loss of generality, we use the last category as the reference category and define $n_i - 1$ dummy variables $X_i^{(1)}, X_i^{(2)}, \dots, X_i^{(n_i-1)}$ such that

$$X_i^{(j)}(x) = \begin{cases} 1, & j = x \\ 0, & j \neq x \end{cases}$$

where $i = 1, \dots, k;\ j = 1, \dots, n_i - 1;\ x = 1, \dots, n_i$ .We model the association between AUC $\pi_{x_1 x_2 \dots x_k}$ and covariates using a logistic model. Such a model specifies that the *logit* of $\pi_{x_1 x_2 \dots x_k}$ is a linear combination of terms that are products of the dummy variables defined above. Specifically,

$$logit\left(\pi_{x_1 x_2 \dots x_k}\right) = \boldsymbol{Z}_{x_1 x_2 \dots x_k}\boldsymbol{\beta}$$

where $\boldsymbol{Z}_{x_1 x_2 \dots x_k}$ is a row vector whose elements are zeroes or ones and are products of $X_1^{(1)}(x_1), \dots, X_1^{(n_i-1)}(x_1), \dots, X_k^{(1)}(x_k), \dots, X_k^{(n_k-1)}(x_k)$, and $\boldsymbol{\beta}$ is a column vector of nonrandom unknown parameters. Now, define a column vector $\boldsymbol{\pi}$ by stacking up $\pi_{x_1 x_2 \dots x_k}$ and define a matrix $\boldsymbol{Z}$ by stacking up $Z_{x_1 x_2 \dots x_k}$, as x_i ranges from 1 to n_i, $i = 1, \dots, k$, so that our final model is

$$logit(\boldsymbol{\pi}) = \boldsymbol{Z}\boldsymbol{\beta} \qquad \textbf{(1)}$$

The reason for us to use a *logit* transformation of the AUC instead of using the original AUC is for variance stabilization. We will illustrate the above general model using examples.

2.2 Examples

Consider as an example with only one covariate X_1 with $n_1 = 3$ categories. The last category is the reference category. The above general formula can be applied to this specific example as follows.

$$logit\ (\pi_i) = \beta_0 + \beta_1 X_1^{(1)}(i) + \beta_2 X_1^{(2)}(i),$$

where $i = 1, 2, 3,$

$$X_1^{(1)}(i) = \begin{cases} 1, & if\ i = 1 \\ 0, & otherwise' \end{cases}$$

$$X_1^{(2)}(i) = \begin{cases} 1, & if\ i = 2 \\ 0, & otherwise' \end{cases}$$

And,

$$Z_i = (1, X_1^{(1)}(i), X_1^{(2)}(i)$$

The matrix form of $\boldsymbol{\pi}$, $\boldsymbol{Z}$, and $\boldsymbol{\beta}$ can be expressed as follows;

$$\boldsymbol{\pi}_{3\times1}=\begin{pmatrix}\pi_1\\ \pi_2\\ \pi_3\end{pmatrix},\qquad \boldsymbol{Z}_{3\times3}=\begin{pmatrix}1 & 1 & 0\\ 1 & 0 & 1\\ 1 & 0 & 0\end{pmatrix},\qquad \boldsymbol{\beta}_{3\times1}=\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix}.$$

From this formulation, we see that $\beta_1 = \beta_2 = 0$ implies no interaction exists between the group variable and covariate X_1. If additionally $\beta_0 = 0$, then there is no difference between two groups in the response variable, Y.

Considering another example where $k = 2, n_1 = 2, n_2 = 3$, and no interaction between two covariates X_1 and X_2 exists, we obtain the following model;

$$logit\left(\pi_{ij}\right) = \beta_0 + \beta_1 X_1^{(1)}(i) + \beta_2 X_2^{(1)}(j) + \beta_3 X_2^{(2)}(j),$$

