ABSTRACT

Beta Regression for Modeling a Covariate-Adjusted ROC

Sarah Stanley, Ph.D.

Mentor: Jack D. Tubbs, Ph.D.

The receiver operating characteristic (ROC) curve is a well-accepted measure of accuracy for diagnostic tests. In many applications, test performance is affected by covariates. As a result, several regression methodologies have been developed to model the ROC as a function of covariate effects within the generalized linear model (GLM) framework. We present an alternative to two existing parametric and semi-parametric methods for estimating a covariate-adjusted ROC. These methods utilize GLMs for binary data with an expected value equal to the probability that the test result for a diseased subject exceeds that of a non-diseased subject with the same covariate values. This probability is referred to as the placement value. Given that the ROC is the cumulative distribution of the placement values, we propose a new method that directly models the placement values through beta regression. We compare the beta regression method to the existing parametric and semiparametric approaches with simulation and a clinical study. Bayesian extensions for the parametric and the beta methods are developed and the performance of these extensions is evaluated through simulation. We apply the proposed beta regression approach and its Bayesian extension to a simple network meta-analysis problem using a Bayesian indicator model selection method.

Beta Regression for Modeling a Covariate-Adjusted ROC

by

Sarah Stanley, B.S., M.S.

A Dissertation

Approved by the Department of Statistical Science

_____

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

_____
Jack D. Tubbs, Ph.D., Chairperson

_____
Corey P. Carbonara, Ph.D.

_____
John W. Seaman, Jr., Ph.D.

_____
James D. Stamey, Ph.D.

_____
Dean M. Young, Ph.D.

Accepted by the Graduate School
May 2018

_____
J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Thank you to my advisor Dr. Jack Tubbs for his consistent guidance and encouragement during the research process. Looking back over the last few years, I could not have asked for a better mentor, and I am grateful for the time and energy he has devoted to this venture. Thank you also to Dr. James Stamey who played an integral role in the latter chapters of this dissertation. Additional thanks go to my committee as a whole for their time investment and feedback. Finally, thank you to my parents who have been faithful encouragers during this journey. Without their love and prayers, I would not have made it to this point.

CHAPTER ONE

Introduction

A well known problem in the testing literature is to determine and control how covariates affect a test's ability to distinguish between two populations. Two widely used measures of accuracy for diagnostic tests are the receiver operating characteristic (ROC) curve and the area under the ROC (AUC). In this dissertation, we propose a new method for estimating a covariate-adjusted ROC and extend application of the new method to a simple network meta-analysis. We begin with defining notation and presenting a motivating example.

1.1 Notation

1.1.1 Receiver Operating Characteristics Curve (ROC)

We briefly introduce the ROC as a measure of test accuracy as well as the notation used for a covariate adjusted ROC. Suppose we have two populations, one non-diseased ($\bar{D}$) and one diseased (D). Let $Y_{\bar{D}}$ denote the test result for an observation from the non-diseased (reference) population and let Y_D denote the test result for an observation from the diseased (comparator) population. Suppose that we classify a subject as being from the diseased population if $Y \geq c$. Then the test's true positive rate is TPR$(c) = \Pr[Y \geq c|D]$. Similarly, the test's false positive rate is FPR$(c) = \Pr[Y \geq c|\bar{D}]$. The ROC curve, defined as the set of all TPR-FPR pairs, quantifies the separation between the diseased and non-diseased populations. The ROC has many forms in the literature. In this dissertation, we restrict our attention to the survival curve and the placement values given by,

$$\begin{aligned} \text{ROC}(t) &= S_D(S_{\bar{D}}^{-1}(t)) \\ &= P[PV_D \leq t], \end{aligned}$$

for $t \in (0, 1)$, where $S_D, S_{\bar{D}}$ are survival functions for the diseased and non-diseased populations respectively, and PV_D represents the placement values for the diseased subjects. The placement value is the probability that a diseased test result exceeds a non-diseased test result given the same covariate value. As illustrated in Figure 1.1.1, one can think of the placement value for a diseased observation as found by mapping the diseased response value on to the reference distribution and calculating the area to the right. Populations which exhibit a high degree of separation will thus yield placement values close to zero.

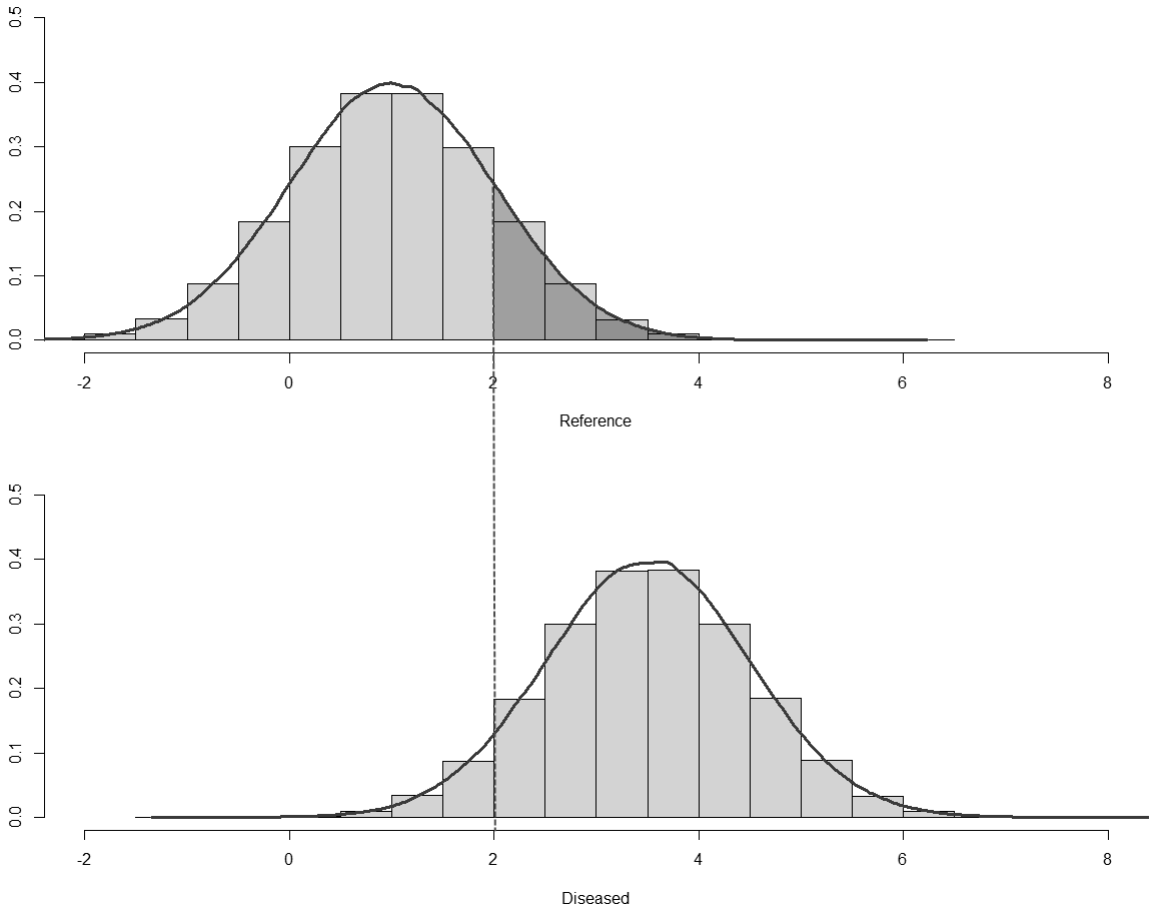

Figure 1.1.1: Illustration of placement value calculation

The area under the curve (AUC) is a common summary measure of the ROC given by $P(Y_D > Y_{\bar{D}})$. The AUC is the probability that a randomly selected subject is classified into the correct population. As illustrated in Figure 1.1.2, populations with a high degree

of overlap will yield a nearly diagonal ROC with an AUC close to 0.5. Those with a high degree of separation will yield an AUC close to 1.

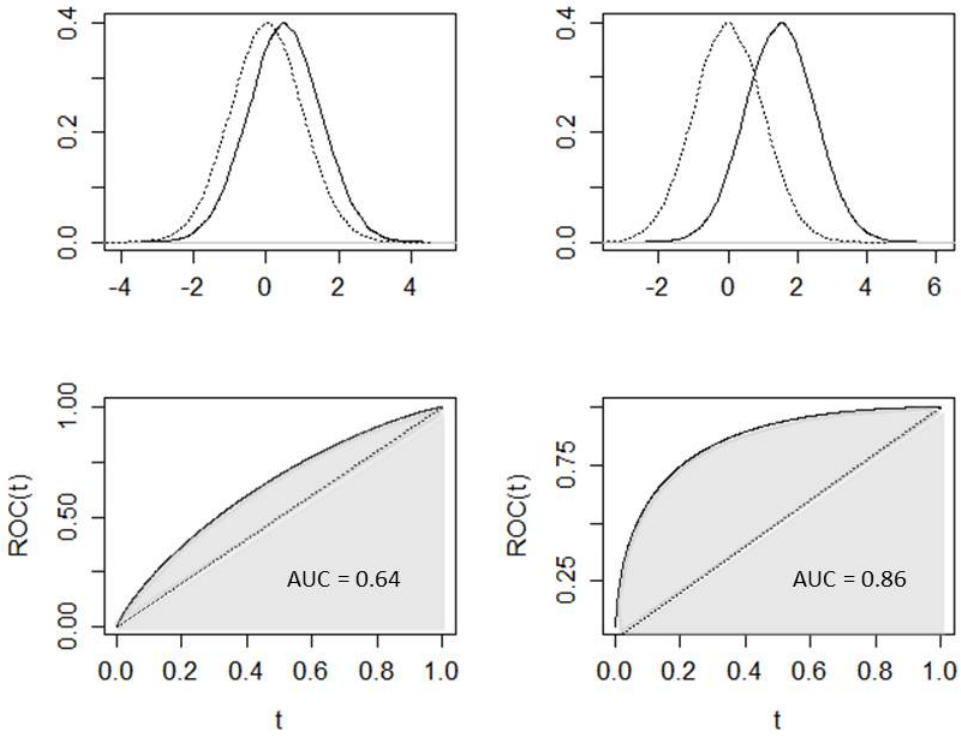

Figure 1.1.2: Illustration of the ROC and the AUC for different population separations

1.1.2 Covariate-Adjusted ROC Curve

Let X denote covariates common to both populations, such as age and BMI. Let X_D denote covariates that are specific to the diseased group, such as disease duration, disease severity, or previous treatment. The covariate-adjusted ROC can then be written as

$$\text{ROC}_{X,X_D}(t) = S_{D,X,X_D}(S^{-1}_{\bar{D},X}(t)), \text{ for } t \in (0,1), \tag{1.1}$$

where $S_{D,X,X_D}(c) = P(Y_D \geq c|X, X_D)$ and $S_{\bar{D},X}(c) = P(Y_{\bar{D}} \geq c|X)$ are survival functions at threshold c. Thus, the $\text{ROC}_{X,X_D}(t)$ is the probability that a test result, Y_D, for a diseased subject is greater than or equal to the t^{th} quantile for the covariate adjusted test results of non-diseased subjects.

1.2 Motivating Example

Suppose we wish to compare a competing treatment to a common reference treatment as in the Protocol I study (Elman et al., 2015) in the Diabetic Retinopathy Clinical Research Network (NCT 00444600). In the study, each patient had been previously diagnosed with either type 1 or type 2 diabetes as well as diabetic macular edema (DME) affecting the center of the macula. DME, the accumulation of fluid in the retina, results from diabetic retinopathy, a condition in which high levels of blood sugar weaken the blood vessels in the eye, causing a build up of pressure and fluid leakage over time. The resulting excess fluid in the retina causes swelling which impedes the function of the macula, the part of the eye which controls visual acuity and sensitivity to light.

The patients were randomized to one of four treatment groups. For the purpose of our example, we will consider two groups: A – a sham injection with laser treatment and B – a 0.5 mg injection of intravitreal ranibizumab along with laser treatment given three to ten days after injection. The primary outcome was visual acuity at one year adjusted for baseline acuity. Visual acuity was measured with Optical Coherence Tomography (OCT) which detects changes in retinal thickness, and the ETDRS test which records the number of letters that a patient can correctly identify. In this context, a favorable result is a decrease in retinal thickness which corresponds to vision improvement. An effective treatment would then result in a lower OCT measurement as compared to a baseline OCT measurement.

We define treatment A (laser therapy alone) as the reference population and treatment B as the comparator population. We define the response of interest to be the amount of decrease in retinal thickness from baseline at one year. If treatment B is effective, the resulting ROC curve should be different from the diagonal line, reflecting a considerable separation in the population densities. We are interested in the effect of covariates on the separation between the populations. Is the separation between the responses affected by a patient's age or the length of time since his diabetes diagnosis. To answer these questions, we apply ROC-regression methodology. We introduce the background of ROC-regression

in the literature review as well as a summary of the work done in this area to date. We then elaborate on three ROC-regression methods and compare performance via simulation in Chapter Two.

1.3 Literature Review

We begin with an overview of existing methods in the literature that provide covariate adjustments for the ROC and the AUC. Dodd and Pepe (2003) proposed a semi-parametric AUC regression method to model a covariate-adjusted AUC and suggested the use of bootstrapping methodology to estimate the standard errors of the regression coefficients. Zhang et al. (2011) used the relationship between the AUC and the Mann-Whitney statistic, a non-parametric unbiased estimate of the AUC to suggest an alternative to Dodd and Pepe's bootstrapping estimation. The alternative was based on the work of DeLong et al. (1988) and used the delta method to estimate the variance of the Mann-Whitney statistic as well as the variance of the parameters. Buros (2015), used the AUC regression model from Dodd and Pepe (2003) with the analytic solution for the standard errors from Zhang et al. (2011) to develop an adjusted Jonckheere Terpstra (JT) test for discrete covariates. Buros et al. (2017) utilized the adjusted JT test and AUC regression to develop a nonparametric multiple comparison procedure involving a monotone alternative hypothesis as often occurs in dose response study. Van Zyl (2017) extended the multiple comparison of Buros et al. (2017) to a zero-dose control model.

Along with AUC regression methodology, procedures for modeling a covariate-adjusted ROC also exist in the testing literature. Pepe (1998) provides a review of three major approaches to ROC regression that account for covariate effects. The first approach was developed by Tosteson and Begg (1988) who used ordinal regression to model the test outcome and then examined covariate effects on the ROC. Beam (1995) and Gatsonis (1995) extended the test outcome regression to random effects models for ordinal test results. The second approach mentioned by Pepe (1998) involves regression models developed by

Thompson and Zucchini (1989) for the area under the ROC curve, a precursor to the semi-parametric regression work of Dodd and Pepe (2003). The third approach proposed by Pepe (2000) directly models the ROC curve using parametric distribution-free methods.

In this dissertation, we direct our attention to the approach that directly models the ROC as opposed to modeling the underlying distributions of test responses for the diseased and non-diseased populations. Advantages to this approach include the accommodation of multiple test types, use of continuous covariates, and the ability to restrict the model to the portion of the ROC that is of interest. When originally proposed, Pepe's direct modeling approach was difficult to implement due to the requirement of special programming. Simplifications have since been made as Pepe (2000) developed a GLM framework for the ROC (ROC-GLM) which eased the previous computation required for parameter estimation. In particular, Pepe (2000) proposed a generalized linear model framework for the ROC given by

$$ROC_X(t) = g(h_0(t) + X'\beta), \tag{1.2}$$

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

Alonzo and Pepe (2002) expanded the utility of the ROC-GLM in (1.2) by specifying a parametric form for $h_0(\cdot)$ and using a binary indicator as an outcome variable. Thus, rather than perform pairwise comparisons between each observation from the diseased and non-diseased samples (as in the Mann-Whitney statistic), Alonzo and Pepe (2002) compared each diseased observation to a specified set of covariate-adjusted quantiles for the non-diseased population. The resultant binary values could then be modeled using a logistic regression approach.

Pepe and Cai (2004) extended the parametric ROC-GLM by allowing a non-parametric form for $h_0(\cdot)$. This semi-parametric approach hinged upon the relationship between the ROC and the placement value. The ROC is the cumulative distribution function of the

placement values PV_D as seen in the following covariate-adjusted notation,

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

Cai (2004) further developed the semiparametric approach by demonstrating that (1.2) is equivalent to $h_0(PV_D) = -X'\beta + \epsilon$, where $h_0(\cdot)$ is unknown and PV_D is the set of placement values for the diseased observations. Implementation of the semiparametric model is dependent upon pairwise comparisons of the placement values to estimate the covariate effects β that are then included as an offset in the estimation of $h_0(\cdot)$ (Rodriguez-Alvarez et al., 2011).

1.4 Plan of the Dissertation

The use of placement values by Cai (2004) motivates the development of an alternative approach to modeling the covariate-adjusted ROC. Given that the ROC is the cumulative distribution of the placement values for the diseased observations, we propose a new method in Chapter Two that directly models the placement values using beta regression. Chapter Two also details the parametric and semiparametric approaches, and we show that the new beta approach is not only easy to implement, but it also removes the need for pairwise comparisons, eliminating the dependency among the binary response variable induced by the parametric and semiparametric methods. We compare the proposed method to the existing models with simulation and an application to the DME study from our motivating example. In Chapter Three, we extend the parametric and beta regression approaches to the Bayesian paradigm using hierarchical modeling. The performance of the Bayesian extensions is compared through simulation study and both methods are applied to the DME study. We introduce an application of the beta approach to a simple network meta-analysis problem in Chapter Four through the use of a Bayesian variable selection method. The performance of the beta approach in conjunction with variable selection is evaluated through

simulation study. Chapter Five summarizes a project done in conjunction with Eli Lilly and is largely unrelated to the work presented in the preceding chapters. We provide an introduction to power priors and the use of the deviance information criterion (DIC) as a guide for choosing the value of the power prior parameter. To evaluate the performance of the DIC for parameter guidance in a generalized linear model context, we perform a simulation study for normal linear regression and logistic regression models.

CHAPTER TWO

ROC Regression Methodology

In this chapter, we introduce the parametric and semiparametric ROC regression methods and propose an alternative model based on beta regression. The ROC models are compared through simulation and applied to a clinical study.

2.1 Parametric Approach

Alonzo and Pepe (2002) proposed a parametric extension of (1.2) as,

$$\text{ROC}_{X,X_D}(t) = g(\gamma_1 h_1(t) + \gamma_2 h_2(t) + \beta X + \beta_D X_D), \tag{2.1}$$

with $\gamma_1, \gamma_2, \beta,$ and β_D as model parameters, $h_1(t) = 1, h_2(t) = \Phi^{-1}(t)$, and $g(\cdot) = \Phi(\cdot)$ where $\Phi(\cdot)$ is the cdf of the standard normal. Alonzo and Pepe's approach is known as a parametric distribution free method because a parametric model is specified for the ROC, but no assumptions are made about the distributions for Y_D and $Y_{\bar{D}}$ (Alonzo and Pepe, 2002).

The parametric model, (2.1), follows from Pepe (2000) where the ROC is written as the expectation of the binary indicator $U_{ij} = I[Y_{D_i} \geq Y_{\bar{D}_j}]$ for all pairs of observations $\{(Y_{D_i}, Y_{\bar{D}_j}), i = 1, \ldots, n_D; j = 1, \ldots, n_{\bar{D}}\}$, with n_D and $n_{\bar{D}}$ denoting the number of observations from the diseased and non-diseased populations, respectively, and I denoting the indicator function.

Alonzo and Pepe (2002) proposed a modification by replacing $Y_{\bar{D}_j}$ with $S^{-1}_{\bar{D},X_i}(t)$, for $t \in T = \{n_T \text{ chosen values of FPRs } \in (0,1)\}$. In this case, the binary indicator becomes $U_{it} = I[Y_{D_i} \geq S^{-1}_{\bar{D},X_i}(t)]$. Note, the expected value of U_{it} satisfies

$$E(U_{it}) = E(I[Y_{D_i} \geq S^{-1}_{\bar{D},X_i}(t)]) = \Pr[S_{\bar{D},X_i}(Y_{D_i}) \leq t] = \Pr[PV_D \leq t],$$

where PV_D is the placement value for the observation Y_{D_i} given the covariate vector X.

An algorithm for (2.1) can be written as

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

(2) Estimate the covariate specific survival function $S_{\bar{D},X_j}$ for the reference population at each $t \in T, j = 1, \ldots, n_{\bar{D}}$ using quantile regression.

(3) For each diseased observation Y_{D_i}, calculate the placement values

$$PV_{D_i} = \hat{S}_{\bar{D},X_i}(Y_{D_i}), i = 1, \ldots, n_D.$$

(4) Calculate the binary placement value indicator $\hat{U}_{it} = I[PV_{D_i} \leq t], t \in T$.

(5) Fit the model $E[\hat{U}_{it}] = g^{-1}[\sum_{k=1}^{K} \gamma_k h_k(t) + X'\beta]$.

In step (1), we specify a set of n_T false positive rates (FPRs), where in practice the FPRs are equally spaced. In step (2), we estimate the covariate-adjusted reference survival curve using quantile regression on the set of FPRs. The quantile regression yields n_T covariate adjusted estimates of the reference survival curve for each $Y_{\bar{D}_j}$. In step (3), we calculate the placement values for each diseased observation Y_{D_i}. The placement values are calculated by evaluating the covariate-adjusted reference survival curve at each Y_{D_i}, resulting in n_D probabilities. We next create a binary indicator $\hat{U}_{it}$ in step (4) by performing n_D to n_T comparisons between the placement values and the set of FPRs. Note that step (4) is similar to the Mann Whitney statistic formed by making n_D to $n_{\bar{D}}$ comparisons from which we can derive the area under the curve (AUC)(Bamber, 1975). In step (5), the covariate adjusted ROC is obtained by modeling the expectation of $\hat{U}_{it}$ using a probit link.

2.2 Semiparametric Approach

Pepe and Cai (2004) extended the parametric approach by proposing a semiparametric method allowing for an arbitrary non-parametric baseline function $h_0(\cdot)$ in (1.2). Their approach required the simultaneous estimation of $h_0(\cdot)$ and β. Cai (2004) introduced a method of estimating parameters for the semiparametric model by demonstrating that (1.2)

is equivalent to $h_0(PV_D) = -X'\beta + \epsilon$, where ϵ is a random variable with known distribution g, $h_0(\cdot)$ is an unspecified increasing function, and PV_D represents placement values for the diseased observations. Cai used pairwise comparison of placement values to estimate β before estimating the baseline function $h_0(\cdot)$. An algorithm for implementing the semiparametric approach is as follows.

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

(2) Estimate the covariate specific survival function $S_{\bar{D},X_j}$ for the reference population at each $t \in T, j = 1, \ldots, n_{\bar{D}}$ using quantile regression.

(3) Calculate the placement values

$$PV_{D_i} = \hat{S}_{\bar{D},X_i}(Y_{D_i}), \quad i = 1, \ldots, n_D.$$

(4) Calculate the binary placement value indicator

$$\hat{U}_{it} = I[PV_{D_i} \leq t], \quad t \in T.$$

(5) For each pair of observations in Y_D, calculate

$$V_{ij} = I[PV_{D_i} \leq PV_{D_j}] \text{ and } x_{ij} = x_{D_i} - x_{D_j}$$

with $i, j = 1, \ldots, n_D, i \neq j$.

(6) Fit the following GLM without an intercept to estimate β

$$g(V) = -X'\beta.$$

(7) Estimate $h_0(\cdot)$ using $g(E[\hat{U}_{it}]) = \text{intercept} + \text{offset}(X'\hat{\beta})$.

Note that steps (1) - (4) are identical to those of the parametric method. The difference between the two approaches appears in step (5), where we create a second binary indicator describing the relationship between each pair of placement values. In this step, we also calculate the pairwise differences for each covariate. We then fit a GLM without an intercept to the binary indicator created in step 5, adjusting for covariates using the pairwise differences. From this model, we obtain an estimate for β. In step (7), we then estimate

$h_0(\cdot)$ by modeling the binary indicator $\hat{U}$ as a function of the intercept and an offset term that accounts for $\hat{\beta}$ (Rodriguez-Alvarez et al., 2011).

2.3 Beta Approach

The parametric and semiparametric approaches to estimating the covariate adjusted ROC given in equation (1.2) were dependent upon a binary random variable defined by the placement values of the diseased response as referenced with the non-diseased population. The use of the binary random variable leads to additional correlation in the model, and the resulting estimates for the standard errors of the regression coefficients are incorrect. To account for the additional correlation, Alonzo and Pepe (2002) proposed a bootstrapping procedure for obtaining the standard errors. In this section, we present an alternative method that models the covariate-adjusted ROC as the cdf of the placement values directly and bypasses the need for a binary random variable. The beta regression model is used in this approach.

A brief introduction to the beta generalized linear model given in Ferrari and Cribari-Neto (2004) is presented here. Suppose that $Z \sim \text{Beta}(a, b)$, in which case,

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

By letting $\mu = \frac{a}{a+b}$ and $\phi = a + b$, we obtain the reparameterized beta distribution with mean and variance

$$E(Z) = \mu, \text{ and } Var(Z) = \tfrac{\mu(1-\mu)}{1+\phi}.$$

Let $z_1, \ldots, z_n$ be independent random variables from a beta density with mean μ_t, t = 1,…,n and scale parameter ϕ. Then the beta regression model can be written as

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

where β is a vector of regression parameters, $x_{t1}, \ldots, x_{tk}$ are observations on k covariates, and g is a monotonic link function. Using the logit link, we have $\mu_t = \dfrac{1}{1 + e^{-x_t'\beta}}$.

Estimates of the original parameters a and b are

$$\hat{a} = \frac{\hat{\phi}}{1 + e^{-x_t'\hat{\beta}}} \text{ and } \hat{b} = \hat{\phi}\left(1 - \frac{1}{1 + e^{-x_t'\hat{\beta}}}\right). \tag{2.2}$$

An algorithm for the proposed method using the beta distribution for the placement values can be written as follows.

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

(2) Estimate the covariate specific survival function $S_{\bar{D},X_j}$ for the reference population at each $t \in T, j = 1, \ldots, n_{\bar{D}}$ using quantile regression.

(3) Calculate the placement values

$$PV_{D_i} = \hat{S}_{\bar{D},X_i}(Y_{D_i}), \quad i = 1, \ldots, n_D.$$

(4) Perform a beta regression on the placement values to obtain estimates of β and ϕ.