where i = 1, 2, j = 1, 2, 3,

$$X_1^{(1)}(i) = \begin{cases}1, & if\ i = 1\\ 0, & otherwise'\end{cases}$$

$$X_2^{(1)}(j) = \begin{cases}1, & if\ j = 1\\ 0, & otherwise'\end{cases}$$

$$X_2^{(2)}(j) = \begin{cases}1, & if\ j = 2\\ 0, & otherwise'\end{cases}$$

And,

$$Z_{ij} = (1, X_1^{(1)}(i), X_2^{(1)}(j), X_2^{(2)}(j)$$

The matrix form of π, Z, and β can be expressed as follows;

$$\boldsymbol{\pi}_{6\times1}=\begin{pmatrix}\pi_{11}\\ \pi_{12}\\ \pi_{13}\\ \pi_{21}\\ \pi_{22}\\ \pi_{23}\end{pmatrix},\qquad \boldsymbol{Z}_{6\times4}=\begin{pmatrix}1 & 1 & 1 & 0\\ 1 & 1 & 0 & 1\\ 1 & 1 & 0 & 0\\ 1 & 0 & 1 & 0\\ 1 & 0 & 0 & 1\\ 1 & 0 & 0 & 0\end{pmatrix},\qquad \boldsymbol{\beta}_{4\times1}=\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\end{pmatrix}.$$

When $k=2, n_1=2, n_2=3$ and there **exists an interaction** between two covariates X_1 and X_2, we obtain the following model

$$logit\left(\pi_{ij}\right) = \beta_0 + \beta_1 X_1^{(1)}(i) + \beta_2 X_2^{(1)}(j) + \beta_3 X_2^{(2)}(j) + \beta_4\{X_1^{(1)}(i) * X_2^{(1)}(j)\} + \beta_5\{X_1^{(1)}(i) * X_2^{(2)}(j)\}$$

where

$$Z_{ij} = (1, X_1^{(1)}(i), X_2^{(1)}(j), X_2^{(2)}(j), X_1^{(1)}(i) * X_2^{(1)}(j), X_1^{(1)}(i) * X_2^{(2)}(j))$$

The matrix form of π, Z, and β can be expressed as follows,

$$\boldsymbol{\pi}_{6\times1} = \begin{pmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{13} \\ \pi_{21} \\ \pi_{22} \\ \pi_{23} \end{pmatrix}, \qquad \boldsymbol{Z}_{6\times6} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \boldsymbol{\beta}_{6\times1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}.$$

2.3 Estimation

First, we denote the number of observations with covariates $X_1 = i_1, \dots, X_k = i_k$ in groups A and B by $N^A_{i_1 \dots i_k}$ and $N^B_{i_1 \dots i_k}$, respectively. We assume both $N^A_{i_1 \dots i_k}$ and $N^B_{i_1 \dots i_k}$ are greater than zero in the following development. An unbiased estimator of $\pi_{i_1 \dots i_k}$ proposed by (Mann & Whitney, 1947) is

$$\hat{\pi}_{i_1 \dots i_k} = \frac{\sum_{l=1}^{N^A_{i_1 \dots i_k}} \sum_{j=1}^{N^B_{i_1 \dots i_k}} I_{lj}}{N^A_{i_1 \dots i_k} N^B_{i_1 \dots i_k}}$$

where

$$I_{i_1 \dots i_k;\, lj} = \begin{cases} 1, & Y^A_{i_1 \dots i_k;\, l} > Y^B_{i_1 \dots i_k;\, j}, \\ \frac{1}{2}, & Y^A_{i_1 \dots i_k;\, l} = Y^B_{i_1 \dots i_k;\, j}, \\ 0, & Y^A_{i_1 \dots i_k;\, l} < Y^B_{i_1 \dots i_k;\, j}. \end{cases}$$

And $Y^A_{i_1 \dots i_k;\, l}$ and $Y^B_{i_1 \dots i_k;\, j}$ are observations with $X_1 = i_1, \dots, X_k = i_k$ in groups A and B, respectively. It is shown by (DeLong, DeLong, & Clarke-Pearson, 1988) that

$$\hat{\pi}_{i_1 \dots i_k} \dot{\sim} N\left(\pi_{i_1 \dots i_k}, \sigma^2_{i_1 \dots i_k}\right)$$

In order to obtain an estimator for $\sigma^2_{i_1 \dots i_k}$, they first computed

$$V^A_{i_1 \dots i_k;\, l} = \frac{1}{N^B_{i_1 \dots i_k}} \sum_{j=1}^{N^B_{i_1 \dots i_k}} I_{lj}, \quad l = 1, \dots, N^A_{i_1 \dots i_k}$$

and

$$V^B_{i_1\dots i_k;j} = \frac{1}{N^A_{i_1\dots i_k}} \sum_{l=1}^{N^A_{i_1\dots i_k}} I_{lj}, \quad j = 1, \dots, N^B_{i_1\dots i_k}.$$

Then, an estimate of the variance of the nonparametric AUC was

$$\hat{\sigma}^2_{i_1\dots i_k} = \frac{\left(s^A_{i_1\dots i_k}\right)^2}{N^A_{i_1\dots i_k}} + \frac{\left(s^B_{i_1\dots i_k}\right)^2}{N^B_{i_1\dots i_k}},$$

where $\left(s^A_{i_1\dots i_k}\right)^2$ and $\left(s^B_{i_1\dots i_k}\right)^2$ were the sample variances of $(V^A_{i_1\dots i_k;\, l}: l = 1, \dots, N^A_{i_1\dots i_k}$ and $V^B_{i_1\dots i_k;j}: j = 1, \dots, N^B_{i_1\dots i_k}$, respectively. Clearly, we need both $N^A_{i_1\dots i_k}$ and $N^B_{i_1\dots i_k}$ are greater than two in order to compute $\hat{\sigma}^2_{i_1\dots i_k}$.

Now, in order to estimate parameters in Model (1), we first derive the asymptotic variance of $\hat{\gamma}_{i_1\dots i_k}$ using the delta method, which results in

$$\hat{\gamma}_{i_1\dots i_k} = logit\left(\hat{\pi}_{i_1\dots i_k}\right) \dot{\sim} N\left\{logit\left(\pi_{i_1\dots i_k}\right), \tau^2_{i_1\dots i_k}\right\},$$

where $\tau^2_{i_1\dots i_k} = \frac{\sigma^2_{i_1\dots i_k}}{\hat{\pi}^2_{i_1\dots i_k}\left(1-\hat{\pi}_{i_1\dots i_k}\right)^2}$ and can be estimated by $\hat{\tau}^2_{i_1\dots i_k} = \frac{\hat{\sigma}^2_{i_1\dots i_k}}{\hat{\pi}^2_{i_1\dots i_k}\left(1-\hat{\pi}_{i_1\dots i_k}\right)^2}$.

Rewriting the above model, we obtain

$$\hat{\gamma}_{i_1\dots i_k} = logit\left(\hat{\pi}_{i_1\dots i_k}\right) = Z_{i_1\dots i_k}\boldsymbol{\beta} + \epsilon_{i_1\dots i_k}$$

where, $\epsilon_{i_1,\dots,i_k} \sim N\left(0, \hat{\tau}^2_{i_1,\dots,i_k}\right)$. Then, by stacking up the $\hat{\gamma}_{1_i,\dots,i_k}$ to be $\widehat{\boldsymbol{\gamma}}$, $Z_{i_1\dots i_k}$ to be $\boldsymbol{Z}$, and $\epsilon_{i_1,\dots,i_k}$ to be $\boldsymbol{\epsilon}$, we have

$$\widehat{\boldsymbol{\gamma}} = logit\, \widehat{\boldsymbol{\pi}} = \boldsymbol{Z\beta} + \boldsymbol{\epsilon}$$

where, $E(\epsilon) = 0$ and $\widehat{\mathrm{T}} = Var(\boldsymbol{\epsilon}) = diag(\hat{\tau}^2_{i_1,\dots,i_k})$ which is a diagonal matrix. Finally, by using the generalized least squares method, we estimate the parameters $\boldsymbol{\beta}$ and its variance-covariance matrix as follows;

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}'\, \widehat{\mathbf{T}}^{-1}\, \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}'\, \widehat{\boldsymbol{T}}^{-1}\, \widehat{\boldsymbol{\gamma}}$$

and

$$\hat{V}\left(\widehat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{Z}'\, \widehat{\boldsymbol{T}}^{-1}\boldsymbol{Z}\right)^{-1}$$