(5) Transform to obtain $a = \mu\phi$ and $b = (1 - \mu)\phi$.

(6) Calculate the cdf of the placement values using the Beta(a,b) distribution found above to obtain the ROC and the AUC.

Steps (1) - (3) are identical to the parametric and semiparametric cases. In step (4), we model the placement values directly using beta regression to obtain estimates of β and ϕ. We then apply equation (2.2) to obtain beta parameters a and b and calculate the cdf of the placement values using the resulting Beta(a, b) distribution that yields an estimate for the ROC. The AUC is obtained by integrating the Beta(a, b) cdf, which results in $b/(a + b)$ by Fubini's theorem (Fubini, 1907).

2.4 Simulation Studies

We compare the parametric, semiparametric and beta ROC regression methods through two simulations, one using normally distributed data and the other using data from an extreme value distribution. Rodriguez-Alvarez et al. (2011) provide a comparison of several indirect and direct ROC regression methods including the parametric (Alonzo and Pepe, 2002) and semiparametric (Cai, 2004) for binormal and extreme value data. The data models in this section are similar to those of Rodriguez-Alvarez et al. (2011). For simplicity, we consider one continuous covariate from a uniform distribution. The models and results follow.

2.4.1 Binormal Data

Suppose that $Y_D \sim N(\mu_D, \sigma_D)$ and $Y_{\bar{D}} \sim N(\mu_{\bar{D}}, \sigma_{\bar{D}})$. Then using $\text{ROC}(t) = S_D(S_{\bar{D}}^{-1}(t)),$ for $t \in (0,1)$, we derive the binormal ROC and AUC,

$$ROC(t) = \Phi[a + b\Phi^{-1}(t)], \text{ and } AUC = \Phi\left[\frac{a}{\sqrt{1+b^2}}\right],$$

where $a = (\mu_D - \mu_{\bar{D}})/\sigma_D$ and $b = \sigma_{\bar{D}}/\sigma_D$. The following models were used for the binormal simulation

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

where $X \sim U(0,1)$ and $\epsilon_D, \epsilon_{\bar{D}} \sim N(0, 1.5^2)$. Given the model, the true ROC and AUC at covariate $X = x_0,\ \ t \in (0,1)$ are

$$ROC(t) = \Phi\left[\frac{0.5 + x_0}{1.5} + \Phi^{-1}(t)\right],$$

and

$$AUC(x_0) = \Phi\left[\frac{0.5 + x_0}{\sqrt{4.5}}\right].$$

We generate 1000 data sets of size $n_D, n_{\bar{D}} = 200$ from which we calculate the ROC and AUC for each of the three methods. We also compute the mean squared error (MSE) for the AUC of each method. Boxplots of the MSE values for each method are given in

Figure 2.4.1. The summary statistics for the MSE of the AUC are given in Table 2.1. We note that the mean MSE and standard deviation for the parametric method are smaller than the corresponding results for the beta and semiparametric methods. The beta mean MSE is, however, within one standard deviation of the parametric mean MSE. Plots of the simulated and true ROC curves are included in Figure 2.4.2 for covariate values $x_0 = \{0.2, 0.5, 0.8\}$. The dotted lines represent plus and minus two standard deviations from the simulated mean ROC. The cdf of a Uniform$(0, 1)$ distribution representing the ROC for identical populations is included for reference. Observe that the AUC increases with an increase in the covariate value.

Table 2.1: Summary of MSEs for binormal

Method	1st.Qu.	Median	3rd.Qu.	Mean	St. Dev.
Beta	0.000521	0.001251	0.002544	0.001819	0.001831
Parametric	0.000383	0.000936	0.001996	0.001398	0.001446
Semiparametric	0.000958	0.001912	0.003369	0.002459	0.002088

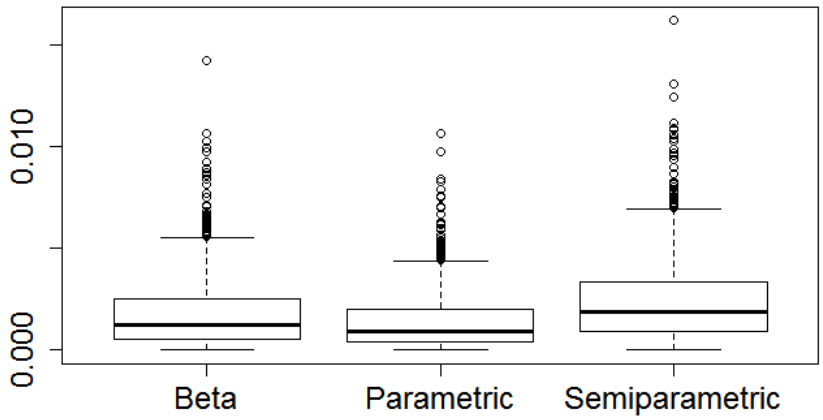

Figure 2.4.1: Boxplots of the estimated MSE for the AUC of each method based on 1000 estimates ($n_D = n_{\bar{D}} = 200$)

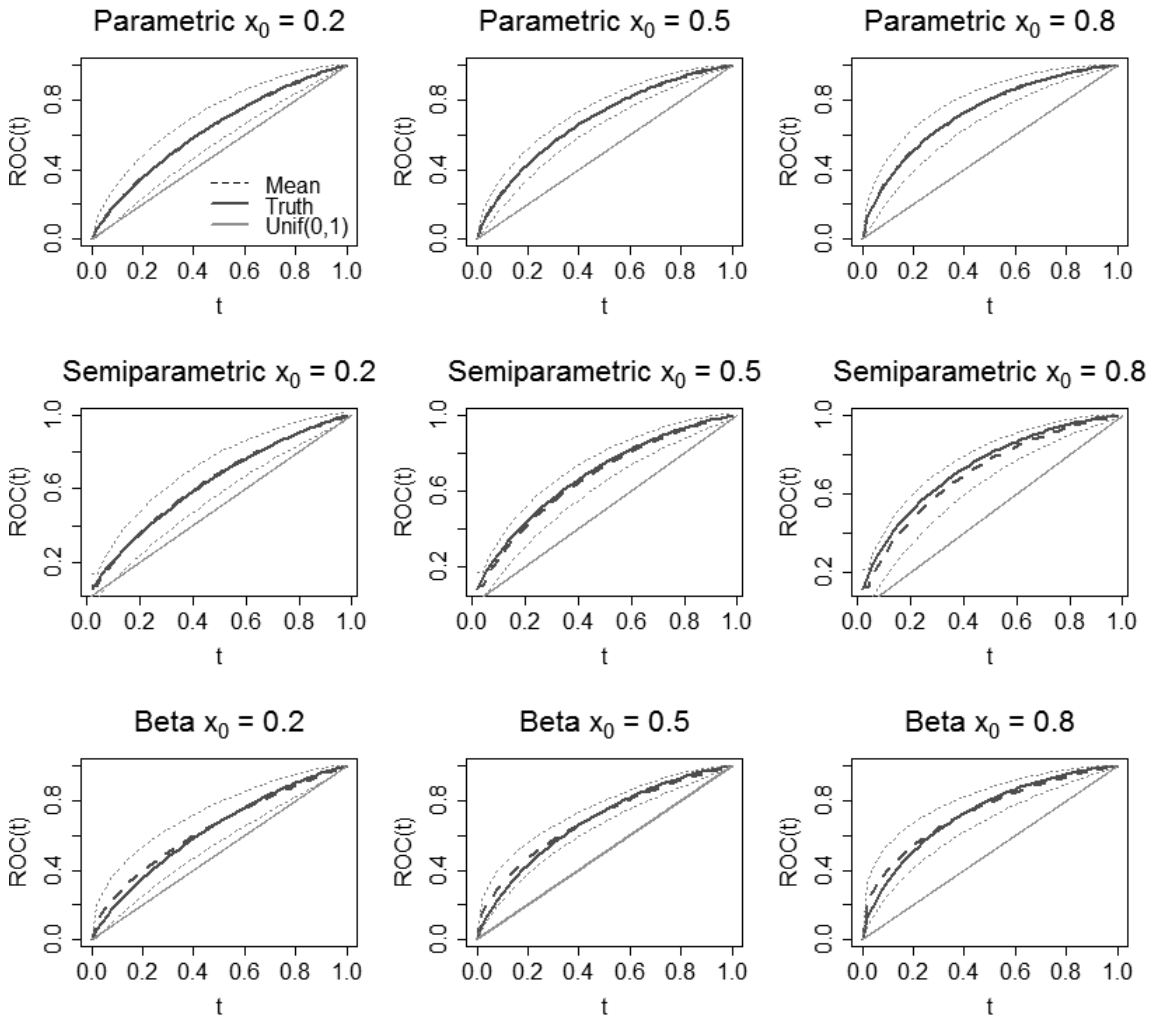

Figure 2.4.2: Comparison of simulated ROC and true ROC for binormal data

2.4.2 Extreme Value Data

The extreme value distribution used in the following models has a cdf of the form $F(x) = \exp\{-\exp[-(x-\mu)/\beta]\}$, where $\mu \in \mathbb{R}, \beta > 0, x \in (-\infty, \infty)$. This extreme value distribution is also known as the Gumbel or double exponential distribution (Balakrishnan and Nevzorov, 2003). In choosing a model for simulation, we note that the extreme value distribution exhibits more sensitivity than the normal distribution to differences in location and scale for the two populations. Highly separated populations will yield an AUC of one

regardless of covariate value. We thus consider scenarios such as the following in which the covariate effect can be assessed.

$$Y_D = 2 + 2.5X + \epsilon_D, \text{ and } Y_{\bar{D}} = 1 + 2X + \epsilon_{\bar{D}},$$

where $X \sim U(0,1)$ and $\epsilon_D, \epsilon_{\bar{D}}$ have an extreme value distribution with $\mu = 0$ and $\beta = 1.5$. The true value of the ROC when $X = x_0$ is

$$\text{ROC}_X(t) = 1 - \exp[-\exp\{\ln[-\ln(1-t)] - \frac{1 + 0.5x_0}{1.5}\}].$$

We approximate the AUC using numerical integration. A plot of the densities for Y_D and $Y_{\bar{D}}$ appears in Figure 2.4.3 as well as the true ROC at covariate values 0, 0.5, and 1.

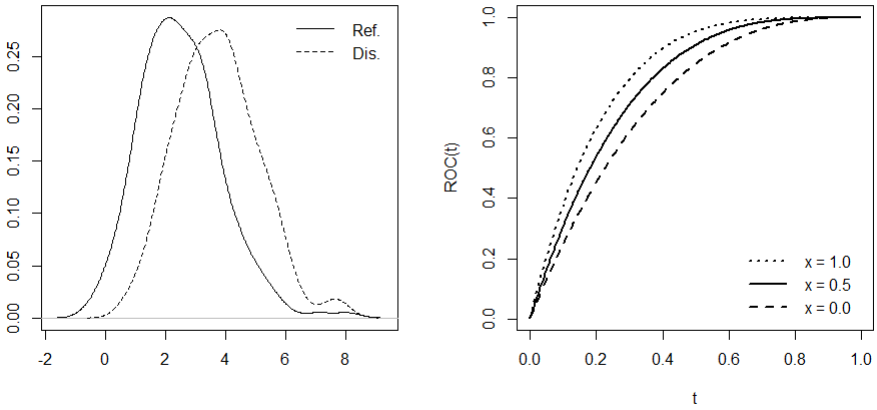

Figure 2.4.3: Density plots and ROCs for extreme value data

As in the binormal simulation, we generate 1000 data sets of size $n_D, n_{\bar{D}} = 200$, calculating the resulting ROC and AUC estimates from each of the parametric, semiparametric, and beta methods and comparing to the truth using the MSE. The summary statistics for the MSE of the AUC are given in Table 2.2. We observe that the beta mean MSE and standard deviation are slightly larger than the corresponding results for the parametric method although the means are within one standard deviation of each other as in the binormal simulation. The semiparametric mean MSE is smaller than that of the beta, but examination of Figure 2.4.4 shows that the beta method provides a better estimation of the

true ROC. Plots of the simulated and true ROC curves are included in Figure 2.4.4 for covariate values $x_0 = \{0.2, 0.5, 0.8\}$. The dotted lines represent plus and minus two standard deviations from the simulated mean ROC.

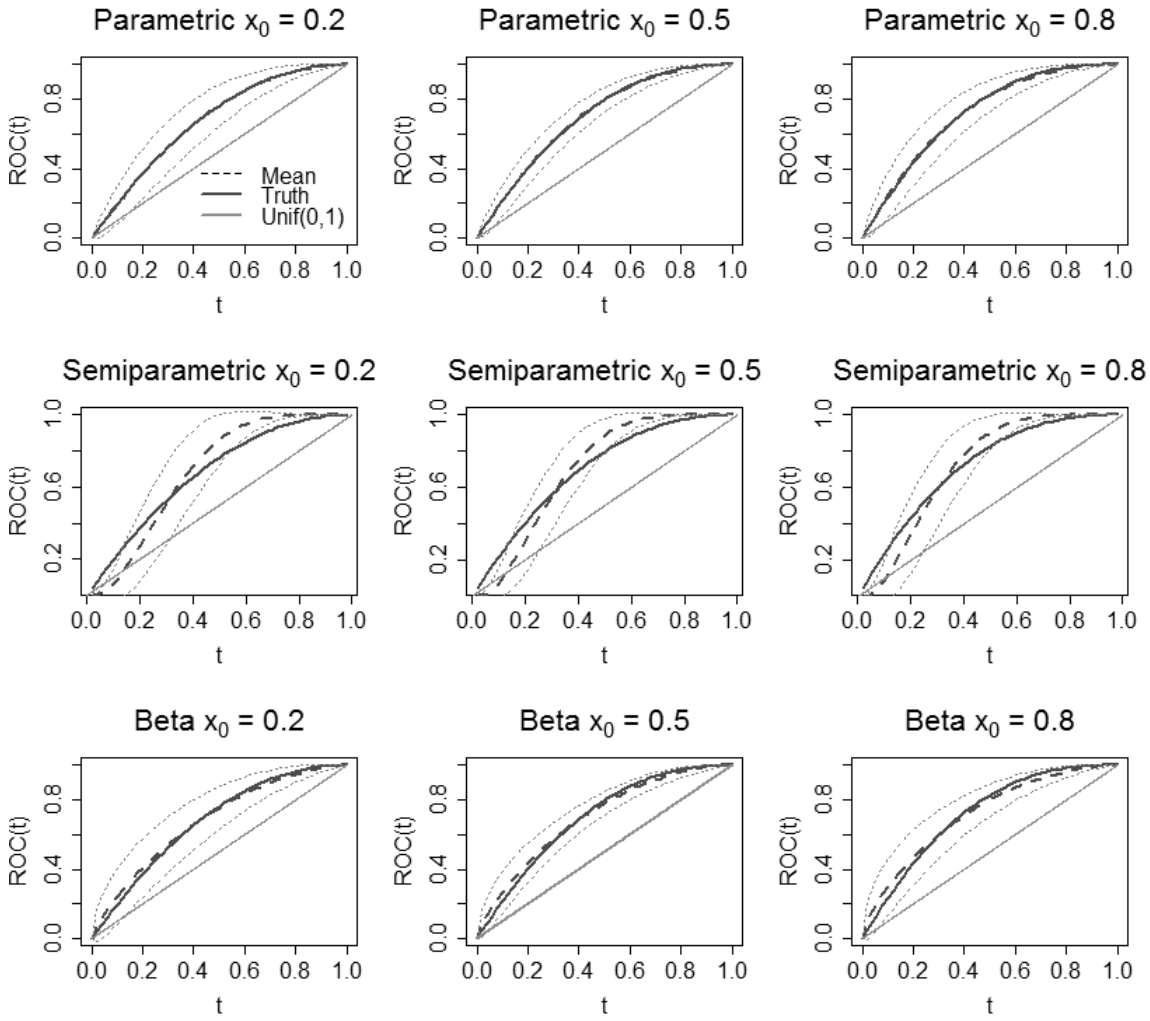

Figure 2.4.4: Comparison of simulated ROC and true ROC for extreme value data

Table 2.2: Summary of MSEs for extreme value

Method	1st.Qu.	Median	3rd.Qu.	Mean	St.Dev.
Beta	0.000528	0.001248	0.002578	0.001878	0.002567
Parametric	0.000406	0.000969	0.001932	0.001449	0.002046
Semi-parametric	0.000432	0.001038	0.002278	0.001712	0.002422

2.4.3 Discussion

The binormal and extreme value simulations provide a comparison of the three ROC regression methods. For both distributions, the beta regression approach yields results comparable to those of the parametric method, and both the beta and parametric ROC estimates closely align with the true covariate-adjusted ROC. Recall that the advantage of the beta regression model over the pre-existing methods is the ability to directly model the placement values without the use of a binary indicator. We have shown through simulation that the beta method is a viable alternative to the parametric and semiparametric models and merits additional exploration. We further illustrate the performance of the parametric and beta approaches with a return to the clinical study from our motivating example.

2.5 Application to a DME Study

In this example, the parametric and beta methods are used for subject-specific data from the Protocol I study (Elman et al., 2015) in the Diabetic Retinopathy Clinical Research Network (NCT 00444600). The study was designed to determine the efficacy of ranibizumab alone and ranibizumab in combination with laser therapy as compared to the efficacy of laser therapy alone in the treatment of diabetic macular edema (DME). In the study, each patient had been previously diagnosed with either type 1 or type 2 diabetes as well as diabetic macular edema affecting the center of the macula. The patients were randomized to one of four treatment groups. For the purpose of our example, we will consider two groups: A – a sham injection with laser treatment and B – a 0.5 mg injection of intravitreal ranibizumab along with laser treatment given three to ten days after injection. The primary outcome was visual acuity at one year adjusted for baseline acuity. Visual acuity was measured with Optical Coherence Tomography (OCT) which detects changes in retinal thickness, and the ETDRS test which records the number of letters that a patient can correctly identify. In this context, a favorable result is a decrease in retinal thickness which corresponds to vision improvement.

We define treatment A (laser therapy alone) to be the reference population and treatment B the comparator population. To investigate the performance of the three ROC regression models, we define the response of interest to be the amount of decrease in retinal thickness from baseline at one year. If treatment B is effective, the amount of decrease in retinal thickness should be higher for patients in the comparator population (treatment B) than for those in the reference population (treatment A). Density plots of the response (decrease in OCT from baseline) and one year OCT values for each treatment group appear in Figure 2.5.1. We note a high degree of overlap in the responses for the two groups, implying that the resulting ROC will be close to the diagonal line.

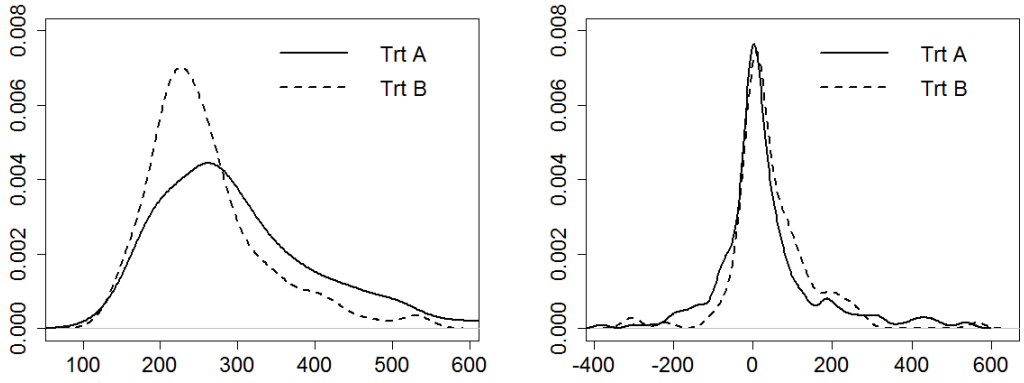

Figure 2.5.1: Density plots of one year OCT measurements on the left and one year decrease in OCT from baseline on the right

We are interested in the effect of covariates on the separation between the populations. Covariates common to both populations are gender and age at enrollment, and for illustrative purposes, we assume that duration of diabetes is a covariate associated with the comparator population. In this example, duration is a binary variable with a value of 1 if the duration is greater than or equal to the median of 17 years, and zero otherwise. A summary for each population and covariates of interest is included in Table 2.3. Boxplots of age, one year OCT, and one year decrease in OCT appear in Figures 2.5.2 and 2.5.3.

Table 2.3: Summary statistics for OCT and age by gender and duration of diabetes

Subset	n	OCT Baseline		OCT One Year		Age	
Trt A		Mean(SD)	Med.	Mean(SD)	Med.	Mean(SD)	Med.
Females	100	323.40(116.22)	309	303.01(112.17)	282	62.84(10.66)	63
Dur(0)	56	336.36(109.34)	326	296.25(85.84)	287	62.14(10.72)	61
Dur(1)	44	306.91(123.72)	285	311.61(139.32)	269	63.73(10.65)	65
Males	141	346.70(134.83)	317	305.37(111.69)	275	62.09(9.98)	63
Dur(0)	81	352.48(133.51)	317	313.23(119.09)	278	61.22(10.01)	62
Dur(1)	60	338.88(137.34)	318	294.75(100.86)	273	63.25(9.89)	64
Trt B		Mean(SD)	Med.	Mean(SD)	Med.	Mean(SD)	Med.
Females	55	297.51(103.69)	281	256.00(87.72)	226	63.02(10.79)	65
Dur(0)	23	300.26(124.60)	281	239.00(87.01)	220	61.04(11.64)	65
Dur(1)	32	295.53(87.74)	285	268.22(87.53)	234	64.44(10.09)	65
Males	64	306.17(93.38)	281	261.45(67.03)	252	60.66(9.81)	61
Dur(0)	29	324.41(106.27)	285	260.69(64.60)	257	58.10(8.75)	58
Dur(1)	35	291.06(79.61)	266	262.09(69.92)	249	62.77(10.25)	64

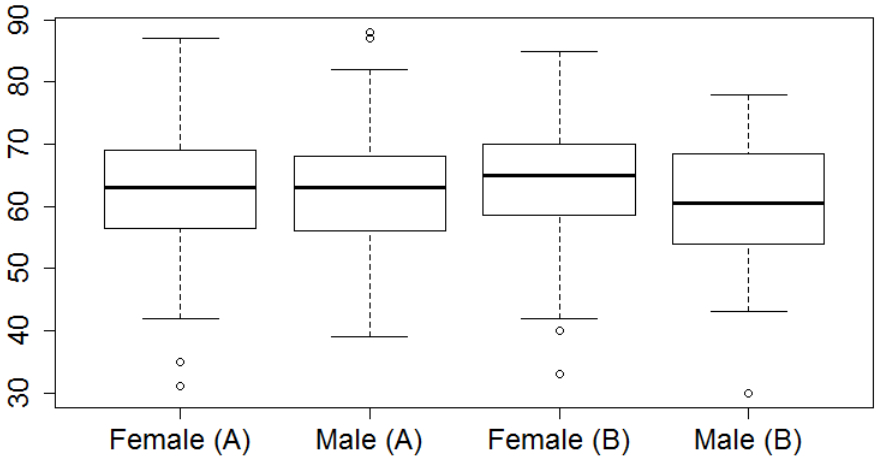

Figure 2.5.2: Boxplots for age by treatment and gender

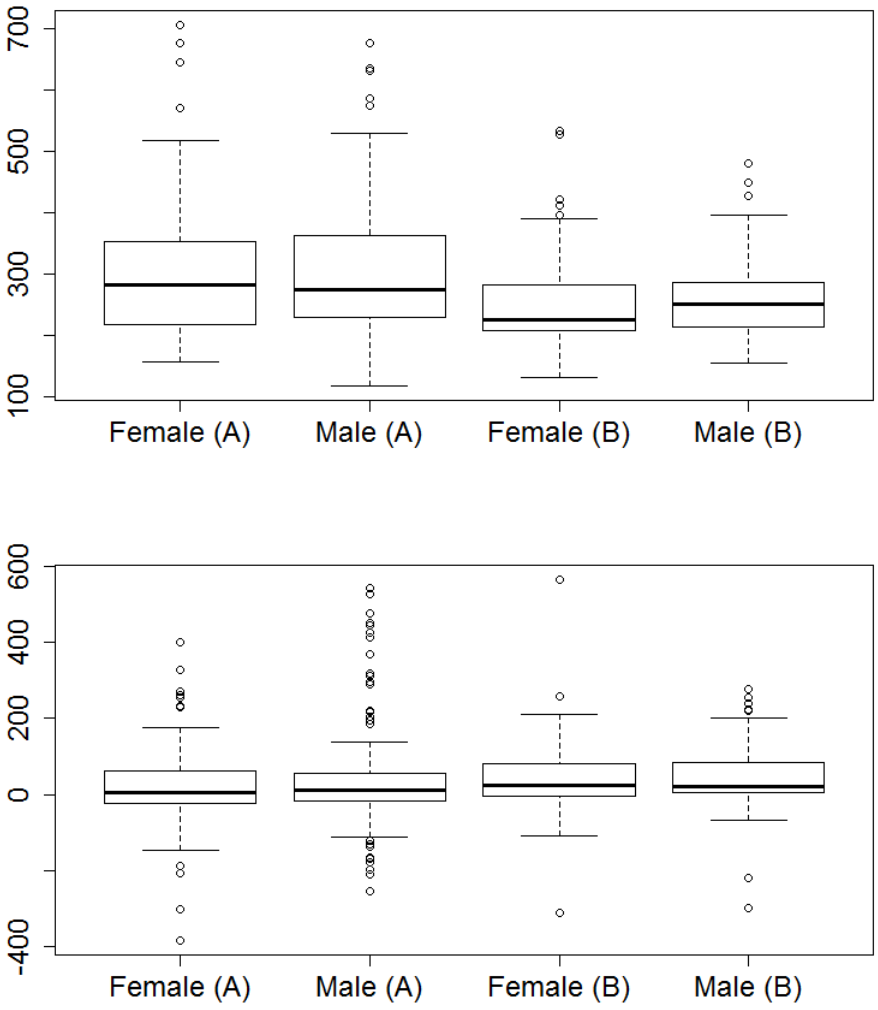

Figure 2.5.3: Boxplots for one year OCT (top) and one year decrease in OCT (bottom) by treatment and gender

For the comparator treatment B, we note a slight difference in median age between males and females. The one year OCT measurements for treatment B are lower than those for treatment A. The amount of decrease in OCT measurements from baseline is slightly higher for those in treatment B and there appears to be very little gender effect. Each of the methods is performed for the following ROC-GLM

$$ROC_X(t) = g(h_0(t) + \beta_1 * \text{age} + \beta_2 * \text{gender} + \beta_3 * \text{duration}).$$

Plots of the resulting ROC curves for different covariate values appear in Figures 2.5.4 and 2.5.5 for the parametric and beta approaches, respectively. The dotted line represents

a Uniform$(0, 1)$ cdf to illustrate the case for identical populations. Note that for both the parametric and beta methods, the AUC increases with age which indicates that the amount of decrease in OCT measurements from baseline was higher for older patients receiving treatment B. As anticipated given the overlap in response densities (Figure 2.5.1), the ROC is nearly diagonal when accounting for covariates.