The above equations can be used to construct a $100(1-\alpha)\%$ Wald confidence intervals for $\boldsymbol{\beta_i}$ using formula

$$\hat{\beta}_i \pm Z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{\beta}_i\right)},$$

Where, $Z_{1-\alpha/2}$ is the $(1-\alpha/2)^{th}$ quantile of the standard normal distribution. Equivalently, we reject $H_0: \beta_i = 0$ if $|\hat{\beta}_i| > Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\beta}_i)}$. The p-value for testing H_0 is $2P\left(|Z| > |\hat{\beta}_i|/\sqrt{\hat{V}(\hat{\beta}_i)}\right)$, where Z is a random variable with the standard normal distribution.

Now, the total number of cells (combinations of covariates $X_1, \ldots, X_k$) is $n_1 n_2 \cdots n_k$. As mentioned earlier, for a cell to be usable in the estimation, the cell needs to have at least two observations from Group A and two observations from Group B. As long as the total number of usable cells is larger than the dimension of $\boldsymbol{\beta}$, then the matrix $\boldsymbol{Z}'\,\hat{\boldsymbol{\Upsilon}}^{-1}\,\boldsymbol{Z}$ is invertible and consequently, $\hat{\boldsymbol{\beta}}$ is computable and model (1) is identifiable.

3. Simulation

We conducted simulation studies to assess the performance of our new model estimation method under various scenarios. For each simulation scenario, we generated 10,000 data sets. For each dataset, we computed the parameter estimates, variance estimates, and 95% CIs for the parameters according to the method derived in the previous chapter. Then, for each parameter, the bias was calculated as the average of the 10,000 parameter estimates subtracting the true parameter value, and the coverage probability was calculated as the proportion of the 10,000 95% CIs that cover the true parameter value.

Two models were used to generate data. One discrete covariate was used for the first model, and two discrete covariates were used for the second model. We aim to generate a response variable and covariates that satisfy model (1). For this purpose, we will use the following proposition: let $Y^A = -\log(u^A) + \text{logit}(\pi)$ and $Y^B = -\log(u^B)$, where u^A and u^B are independent random variables with an exponential(1) distribution, then $P(Y^A > Y^B) = \pi$.

For the first model, we considered one covariate with three levels. The general model (1) in this case can be written as $logit(\pi_i) = \beta_0 + \beta_i$, where $i = 1, 2, 3$. In order to generate data satisfying this model, by the proposition above, we generated $Y_{i;j}^A = -\log\left(u_{i;j}^A\right) + logit(\pi_i)$ and $Y_{i;l}^B = -\log\left(u_{i;l}^B\right)$, where $i = 1, 2, 3$; $j = 1, \ldots, n^A$; $l = 1, \ldots, n^B$, and $\{u_{i;j}^A, u_{i;l}^B : i = 1, 2, 3;\ j = 1, \ldots, n^A;\ l = 1, \ldots, n^B\}$ was a set of mutually independent random variables with exponential(1) distribution. For numerical values of the parameters, we set β_1 to be zero so that level 1 was the reference level. Two sets of parameter values: (1) $\beta_0 = 0.15$, $\beta_2 = 0.50$, $\beta_3 = 1.00$, (2) $\beta_0 = 0.10$, $\beta_2 = 0.30$, $\beta_3 = 1.20$, and three sets of sample sizes: (1) n_A=14, n_B=16, (2) n_A=36, n_B=30, and (3) n_A=100, n_B=120, were used in the simulations for the first model.

For the second model, we considered two discrete covariates one with two levels and the other with three levels. The general model (1) in this case can be written as $logit\left(\pi_{ij}\right) = \beta_0 + \beta_i + \beta_j$ and examples include $logit(\pi_{11}) = \beta_0 + \beta_1 + \beta_2$, $logit(\pi_{12}) = \beta_0 + \beta_1 + \beta_3$, $logit(\pi_{21}) = \beta_0 + \beta_2$, and $logit(\pi_{23}) = \beta_0$. Note that this model regards last level of both covariates as the reference level. In order to generate data satisfying this model, by the

proposition above, we generated $Y_{im;j}^{A} = -\log\left(u_{im;j}^{A}\right) + logit(\pi_{im})$ and $Y_{im;l}^{B} = -\log\left(u_{im;l}^{B}\right)$, where $i = 1,2$; $m = 1,2,3$; $j = 1,\ldots,n^{A}$; $l = 1,\ldots,n^{B}$, and $\{u_{im;j}^{A}, u_{im;l}^{B} : i = 1,2; m = 1,2,3;\ j = 1,\ldots,n^{A};\ l = 1,\ldots,n^{B}\}$ was a set of mutually independent random variables with exponential(1) distribution. Two sets of parameter values: (1) $\beta_0 = 0.15$, $\beta_1 = 0.70$, $\beta_2 = 0.50$, $\beta_3 = 1$, (2) $\beta_0 = 0.10$, $\beta_1 = 0.40$, $\beta_2 = 0.80$, $\beta_3 = 1.5$, and three sets of sample sizes: (1) n_A=25, n_B=30, (2) n_A=50, n_B=60, and (3) n_A=100, n_B=120, were used in the simulations for the second model.

Tables 1 and 2 present the true parameter values, parameter estimates (calculated as the average of the 10,000 simulations), biases, and coverage probabilities of the 95% CIs for the two models, respectively. For both models with various sample sizes and different sets of true parameter values, our estimation procedure produced estimates with small biases. In addition, the 95% CIs have coverage probabilities close to the nominal level. As expected, the biases decrease and the coverage probabilities get closer to the nominal level as the sample sizes increase. In addition, as sample sizes increase, the distribution of parameter estimates gets closer to normal distribution.

Table 1: Parameter estimates, biases, and 95% CIs for the model with one discrete covariate

n	Parameter	True Value (β_i)	Estimate $(\bar{\beta}_i)$	Bias $(\bar{\beta}_i - \beta_i)$	Coverage 95% CI
n_A=14,	β_0	0.15	0.1575	0.0075	0.9640
n_B=16	β_2	0.5	0.5351	0.0351	0.9601
	β_3	1.0	1.0559	0.0559	0.9582
n_A=36,	β_0	0.15	0.1563	0.0063	0.9571
n_B=30	β_2	0.5	0.5079	0.0079	0.9543
	β_3	1.0	1.0249	0.0249	0.9543
n_A=100,	β_0	0.15	0.1517	0.0017	0.9517
n_B=120	β_2	0.5	0.5036	0.0036	0.9494
	β_3	1.0	1.0099	0.0099	0.9539
n_A=14,	β_0	0.1	0.1043	0.0043	0.9613
n_B=16	β_2	0.3	0.3183	0.0183	0.9565
	β_3	1.2	1.2833	0.0833	0.9570
n_A=36,	β_0	0.1	0.0994	-0.0006	0.9549
n_B=30	β_2	0.3	0.3106	0.0106	0.9544
	β_3	1.2	1.2370	0.0370	0.9541
n_A=100,	β_0	0.1	0.1002	0.0002	0.9501
n_B=120	β_2	0.3	0.3029	0.0029	0.9505
	β_3	1.2	1.2112	0.0112	0.9492

Table 2: Parameter estimates, biases, and 95% CIs for the model with two discrete covariates.

n	Covariate	Parameter	True Value (β_i)	Estimate $(\bar{\beta}_i)$	Bias $(\bar{\beta}_i - \beta_i)$	Coverage 95% CI
n_A=25,		β_0	0.15	0.1540	0.0040	0.9597
n_B=30	X_1	β_1	0.7	0.7076	0.0076	0.9594
	X_2	β_2	0.5	0.5021	0.0021	0.9549
		β_3	1.0	1.0166	0.0166	0.9554
n_A=50,		β_0	0.15	0.1534	0.0034	0.9566
n_B=60	X_1	β_1	0.7	0.7071	0.0071	0.9524
	X_2	β_2	0.5	0.4977	-0.0023	0.9534
		β_3	1.0	1.0041	0.0041	0.9541
n_A=100,		β_0	0.15	0.1505	0.0005	0.9510
n_B=120	X_1	β_1	0.7	0.6994	-0.0006	0.9497
	X_2	β_2	0.5	0.5014	0.0014	0.9515
		β_3	1.0	1.0028	0.0028	0.9532
n_A=25,		β_0	0.1	0.1032	0.0032	0.9595
n_B=30	X_1	β_1	0.4	0.4036	0.0036	0.9577
	X_2	β_2	0.8	0.8087	0.0087	0.9575
		β_3	1.5	1.5242	0.0242	0.9560