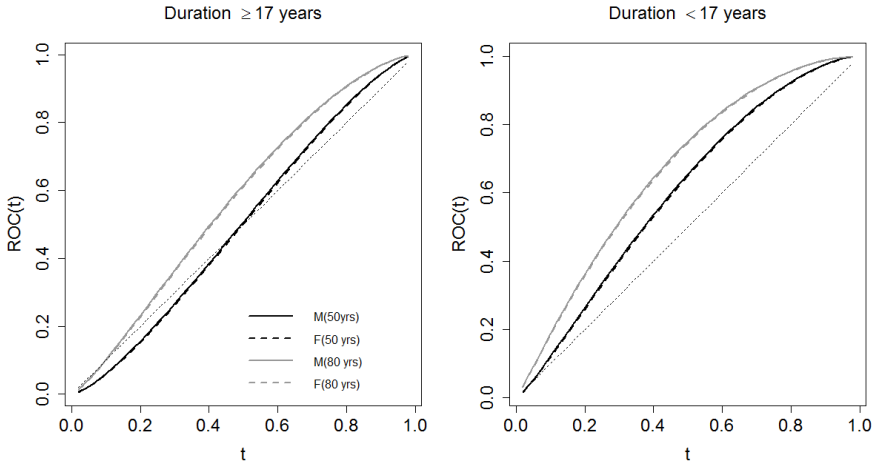

Figure 2.5.4: Covariate-adjusted ROC curves from the parametric method for males and females at ages 50 and 80

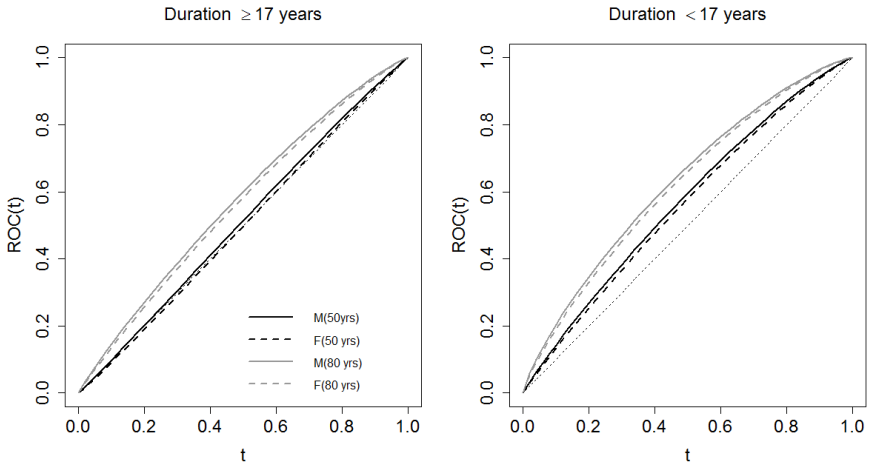

Figure 2.5.5: Covariate-adjusted ROC curves from the beta method for males and females at ages 50 and 80

CHAPTER THREE

Bayesian Approaches to ROC Regression

In this chapter, we introduce Bayesian extensions of the parametric ROC regression method (Alonzo and Pepe, 2002) and the newly proposed beta approach. Both of these methodologies lend themselves well to the Bayesian paradigm, in that the binary and beta regression models can be easily written as Bayesian hierarchical models. We observe the performance of the Bayesian extensions through simulation study.

3.1 Bayesian Parametric Approach

Recall that the parametric model proposed by Alonzo and Pepe (2002) is given by

$$\text{ROC}_{X,X_D}(t) = g(\gamma_1 h_1(t) + \gamma_2 h_2(t) + \boldsymbol{\beta} X + \boldsymbol{\beta}_D X_D),$$

with $\gamma_1, \gamma_2, \boldsymbol{\beta}$, and $\boldsymbol{\beta}_D$ as model parameters, $h_1(t) = 1, h_2(t) = \Phi^{-1}(t)$, and $g(\cdot) = \Phi(\cdot)$ where $\Phi(\cdot)$ is the cdf of the standard normal. This approach is distribution free because a parametric model is specified for the ROC, but no assumptions are made about the distributions for Y_D and $Y_{\bar{D}}$. As in the ROC-GLM framework proposed by Pepe (2000), the parametric approach expresses the ROC as the expectation of a binary indicator with a modification to ease computational intensity. When originally proposed by Pepe (2000), the binary indicator U_{ij} was calculated for all pairs of observations $\{(Y_{D_i}, Y_{\bar{D}_j}), i = 1, \ldots, n_D; j = 1, \ldots, n_{\bar{D}}\}$, with n_D and $n_{\bar{D}}$ denoting the number of observations from the diseased and non-diseased populations, respectively. Alonzo and Pepe (2002) advocated replacing $Y_{\bar{D}_j}$ in the binary indicator $U_{ij} = I[Y_{D_i} \geq Y_{\bar{D}_j}]$ with $S^{-1}_{\bar{D},X_i}(t)$, for $t \in (0, 1)$. Given the modified indicator, the expected value of U_{it} satisfies

$$E(U_{it}) = E(I[Y_{D_i} \geq S^{-1}_{\bar{D},X_i}(t)]) = \Pr[S_{\bar{D},X_i}(Y_{D_i}) \leq t] = \Pr[PV_D \leq t], \tag{3.1}$$

where PV_D is the placement value for the observation Y_{D_i} given the covariate vector X.

The covariate adjusted ROC is obtained by modeling

$$E[\hat{U}_{it}] = g^{-1}(\sum_{k=1}^{K} \gamma_k h_k(t) + X'\boldsymbol{\beta}), \tag{3.2}$$

using a probit link. In the Bayesian extension of the parametric model, we write (3.2) as a hierarchical model and apply Bayesian binary regression. A brief introduction to Bayesian binary regression follows.

3.1.1 Introduction to Bayesian Binary Regression

Suppose we have independent binary random variables $Z_i, ..., Z_N \sim$ Bernoulli(θ_i) where the probability of success θ_i is dependent on a covariate vector $\boldsymbol{x_i'} = (x_{i1}..., x_{iN})$. The binary regression model is defined as $\theta_i = g^{-1}(\boldsymbol{x_i'\beta})$, where $\boldsymbol{\beta}$ is a $(k \times 1)$ vector of regression parameters and $g^{-1}(\cdot)$ is a link function, conventionally either the logit or probit. We obtain the probit model by specifying $g^{-1}(\cdot) = \Phi(\cdot)$, where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. The Bayesian probit regression model is formed by specifying a prior distribution $\pi(\boldsymbol{\beta})$ for the regression parameters yielding the following structure (Albert and Chib, 1993).

$$Z_i \sim \text{Bernoulli}(\Phi(\eta_i))$$

$$\eta_i = \boldsymbol{x_i'\beta}$$

$$\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta})$$

Posterior inferences for the regression parameters are easily obtained using MCMC simulation in OpenBUGS or JAGS (Hornik et al., 2003).

3.1.2 Algorithm for Bayesian Parametric Method

An algorithm for performing the Bayesian Parametric method can be written as

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

(2) Estimate the covariate specific survival function $S_{\bar{D},X_j}$ for the reference population at each $t \in T, j = 1, \ldots, n_{\bar{D}}$ using quantile regression.

(3) For each diseased observation Y_{D_i}, calculate the placement values

$$PV_{D_i} = \hat{S}_{\bar{D},X_i}(Y_{D_i}), i = 1, \ldots, n_D.$$

(4) Calculate the binary placement value indicator $\hat{U}_{it} = I[V_{D_i} \leq t], t \in T$.

(5) Fit the model $E[\hat{U}_{it}] = g^{-1}[\sum_{k=1}^{K} \gamma_k h_k(t) + X'\boldsymbol{\beta}]$ using Bayesian binary regression such that

$$\hat{U}_{it} \sim \text{Bernoulli}(\theta_i)$$

$$\text{probit}(\theta_i) = \sum_{k=1}^{K} \gamma_k h_k(t) + X'\boldsymbol{\beta}$$

$$\gamma_k \sim \pi(\gamma_k)$$

$$\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta})$$

Note that in the following examples and simulations, we choose to use flat priors for the regression coefficients, giving heavy influence to the data in order to compare between the Bayesian and non-Bayesian parametric methods.

3.1.3 Example with Binormal Data

To illustrate the performance of the Bayesian parametric extension as compared to the parametric method (Alonzo and Pepe, 2002), we consider an example using binormal data. The data are generated from the following models

$$Y_D = 2 + 4X + \epsilon_D, \text{ and } Y_{\bar{D}} = 1.5 + 3X + \epsilon_{\bar{D}}. \tag{3.3}$$

where $X \sim U(0,1)$ and $\epsilon_D, \epsilon_{\bar{D}} \sim N(0, 1.5^2)$. In which case, the true ROC at $X = x_0,\ t \in (0,1)$ is $ROC(t) = \Phi[(0.5 + x_0)/1.5 + \Phi^{-1}(t)]$ We generate a single data set of size $n_D, n_{\bar{D}} = 200$ from which we calculate the ROC for the parametric method and the Bayesian parametric method. The results for the Bayesian parametric approach were obtained from MCMC simulation using Jags (5000 burn-in, 10000 iterations) with Normal(0, .01) priors specified for the regression coefficients. A plot of the resulting ROCs appears in Figure 3.1.1 for three covariate values. As expected with the use of diffuse priors, we

observe that the Bayesian parametric extension is nearly identical to the non-Bayesian parametric method. Convergence diagnostic plots as well as plots of posterior densities for the regression coefficients are included in the appendix.

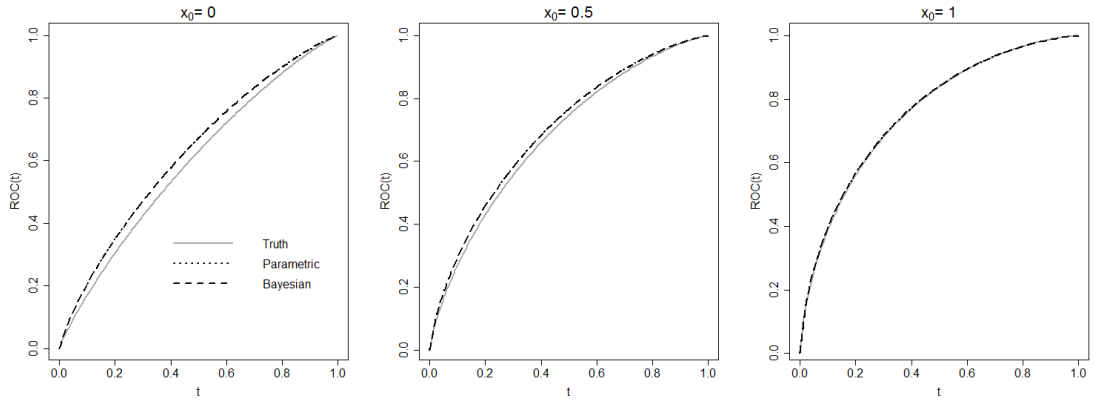

Figure 3.1.1: Comparison of parametric, Bayesian parametric, and true ROC curves for binormal data

3.2 Bayesian Extension of the Beta Approach

The beta regression approach described in Chapter Two is also easily extended to the Bayesian paradigm through a hierarchical model. Bayesian beta regression was proposed by Branscum et al. (2007) and we include a brief introduction to the method here. Given $Z \sim \text{Beta}(a, b)$, the choice of parameters a and b determines a wide range of shapes for the density function

$$f(z|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} z^{a-1}(1-z)^{b-1}, \text{ for } z \in (0,1) \text{ and } a, b > 0.$$

The beta density's flexibility makes it a natural choice for modeling continuous data restricted to the (0,1) interval. Given $Z \sim \text{Beta}(a, b)$, the expected value is $E(Z) = a/(a+b)$ and the variance is $Var(Z) = ab/[(a+b)^2(a+b+1)]$. To incorporate covariate information, we reparameterize the beta density by letting $\mu = \frac{a}{a+b}$ and $\phi = a + b$. The reparameterized mean and variance are thus $E(Z) = \mu$, and $Var(Z) = \mu(1-\mu)/(1+\phi)$. The form of the beta regression model is as follows. Let $z_1, \ldots, z_n$ be independent random

variables from a beta density with mean μ_t, t = 1,. . . ,n and scale parameter ϕ. Then the beta regression model can be written as

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

where $\boldsymbol{\beta}$ is a vector of regression parameters, $x_{t1}, \ldots, x_{tk}$ are observations on k covariates, and g is a monotonic link function. Using the logit link, we have $\mu_t = \dfrac{1}{1 + e^{-\mathbf{x}_t'\boldsymbol{\beta}}}$. Estimates of the original parameters a and b are

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

The Bayesian beta regression model as proposed by Branscum et al. (2007) is

$$\begin{aligned} z_t|\mu_t, \phi &\sim \text{Beta}(\mu_t\phi, \phi(1-\mu_t)) \\ \mu_t \equiv \mu_t(\mathbf{x}_t) &= g^{-1}(\mathbf{x}_t'\boldsymbol{\beta}) \\ \pi(\boldsymbol{\beta}, \phi) &= \pi(\boldsymbol{\beta})\pi(\phi), \end{aligned}$$

where g is a link function. In this discussion, we specify the logit, although other link functions such as the probit or complimentary log-log links could be used. The likelihood function for independent data is

$$L(\boldsymbol{\beta}, \phi) = \prod_{t=1}^{n} \frac{\Gamma(\phi)}{\Gamma(g^{-1}(\mathbf{x}_t'\boldsymbol{\beta})\phi)\Gamma(\phi(1 - g^{-1}(\mathbf{x}_t'\boldsymbol{\beta})))} z_t^{g^{-1}(\mathbf{x}_t\boldsymbol{\beta})\phi - 1}(1 - z_t)^{\phi(1-g^{-1}(\mathbf{x}_t'\boldsymbol{\beta}))-1}.$$

The posterior distribution is given by

$$\begin{aligned} \pi(\boldsymbol{\beta}, \phi|\boldsymbol{z}) \propto &\prod_{i=1}^{n} (\Gamma(g^{-1}(\mathbf{x}_t'\boldsymbol{\beta})\phi)\Gamma(\phi(1 - g^{-1}(\mathbf{x}_t'\boldsymbol{\beta}))))^{-1} z_t^{g^{-1}(\mathbf{x}_t'\boldsymbol{\beta})\phi}(1 - z_t)^{\phi(1-g^{-1}(\mathbf{x}_t'\boldsymbol{\beta}))-1} \\ &\times \pi(\boldsymbol{\beta}, \phi)\Gamma(\phi)^n. \end{aligned}$$

We use Gibbs sampling to iteratively sample from the full conditional distributions $\pi(\boldsymbol{\beta}|\phi, \boldsymbol{z} \propto L(\boldsymbol{\beta}, \phi)\pi(\boldsymbol{\beta})$ and $\pi(\phi|\boldsymbol{\beta}, \boldsymbol{z})\pi)(\phi)$ to generate a Monte Carlo sample from the posterior $\pi(\boldsymbol{\beta}, \phi|\boldsymbol{z})$. Posterior inferences for the mean response $\mu(\boldsymbol{z})$ and the regression parameters $\boldsymbol{\beta}$ and ϕ are easily obtained in OpenBUGS or JAGS.

3.2.1 Algorithm

An algorithm can be written as follows.

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

(2) Estimate the covariate specific survival function $S_{\bar{D},X_j}$ for the reference population at each $t \in T, j = 1, \ldots, n_{\bar{D}}$ using quantile regression.

(3) Calculate the placement values

$$PV_{D_i} = \hat{S}_{\bar{D},X_i}(Y_{D_i}), i = 1, \ldots, n_D.$$

(4) Perform a Bayesian beta regression on the placement values to obtain estimates of $\boldsymbol{\beta}$ and ϕ.

(5) Transform to obtain $a = \mu\phi$ and $b = (1 - \mu)\phi$.

(6) Calculate the cdf of the placement values using the Beta(a,b) distribution found above to obtain the ROC and the AUC.

Steps (1) - (3) are identical to the parametric and semi-parametric cases. In step (4), we model the placement values directly using Bayesian beta regression to obtain estimates of $\boldsymbol{\beta}$ and ϕ. We then apply a transformation to return to the original beta parameters a and b and calculate the cdf of the placement values using the resulting Beta(a, b) distribution which yields an estimate for the ROC. The AUC is obtained by implementing the trapezoid rule to calculate the area under the ROC.

3.2.2 Example with Binormal Data

The performance of the Bayesian beta extension as compared to the non-Bayesian beta method is considered with a binormal example. The data are generated as in Model (3.3). We generate a single data set of size $n_D, n_{\bar{D}} = 200$ from which we calculate the ROC for the beta method and the Bayesian beta method. The results for the Bayesian beta approach were obtained from MCMC simulation using Jags (5000 burn-in, 10000

iterations) with Normal$(0, .01)$ priors specified for the regression coefficients and a diffuse Gamma$(.01, .01)$ for the scale parameter. A plot of the resulting ROCs appears in Figure 3.2.1 for three covariate values. The ROC curve resulting from the Bayesian extension is nearly identical to non-Bayesian beta ROC curve, as expected with the use of diffuse priors. Convergence diagnostic plots for this example as well as plots of posterior densities for the regression coefficients are included in the appendix.

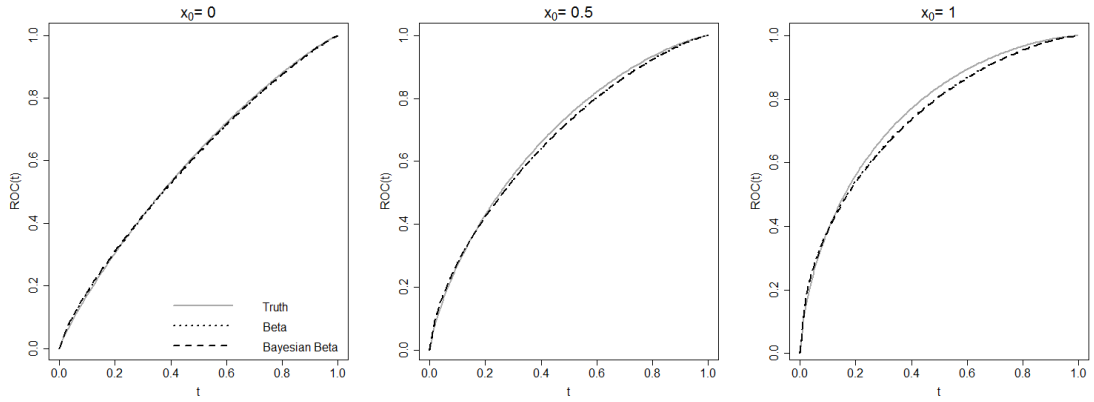

Figure 3.2.1: Comparison of ROCs for beta, Bayesian beta, and truth for binormal data

3.3 Simulation Study

We compare the Bayesian parametric and beta ROC regression methods through two simulations, one using normally distributed data and the other using data from an extreme value distribution. The data models in this section are similar to those of Chapter Two. For simplicity, we consider one continuous covariate from a uniform distribution (Rodriguez-Alvarez et al., 2011). The models and results follow.

3.3.1 Binormal Data

We compare the Bayesian parametric and Bayesian beta approaches using a simulation with binormal data. The data are generated from Model (3.3 and the true ROC at $X = x_0,\ t \in (0,1)$ is $ROC(t) = \Phi[(0.5 + x_0)/1.5 + \Phi^{-1}(t)]$. We generate 500 data sets of size $n_D, n_{\bar{D}} = 200$ from which we calculate the ROC and AUC for the Bayesian

parametric and Bayesian beta methods. For the Bayesian beta approach, the regression parameters were given Normal$(0, .01)$ priors and the scale parameter was given a diffuse Gamma$(.001, .001)$. We also compute the mean squared error (MSE) for the ROC estimates for both the parametric and the beta. Boxplots of the MSE values for the AUC are given in Figure 3.3.1. The summary statistics for the MSE are given in Table 3.1. As expected, the mean and standard deviation for the parametric method are smaller than those resulting from the beta method given that the parametric assumption (Alonzo and Pepe, 2002) performs well with binormal data. Note, however, that the mean for the beta method is within one standard deviation of the mean for the parametric method. Plots of the simulated and true ROC curves are included in Figure 3.3.2 for covariate values $x_0 = \{0.2, 0.5, 0.8\}$. The dotted lines represent the average 2.5% and 97.5% quantiles of the simulated ROC estimates. Observe that the AUC increases with an increase in the covariate value.

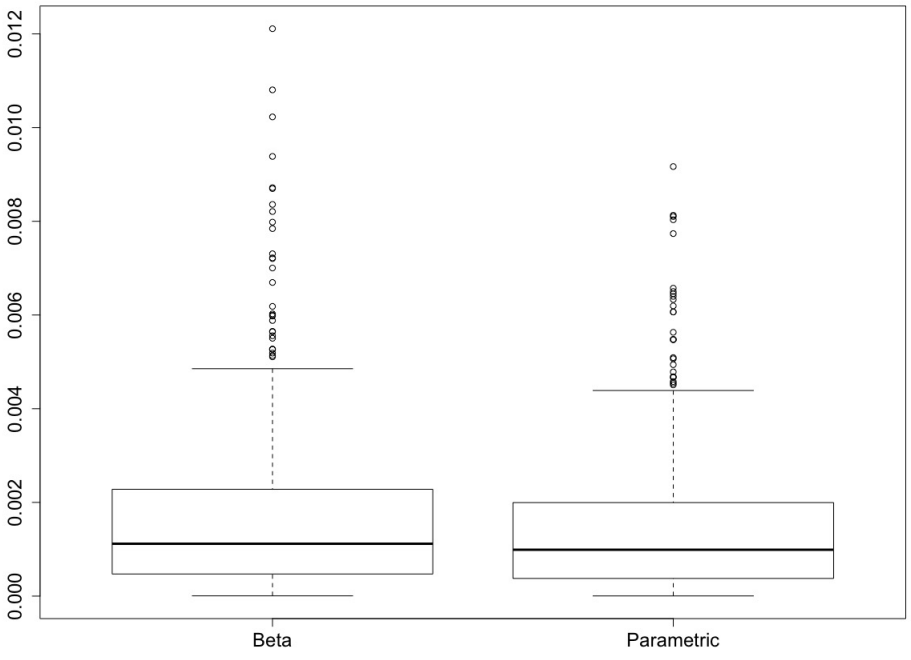

Figure 3.3.1: Boxplots of the estimated MSEs for the ROC resulting from the Bayesian parametric and Bayesian beta methods based on 500 estimates ($n_D = n_{\bar{D}} = 200$)

Table 3.1: Summary of MSEs for binormal data

Method	1st.Qu.	Median	3rd.Qu.	Mean	St.Dev.
Beta	0.000482	0.001112	0.002297	0.001696	0.001838
Parametric	0.000385	0.000979	0.002011	0.001451	0.001515

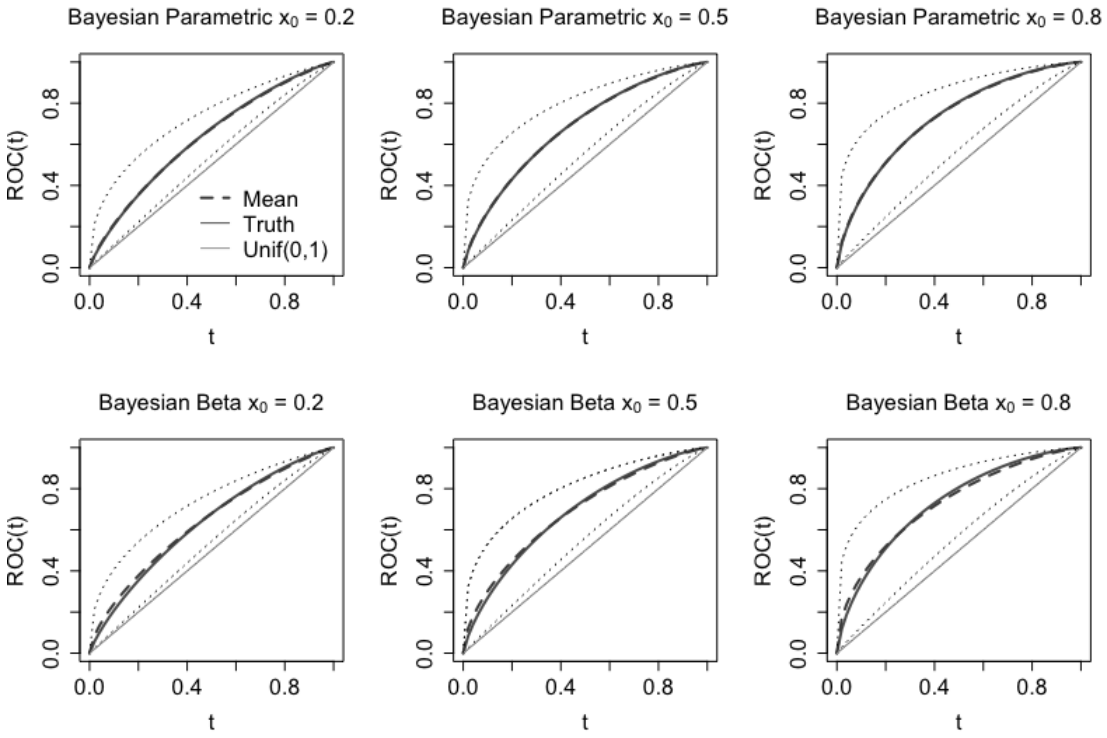

Figure 3.3.2: Comparison of simulated ROC and true ROC for binormal data

3.3.2 Extreme Value Data

A second simulation to compare the Bayesian parametric and Bayesian beta approaches is performed using data from the extreme value distribution. The data are generated from $Y_D = 2 + 2.5X + \epsilon_D$ and $Y_{\bar{D}} = 1 + 2X + \epsilon_{\bar{D}}$, where $X \sim U(0,1)$ and $\epsilon_D, \epsilon_{\bar{D}}$ have an extreme value distribution with mean 0 and standard deviation 1.5. The true value of the ROC when $X = x_0$ is

$$\text{ROC}_X(t) = 1 - \exp\left\{ -\exp\{\ln[-\ln(1-t)] - \frac{1 + 0.5x_0}{1.5}\}\right\}.$$

For the Bayesian parametric approach, the regression coefficients were each given a diffuse Normal $(0, .01)$ prior. For the Bayesian beta approach, the regression parameters were given Normal$(0, .01)$ priors and the scale parameter was given a diffuse Gamma$(.001, .001)$. As in the binormal simulation, we generate 500 data sets of size $n_D, n_{\bar{D}} = 200$ from which we calculate the ROC and AUC for the Bayesian parametric and Bayesian beta methods through MCMC simulation using JAGS (20000 iterations, 5000 burn-in). We also compute the mean squared error (MSE) for the AUC estimates for both the parametric and the beta. The summary statistics for the MSE are given in Table 3.2. Note that the mean MSE and standard deviation for the Bayesian beta method are slightly higher than the analogous results for the Bayesian parametric method, but the means are less than one standard deviation apart. Plots of the simulated and true ROC curves are included in Figure 3.3.3 for covariate values $x_0 = \{0.2, 0.5, 0.8\}$. The dotted lines represent the average 2.5% and 97.5% quantiles of the simulated ROC estimates.