n_A=50,		β_0	0.1	0.1008	0.0008	0.9548
n_B=60	X_1	β_1	0.4	0.4050	0.0049	0.9524
	X_2	β_2	0.8	0.8081	0.0081	0.9557
		β_3	1.5	1.5158	0.0158	0.9537
n_A=100,		β_0	0.1	0.0996	-0.0005	0.9514
n_B=120	X_1	β_1	0.4	0.4013	0.0013	0.9502
	X_2	β_2	0.8	0.8030	0.0030	0.9523
		β_3	1.5	1.5089	0.0089	0.9535

4. Example

To illustrate how to apply the proposed method, we obtained data from a randomized and controlled clinical trial, which was designed to increase knowledge and awareness to prevent Fetal Alcohol Spectrum Disorders (FASDs) in children through the development of printed materials targeting women of childbearing age in Russia. The study objective was to evaluate effects of FASDs education brochures with different types of information and visual images on knowledge, attitudes, and alcohol consumption among women of childbearing age. The study was conducted in two regions in Russia including St. Petersburg (SPB) and the Nizhny Novgorod Region (NNR) in 2007-2008.

A total of 422 women were recruited from 20 women's clinics in St. Petersburg and the Nizhny Novgorod Region in Russia and were randomly assigned to one of three groups (defined by the GROUP variable): (1) a printed FASD prevention brochure with positive images and information stated in a positive way, positive group (PG) (n=141), (2) a FASD education brochure with negative messages and vivid images, negative group (NG) (n=141), and (3) a general health material, control group (CG) (n=140). Each study participant received one of three printed education brochures in accordance with her group assignment. For the purpose of the analysis in this article, only women in the PG and CG were included. A structured interview to assess women's alcohol consumption, knowledge about prenatal effects of alcohol, FASDs, and attitudes to drinking during pregnancy was designed for the study at baseline. At a one-month follow-up, women in all three groups completed a face-to-face structured interview of self-reported alcohol consumption, knowledge about prenatal effects of

alcohol and FASD, and attitudes to drinking during pregnancy. Data were obtained from the study principal investigators. The response variable was the change in the number of drinks per day (CHANGE_DRINK=number of drinks after-number of drinks before) on average in the last 30 days from one-month follow-up to baseline. Two covariates considered for the proposed method were "In the last 30 days, have you smoked cigarettes?" (SMOKE) and "In the last 30 days, did you take any other vitamins?" (OVITAMIN). Both covariates had "Yes" or "No" as the two levels. The question of interest here was to assess the joint predictive effects of SMOKE and OVITAMIN on whether the participants reduced the number of drinks per day from baseline to one month follow-up period. A total of 210 women with no missing data on any of the CHANGE_DRINK, SMOKE, GROUP, and OVITAMIN were included in the analysis.

The response variable CHANGE_DRINK was heavily skewed and not normally distributed in each group (Shapiro-Wilk $p<0.001$). Therefore, we decided to use the AUC regression model to analyze the data. In the AUC regression model we define π=P(Y_CG>Y_PG). Note that a π of greater than .5 means that women in the PG had a greater reduction of alcohol drinks than those in the CG. For statistical results, all p-values < .05 were considered statistically significant and 95% CIs were presented.

We first fit an AUC regression model including both main effects of the covariates. Note that the main effects of the covariates in fact represented their interactions with the GROUP variable, which is different than the linear or generalized linear model frame. The reason is that the GROUP variable is involved in defining the AUC. Table 3 presents the parameter estimates, SEs, p-values, and 95% CIs. Because parameter β_2 was not significantly different from 0, we dropped OVITAMIN and fit another model including only the SMOKE main effect.Table 4 shows a significant interaction between SMOKE and GROUP because the SMOKE was statistically significant (95% CI: (0.056, 1.478)). Therefore, the final model was logit(π(SMOKE))=$\hat{\beta}\_0+\hat{\beta}\_1$ I(SMOKE=Yes). Because the interaction between SMOKE and GROUP was significant, we need to use AUC as a measure of the GROUP effect on CHANGE_DRINK for smokers and non-smokers separately. Specifically, the AUCs were 0.537 (insignificant) and 0.713 (significant) for non-smokers and smokers, respectively. This implies that the effect of positive and control brochures were similar for nonsmokers; however, for smokers, the probability that the positive brochure had a better effect than the control brochure in terms of alcohol reduction is 71.30%, indicating the positive brochure is a better option than the control brochure.

Table 3: Parameter estimates and 95% confidence interval obtained using the proposed method for the FAS clinical trial example data* (n=210) with SMOKE and OVITAMIN as main effects.

Parameter	Level	Estimate	SE	95% CI	*p*
$\hat{\beta}_0$	Intercept	0.103	0.197	(-0.284, 0.490)	0.600
$\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$	**SMOKE = Yes**	**0.743**	**0.369**	**(0.021, 1.466)**	**0.044**
$\hat{\beta}_2$	OVITAMIN = Yes	0.219	0.338	(-0.444, 0.881)	0.517

***Development of Education Materials for Prevention of FAS in Russia was supported by the Centers for Disease Control and Prevention (CDC), National Center on Birth Defects and Developmental Disabilities (NCBDDD) through a cooperative agreement with the Association of University Centers on Disabilities (AUCD), Grants Number AUCD RTOI 2005-999-01 and RTOI 2005-999-01 to Barbara Bonner, PhD and Tatiana Balachova, PhD at OUHSC.**

Table 4: Parameter estimates and 95% confidence interval obtained using the proposed method for the FAS clinical trial example data (n=210) with SMOKE as main effect.

Parameter	Level	Estimate	SE	95% CI	$\hat{\pi}$* (95 % CI)	*p*
$\hat{\beta}_0$	Intercept	0.143	0.168	(-0.185, 0.471)	$\hat{\pi}_{NS} = 0.537$ (0.454, 0.616)	0.393
$\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$	**SMOKE = Yes**	**0.767**	**0.363**	**(0. 056, 1.478)**	$\widehat{\boldsymbol{\pi}}_{\boldsymbol{S}} = \mathbf{0.713}$ **(0.514, 0.814)**	**0.035**

***NS = Non-smokers; S = Smokers**

5. Conclusion

In this article, we conducted research on AUC regression on a non-normally distributed response variable while adjusting for discrete covariates. We derived explicit and non-iterative formulas for point and interval parameter estimation. Simulation studies showed that our method produced parameter estimates with small biases and confidence intervals with close to

nominal coverage. We illustrated our method using a pediatric data example and demonstrated that the results from our modeling can be easily interpreted.

In classic linear models with discrete covariates comparing two groups, the differences in mean response are first computed at each combination (which we define as cell) of the discrete covariates. Then, the overall comparison between the two groups is calculated as weighted averages of these differences computed at the cellular level. Our AUC regression works in a similar way. Therefore, in order for a cell to be usable, at least two observations from each group are required. Moreover, users of our method need to be careful when the total sample size is small and there are too many covariates or some of the covariates have too many levels. It is prudent for the users to ensure that the model they fit is identifiable. Additionally, similar to linear models, our AUG regression model may not be appropriate for sparse data.

Our method has a broad applicability even though it does not handle continuous covariates. This is because only discrete covariates are of interest in many applications. Even when there are continuous covariates, often times these covariates can be categorized. In fact, practitioners sometimes prefer to use these transformed discrete covariates due to the ease of interpretation. In applications where adjusting for continuous covariates is necessary, we are currently working on new methodologies to cope with this type of data.