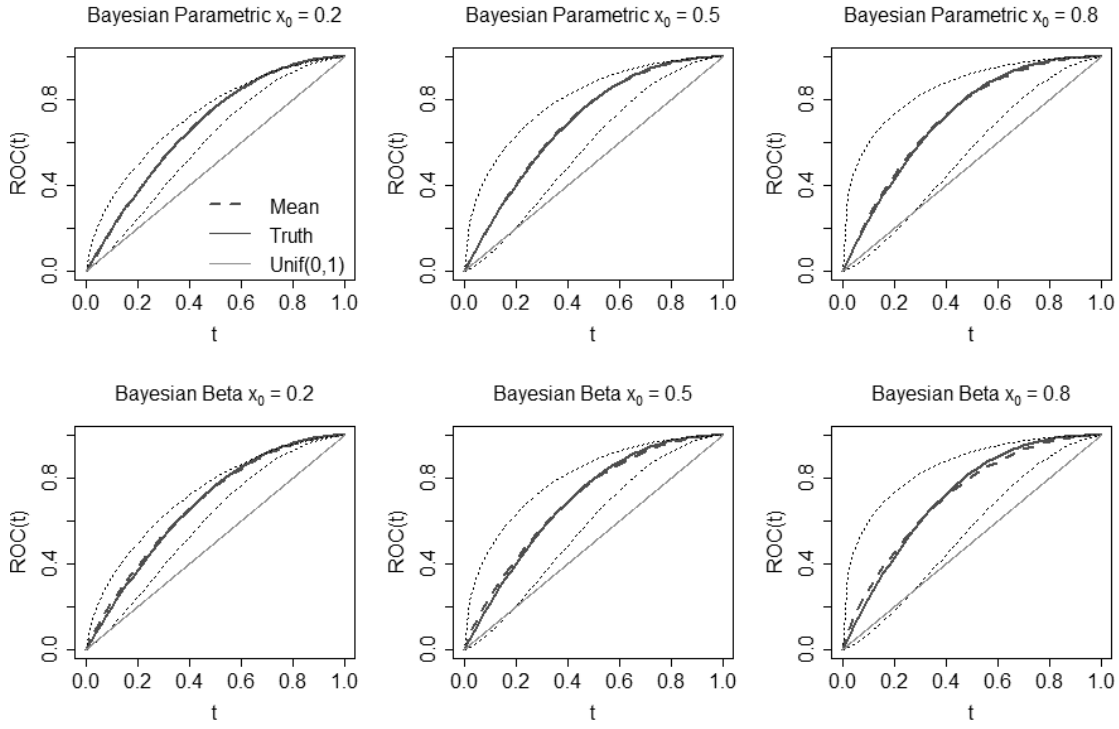

Figure 3.3.3: Comparison of simulated ROC and true ROC for extreme value data

Table 3.2: Summary of MSEs for extreme value data

Method	1st.Qu.	Median	3rd.Qu.	Mean	St.Dev.
Beta	0.000489	0.001110	0.002185	0.001733	0.003010
Parametric	0.000396	0.000928	0.001892	0.001470	0.002525

3.4 DME Application

We return to the DME example from Chapter Two as an illustration of the Bayesian extensions of the parametric and beta approaches. As before, we define treatment A (laser therapy alone) to be the reference population and we define treatment B as the comparator population. For the purpose of investigating the performance of the three methods, we define the response of interest to be the amount of decrease in retinal thickness from baseline at one year. If treatment B is effective, the amount of decrease in retinal thickness from baseline should be higher than for patients in the comparator population (treatment B) than for those in the reference population (treatment A).

We are interested in the effect of covariates on the separation between the populations. Covariates common to both populations are gender, age at enrollment, and duration of diabetes is a covariate associated with the diseased group. In this example, duration is a binary variable with a value of 1 if the duration is greater than or equal to 20 years, and zero otherwise. Each of the methods is performed for the following ROC-GLM

$$ROC_X(t) = g(h_0(t) + \beta_1 * \text{age} + \beta_2 * \text{gender} + \beta_3 * \text{duration}).$$

Plots of the resulting ROC curves for different covariate values appear in Figures 3.4.1 and 3.4.2 for the parametric and beta approaches, respectively. Note that for both the Bayesian parametric and beta methods, the AUC increases with age which indicates that the amount of decrease in OCT measurements from baseline was higher for older patients receiving treatment B. As noted in Chapter Two, given the overlap in response densities (Figure 2.5.1), the ROC is nearly diagonal (dotted line) when accounting for covariates.

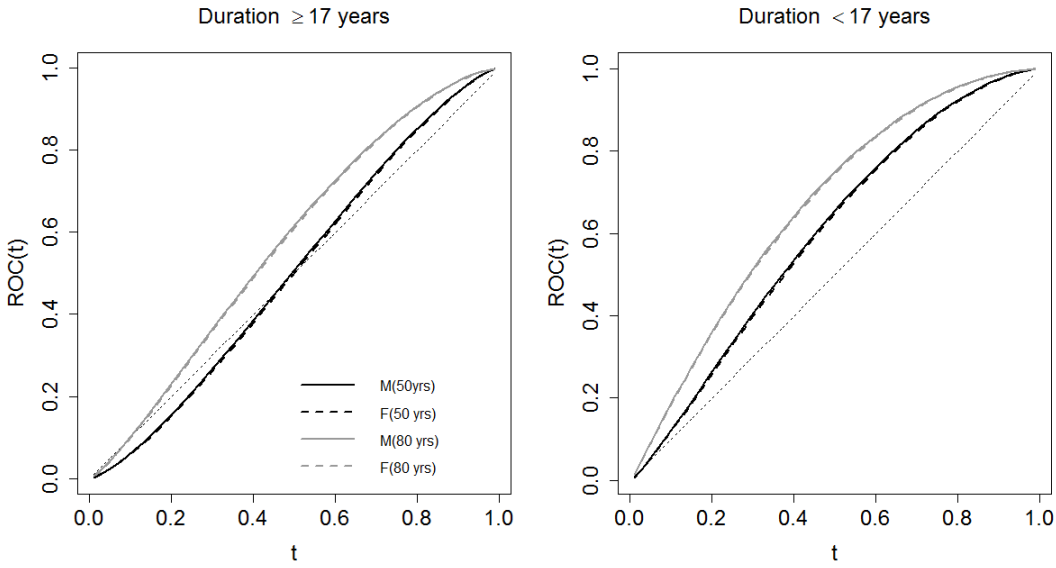

Figure 3.4.1: Covariate-adjusted ROC curves from the Bayesian parametric method

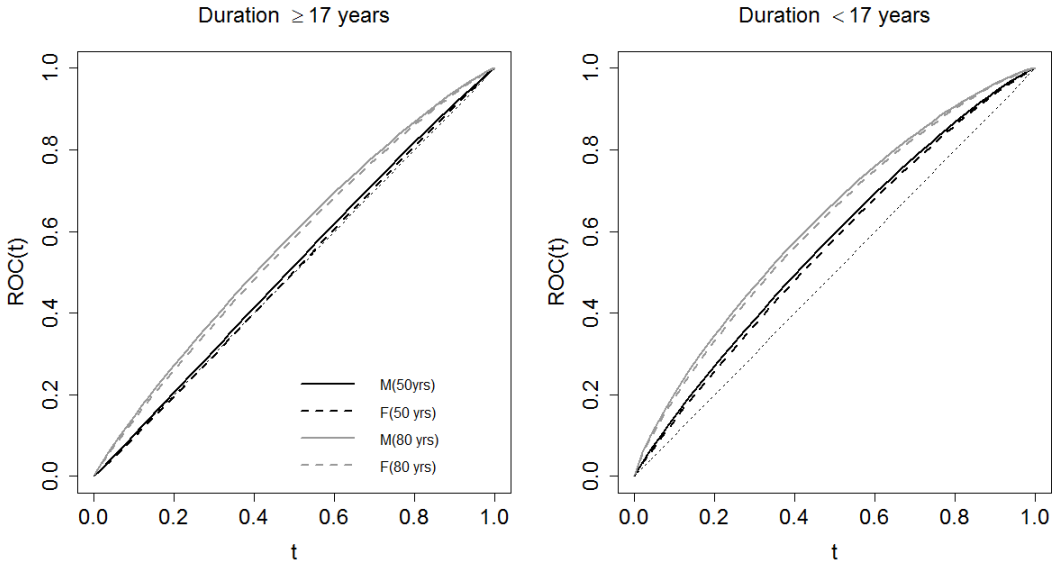

Figure 3.4.2: Covariate-adjusted ROC curves from the Bayesian beta method

CHAPTER FOUR

Indirect Comparison of ROC Curves

Having introduced ROC regression methodology and explored the performance of three methods, we now apply the beta approach to a simple network meta-analysis. Suppose we have two studies and three treatments A, B, and C. The first study compares treatment A to treatment B and the second study compares treatment A to treatment C. Given that treatment A is common to both studies, we would like to explore the ability to draw an indirect comparison between treatments B and C. The described scenario represents a simple network-meta analysis that can be summarized in Figure 4.0.1, where the solid lines indicate observed comparisons and the dashed line indicates the indirect comparison.

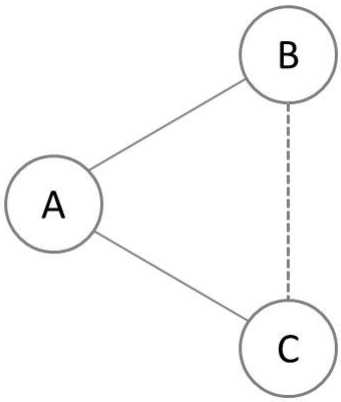

Figure 4.0.1: Simple Network Meta-Analysis Diagram

The beta ROC regression methodology introduced in Chapter Two and extended to the Bayesian paradigm in Chapter Three offers an approach for drawing such an indirect comparison. Essentially, we obtain an ROC curve from each of Studies 1 and 2 using the beta approach. We thus have a covariate-adjusted ROC curve that compares treatment A to treatment B and a second covariate-adjusted ROC curve that compares treatment A to treatment C. To draw an indirect comparison between B and C, we apply a Bayesian

variable selection method to assess the relationship between the two ROC curves. We begin with defining notation and then describe the variable selection procedure for the indirect comparison.

4.1 Notation

Suppose we have two studies (Study 1 and Study 2) and three treatments A,B, and C such that Study 1 compares A to B and Study 2 compares A to C. In each study, we define treatment A to be the reference group and denote the reference responses by Y_{A1} and Y_{A2} for Study 1 and Study 2, respectively. The responses for the comparator treatments B and C are denoted Y_B and Y_C. We perform the Bayesian beta ROC regression approach and obtain a covariate-adjusted ROC curve for each of Study 1 and Study 2. We use ROC_{AB} to denote the ROC curve from Study 1 and ROC_{AC} to denote the ROC curve from Study 2. If $ROC_{AB} \neq ROC_{AC}$ then treatments B and C are different in terms of how they relate to the reference treatment A. In contrast, if $ROC_{AB} = ROC_{AC}$ then treatments B and C are similar in relation to treatment A.

4.2 Bayesian Variable Selection

To determine whether the resulting ROC_{AB} and ROC_{AC} are the same or different, we borrow an approach from Bayesian variable selection methodology. Bayesian variable selection was developed as a solution to the familiar regression problem of identifying a subset from a large number of explanatory variables that explains a large proportion of the response variation. (O'Hara et al., 2009) provide an overview of Bayesian variable selection methods, among which is an indicator selection model which we will apply in the simple network example. We illustrate the idea of a Bayesian indicator selection as follows.

Suppose we construct a regression model to explain an outcome $y_i, i = 1, ..., N$ for an individual i and let $x_{i,j}, j = 1, ..., p$ represent covariate values. Given a vector of

parameters $\boldsymbol{\beta}$ a linear regression model for the response y_i is

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + e_i,$$

where β_0 is the intercept and the errors e_i are from a Normal$(0, \sigma^2)$ distribution. Assuming that y_i is from an exponential family yields a GLM written as

$$E[g(y_i)] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j},$$

where $g(\cdot)$ is a link function. Bayesian variable selection specifies spike and slab priors (Miller, 2002) on each β_j which are used in determining which of the β_js are equal to zero. The probability mass (the spike) is set at zero and the slab is set elsewhere. An indicator variable I_j denotes whether the variable j is in the spike ($I_j = 0$) or the slab ($I_j = 1$) portion of the prior. That is, if $I_j = 0$ the covariate j is absent from the model. Note that I_j is an auxiliary variable.

4.2.1 Indicator Model Selection

In indicator model selection, the spike $\theta_j|(I_j = 0)$ is set equal to zero and the slab $\theta_j|(I_j = 1)$ is set to β_j, where θ_j is an auxiliary variable defined such that $\theta_j = I_j\beta_j$. Kuo and Mallick (1998) proposed an indicator model selection method that assumes that θ_j and I_j are independent *a priori*. Independent priors are placed on each of I_j and β_j such that $\pi(I_j, \beta_j) = \pi(I_j)\pi(\beta_j)$. The model is fit using MCMC where the variable selection portion of the model relies on estimation of I_j and β_j. The mean value of the indicator I_j *a posteriori* represents the probability that variable j is in the model. When $I_j = 0$, the MCMC algorithm updates the value of θ_j using the full conditional distribution which is $\pi(\theta_j)$.

4.2.2 Application to Simple Network

We explain the application of the indicator model selection method (Kuo and Mallick, 1998) to the simple network with the aid of a diagram. Following the algorithm for the Bayesian beta approach, we begin with the responses from Study 1 and estimate a

covariate-adjusted survival curve for the reference treatment A. We then calculate the placement values for the treatment B responses (denoted PV_B in the diagram) and proceed with the Bayesian beta regression model to obtain estimates of $a_1 = \mu_1\phi_1$ and $b_1 = (1-\mu_1)\phi_1$. We then estimate the placement values PV_C for treatment C using the covariate-adjusted reference survival from Study 2. The indicator model selection method is incorporated in the Bayesian beta regression model for PV_C. Note in the diagram that each of β_{02} and β_{12} is multiplied by an indicator variable ω. This corresponds to the indicator I_j in Kuo and Mallick's algorithm. The indicator ω is given a spike and slab prior so that if $\omega = 0$, then β_{02} and β_{12} are absent from the model. That is, the placement values PV_C come from the same beta distribution as the placement values PV_B and the resulting ROC_{AB} and ROC_{AC} are equivalent. If $\omega = 1$, then both β_{02} and β_{12} are present in the model and the resulting ROCs from each study are said to be different. The posterior mean of ω thus represents the probability that β_{02} and β_{12} are needed in the model and that the resulting ROCs are not the same. We illustrate the performance of indicator selection in conjunction with the Bayesian beta ROC regression method through simulation study.

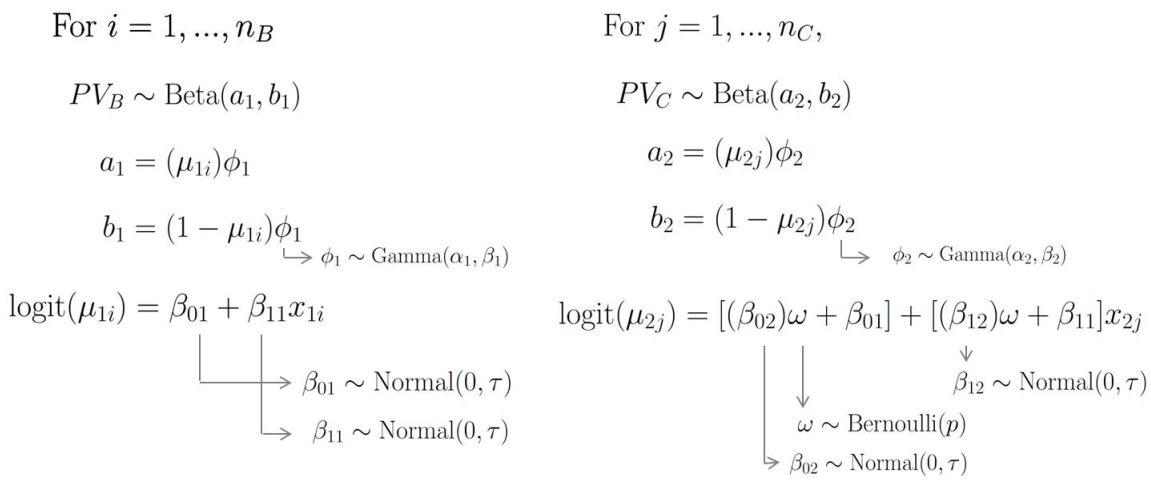

Figure 4.2.1: Diagram of indirect model selection for beta regression model

4.3 Simulation Study

The simulations included in this section are intended to serve as a proof of concept for using a Bayesian indicator model selection method in conjunction with Bayesian beta regression. While certainly not exhaustive, the simulations show how the model selection method performs for various degrees of separations among treatments in a simple meta-analysis network. We consider data from both normal and extreme value distributions.

4.3.1 Normal Data

Normal data are generated for each of Studies 1 and 2 according to the simulation schemes in Table 4.1 with variance of one and covariate distributed Uniform$(0, 1)$. Study 1 compares treatments A and B, and Study 2 compares treatments A and C. Density plots for each treatment are given in the upper section of Figure 4.3.1. The corresponding ROC curves at covariate value $x = 0.5$ are given in the lower section of Figure 4.3.1, where the dotted lines represent a Unif$(0, 1)$ cdf to illustrate the case of identical populations. Note that the reference treatments for Studies 1 and 2 are generated independently from the specified normal distribution, although for plot readability only one reference density appears in Figure 4.3.1. We generate 500 datasets from each simulation scheme with sample sizes $n_{D_1}, n_{\bar{D}_1} = 200$ and $n_{D_2}, n_{\bar{D}_2} = 200$ for Studies 1 and 2 respectively. We apply the Bayesian indicator model selection method to each dataset and record the posterior mean estimate of ω. Because we assume no prior knowledge regarding the necessity of a second model for Study 2, we use a Bernoulli(.5) prior for the indicator ω. Flat normal priors were given to the beta regression coefficients $\beta_{01}, \beta_{02}, \beta_{11}, \beta_{12}$, and a diffuse Gamma$(.001, .001)$ prior was specified for the scale parameter ϕ. MCMC simulations were run in JAGS with 10000 iterations and 5000 burn-in. Diagnostic plots are included in the appendix.

Table 4.1: Covariate dependent simulation means for binormal data

Simulation Scheme	A_1, A_2	B	C
I: $A_1 = A_2 = B = C$	1.5 + 3x	1.5 + 3x	1.5 + 3x
II: $A_1 = A_2 = B < C$	1.5 + 3x	1.5 + 3x	2 + 3.5x
III: $A_1 = A_2 < B = C$	1.5 + 3x	2 + 4x	2 + 4x
IV: $A_1 = A_2 < B < C$	1.5 + 3x	2 + 3.5x	2.5 + 4x

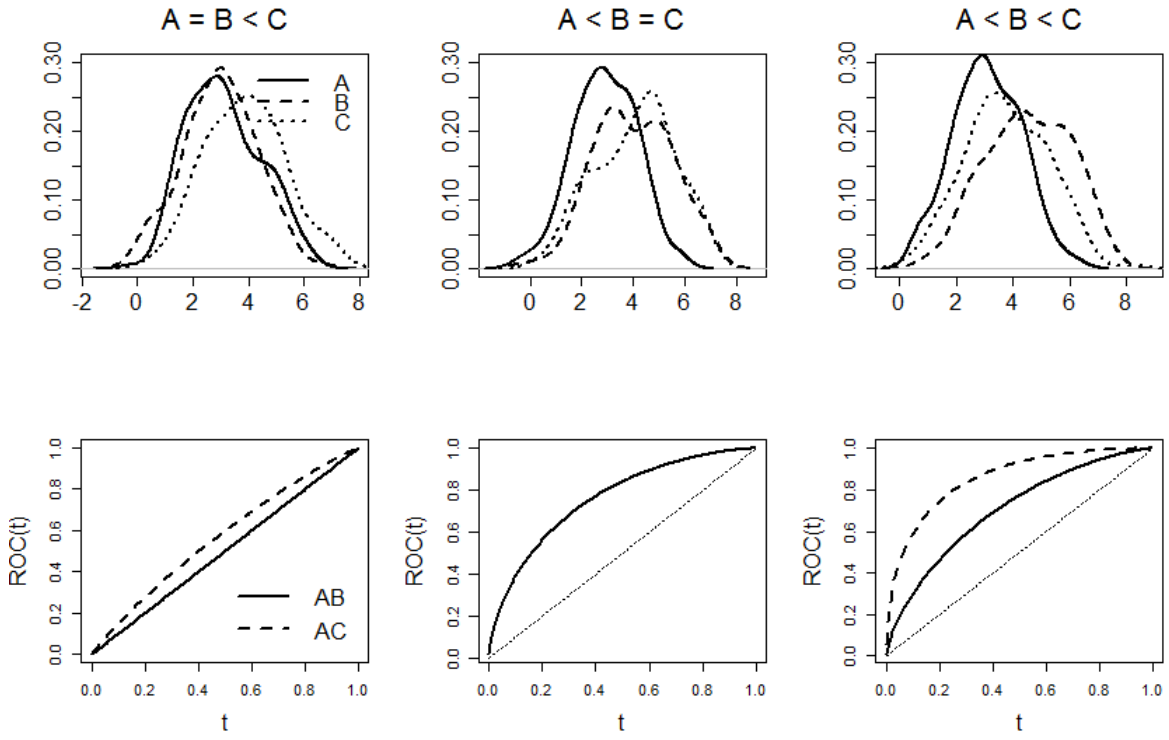

Figure 4.3.1: Top: Densities for simulation schemes II, III, and IV, Bottom: True binormal ROC curves at covariate value 0.50

4.3.1.1 Results. Scenario I represents the null case and will be used to determine a suitable cut off value for the inclusion probability ω to be used in determining whether a separate beta regression model is needed for Study 2. Because the B and C populations are independently generated from a common normal distribution, we expect the resulting ROC_{AB} to be the same as ROC_{AC} and the indicator model selection method should yield a posterior mean for ω that is close to zero. We generate 500 data sets from simulation

scheme I (see Table 4.1) with a variance of one and $n_D = n_{\bar{D}} = 200$. A summary of the resulting posterior estimates for ω at several quantiles is included in Table 4.2. The results are ranked by mean. Note that choosing the 95^{th} percentile (0.8920) as a cut-off probability yields a type I error rate of 5% under the null hypothesis that $\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$. For the remaining simulations schemes, we would thus conclude that a second model is needed for Study 2 if the posterior mean ω exceeds 0.8920.

Table 4.2: Summary of posterior estimates for ω with $n_D, n_{\bar{D}} = 200$.

Quantile	Mean	SD	2.5%	25%	50%	75%	97.5%
0.01	0.0000	0.0000	0	0	0	0	0
0.25	0.0090	0.0945	0	0	0	0	0
0.50	0.0210	0.1435	0	0	0	0	0
0.75	0.0820	0.2745	0	0	0	0	1
0.80	0.1530	0.3602	0	0	0	0	1
0.85	0.3190	0.4663	0	0	0	1	1
0.90	0.5740	0.4947	0	0	1	1	1
0.95	0.8920	0.3105	0	1	1	1	1

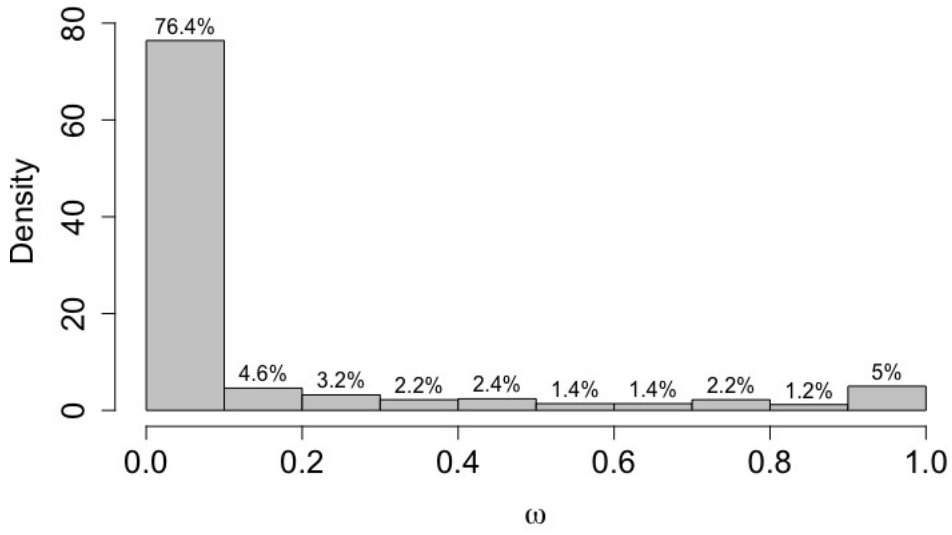

Figure 4.3.2: Histogram of posterior means for ω from Scenario I

In Scenario II, both reference treatments as well as treatment B are independently generated from the same normal distribution. Treatment C is generated from a normal distribution with a higher mean (Table 4.1). Note that the covariate effect for treatment C is higher than the covariate effect for the other treatments, which implies that the probability of inclusion ω should be greater than the cut-off 0.8920 established in the null case. We generate 500 data sets from simulation scheme II with a variance of one and sample sizes $n_D, n_{\bar{D}} = 200$. A histogram of the posterior means for ω appears in Figure 4.3.3. All but eight (1.6 %) of the the posterior means for ω are greater than 0.8290 which indicates that a second model is needed for Study 2 as anticipated.

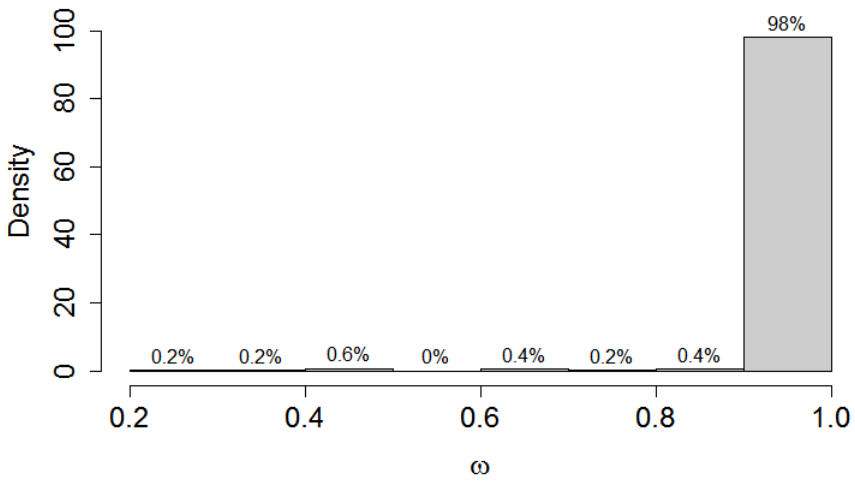

Figure 4.3.3: Histogram of posterior means for ω from Scenario II

For Scenario III, the reference distributions for both studies are the same. Treatments B and C are generated from a normal distribution with mean higher than that of the reference (Table 4.1). We expect $ROC_{AB} = ROC_{AC}$ which implies that the probability of inclusion ω should be less than 0.8920 as established in the null case. We generate 500 data sets from simulation scheme III with a variance of one and sample sizes $n_D, n_{\bar{D}} = 200$. A summary of the resulting ranked posterior estimates for ω is included in Table 4.3 and a

histogram of the posterior means for ω appears in Figure 4.3.4. Note that roughly 10% of the 500 posterior means for ω exceeded the cut-off probability of 0.8920.

Table 4.3: Summary of posterior estimates for ω from Scenario III

Quantile	Mean	SD	2.5%	25%	50%	75%	97.5%
0.01	0.0010	0.0316	0	0	0	0	0
0.25	0.0100	0.0995	0	0	0	0	0
0.50	0.0350	0.1839	0	0	0	0	1
0.75	0.1810	0.3852	0	0	0	0	1
0.80	0.3010	0.4589	0	0	0	1	1
0.85	0.4790	0.4998	0	0	0	1	1
0.90	0.7490	0.4338	0	0	1	1	1
0.95	0.9790	0.1435	1	1	1	1	1

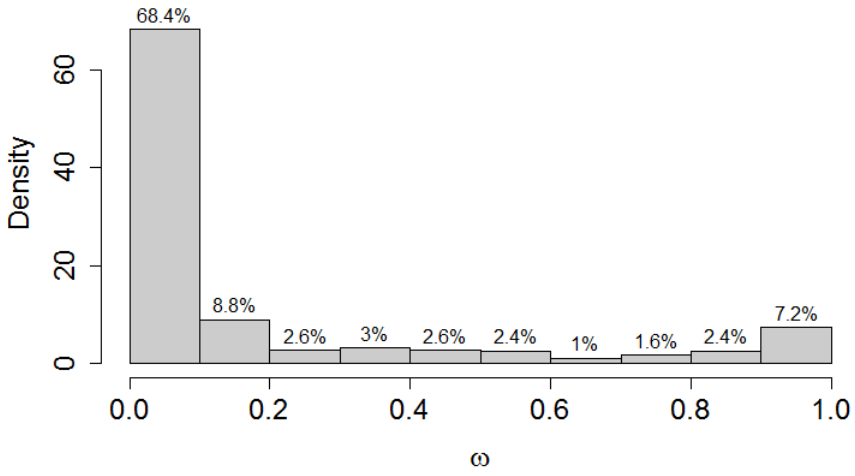

Figure 4.3.4: Histogram of posterior means for ω from Scenario III

In Scenario IV, the mean for treatment C is greater than the mean for treatment B, and both are greater than the means for the references. Using simulation scheme IV, we generate 500 data sets with a variance of one and sample sizes $n_D, n_{\bar{D}} = 200$. A summary of the resulting ranked posterior estimates for ω is included in Table 4.4 and a histogram of the posterior means for ω appears in Figure 4.3.5. Note that 98% of the the posterior means for ω are greater than 0.8920 which indicates that a second model is needed for Study 2.

Table 4.4: Summary of posterior estimates for ω from Scenario IV

1	0.01	0.1290	0.3354	0	0	0	0	1
2	0.25	1.0000	0.0000	1	1	1	1	1
3	0.50	1.0000	0.0000	1	1	1	1	1
4	0.75	1.0000	0.0000	1	1	1	1	1
5	0.80	1.0000	0.0000	1	1	1	1	1
6	0.85	1.0000	0.0000	1	1	1	1	1
7	0.90	1.0000	0.0000	1	1	1	1	1
8	0.95	1.0000	0.0000	1	1	1	1	1

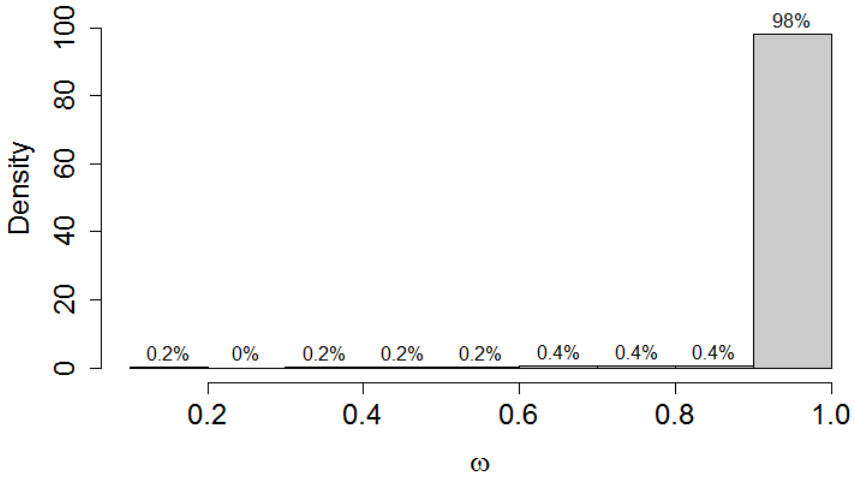

Figure 4.3.5: Histogram of posterior means for ω from Scenario IV

4.3.2 Extreme Value Data

Extreme value data are generated for each of Studies 1 and 2 according to the simulation schemes in Table 4.5 with scale of one and covariate distrbution Uniform$(0, 1)$. Study 1 compares treatments A and B, and Study 2 compares treatments A and C. Density plots for each treatment are given in the upper section of Figure 4.3.6. The corresponding ROC curves at covariate value $x = 0.5$ are given in the lower section of Figure 4.3.6, where the dotted lines represent a Unif$(0, 1)$ cdf to illustrate the case of identical populations. Note that the reference treatments for Studies 1 and 2 are generated independently from the specified normal distribution, although for plot readability only one reference density appears in Figure 4.3.6. We generate 500 datasets from each simulation scheme with sample

sizes $n_{D_1}, n_{\bar{D}_1} = 200$ and $n_{D_2}, n_{\bar{D}_2} = 200$ for Studies 1 and 2, respectively. We apply the Bayesian indicator model selection method to each dataset and record the posterior mean estimate of ω. Because we assume no prior knowledge regarding the necessity of a second model for Study 2, we use a Bernoulli(.5) prior for the indicator ω. Flat normal priors were given to the beta regression coefficients $\beta_{01}, \beta_{02}, \beta_{11}, \beta_{12}$, and a diffuse Gamma$(.001, .001)$ prior was specified for the scale parameter ϕ. MCMC simulations were run in JAGS with 10000 iterations and 5000 burn-in. Diagnostic plots are included in the appendix.

Table 4.5: Covariate dependent simulation means for extreme value data

Simulation Scheme	A_1, A_2	B	C
I: $A_1 = A_2 = B = C$	1 + 2x	1 + 2x	1 + 2x
II: $A_1 = A_2 = B < C$	1 + 2x	1 + 2x	1 + 3x
III: $A_1 = A_2 < B = C$	1 + 2x	1 + 3x	1 + 3x
IV: $A_1 = A_2 < B < C$	1 + 2x	1 + 3x	2 + 3x

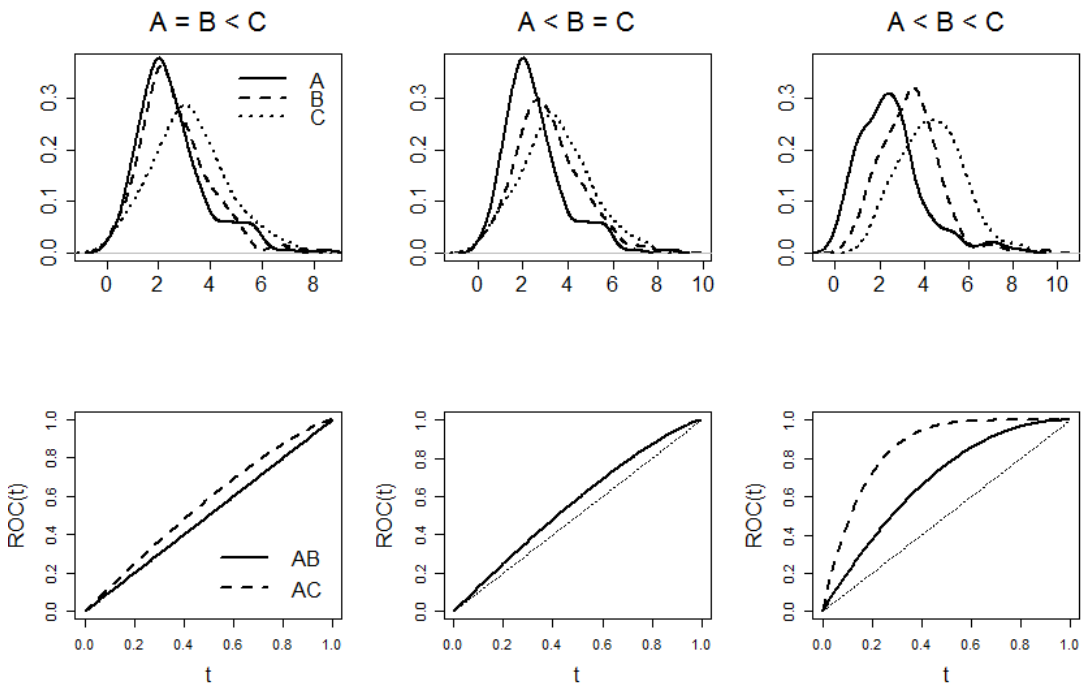

Figure 4.3.6: Top: Densities for simulation schemes II, III, and IV, Bottom: True extreme value ROC curves at covariate value 0.50

4.3.2.1 Results. Scenario I represents the null case and will be used to determine a suitable cut off value for the inclusion probability ω to be used in determining whether a separate beta regression model is needed for Study 2. Because the B and C populations are independently generated from a common extreme value distribution, we expect the resulting ROC_{AB} to be the same as ROC_{AC} and the indicator model selection method should yield a posterior mean for ω that is close to zero. We generate 500 data sets from simulation scheme I (see Table 4.5) with a scale of one and $n_D = n_{\bar{D}} = 200.$. A summary of the resulting ranked mean posterior estimates for ω at several quantiles is included in Table 4.6. Note that choosing the 95^{th} percentile (0.8040) as a cut-off probability yields a type I error rate of 5% under the null hypothesis that $\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$. For the remaining simulations, we would thus conclude that a second model is needed for Study 2 if the posterior mean ω exceeds 0.8040.

Table 4.6: Summary of posterior estimates for ω from Scenario I with extreme value data

Quantile	Mean	SD	2.5%	25%	50%	75%	97.5%
0.01	0.0000	0.0000	0	0	0	0	0
0.25	0.0090	0.0945	0	0	0	0	0
0.50	0.0250	0.1562	0	0	0	0	0
0.75	0.1080	0.3105	0	0	0	0	1
0.80	0.1480	0.3553	0	0	0	0	1
0.85	0.2510	0.4338	0	0	0	1	1
0.90	0.4610	0.4987	0	0	0	1	1
0.95	0.8040	0.3972	0	1	1	1	1

In Scenario II, both reference treatments as well as treatment B are independently generated from the same normal distribution. Treatment C is generated from a normal distribution with a higher mean (Table 4.1). Note that the covariate effect for treatment C is higher than the covariate effect for the other treatments, which implies that the probability of inclusion ω should be greater than the cut-off 0.8040 established in the null case.

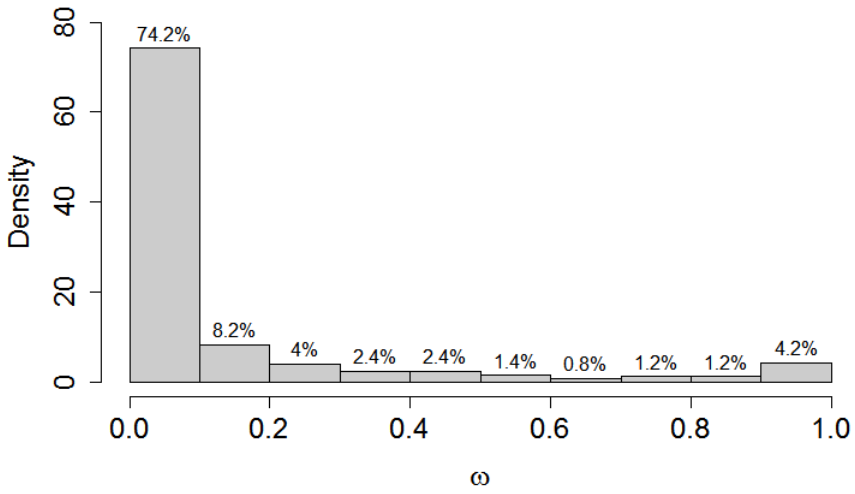

Figure 4.3.7: Histogram of posterior means for ω from Scenario I with extreme value data

We generate 500 data sets from simulation scheme II with a variance of one and sample sizes $n_D, n_{\bar{D}} = 200$. A summary of the resulting ranked posterior estimates for ω is included in Table 4.7 and a histogram of the posterior means for ω appears in Figure 4.3.8. Note that 34% of the posterior means for ω are less than the cut-off of 0.8404. The remaining 66% support the need for a second model for Study 2.

Table 4.7: Summary of posterior estimates for ω from Scenario II with extreme value data

Quantile	Mean	SD	2.5%	25%	50%	75%	97.5%
0.01	0.0010	0.0316	0	0	0	0	0
0.25	0.4720	0.4995	0	0	0	1	1
0.50	0.9990	0.0316	1	1	1	1	1
0.75	1.0000	0.0000	1	1	1	1	1
0.80	1.0000	0.0000	1	1	1	1	1
0.85	1.0000	0.0000	1	1	1	1	1
0.90	1.0000	0.0000	1	1	1	1	1
0.95	1.0000	0.0000	1	1	1	1	1

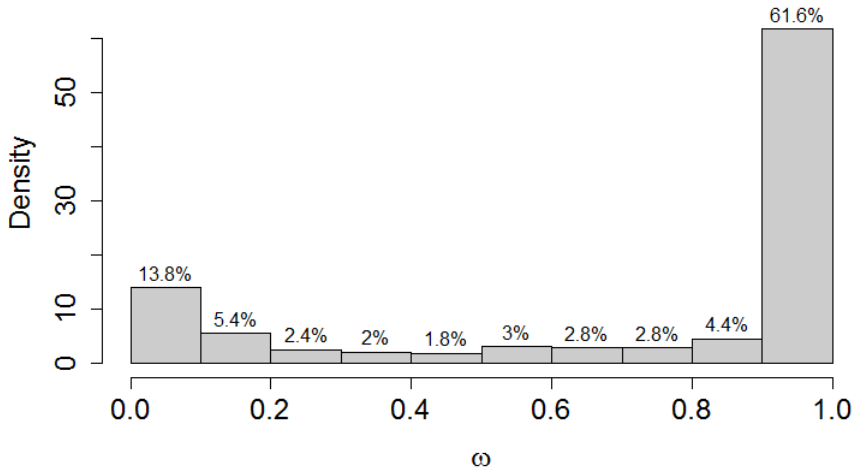

Figure 4.3.8: Histogram of posterior means for ω from Scenario II with extreme value data

For Scenario III, the reference distributions for both studies are the same. Treatments B and C are generated from an extreme value distribution with a location parameter higher than that of the reference (Table 4.5). We expect $ROC_{AB} = ROC_{AC}$ which implies that the probability of inclusion ω should be less than 0.8040 as established in the null case. We generate 500 data sets from simulation scheme III with a scale parameter of one and sample sizes $n_D, n_{\bar{D}} = 200$. A summary of the resulting ranked posterior estimates for ω is included in Table 4.8 and a histogram of the posterior means for ω appears in Figure 4.3.9. Note that approximately 7.8% of the 500 posterior means for ω exceeded 0.8040.

Table 4.8: Summary of posterior estimates for ω from Scenario III with extreme value data

Quantile	Mean	SD	2.5%	25%	50%	75%	97.5%
0.01	0.0010	0.0316	0	0	0	0	0
0.25	0.0110	0.1044	0	0	0	0	0
0.50	0.0340	0.1813	0	0	0	0	1
0.75	0.1610	0.3677	0	0	0	0	1
0.80	0.2510	0.4338	0	0	0	1	1
0.85	0.4430	0.4970	0	0	0	1	1
0.90	0.7020	0.4576	0	0	1	1	1
0.95	0.9880	0.1089	1	1	1	1	1

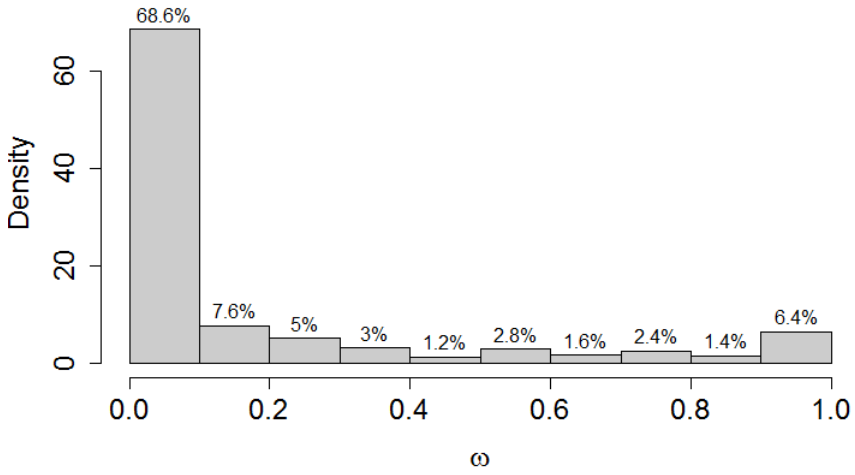

Figure 4.3.9: Histogram of posterior means for ω from Scenario III with extreme value data

In Scenario IV, the location parameter for treatment C is greater than the location parameter for treatment B, and both are greater than those for the references. Using simulation scheme IV, we generate 500 data sets with a scale parameter of one and sample sizes $n_D, n_{\bar{D}} = 200$. All of the posterior means for ω were equal to one, a result expected from the density plots in Figure 4.3.6.

4.3.3 Discussion

The preceding simulations provide a preliminary exploration of the performance of Bayesian indicator model selection used in conjunction with Bayesian beta regression. Opportunities for further investigation are numerous, including exploring the effect of different variances, examining different magnitudes of covariate effect across the treatments, and identifying an amount of separation among the treatments that might yield contradictory results in terms of the inclusion probability ω. In the null cases of the normal and extreme value simulations, we noted that sample size is highly influential in determining a cut-off probability for ω. Given a simulation context, we have the advantage of choosing the distribution, sample size, and variance. Extension of the methods introduced in this chapter to real data poses a few additional challenges. To establish a cut-off value, one would need

to set up a null case simulation, requiring an estimate of the variances and an initial idea of the separation between the treatments. Future work will examine these ideas.

CHAPTER FIVE

The Deviance Information Criteria and Power Prior Specification

In this chapter, we summarize the findings of a simulation project that investigates the role of the deviance information criterion (DIC) in choosing the value of the power prior parameter. We provide an introduction to power priors and a brief overview of their appearance in the literature. This chapter focuses specifically on the use of power priors in a generalized linear model (GLM) context. We examine the performance of the DIC as a guide for power parameter specification, through simulation studies of normal linear regression and logistic regression models.

5.1 Introduction to Power Priors

In the Bayesian paradigm, informative prior elicitation remains a widely studied and important topic. The task of quantifying prior information and building a suitable prior distribution often proves difficult, particularly in settings that involve large amounts of historical data. The power prior first formalized by Ibrahim and Chen (2000) offers a systematic procedure for building an informative prior in the presence of historical data. With its convenient theoretical properties and relative ease of construction and computation, the power prior has gained popularity as a suitable general class of priors that can easily be applied to various regression models including the GLM. Given historical data D_0 and current data D, the power prior is constructed by raising the likelihood of D_0 to the power a_0, where $0 \leq a_0 \leq 1$ is known as the power parameter. The value of a_0 controls the degree of borrowing from the historical data with $a_0 = 0$ yielding no borrowing, and $a_0 = 1$, full borrowing. That is, the power parameter quantifies the amount of heterogeneity between the historical and current data.

Advocates of the power prior approach argue that the method provides an objective way to elicit an informative prior, given that the degree of informativeness is largely dictated by the historical data. However, the process of choosing the value of the power parameter a_0 is highly subjective and remains a topic of debate. Ibrahim et al. (2015) suggest several methods for choosing the value of the hyperparameter a_0, including specifying a proper prior distribution as in a hierarchical model. Establishing a hyperprior, however, increases computational difficulty and closed forms are no longer possible. As an alternative Ibrahim et al. (2015) propose fixing a_0 and using a model selection criteria such as the deviance information criterion (DIC) to observe sensitivity to the choice of a_0. In this chapter, we explore through simulation the use of the DIC as a guide for choosing an appropriate value of a_0 for analysis of generalized linear models.

5.1.1 Formulation

We begin with the basic formulation of the power prior as proposed by Ibrahim and Chen (2000). Given historical data D_0, current data D, and a vector of parameters θ, we have that the corresponding likelihoods are $L(\theta|D_0)$ and $L(\theta|D)$, respectively. The power prior is then constructed by

$$\pi(\theta|D_0, a_0) \propto L(\theta|D_0)^{a_0}\pi_0(\theta), \tag{5.1}$$

where $\pi_0(\theta)$ is the prior specified for θ before observing D_0, and $a_0 \in [0, 1]$. The posterior distribution for θ is

$$\pi(\theta|D, D_0, a_0) \propto L(\theta|D)L(\theta|D_0)^{a_0}\pi_0(\theta). \tag{5.2}$$

Note that the power parameter a_0 controls the heaviness of the tails of the prior for θ. Smaller values of a_0 yield heavier tails and give less weight to the historical data. Taking $a_0 = 1$ reduces (5.1) to Bayes' theorem with the resulting posterior distribution acting as the prior for the current data. Setting $a_0 = 0$ yields $\pi(\theta|D_0, a_0) \equiv \pi_0(\theta)$ and thus removes all influence of historical data on the prior for θ. The ability to control a_0 is especially important when there exists heterogeneity between the historical and current studies or a

notable difference in sample sizes. Given a proper initial prior for $\pi_0(\theta)$, the power prior given in (5.1) will also be proper. Ibrahim et al. (2015) note that the power prior has in advantage over other priors in that it shares the properties of the likelihood function. In particular, likelihood theory can be used to obtain propriety of the power prior for various models including logistic regression models (Ibrahim et al., 1998) and GLMs in general (Chen et al., 2000).

5.1.2 Power Priors for Generalized Linear Models

The GLM power prior derived by Ibrahim et al. (2015) is formulated as follows. Suppose y_i for $i = 1, ..., n$ is the response variable for the current study and $\mathbf{x}_i$ is a p-dimensional vector of covariates. Then the current study data can be written as $D = \{(y_i, \mathbf{x}_i), i = 1, ..., n\} \equiv (n, \mathbf{y}, X)$, where $y = (y_1, ..., y_n)'$ and $X = (\mathbf{x}_1', ..., \mathbf{x}_n')'$. In this chapter, following the notation of Ibrahim et al. (2015), we assume a GLM for the response with density of exponential class for y_i given $\mathbf{x}_i$ such that

$$f(y_i|\mathbf{x}_i, \theta_i, \tau) = \exp\{\alpha_i^{-1}(\tau)(y_i\theta_i - \psi(\theta_i)) + \phi(y_i, \tau)\} \text{ for } i = 1, ..., n,$$

where θ_i is the GLM canonical parameter, τ is the scale parameter, and ψ, ϕ determine the specific exponential family such as the normal or Poisson. Note that $\alpha_i(\tau)$ is often expressed as a function of know weights w_i such that $\alpha_i(\tau) = \tau^{-1}w_i^{-1}$. In the GLM formulation, the $\theta_i's$ are assumed to satisfy $\theta_i = h(\eta_i)$ and $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$, where h is the link function and $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients. We can then rewrite $f(y_i|\mathbf{x}_i, \theta_i, \tau)$ as

$$f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \exp\{\alpha_i^{-1}(\tau)[y_i h(\mathbf{x}_i'\boldsymbol{\beta}) - \psi(h(\mathbf{x}_i'\boldsymbol{\beta}))] + \phi(y_i, \tau)\} \text{ for } i = 1, ..., n.$$

The likelihood of the current data is thus given by $L(\boldsymbol{\beta}|D) = \prod_{i=1}^{n} f(y_i|\mathbf{x}_i, \boldsymbol{\beta})$.