Acknowledgment

Research reported in this publication was partially supported by the Centers for Disease Control and Prevention (CDC), National Center on Birth Defects and Developmental Disabilities (NCBDDD) through a cooperative agreement with the Association of University Centers on Disabilities (AUCD), Grants Number AUCD RTOI 2005-999-01 and RTOI 2005-999-01 and the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the Fogarty International Center (Brain Disorders in the Developing World: Research Across the Lifespan) of the National Institutes of Health (NIH) under Award R01AA016234 to Tatiana Balachova, PhD and Barbara Bonner, PhD at OUHSC. The content is solely the responsibility of the authors and does not necessarily represent the official views of the CDC, AUCD, and NIH.

References

[1] Branscum, A. J., Johnson, W. O., Hanson, T. E., & Baron, A. T. (2015). Flexible regression models for ROC and risk analysis, with or without a gold standard. *Statistics in Medicine, 34*(30), 3997-4015. doi:10.1002/sim.6610

[2] Brumback, L. C., Pepe, M. S., & Alonzo, T. A. (2006). Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in Medicine, 25*(4), 575-590. doi:10.1002/sim.2345

[3] Branscum, A. J., Johnson, W. O., Hanson, T. E., & Baron, A. T. (2015). Flexible regression models for ROC and risk analysis, with or without a gold standard. *Statistics in Medicine, 34*(30), 3997-4015. doi:10.1002/sim.6610

[4] Brumback, L. C., Pepe, M. S., & Alonzo, T. A. (2006). Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in Medicine, 25*(4), 575-590. doi:10.1002/sim.2345

[5] Buros, A., Tubbs, J. D., & Van Zyl, J. S. (2016). Application of AUC Regression for the Jonckheere Trend Test. *Statistics in Biopharmaceutical Research*, 0-0. doi:10.1080/19466315.2016.1265581

[6] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics, 44*(3), 837-845. doi:10.2307/2531595

[7] Dodd, L. E., & Pepe, M. S. (2003). Semiparametric regression for the area under the receiver operating characteristic curve. *Journal of the American Statistical Association, 98*(462), 409-417.

[8] Gao, F., Xiong, C., Yan, Y., Yu, K., & Zhang, Z. (2008). Estimating optimum linear combination of multiple correlated diagnostic tests at a fixed specificity with Receiver Operating Characteristic Curves. *Journal of Data Science, 6*(1), 105-123.

[9] Inacio de Carvalho, V., Jara, A., E. Hanson, T., & de Carvalho, M. (2013). Bayesian Nonparametric ROC Regression Modeling. 623-646. doi:10.1214/13-BA825

[10] Liu, H., Li, G., Cumberland, W. G., & Wu, T. (2005). Testing statistical significance of the area under a receiving operating characteristics curve for repeated measures design with bootstrapping. *Journal of Data Science, 3*(3), 257-278.

[11] Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. 50-60. doi:10.1214/aoms/1177730491

[12] Rajan, K. B., & Zhou, X. H. (2012). Semi-parametric area under the curve regression method for diagnostic studies with ordinal data. *Biom J, 54*(1), 143-156. doi:10.1002/bimj.201100025

[13] Rodríguez-Álvarez, M., Roca-Pardiñas, J., & Cadarso-Suárez, C. (2011). ROC curve and covariates: extending induced methodology to the non-parametric framework. *Statistics and Computing, 21*(4), 483-499. doi:10.1007/s11222-010-9184-1

[14] Rodríguez, A., & Martínez, J. C. (2014). Bayesian semiparametric estimation of covariate-dependent ROC curves. *Biostatistics (Oxford, England), 15*(2), 353-369. doi:10.1093/biostatistics/kxt044

[15] Van Elteren, P. H. (1960). On the combination of independent two sample tests of Wilcoxon. *Bulletin of the Institute of International Statistics, 37*, 351-361.

[16] Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin, 1*(6), 80-83. doi:10.2307/3001968

[17] Zhang, L., Zhao, Y. D., & Tubbs, J. D. (2011). Inference for Semiparametric AUC Regression Models with Discrete Covariates. *Journal of Data Science, 9*, 625-637.