Similarly, let y_{0i} for $i = 1, ..., n_0$ be the response variable for the historical study and $\mathbf{x}_{0i}$ be a p-dimensional vector of covariates. Then the historical data can be written as $D_0 = \{(y_{0i}, \mathbf{x}_{0i}), i = 1, ..., n_0\} \equiv (n_0, \mathbf{y}_0, X_0)$, where $y_0 = (y_{01}, ..., y_{0n})'$ and $X = (\mathbf{x}_{01}', ..., \mathbf{x}_{n0}')'$.

We assume a GLM for the response such that

$$f(y_{0i}|\mathbf{x}_{0i}, \boldsymbol{\beta}) = \exp\{\alpha_{0i}^{-1}(\tau)[y_{0i}h(\mathbf{x}_{0i}'\boldsymbol{\beta}) - \psi(h(\mathbf{x}_{0i}'\boldsymbol{\beta}))] + \phi(y_{0i}, \tau)\} \text{ for } i = 1, ..., n_0,$$

where $\alpha_{0i} = \tau^{-1}w_{0i}^{-1}$. The likelihood for the historical data is $L(\boldsymbol{\beta}|D_0) = \prod_{i=1}^{n} f(y_i|\mathbf{x}_i, \boldsymbol{\beta})$. We thus have that the power prior for the GLM with fixed a_0 is

$$\pi(\boldsymbol{\beta}|D_0, a_0) \propto \{L(\boldsymbol{\beta}|D_0)\}^{a_0}\pi_0(\boldsymbol{\beta}) \text{ for } a \in [0, 1],$$

where $\pi_0(\boldsymbol{\beta})$ is the initial prior for $\boldsymbol{\beta}$. The conditional power prior for the GLM with a random a_0 is given by

$$\pi(\boldsymbol{\beta}|D_0, a_0) = \frac{[L(\boldsymbol{\beta}|D_0)]^{a_0}\pi_0(\boldsymbol{\beta})}{\int [L(\boldsymbol{\beta}|D_0)]^{a_0}\pi_0(\boldsymbol{\beta})d\boldsymbol{\beta}}.$$

5.1.3 Specifying the Power Parameter

Choosing a value for a_0 remains an important issue in the use of the power prior. Specifying a proper prior distribution for a_0 such as a beta distribution, presents an easy solution but is more intensive computationally than choosing a fixed value of a_0. As an alternative, Ibrahim et al. (2015) suggest fixing a_0 and using model selection criterion such as the deviance information criterion (DIC) to establish a starting value for analysis, noting that several sensitivity studies should be performed once the starting value has been determined (Ibrahim et al., 2003). The DIC for the generalized linear model is based on the deviance function for GLMs defined as

$$\begin{aligned}\text{Dev}(\boldsymbol{\beta}) &= -2\sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) \\ &= -2\sum_{i=1}^{n} \{\alpha_i^{-1}(y_i h(\mathbf{x}_i'\boldsymbol{\beta}) - \psi(h(\mathbf{x}_i'\boldsymbol{\beta})) + \phi(y_i)\}.\end{aligned}$$

From Spiegelhalter et al. (2002), the DIC for a_0 is $\text{DIC}(a_0) = \text{Dev}(\tilde{\beta}) + 2p_D(a_0)$, where $\tilde{\boldsymbol{\beta}} = E[\boldsymbol{\beta}|D, D_0, a_0]$ and $p_D(a_0) = E[\text{Dev}(\boldsymbol{\beta})|D, D_0, a_0 - \text{Dev}(\tilde{\beta})$. Ibrahim et al. (2015) show that the optimal value of a_0 is then

$$a_{0,DIC}^{opt} = \underset{0 \le a_0 \le 1}{\arg\min}\, \text{DIC}(a_0).$$

The performance of the DIC as a guide in choosing the value of a_0 for a GLM power prior is evaluated through simulation study.

5.2 Simulation Study

Ibrahim et al. (2015) examine the empirical performance of the normal linear regression power prior with random a_0 versus fixed a_0 considering two scenarios, one in which the historical and current data are similar and one in which they are different. We follow a similar scheme and examine the relationship between the DIC and a_0 for a normal linear regression model and a logistic regression model. Before considering power priors for GLMs, we provide a simple example using data from an exponential distribution and we show how the value of a_0 influences the posterior distribution of the exponential parameter.

5.2.1 Single Exponential Sample

The exponential model for the historical is $y_{0_i} \sim \text{Exp}(\lambda_0)$ for $i = 1, ..., n_0$ and the model for the current data is $y_j \sim \text{Exp}(\lambda)$ for $j = 1, ..., n$. The two scenarios considered are summarized in Table 5.1. Density plots of the historical and current data for each scenario are included in Figure 5.2.1.

Table 5.1: Exponential Simulation Scenarios

Scenario	Historical	Current	Sample Size
I	$\lambda_0 = 2.5$	$\lambda = 2$	$n = 250, n_0 = 500$
II	$\lambda_0 = 6$	$\lambda = 2$	$n = 250, n_0 = 500$

We generate 500 simulated datasets for each scenario and calculate the posterior means and standard deviations for λ using the power prior with fixed values of $a_0 = 0.10, 0.25, 0.50, 0.75, 0.90$. We also recorded the DIC for each of the 500 models. For Scenario I, little heterogenity exists between the historical and current data though there is a difference in sample size. Given the similarity, we might expect the DIC to indicate a higher value of a_0 as a starting point for analysis. In Scenario II, the historical data is much

different from the current, and we anticipate that the lowest DIC will be associated with a small value of a_0.

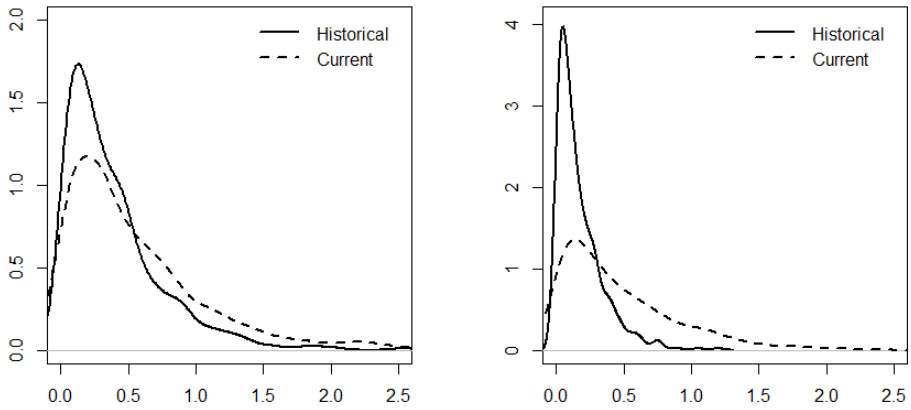

Figure 5.2.1: Density plots of historical and current exponential data for Scenario I (left) and Scenario II (right)

For both scenarios, the average DIC for the 500 models is reported in Table 5.2 for each value of a_0 along with the posterior means for λ. Recall that a lower DIC value indicates better fit of the model. The lowest average DIC for Scenario I is associated with $a_0 = .10$, which indicates that the historical data should have little influence through the power prior despite the similarity between the historical and current data. Note that the average DIC values for Scenario I lie in a much narrower range than the average DIC values for Scenario II. As anticipated given the distinct difference between historical and current data in Scenario II, the lowest DIC value is associated with $a_0 = .10$. Posterior densities for the exponential parameter λ are given in Figure 5.2.2. We observe that posterior distribution of λ is highly influenced by the value of a_0.

Table 5.2: Average values of the DIC and posterior means for Scenarios I and II

Scenario I			Scenario II		
a_0	DIC	λ	a_0	DIC	λ
0.10	152.264	2.084	0.10	155.540	2.267
0.25	152.928	2.156	0.25	168.742	2.589
0.50	154.403	2.235	0.50	198.385	3.019
0.75	155.736	2.285	0.75	228.787	3.353
0.90	156.420	2.307	0.90	245.980	3.520

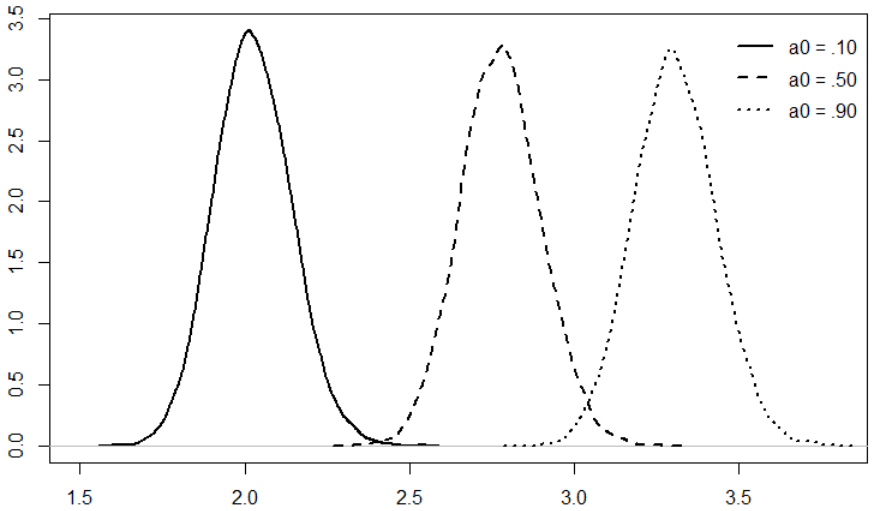

Figure 5.2.2: Posterior densities of λ by value of a_0

5.2.2 Normal Linear Regression

To illustrate performance of the DIC in the GLM context, we turn our attention to normal linear regression. We specify the models for the current and historical data as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$y_{i0} = \beta_{00} + \beta_{10} x_{i0} + \epsilon_{i0}$$

where $\epsilon_i \overset{iid}{\sim} N(0,1), i = 1, ..., n$, and $\epsilon_{i0} \overset{iid}{\sim} N(0,1), i = 1, ..., n_0$. Note that ϵ_i is assumed independent of the ϵ_{i0}. The covariates x_i and x_{i0} are assumed to be independently generated from a standard normal distribution. The scenarios considered appear in Table 5.3. For each case, we generated 500 datasets, and for each dataset, we calculated the posterior

means and standard deviations for β_0 and β_1 using the power prior with fixed values of $a_0 = 0.10, 0.25, 0.50, 0.75, 0.90$. We also recorded the DIC for each of the 500 models.

For Scenario I, the historical and current datasets are quite similar, β_{10} is slightly lower than β_1 and the sample sizes are identical. Given the similarity between the datasets, we might reasonably expect the DIC to indicate a higher value of a_0 as a suggestion for the power parameter. The average DIC for the 500 models is reported in Table 5.4 for each

Table 5.3: Scenarios for Normal Linear Regression Simulations

Scenario	Historical		Current		Sample Size
I	$\beta_{00} = 1$	$\beta_{10} = 1.75$	$\beta_0 = 1$	$\beta_1 = 2$	$n = n_0 = 500$
II	$\beta_{00} = 1$	$\beta_{10} = 1.75$	$\beta_0 = 1$	$\beta_1 = 2$	$n = 250, n_0 = 500$
III	$\beta_{00} = 1$	$\beta_{10} = 2.25$	$\beta_0 = 1$	$\beta_1 = 2$	$n = n_0 = 500$
IV	$\beta_{00} = 1$	$\beta_{10} = 2.25$	$\beta_0 = 1$	$\beta_1 = 2$	$n = 250, n_0 = 500$

value of a_0 along with the posterior means for β_0, β_1, and σ. Observe that the lowest average DIC occurs for $a_0 = 0.10$, which is contrary to what we expected. For further investigation, we include the distribution of the DIC values for each value of a_0 in Figure 5.2.3. We note that the distributions exhibit no extreme behavior across the a_0 values.

Table 5.4: Average values of the DIC and posterior means for Scenarios I and II

Scenario I					Scenario II				
a_0	DIC	β_0	β_1	σ	a_0	DIC	β_0	β_1	σ
0.10	1426.036	1.049	1.996	0.954	0.10	720.522	1.104	2.000	0.910
0.25	1447.186	1.124	1.996	0.888	0.25	770.525	1.255	1.999	0.790
0.50	1536.926	1.249	1.996	0.791	0.50	1066.970	1.505	1.999	0.609
0.75	1728.800	1.373	1.998	0.703	0.75	2730.522	1.756	1.998	0.407
0.90	1927.206	1.450	1.996	0.648	0.90	1513.468	1.224	1.997	0.809

In Scenario II, the historical and current data sets are again similar as in Scenario I, but the sample size for the current dataset is half that of the historical. The average DIC for the 500 models is reported in Table 5.4 for each value of a_0 along with the posterior means for β_0, β_1, and σ. Again, we observe that the lowest average DIC occurs for $a_0 = 0.10$,

which is unexpected given the similarity in the data. We include posterior density plots of the regression coefficients β_0 and β_1 to illustrate the influence of a_0 (Figure 5.2.4).

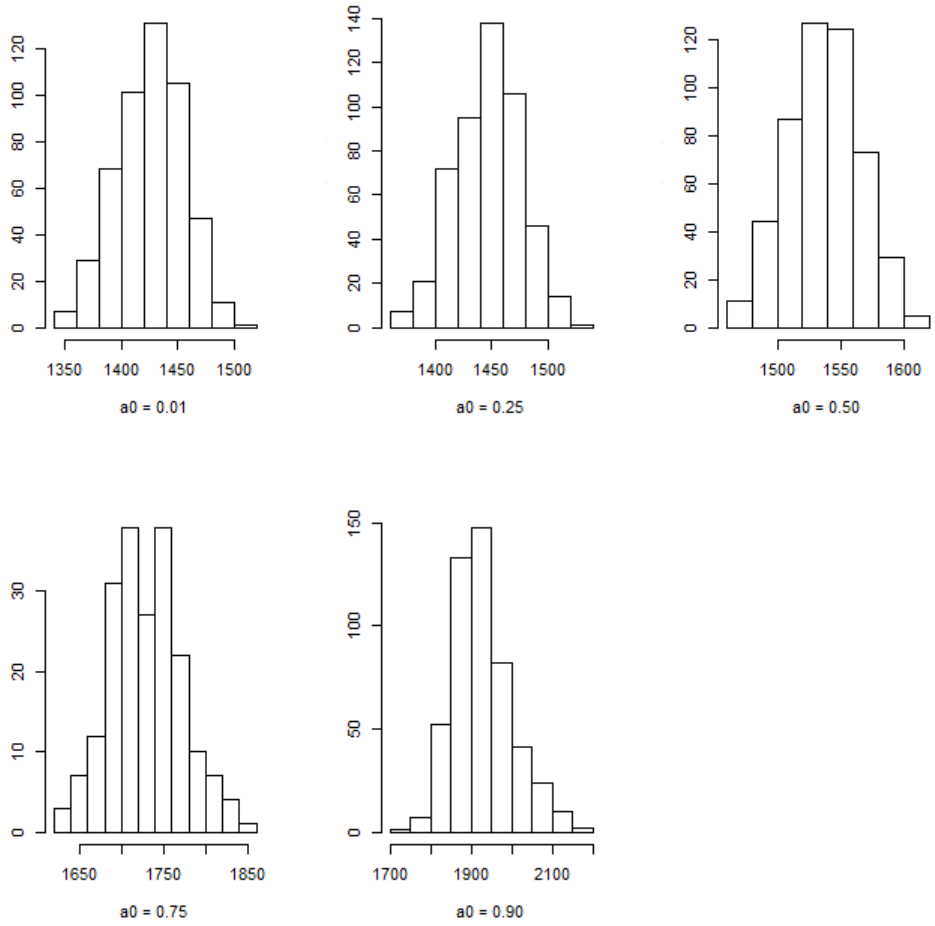

Figure 5.2.3: Distribution of DIC values for Scenario I

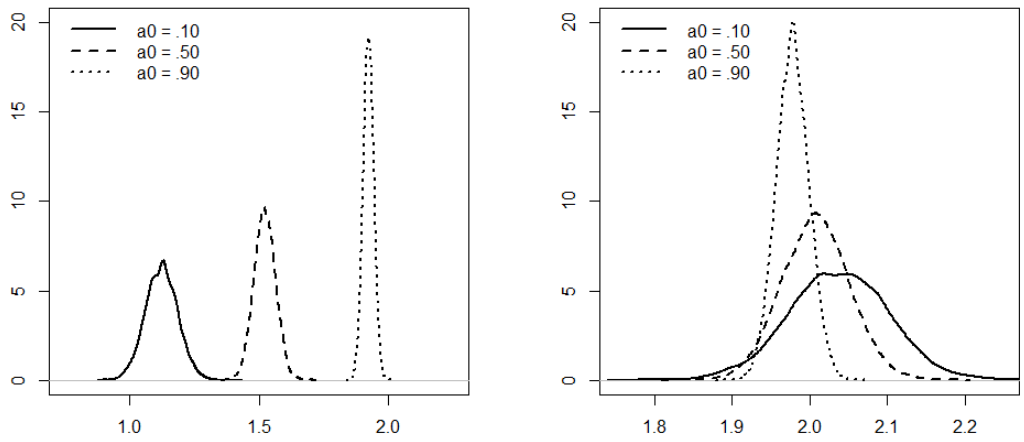

Figure 5.2.4: Posterior densities of β_0 (left) and β_1 (right) for different a_0 values

Scenario III provides an example in which the historical and current datasets are again similar, but the historical β_{10} is now higher than the current β_1 (Table 5.5). The sample sizes are identical. Again the lowest average DIC occurs at $a_0 = 0.10$, with the highest average DIC occurring for $a_0 = 0.90$. Scenario IV is similar to III, with the exception of the difference in sample sizes. Note that the lowest average DIC value again occurs at $a_0 = 0.10$ (Table 5.6).

Table 5.5: Average values of the DIC and posterior means for Scenario III

a_0	DIC	β_0	β_1	σ
0.10	1426.042	1.049	1.996	0.954
0.25	1447.204	1.124	1.996	0.888
0.50	1536.822	1.249	1.996	0.791
0.75	1729.820	1.375	1.996	0.701
0.90	1926.540	1.450	1.996	0.648

In the preceding scenarios, the data is simulated such that the error satisfies $\epsilon_i \overset{iid}{\sim} N(0,1), i = 1, ..., n$, and $\epsilon_{i0} \overset{iid}{\sim} N(0,1), i = 1, ..., n_0$. We now briefly explore the effects of increasing the error variance which we have been denoting as σ. Continuing with Scenario IV, we change the error variance so that $\sigma = 1.75$ and observe the behavior of

Table 5.6: Average values of the DIC and posterior means for Scenario IV

a_0	DIC	β_0	β_1	σ
0.10	720.527	1.104	2.000	0.910
0.25	770.568	1.255	1.999	0.790
0.50	1067.309	1.505	1.999	0.609
0.75	2753.956	1.756	1.998	0.407

the DIC. As seen in Table 5.7, the lowest average DIC still occurs at $a_0 = 0.10$. We next increase the error to $\sigma = 3$. Again we observe that $a_0 = 0.10$ yields the lowest average DIC (Table 5.7). We thus note that increasing the variance in the data appears to have no effect on the suggested a_0 value as indicated by the DIC.

Table 5.7: Average DIC values and posterior means for different values of σ

Scenario IV ($\sigma = 1.75$)					Scenario IV ($\sigma = 3$)				
a_0	DIC	β_0	β_1	σ	a_0	DIC	β_0	β_1	σ
0.10	1000.306	1.095	1.976	1.606	0.10	1267.933	1.098	1.977	2.751
0.25	1028.092	1.245	1.977	1.423	0.25	1289.640	1.248	1.977	2.451
0.50	1138.330	1.496	1.977	1.192	0.50	1360.307	1.499	1.977	2.098
0.75	1374.210	1.747	1.978	1.004	0.75	1475.673	1.750	1.977	1.841
0.90	1625.540	1.898	1.979	0.900	0.90	1570.947	1.901	1.977	1.713

5.3 Logistic Regression

As a second example of GLM power priors and the DIC, we consider a logistic regression model with two binary covariates. We specify the models for the historical and current data as

$$\text{logit}(y_{i0}) = \beta_{00} + \beta_{10}x_{10i} + \beta_{20}x_{20i}, \text{ for } i = 1, ..., n_0$$

$$\text{logit}(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \text{ for } i = 1, ..., n,$$

where the intercepts β_{00} and β_0 are taken to be zero. The covariates x_{10} and x_1 are from a Bernoulli(0.5) distribution, while x_{20} and x_2 are from a Bernoulli (0.35) distribution. We consider the models as summarized in Table 5.8. For each scenario, we generated 500 simulated datasets, and for each dataset, we calculated the posterior means

and standard deviations for $\beta_0, \beta_1,$ and β_2 using the power prior with fixed values of $a_0 = 0.10, 0.25, 0.50, 0.75, 0.90$. We also recorded the DIC for each of the 500 models.

Table 5.8: Scenarios for Logistic Regression Simulations

Scenario	Historical		Current		Sample Size
I	$\beta_{10} = 0$	$\beta_{20} = 0.21$	$\beta_1 = 1$	$\beta_2 = 0.25$	$n_0 = 500, n = 100$
II	$\beta_{10} = 1.5$	$\beta_{20} = 0.21$	$\beta_1 = 1$	$\beta_2 = 0.25$	$n_0 = 500, n = 100$
III	$\beta_{10} = 1.0$	$\beta_{20} = 0.25$	$\beta_1 = 1$	$\beta_2 = 0.25$	$n_0 = 500, n = 100$

For Scenario I, the current and historical data sets differ in that the first covariate has no effect in the historical data. From Table 5.9, we observe that $a_0 = 0.10$ yields the lowest average DIC value of the a_0 values tested, although there is little difference in the average DICs across the range of a_0. At the iteration level, we have that $a_0 = 0.10$ produced the lowest individual DIC value for 92% of the simulated data sets. In Figure 5.3.1, we include plots of the posterior densities for β_1 and β_2 to illustrate the influence of a_0.

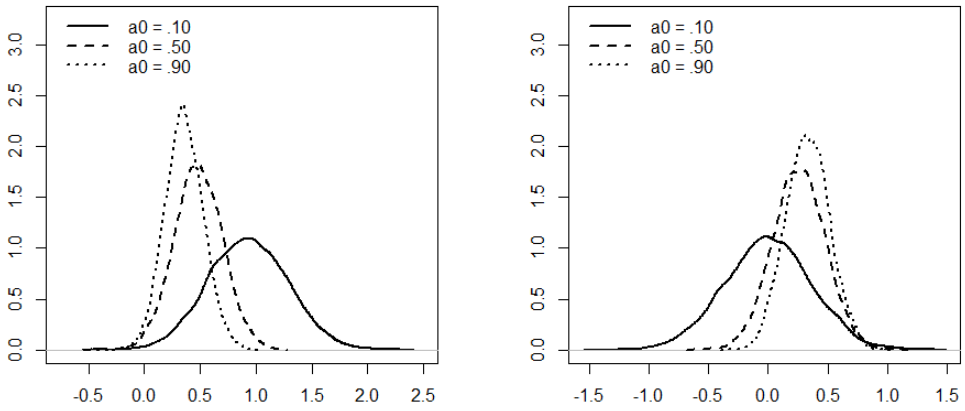

Figure 5.3.1: Posterior densities of β_1 (left) and β_2 (right) for different values of a_0

Table 5.9: Average values of the DIC and posterior means for Scenario I

a_0	DIC	β_0	β_1	β_2
0.10	630.288	0.007	0.900	0.238
0.25	631.855	0.007	0.776	0.231
0.50	635.836	0.007	0.635	0.224
0.75	639.893	0.008	0.537	0.220
0.90	642.179	0.008	0.492	0.218

For Scenario II, the lowest DIC values is associated with $a_0 = 0.25$ (Table 5.10). Given the difference in historical and current data for Scenario III, we might have expected the DIC to recommend a higher value of a_0. Note that at the iteration level, $a_0 = 0.25$ yields the lowest DIC only 33% of the 500 datasets with $a_0 = 0.50$ coming in second at 24%. In Scenario III, the historical and current data are generated from the same distributions. We thus anticipate that the lowest DIC will be associated with $a_0 = 0.10$. From Table 5.10, we note the opposite. The lowest average DIC occurs for $a_0 = 0.90$, and at the iteration level, $a_0 = 0.90$ yields the lowest average DIC for 90% of the 500 datasets.

Table 5.10: Average values of the DIC and posterior means for Scenarios II and III

Scenario II					Scenario III				
a_0	DIC	β_0	β_1	β_2	a_0	DIC	β_0	β_1	β_2
0.10	629.868	0.006	1.049	0.241	0.10	629.777	0.006	1.009	0.245
0.25	629.683	0.006	1.097	0.236	0.25	629.307	0.006	1.007	0.244
0.50	629.977	0.006	1.159	0.230	0.50	628.917	0.006	1.006	0.244
0.75	630.495	0.007	1.203	0.227	0.75	628.768	0.006	1.005	0.244
0.90	630.850	0.007	1.225	0.225	0.90	628.734	0.006	1.005	0.244

5.4 Discussion

The preceding simulations show that the DIC alone does not appear to provide a reliable method for choosing the value of the power parameter a_0, particularly in the GLM context. Given the influence of a_0 on the posterior distributions of the regression coefficients, the choice of a_0 is highly influential upon the estimates resulting from the model. Further exploration may reveal specific instances in which the DIC performs well in terms

of proposing a value for a_0. For those who utilize the GLM power prior, we reiterate that the use of model selection criterion as a guide for selecting a_0 should be followed by sensitivity analysis to ensure that the value of a_0 reflects the heterogeneity in the data.

CHAPTER SIX

Conclusion

In this dissertation, we considered ROC regression methods to determine the effect of covariates on a test's ability to distinguish between two populations. We examined the parametric and semi-parametric ROC regression approaches which utilize GLMS for binary data based on the placement value. The use of placement values in the pre-existing approaches along with expression of the ROC as the cdf of the placement values motivated the proposal of a new beta ROC regression method. We developed the beta approach in Chapter Two and showed that it is not only easy to implement, but also removes the dependency on a binary variable. The beta method was compared to the parametric and semiparametric models with a simulation study using data from both the normal and extreme value distributions. We also examined performance of the parametric and beta methods through application to a DME study. In Chapter Three, we extended the parametric and beta regression approaches to the Bayesian paradigm using hierarchical modeling. The performance of the Bayesian extensions was compared through simulation study for both the normal and extreme value distributions and through application to the DME study from Chapter Two. In Chapter Four, we introduced an application of the beta approach to a simple network meta-analysis problem through the use of a Bayesian variable selection method. We developed a variable selection approach to be used in conjunction with the beta ROC-regression methodology and evaluated its performance through simulation study. Chapter Five provided an introduction to power priors and examined the role of the deviance information criterion as a guide for choosing the value of the power parameter. We were particularly interested in the GLM context and showed through simulation of normal linear and logistic regression models that the DIC is not always a reliable indicator of an appropriate power parameter value.

APPENDICES

APPENDIX A

Convergence Diagnostics for Bayesian Methods

A.1 Diagnostics for Chapter Three

1.1.1 Bayesian Beta Regression

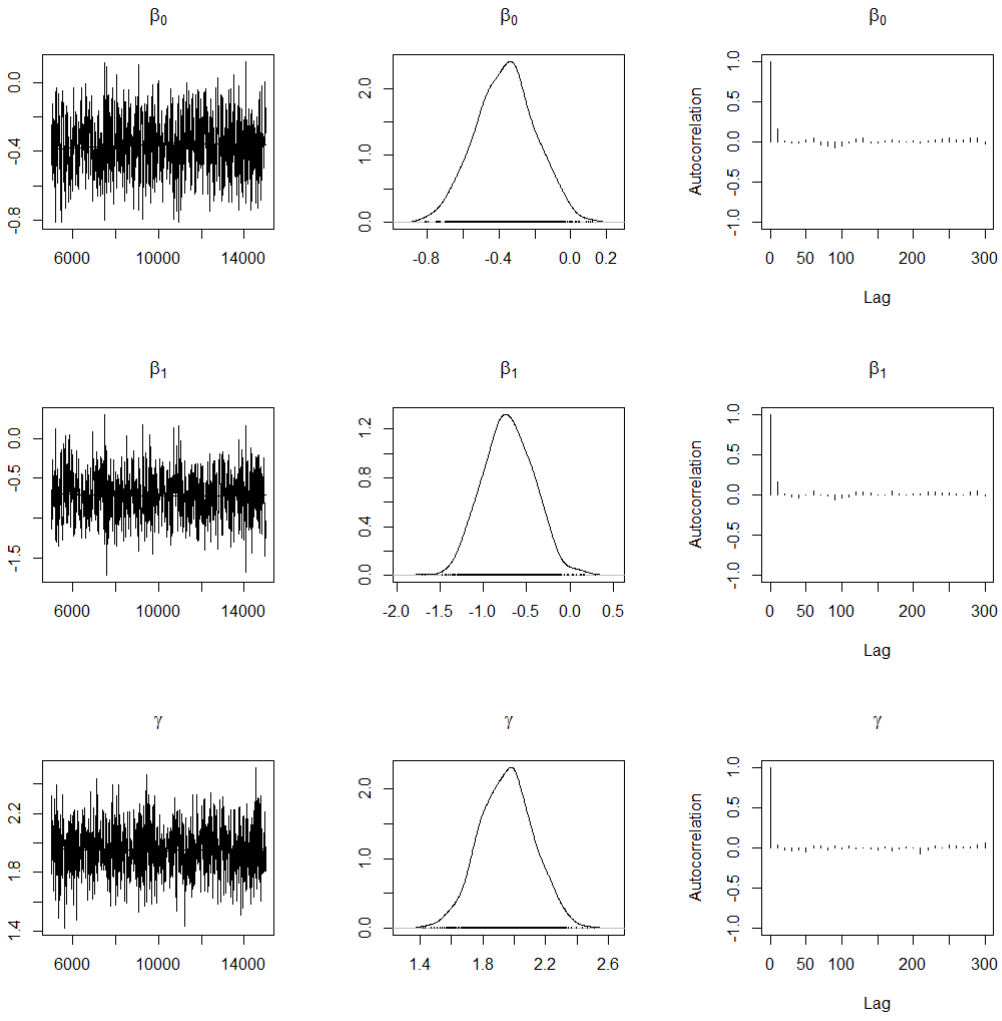

Figure A.1.1: Diagnostic plots from binormal data example for Bayesian beta ROC regression.

1.1.2 Bayesian Parametric Regression (Diffuse Priors)

The following diagnostic plots result from MCMC simulation in JAGS. Normal(0, .01) priors were used for all regression coefficients. The trace plots for both γ_1 and γ_2 exhibit a spike for the first iteration before leveling for the remainder of the plot. Simulations with Normal(0, .1) priors for all regression coefficients were also performed and the resulting trace plots exhibited similar behavior. Minimal difference in the posterior means existed for the regression coefficients for the Normal(0, .01) versus Normal(0, .1) priors.

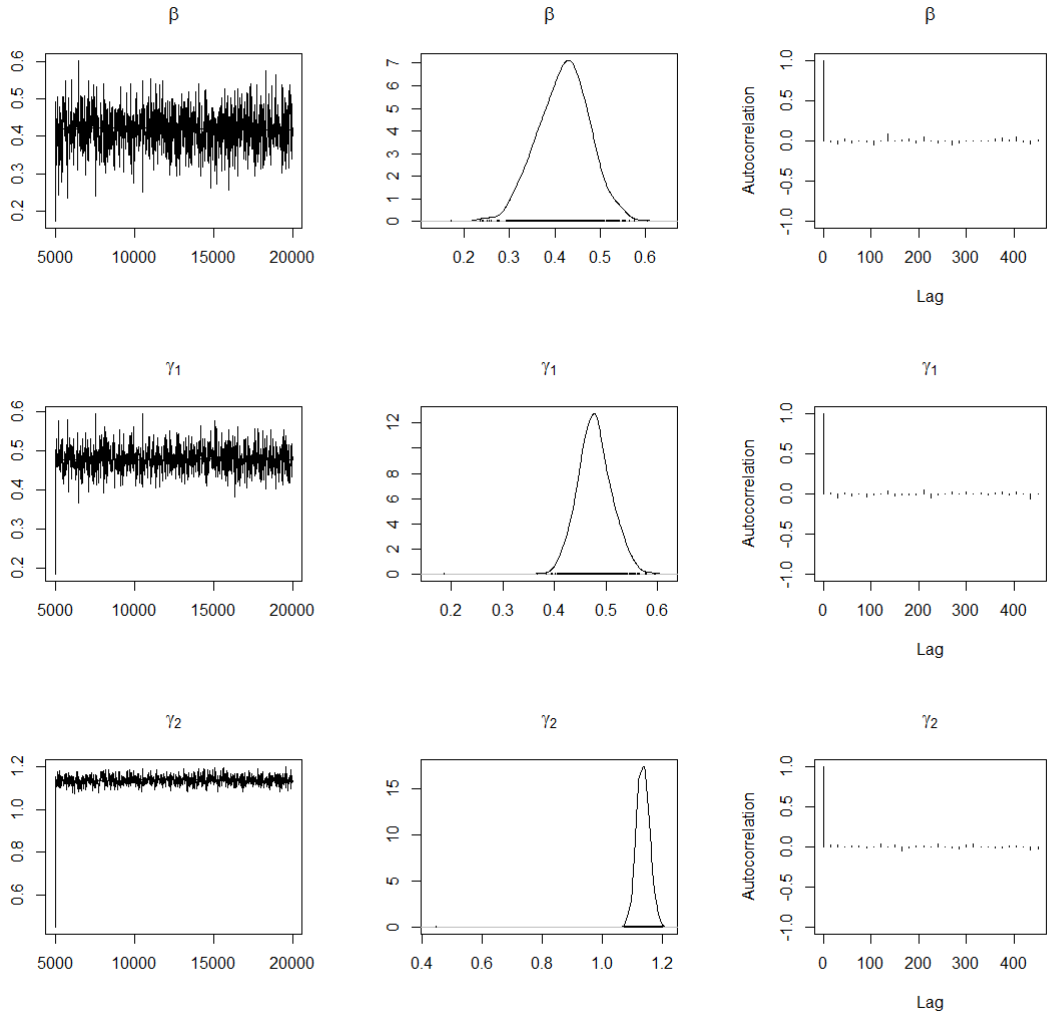

Figure A.1.2: Diagnostic plots from binormal data example for Bayesian parametric ROC regression.

1.1.3 Bayesian Parametric Regression (More Informative Priors)

Using slightly more informative priors, $\beta \sim \text{Normal}(1, .1), \gamma_1 \sim \text{Normal}(0, .1), \gamma_2 \sim \text{Normal}(1, .1)$, we no longer observe a spike in the trace plots. For analysis, we continue with the more diffuse $\text{Normal}(0, .01)$ priors given that the calculation of the ROC is based on the posterior means of β, γ_1, and γ_2 which do not change drastically for diffuse versus more informative priors, and that we wish the data to have heavy influence on the posterior for this example. Future work will investigate specification of informative priors.

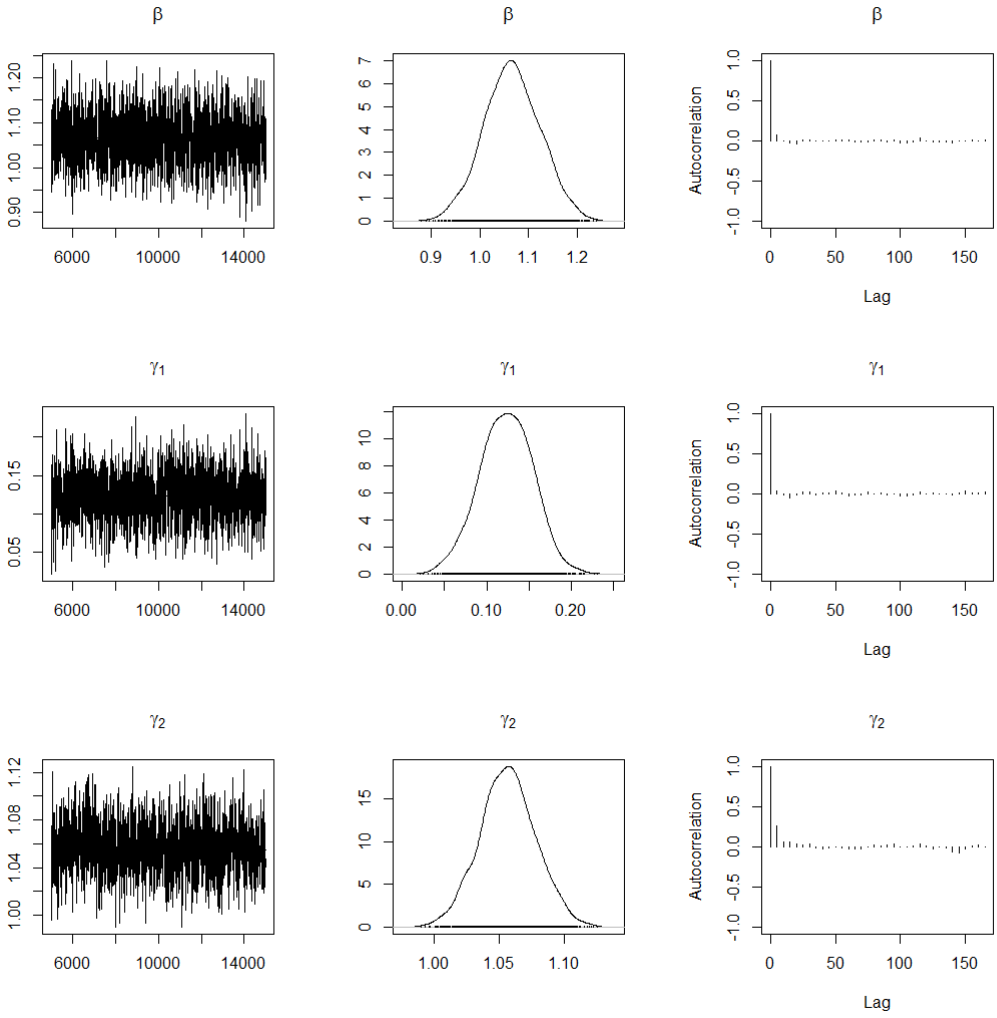

Figure A.1.3: Diagnostic plots from binormal data example for Bayesian parametric ROC regression.

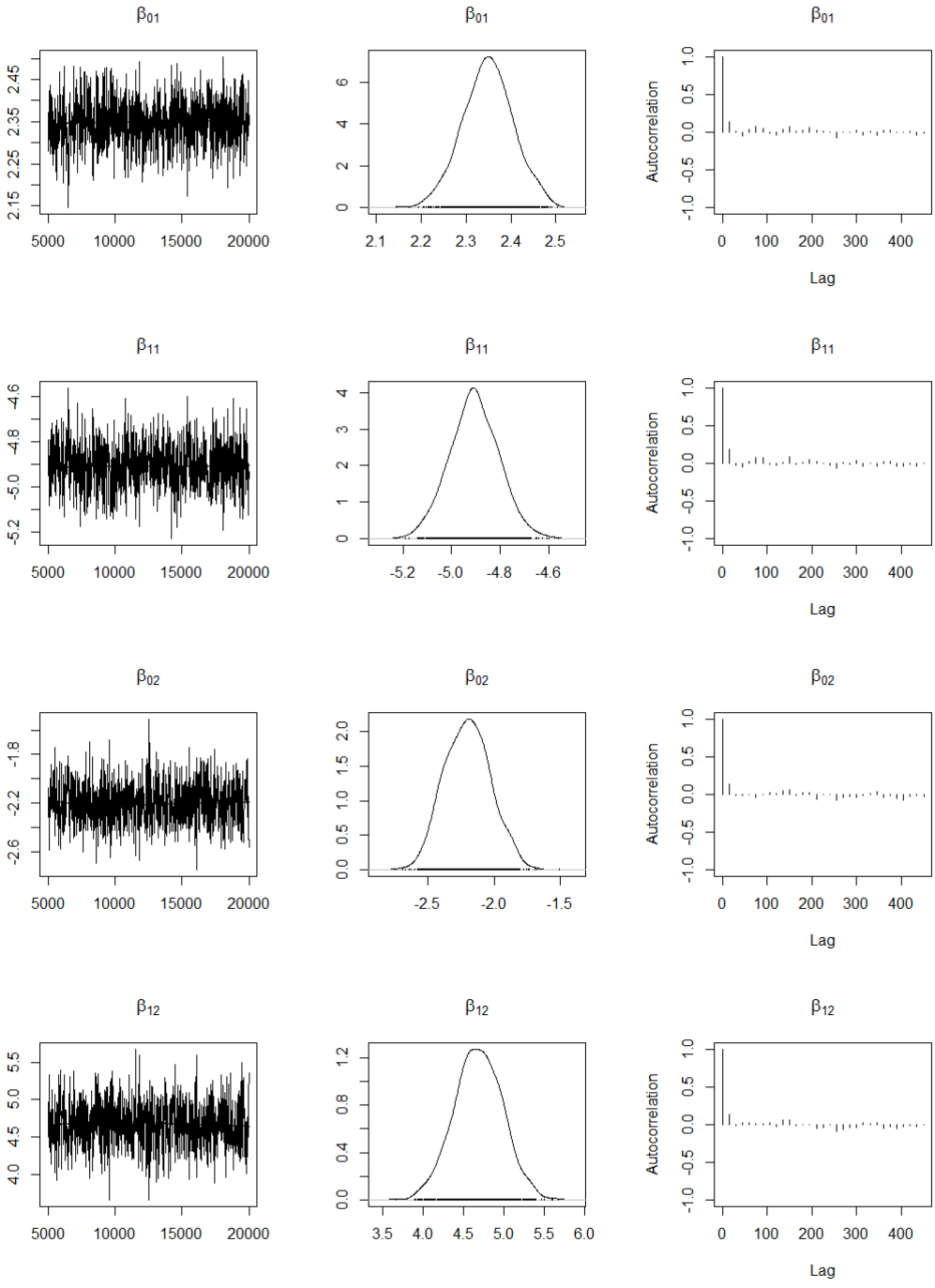

Figure A.2.1: Diagnostic plots from extreme value data Scenario I for Bayesian model selection.

1.3.1 Normal Linear Regression

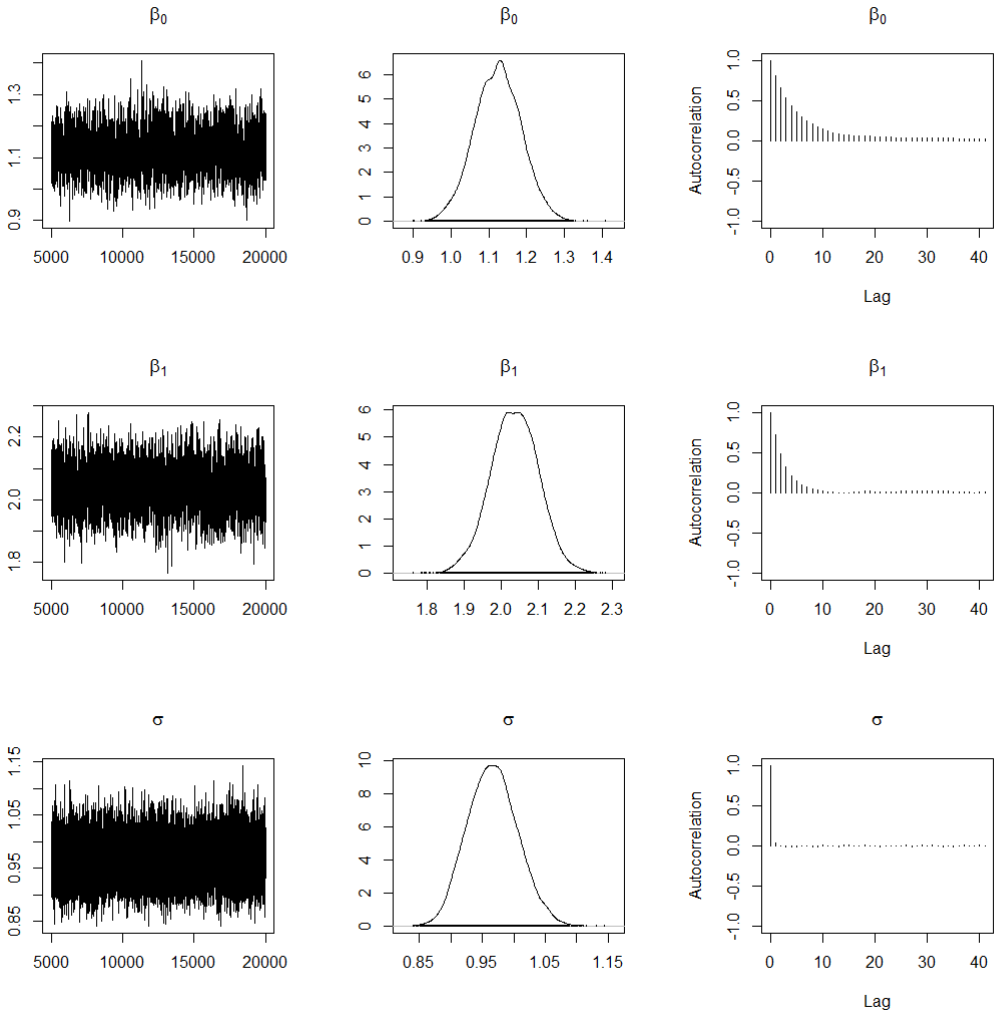

Figure A.3.1: Diagnostic plots from normal linear regression Scenario II with $a_0 = 0.1$

1.3.2 Logistic Regression

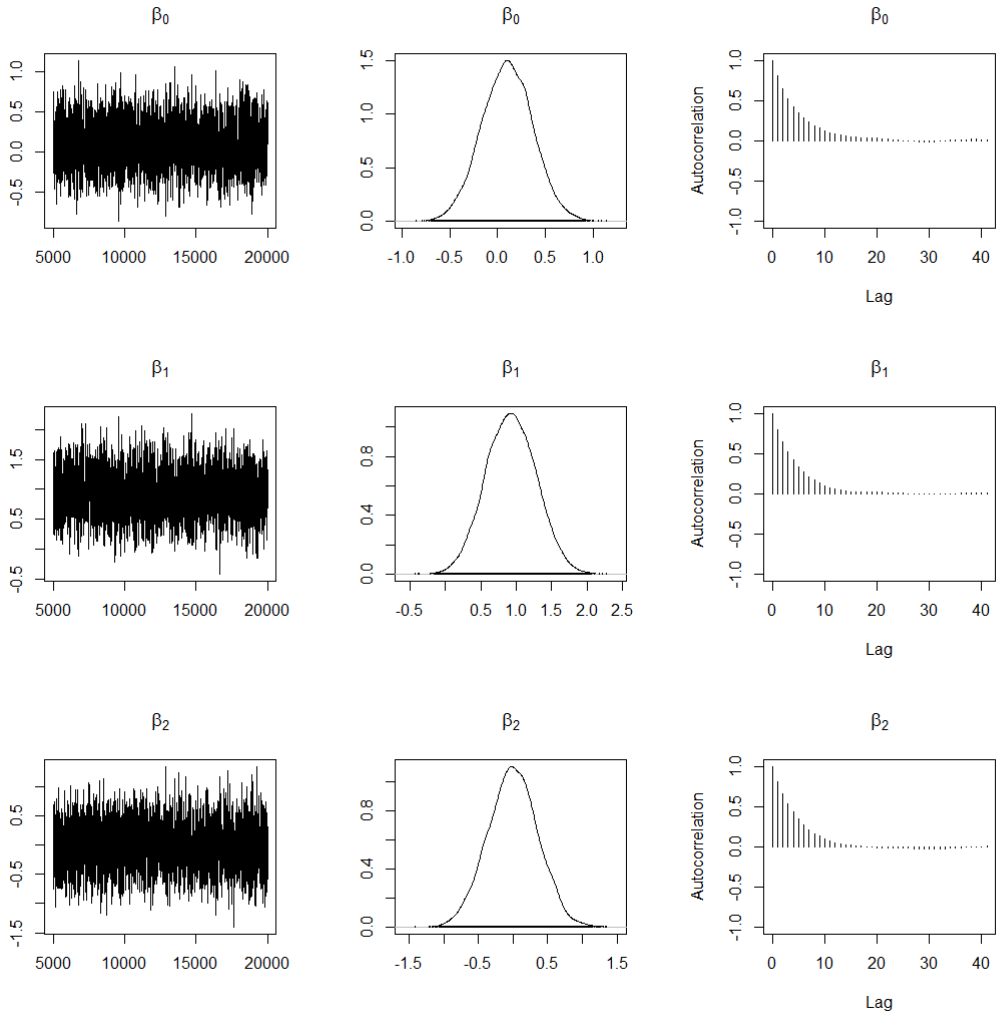

Figure A.3.2: Diagnostic plots from logistic regression Scenario I with $a_0 = 0.1$

APPENDIX B

R-Code

B.1 Parametric ROC Regression Code

```
#### Binormal data generation
z <- runif(200,0,1)  # uniform covariate

yDis <- sapply(z, function(x) rnorm(1, 2 + 4*x, 1.5))
yRef <- sapply(z, function(x) rnorm(1, 1.5 + 3*x, 1.5))

# disease indicator (1 = dis, 0 = ref)
yInd <- rep(c(1,0), each = length(z))

#data1 <- data.frame(cbind(yInd, c(yDis, yRef), z ))

binorm.ref <- data.frame(cbind("y" = yRef, "x" = z))
binorm.dis <- data.frame(cbind("y" = yDis,"x" = z))


# specifying set of false positive rates
FPR <- seq(.02, .98, by = .02)

# quantile regression to estimate reference survival
qr1 <- rq(y ~ x, data=binorm.ref, tau = rev(FPR))
pred1 <- predict.rq(qr1, newdata = binorm.dis)

Inv.t <- qnorm(FPR)     # inverse normal of FPRs
nq <- length(Inv.t)     # number of quantiles
nd <- nrow(binorm.dis)  # number of diseased points


# reshaping data for probit regression
trans.pred1 <- t(pred1)
col.pred1 <- c(trans.pred1)
col.t <- rep(Inv.t, nd)
col.ydis <- rep(binorm.dis$y, each = nq)
col.x <- rep(binorm.dis$x, each = nq)
col.uit <- as.numeric(col.ydis >= col.pred1)

probitData <- data.frame( "fdbar" = col.pred1,
```

```
                            "phiInv" =  col.t,
                            "disRes" = col.ydis,
                            "covX" = col.x,
                            "uit" = col.uit)

probitMod <- glm(uit ~ phiInv + covX,
                 family = binomial(link = "probit"),
                 data = probitData)

####### Calculating the ROC
p <- seq(0, 1, by = .02)
data.ROC <- data.frame(z = seq(min(z), max(z), l = 100))
h <- qnorm(p)  # if logistic link use qlogis

ROCParam <- sapply(1:length(h), function(s){
    pnorm(as.matrix(cbind(1, h[s],
                          data.ROC))%*%
            probitMod$coefficients)
})

####### Calculating the AUC
intFunc <- function(x){
  obj <- function(t){
    pnorm(as.matrix(cbind(1, qnorm(t), x))%*%
            probitMod$coefficients)
  }
  integrate(obj, upper = 1, lower = 0)$value
}

newZ <- seq(min(data421$z), max(data421$z), l = 100)
AUCparam <- sapply(newZ,intFunc)
```

B.2 Beta ROC Regression Code

```
#### Data generation as in Parametric Code

#### Calculating Placement Values
pv <- rep(.0001, nd)  # initializing placement value vector

for (i in 1:nd){
  for(j in 1:(nq - 1)){
    if( binorm.dis$y[i] > rev(pred1[i, ])[j] &&
        binorm.dis$y[i] <= rev(pred1[i,])[j+1]){
      pv[i] <- 1 - FPR[j]
    }
    if(binorm.dis$y[i] < min(pred1)){
      pv[i] <- .9999 }
  } #end j
} # end i

temp <-data.frame(cbind(binorm.dis, pv))

BetaModel <- betareg(pv ~ x, data = temp,
                     link.phi = "identity", link = "logit")

### Extracting coefficients for ROC calculation
intercept <- BetaModel$coefficients$mean[1]
cov1 <- BetaModel$coefficients$mean[2]
scale <- BetaModel$coefficients$precision

# specifying FPRs and covariate values for ROC
p <- seq(0, 1, by = .02)
newZ <- seq(min(z), max(z), l = 100)

### parameters of Beta(a,b) density from beta regression
aVec <- (1/(1 + exp(-intercept - cov1*newZ)))*scale
bVec <- (1 - (1/(1 + exp(-intercept - cov1*newZ))))*scale

### Calculating the ROC and the AUC
ROCbeta <- sapply(1:length(p), function(s){
  pbeta(p[s], (1/(1 + exp(-intercept - cov1*newZ)))*scale,
        (1 - (1/(1 + exp(-intercept - cov1*newZ))))*scale )
})
AUCbeta <- aVec/(aVec + bVec)
```

```
# for  data generation, quantile regression,
# and calculation of placement values see Parametric section

#### Calculating B.hat for expectation of binary indicator

# reshaping data
col.pv <- rep(pv, each = length(FPR))
col.t <- rep(FPR, nd)
col.Bhat <- as.numeric(col.pv <= col.t)

B.hat <- matrix(, nrow = nd, ncol = nq)
for (j in 1:nq){
  for(i in 1:nd){
    B.hat[i,j] <- (pv[i] <= FPR[j])*1
  } #end i
} #end j

##########################################
temp <-data.frame(cbind(binorm.dis, pv))

####### Pairwise differences of placement values
# "combn" function calculates row 2 - row 1, so take
#  negative to get row 1 - row2; store in pvDiff
pvComb <- combn(pv,2)
pvDiff <- -combn(pv, 2, diff)

# Calculating covariate differences
xDiff <- -combn(temp$x, 2, diff)

# if difference <= 0, we have a 1
BinPV <- (pvDiff <= 0)*1

# storing differences in new data frame
temp2 <- data.frame(cbind(BinPV, xDiff))

# First call to GLM to estimate the Betas (No intercept)
probitMod1 <- glm((BinPV) ~ xDiff - 1 ,
                  family = binomial(link = "probit"),
                  data = temp2)

beta.hat <- probitMod1$coefficients
```

```
# Second call to GLM to estimate h0(t) -- note the offset term
xPbeta <- beta.hat*binorm.dis$x
h0 <- apply(B.hat,2,function(s){
  #applied to the columns of B.hat
  glm(s ~ + offset(xPbeta),
      family = binomial(link = "probit"),
      control = list(maxit = 150))$coefficient[1]
})

######### Calculating the ROC
# Compute.ROC takes covariates values of interest, FPR,
# a set of t values, and the betaHat coefficient from glm

Compute.ROC <- function(cov.data = seq(0,1, by = .2),
                        FPRvec = FPR, tVec = FPR,
                        betaCoef = beta.hat){

  xPbetaROC <- cov.data * (betaCoef)

  #Can evaluate function at any set of t's, we just chose FPR
  h <- approxfun(FPRvec, h0)(tVec)
  ROC <- sapply(1:length(tVec),
                function(s) pnorm(h[s] + xPbetaROC))
}

ROCsemi <- Compute.ROC(cov.data = newZ,
                       FPRvec = FPR,
                       tVec = seq(0.02, 0.98, by = .02),
                       betaCoef = beta.hat)

########### Calculating the AUC
# reshaping the data
tVec = seq(0.02,.98, by = .02)
xseq <- newZ # should match the cov.data from Compute.ROC

newROC <- t(ROC.727[[2]])
newROC.long <- c(newROC)
t.long <- rep(tVec, length(xseq))
x.long <- rep(xseq, each = length(tVec))

ROCdata <- data.frame(cbind(ROC = newROC.long,
                            t = t.long,
                            x = x.long,
                            factor.x = as.factor(x.long)))
```

```
auc2 <- function(data, factor){
  #data <- dfParam
  #factor = .2
  data2 <- subset(data, x == factor)
  meanROC <- NULL; dt <- NULL; pAUC <- NULL; AUC <- NULL;
  t <- data2$t
  dt[1] <- t[1]
  meanROC[1] <- data2$ROC[1]/2
  pAUC[1] <- dt[1]*meanROC[1]
  AUC[1] <- pAUC[1]

  for (i in 2:nrow(data2)){
    dt[i] <-  t[i] - t[i-1]
    meanROC[i] <- (data2$ROC[i] + data2$ROC[i-1])/2
    pAUC[i] <- dt[i]*meanROC[i]
    AUC[i] <- AUC[i-1] + pAUC[i]
  }
  (AUC_PVest <- AUC[nrow(data2)])

}

# needs to match covariate data in compute.roc function
covVec <- newZ

AUCsemi <- sapply(covVec, function(x) auc2(ROCdata, x))
```

```
#########################################################
#  Data generation and construction of binary variables
#  follow from the Parametric ROC Regression code.  The
#  call to "glm" for the probit regression is replaced
#  by the call to JAGS for the given model.
#########################################################

model1 <- function(){
      for (i in 1:n) {
             probit(p[i]) <- a0 + phiInv*y[i] + covX*x[i]
             uit[i] ~ dbern(p[i])
      }
    a0 ~ dnorm(0, .1)      # Prior for intercept
    phiInv ~ dnorm(1, .1) # Prior for phiInv
    covX ~ dnorm(1,.1)    # Prior for covariate
  }

data2 <- list("uit" = probitData$uit, "y"= probitData$phiInv,
               "x" = probitData$covX,
               "n" = length(probitData$uit))

parameters <- c("a0", "phiInv", "covX")

BayesModAlonzo <- jags(
    data = data2,
    inits = NULL,
    parameters.to.save = parameters,
    model.file=model1,
    n.burnin = 5000, n.iter = 15000,
    n.thin = 5, n.chains=1)

BayesSum <- BayesModAlonzo$BUGSoutput$summary

########### Calculation of ROC #################
modCoef <- BayesSum[c(1,4,2),1] # intercept, phiInv, covX

ROCparam <- sapply(1:length(h), function(s){
        pnorm(as.matrix(cbind(1, h[s], data.ROC))%*%modCoef)
  })
```

```
#########################################################
#  Data generation and calculation of placement values
#  follow from the Beta ROC Regression code.  The call
#  to "betareg" is replaced by the call to JAGS for the
#  given model.
#########################################################

model1 <- function(){
    for(i in 1:n){
        y[i] ~ dbeta(a[i], b[i])
        a[i] <- mu[i] * gamma
        b[i] <- (1 - mu[i])*gamma
        logit(mu[i]) <- beta0 + beta1 * x[i]
        }
      gamma ~ dgamma(.001,.001)
      beta0 ~ dnorm(0,.1)
      beta1 ~ dnorm(0,.1)
}

data1 <- list("y" = tempQ$pvQuant, "x" = tempQ$z,
              "n" = length(tempQ$pvQuant))

parameters <- c("beta0", "beta1", "gamma")

BayesMod <- jags(
  data = data1,
  inits = NULL,
  parameters.to.save = parameters,
  model.file= model1,
  n.burnin = 5000, n.iter = 10000,  n.chains=1)

BayesSum <- BayesMod$BUGSoutput$summary

########### Calculation of ROC #################
intercept <- BayesSum[1,1]
cov1 <- BayesSum[2,1]
scale <- BayesSum[4,1]

ROCbeta <- sapply(1:length(p), function(s){
    pbeta(p[s], (1/(1 + exp(-intercept - cov1*newZ)))*scale,
          (1 - (1/(1 + exp(-intercept - cov1*newZ))))*scale )
```

```
#########################################################
#  Generate two sets of data following the Beta ROC
#  regression code and calculate the corresponding
#  placement values for each.  Instead of calling
#  "betareg" for each set of placement values, run the
#  following JAGS model to perform Bayesian model
#  selection. The posterior summary of z yields to
#  probability of inclusion.
#########################################################

model1 <- function(){
    for(i in 1:n){
      y1[i] ~ dbeta(a1[i], b1[i])
      a1[i] <- mu1[i] * gamma1
      b1[i] <- (1 - mu1[i])*gamma1
      logit(mu1[i]) <- beta01 + beta11 * x1[i]
    }

    for(j in 1:m){
      y2[j] ~ dbeta(a2[j], b2[j])
      a2[j] <- mu2[j] * gamma2
      b2[j] <- (1 - mu2[j])*gamma2
      logit(mu2[j]) <- beta02*z+beta01 +
                       (beta12*z+beta11) * x2[j]
    }

    gamma1 ~ dgamma(.001,.001)
    beta01 ~ dnorm(0,.1)
    beta11 ~ dnorm(0,.1)

    gamma2 ~ dgamma(.001,.001)
    beta02 ~ dnorm(0,.1)
    beta12 ~ dnorm(0,.1)

    z ~ dbern(.05)
  }

data1 <- list("y1" = tempQ$pvQuant, "x1" = tempQ$z,
                "n" = length(tempQ$pvQuant),
                "y2" = tempQ2$pvQuant2, "x2" = tempQ2$z,
                "m" = length(tempQ2$pvQuant2))
```

```
parameters <- c("beta01", "beta11", "gamma1",
                "beta02", "beta12", "gamma2", "z")

BayesMod1 <- jags(
    data = data1,
    inits = NULL,
    parameters.to.save = parameters,
    model.file= model1,
    n.burnin = 5000, n.iter = 20000,  n.chains=1)

BayesMod1$BUGSoutput$summary
```

2.7.1 Normal Linear Regression

```
model <- function()
{
  for( i in 1 : n ) {

    # Current Data
    y[i] ~ dnorm(mu[i], tau)
    mu[i] <- beta0 + beta1*x1[i]
  }
  # Historical Data (Use Zeros Trick)

  for( i in 1:n0) {
    mu0[i] <- beta0 + beta1*x10[i]
    l[i] <- (-a0/2)*(log(2*3.1416/tau)+ (y0[i] - mu0[i])*tau)

    dummy[i]<-0
    dummy[i] ~ dloglik(l[i])
  }

  # Priors
  beta0~dnorm(0, .001)
  beta1~dnorm(0, .001)
  tau ~ dgamma(.01,.01)
  sigma <- 1/sqrt(tau)

  a0 <- .1  # Fixed a0
}
############  End of Model

# Initial values
inits <- list(tau = 0, beta0 = 0, beta1 = 0)

n0 <- 500; n <- 250
B0 <- 1; B1 <- 2
B00 <- 1; B10 <-1.75

x1 <-rnorm(n, 0, 1)
y <- rnorm(n, B0 + B1*x1, 1)

x10 <- rnorm(n0,0,1)
y0 <- rnorm(n0, B00 + B10*x10, 1)
```

```
data <- list("n" = n, "n0" = n0, "y" = y,  "y0" = y0,
               "x1" = x1,"x10" = x10)

parameters <- c("beta0",  "beta1", "sigma")

ss.sima01 <- bugs(
    data = data,
    inits = inits,
    parameters.to.save = parameters,
    model.file=model,
    working.directory = "C:/Users/sarah_stanley/Documents/
    Normal_powerPrior",
    n.burnin = 5000,
    n.iter = 20000,
    n.chains=3,
    DIC = TRUE, debug = FALSE)

LOG <- bugs.log("C:/Users/sarah_stanley/Documents/
Normal_powerPrior/log.txt")
yDICvec <- rep(LOG$DIC[2,3], 4)
DICvec <- rep(ss.sim1$DIC, 4)
exp <- as.data.frame(ss.sim1$summary[1:4 ,c(1:3,7)])
SimSummary <- cbind(exp, DICvec, yDICvec)
```

2.7.2 Logistic Regression

```
library(coda)
library(R2OpenBUGS)

model <- function()
{
  for( i in 1 : n ) {

    # Current Data
    y[i] ~ dbern(p[i])
    logit(p[i]) <- beta0+beta1*x1[i]+beta2*x2[i]
  }

  # Historical Data (Use Zeros Trick)
    for( i in 1:n0) {
    logit(p0[i])<-  beta0+beta1*x10[i]+beta2*x20[i]

    # log likelihood
    l[i]<-a0*(y0[i]*log(p0[i])+(1-y0[i])*log(1-p0[i]))

    dummy[i]<-0
    dummy[i] ~ dloglik(l[i])
 }

  # Priors
  beta0~dnorm(0, .1)
  beta1~dnorm(0, .1)
  beta2~dnorm(0, .1)

  a0 <- .1   # Fixed a0
}
###########  End of Model

# Initial values
inits <- list(beta0=0, beta1=0, beta2=0)

  # Data generation
  set.seed(m)
  n <- 100; n0 <- 500
  B0 <- 0; B1 <- 1; B2 <- .25
  B00 <- 0; B10 <- 0; B20 <- 0.21

  x10 <- as.numeric(rbinom(n0, 1,0.5))
```

```
  x20 <- as.numeric(rbinom(n0, 1, 0.35))
  y0 <- as.numeric(rbinom(n0, 1, 1/(1 +
              exp(-(cbind(1,x10, x20) %*%
              c(B00, B10, B20))))))

  x1 <- as.numeric(rbinom(n, 1, 0.5))
  x2 <- as.numeric(rbinom(n, 1, 0.35))

  p=1/(1+exp(-cbind(1,x1, x2) %*% c(B0, B1, B2)))
  y <- as.numeric(rbinom(n, 1, p))

  data <- list("n" = n, "n0" = n0, "x1" = x1, "x2" = x2,
               "y" = y, "x10" = x10, "x20" = x20, "y0" = y0)

  parameters <- c("beta0", "beta1", "beta2")

  ########################

log.sima01 <- bugs(
    data = data,
    inits = inits,
    parameters.to.save = parameters,
    model.file=model,
    n.burnin = 5000,
    n.iter = 20000,
    n.chains=3,
    DIC = TRUE, debug = F)

log.sima01$summary
```

APPENDIX C

SAS-Code

C.1 Parametric ROC Regression Code

```
data inparm; seed0=12345; seed1=67891;

*** Normal data generation ***;
do n = 1 to 50;
  x0 = ranuni(seed0);
  x1 = ranuni(seed1);

  LA=0;
  m0 = 0.0;
  s0 = 1.0;
  y0 = m0 + 3*x0 + sqrt(s0)*rannor(seed0);

  m1 = 0;
  s1 = 1.0;
  y1 = m1 + 4*x1 + sqrt(s1)*rannor(seed1);

  la1A = (m1-m0+(x1))/s1;
  la2A = s0**2/s1**2;
  AUC_true = cdf("normal", (la1A/sqrt(1+la2A)), 0, 1);
  output;
  end;
  run;


proc print data=inparm; where n <= 10; run;

data binorm; set inparm;
  y=y0; x=x0; LA=0; output;
  y=y1; x=x1; LA=1; output;

  keep LA y x;
run;

*++++++++++++++++++++++++++++++++++++++++++++++++++;
```

```
proc quantreg data=binorm ; where LA=0;
  ods select ParameterEstimates;
  ods output ParameterEstimates = parms;
  model y = x /quantile=(.01 to .99 by .01) nosummary;
run;

data c; set binorm; where LA=1; run;

proc iml;
  use parms;
  read all into dataRef;
  use c; read all into dataDis;
  vec1 = do(1, 197, 2);
  vec2 = do(2, 198, 2);

  intRefcoef = dataRef[vec1,3];
  xRefcoef = dataRef[vec2,3];

  **initializing vectors/matrices;
  nd = nrow(dataDis);
  nq = dimension(vec1)[1,2];
  FPR = do(.01, .99, .01);
  predSurv = j(nd, nq, 1);
  qq = j(nq, 1, 1);
  t = j(nd, nq, 0);
  xCov = j(nd, nq, 0);
  yDis = j(nd, nq, 0);
  k = nd*nq;
  colUit = j(k, 1, 99);

  **calculating predicted reference survival curve;
  **prepping data for probit regression;

  do i = 1 to nd;
  do j = 1 to nq;
    predSurv[i, j] = intRefcoef[j] + dataDis[i,3]*xRefcoef[j];
    qq[j] = quantile("normal", FPR[j]);
    t[i,j] = qq[j];
    xCov[i,j] = dataDis[i,3];
    yDis[i,j] = dataDis[i,2];
  end;
  end;

  transPredSurv = predSurv`;
```

```
  transPredSurv2 = transPredSurv[nrow(transPredSurv):1,];
  colPred = colvec(transPredSurv2`);
  colX = colvec(xCov);
  colyDis = colvec(yDis);
  colT = colvec(t);
  colUIT = colyDis <= colPred;

  new = colUit||colT||colX;
  create f from new;
  append from new;
run;

data temp; set f;  uit = col1; phiInv = col2;
    x = col3; drop col1-col3; run;

proc probit data = temp plot = predpplot;
  ods select parameterEstimates;
  ods output parameterEstimates = parms2;
  model uit = phiInv x;
run;

proc transpose data=parms2 out=out_parms; run;

data out_parms; set out_parms;
  if _NAME_ = 'Estimate';run;

*** calculating the ROC ***;
data temp1; set out_parms;
  alpha_hat = col1;
  beta_hat = col2;
  theta = col3;

  do x = .5 to 1 by .1;
  do s = 0.001 to 0.999 by .005;
    quant = quantile('normal', s, 0, 1);
    ROC = cdf('normal', +1*alpha_hat + beta_hat*quant + theta*x, 0,1);
  output; end;end;

  keep alpha_hat beta_hat s theta x ROC ;
run;

*** plotting the ROC ***;
proc sgpanel data=temp1;
  panelby x;
```

```
  series y=ROC x=s;
run;

*** calculating the AUC ***;
%macro loop(dsn= , cov=, title= );
  data temp; set &dsn; where x=&cov; keep x s roc;run;
  title &title;
  proc sgplot data=temp;
  series y=roc x=s;
  run;

  proc iml;
    use temp;
    read all into data;
    x   = data[1,1];
    t   = data[,2];
    roc = data[,3];
    np     = nrow(data);
    dt = t;
    auc=roc; meanroc=roc; pAUC=AUC;
    dt[1] = t[1];
    meanROC[1] = ROC[1]/2;
    pAUC[1] = dt[1]*meanROC[1];
    AUC[1]= pAUC[1];

    do i=2 to np by 1;
    dt[i] = t[i] - t[i-1];
    meanROC[i] = (ROC[i] + ROC[i-1])/2;
    pAUC[i] = dt[i]*meanROC[i];
    AUC[i] = AUC[i-1] + pAUC[i];
    end;

    AUC_PVest = AUC[np];
    print x AUC_PVest;
    quit;
%mend;

%loop(dsn = temp1, cov = 0.5, title = 'x = 0.5');
%loop(dsn = temp1, cov = 0.6, title = 'x = 0.6');
%loop(dsn = temp1, cov = 0.7, title = 'x = 0.7');
```

```
data inparm; seed0=12345; seed1=67891;

*** Normal data generation ***;
do n = 1 to 50;
  x0 = ranuni(seed0);
  x1 = ranuni(seed1);

  LA=0;
  m0 = 0.0;
  s0 = 1.0;
  y0 = m0 + 3*x0 + sqrt(s0)*rannor(seed0);

  m1 = 0;
  s1 = 1.0;
  y1 = m1 + 4*x1 + sqrt(s1)*rannor(seed1);

  la1A = (m1-m0+(x1))/s1;
  la2A = s0**2/s1**2;
  AUC_true = cdf("normal", (la1A/sqrt(1+la2A)), 0, 1);
  output;
  end;
  run;


proc print data=inparm; where n <= 10; run;

data binorm; set inparm;
  y=y0; x=x0; LA=0; output;
  y=y1; x=x1; LA=1; output;

  keep LA y x;
run;

*++++++++++++++++++++++++++++++++++++++++++++++++++;

** Quantile Regression;
proc quantreg data=binorm ; where LA=0;
  ods select ParameterEstimates;
  ods output ParameterEstimates = parms;

  model y = x /quantile=(.01 to .99 by .01) nosummary;
run;
```

```
data c; set binorm; where LA=1; run;

proc iml;
  use parms;
  read all into dataRef;
  use c; read all into dataDis;
  vec1 = do(1, 197, 2);
  vec2 = do(2, 198, 2);

  intRefcoef = dataRef[vec1,3];
  xRefcoef = dataRef[vec2,3];

  ** initializing vectors/matrices;
  nd = nrow(dataDis);
  nq = dimension(vec1)[1,2];
  FPR = do(.01, .99, .01);
  predSurv = j(nd, nq, 1);
  predSurvRev = j(nd, nq, 1);

 **calculating predicted reference survival curve;
  do i = 1 to nd;
  do j = 1 to nq;
  predSurv[i, j] = intRefcoef[j] + dataDis[i,3]*xRefcoef[j];
  end;
  end;

* calculating placement values;
  pv = j(nd, 1, .0001);
  do i = 1 to nd;
  do j = 1 to 98;
    if dataDis[i,2] > predSurv[i,j] &
       dataDis[i,2] <= predSurv[i,j+1] then
       pv[i] = 1 - FPR[j];
    if dataDis[i,2] < min(predSurv) then pv[i] = .9999;
  end;
  end;

  new = pv||dataDis[,3];
    create f from new;
    append from new;
run;

data temp; set f;  pv = col1; x = col2;
  drop col1-col2; run;
```

```
proc glimmix data=temp;
  ods select ParameterEstimates;
  ods output ParameterEstimates=parms;
  model pv =  x/dist=beta solution ;
run;

proc transpose data=parms out = mn;
run;

data mn (keep= Intercept e_A1 scale mu omega  tau x roc s);
  set mn;
  if _NAME_='Estimate'
  then do;
  Intercept = col1;
  e_A1 = col2;
  scale = col3;

  do x = .5 to 1.0 by .1;
    mu = 1/(1+exp(-Intercept - e_A1*x));
    omega = mu*scale;
    tau = (1 - mu)*scale;
  do s = 0.001 to 0.999 by .005;
    ROC = cdf("beta",s, omega, tau);
  output; end; end; end;
run;


proc sgpanel data=mn;
  panelby x;
  series y=roc x=s;
run;

data mn2; set mn; keep  x s roc; run;

%macro loop(dsn= , cov=, title= );
  data temp; set &dsn; where x=&cov; keep x s roc;run;
    title &title;
    proc sgplot data=temp;
    series y=roc x=s;
  run;

  proc iml;
    use temp;
    read all into data;
```

```
    x   = data[1,1];
    t   = data[,2];
    roc = data[,3];
    np = nrow(data);
    dt = t;
    auc=roc; meanroc=roc; pAUC=AUC;
    dt[1] = t[1];
    meanROC[1] = ROC[1]/2;
    pAUC[1] = dt[1]*meanROC[1];
    AUC[1]= pAUC[1];

    do i=2 to np by 1;
    dt[i] = t[i] - t[i-1];
    meanROC[i] = (ROC[i] + ROC[i-1])/2;
    pAUC[i] = dt[i]*meanROC[i];
    AUC[i] = AUC[i-1] + pAUC[i];
    end;

    AUC_PVest = AUC[np];
  print x AUC_PVest;
  quit;
%mend;

%loop(dsn = mn2, cov = 0.5, title = 'x = 0.5');
%loop(dsn = mn2, cov = 0.6, title = 'x = 0.6');
%loop(dsn = mn2, cov = 0.7, title = 'x = 0.7');
```

BIBLIOGRAPHY

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679.

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

Balakrishnan, N. and Nevzorov, V. (2003). *A Primer on Statistical Distributions*. Wiley, New Jersey.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, 12(4):387–415.

Beam, C. (1995). 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:S4–13.

Branscum, A. J., Johnson, W. O., and Thurmond, M. C. (2007). Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian & New Zealand Journal of Statistics*, 49(3):287–301.

Buros, A. (2015). *Semiparametric AUC Regresssion for Ordered Treatment Effects*. PhD thesis, Baylor University.

Buros, A., Tubbs, J. D., and Van Zyl, J. S. (2017). Auc regression for multiple comparisons. *Statistics in Biopharmaceuticial Research (to appear)*.

Cai, T. (2004). Semi-parametric ROC regression analysis with placement values. *Biostatistics*, 5(1):45–60.

Chen, M.-H., Ibrahim, J. G., and Shao, Q.-M. (2000). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference*, 84(1-2):121–137.

DeLong, E., DeLong, D., and Clarke-Pearson, D. (1988). Comparing the areas under two or more correlated receiver operating characteristics curves: a nonparametric approach. *Biometrics*, 98:837–844.

Dodd, L. and Pepe, M. (2003). Semiparametric regression for the area under the receiver operating characteristic curve. *Journal of the American Statistical Association*, 98:409–417.

Elman, M. J., Ayala, A., Bressler, N. M., Browning, D., Flaxel, C. J., Glassman, A. R., Jampol, L. M., and Stone, T. W. (2015). Intravitreal ranibizumab for diabetic macular edema with prompt versus deferred laser treatment: 5-year randomized trial results. *Ophthalmology*, 122(2):375–381.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Fubini, G. (1907). Sugli integrali multipli. *Rend. Acc. Naz. Lincei*, 16:608–614.

Gatsonis, C. (1995). Random-effects models for diagnostic accuracy data. *Academic Radiology*, 2:S14–21.

Hornik, K., Leisch, F., and Zeileis, A. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of DSC*, volume 2, pages 1–1.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, pages 46–60.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). The power prior: theory and applications. *Statistics in medicine*, 34(28):3724–3749.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2003). On optimality properties of the power prior. *Journal of the American Statistical Association*, 98(461):204–213.

Ibrahim, J. G., Ryan, L. M., and Chen, M.-H. (1998). Using historical controls to adjust for covariates in trend tests for binary data. *Journal of the American Statistical Association*, 93(444):1282–1293.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Miller, A. (2002). *Subset selection in regression*. CRC Press.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117.

Pepe, M. and Cai, T. (2004). The analysis of placement values for evaluating discriminatory measures. *Biometrics*, 60(2):528–535.

Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, pages 124–135.

Pepe, M. S. (2000). An interpretation for the ROC curve and inference using GLM procedures. *Biometrics*, 56(2):352–359.

Rodriguez-Alvarez, M. X., Tahoces, P. G., Cadarso-Suarez, C., and Lado, M. J. (2011). Comparative study of roc regression techniques – applications for the computer-aided diagnostic system in breast cancer detection. *Computational Statistics and Data Analysis*, 55(1):888–902.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Thompson, M. L. and Zucchini, W. (1989). On the statistical analysis of roc curves. *Statistics in Medicine*, 8(10):1277–1290.

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

Van Zyl, J. S. (2017). *Evaluating Treatment Efficacy Using AUC Modeling*. PhD thesis, Baylor University.

Zhang, L., Zhao, Y. D., and Tubbs, J. D. (2011). Inference for semiparametric auc regression models with discrete covariates. *Journal of Data Science*, 9(4):625–637.