ABSTRACT

Semiparametric AUC Regression for Ordered Treatment Effects

Amy Buros, Ph.D.

Chairperson: Jack D. Tubbs, Ph.D.

We investigated distribution free methods for testing covariate adjusted treatment effects when the researchers believe that these effects are ordered. Dodd and Pepe (2003) proposed a semi-parametric logistic regression model for the area under the ROC curve (AUC). Their approach was motivated by the observation that the Mann-Whitney statistic is a non-parametric estimate of the AUC. Their results allow one to test hypotheses using distribution free methods when the covariates are discrete, however, the standard errors generated using standard GLM software are not correct since the Bernoulli data generated by the Mann-Whitney statistic are correlated. They used the bootstrap method to estimate the standard errors for the AUC regression parameters. Zhang (2008) and Zhang et. al (2011) considered an analytical method for estimating the standard errors based on a modification of a method by DeLong et. al (1988), as an alternative to the bootstrap procedure. In Chapter Two, we compare the DeLong method to two alternative analytical methods for estimating the standard errors. In Chapter Three, we extend the AUC regression model, with and without discrete covariates, to the situation where there are $k > 2$ ordered treatment levels as the alternative hypothesis. This approach extends the Jonckheere-Terpstra statistic (Jonckheere (1954) and Terpstra (1952)) to allow for covariates. In Chapter Four, we introduce a multiple comparison method for

the Jonckheere-Terpstra statistic. Chapter Five gives a summary of the results and describes future work.

Semiparametric AUC Regression for Ordered Treatment Effects

by

Amy Buros, 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

Steve Green, Ph.D.

John W. Seaman Jr., Ph.D.

James D. Stamey, Ph.D.

Dean M. Young, Ph.D.

Accepted by the Graduate School
August 2014

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

First, I must express my immense gratitude to Dr. Tubbs for his guidance through this entire process. Whether you were encouraging me or checking on my progress, you were always there to help me along the way. I have learned so much and I am truly honored to have been able to work with you.

Next, I would like to thank my family for every bit of unwavering love and support that they have given me throughout the years. Mom and Dad, I would never have made it to this point without you. You are the solid foundation on which I have built my own life. Words cannot express how much I care about you, but I hope you always know how much I love and appreciate you. And to my sister, Sara, thank you for always being there for me no matter what. I am so glad to have you as a sister and a friend.

To the entire faculty in the Department of Statistical Science, I appreciate all of the support and opportunities you have provided. To Dr. Stamey, Dr. Seaman, Dr. Harvill, Dr. Johnston, Dr. Young, and Dr. Maddox, thank you for taking the time to teach me. You were always willing to answer questions and offer your knowledge. I would like to thank Dr. Hill for your passion for teaching statistics, I would not have joined this program without you. Finally, I would especially like to thank Dr. Tom Bratcher, who persuaded me to embark on this journey. You are greatly missed.

To my fellow students in the department, thank you for you help, support, and care. Thank you for the laughter, help on homework, and much needed breaks from staring at a computer screen. The last four years would not have been possible without you all. I was incredibly lucky to be in a cohort with such smart, talented, and considerate people.

Many thanks go to my best friends, Laura Emamian and Courtney Johnson. Thank you for sharing the ups and downs, the laughter and the tears, and for always lending an ear.

Lastly, I give my greatest gratitude to my love, Caleb, who has been through this entire journey with me. Thank you for the strength when I was failing, confidence when I wavered, and always maintaining the never ending faith in me that I could succeed. I am so excited to start the next chapter with you.

CHAPTER ONE

Introduction

1.1 Overview of the Problem

This dissertation investigates non-parametric methods that allow for the inclusion of discrete covariates for testing hypotheses when there are greater than two treatment arms. The problem was motivated by the ability of investigators to draw confirmatory conclusions from a clinical trial that consists of a control group and multiple treatment arms. The inclusion of two or more treatment groups can enhance the productivity of the clinical trial by providing supplemental information about a different characteristic of the active therapy. For example, increasing dosage studies to ascertain the minimum effective dose for a drug; or combination drug therapies, illustrating if the combination of two or more drugs is preferable to each drug separately. Specifically, we are interested in clinical trials when the primary response variable is non-normal and treatment medians are ordered. In this dissertation, we consider a distribution-free approach to this problem instead of using normalizing transformation or a generalized linear model based upon specifying an exponential family distribution. Nonparametric methods have been extensively studied when testing for the treatment medians without considering covariates (Lehmann (1975)). However, clinical trials need to examine consistency of the treatment effects across subgroups by adjusting for covariates which could have an impact upon the effectiveness of the therapy (Zhang (2008)). Our objective is to investigate the inclusion of covariates in the ordered multiple treatment arm scenario.

This chapter presents a nonparametric test for population medians using the Area Under the ROC Curve (AUC) Regression proposed by Dodd and Pepe (2003). Their approach was motivated by the relationship between the Mann-Whitney U statistic and the nonparametric estimate of AUC using a generalized linear model.

The model allows for the inclusion of covariates when estimating the AUC, hence, the Mann-Whitney statistic using standard logistic regression software. However, since these estimates are based on dependent Bernoulli data the associated standard errors are incorrect. They suggested using the bootstrap when computing standard errors. Zhang (2008) modified their approach by using a computationally simpler method proposed by DeLong et al. (1988). Section 1.2 contains an introduction and background to methods pertaining to AUC regression. Section 1.3 provides a summary of the method proposed by Dodd and Pepe (2003). An outline of the dissertation is given in section 1.4.

1.2 Background

Testing for differences in multiple treatment levels is increasingly common in clinical studies. For non-normal data, the Kruskal Wallis test (Kruskal and Wallis (1952)) has been extensively examined as a non-parametric method for testing whether two or more samples originate from the same distribution. However, if the treatments (three or more) are ordered a priori then the Jonckheere Trend test is more powerful for testing for median differences among the treatments groups. In most clinical investigations, one would like to investigate the differences in treatments while accounting for covariates. Dodd and Pepe extended the use of the Mann-Whitney statistic to a GLM context by exploiting the relationship between the Mann-Whitney statistic and the AUC for the ROC curve.

1.2.1 Area under the ROC curve

The Reciever Operating Characteristics curve (ROC) provides a graphical measure for the accuracy of a binary classifier system to discriminate between two populations. The ROC is a plot of the true positive rate (sensitivity) versus the false positive rate (1 - specificity). The curve can be applied to diagnostic testing where a continuous variable, Y, is used to classify subjects into either diseased (D) or non-diseased $\left(\bar{D}\right)$ groups according to the classification rule of assigning a subject to the

disease group when $Y > c$ for a specified threshold level c. The ROC curve plots $\Pr[Y > c|D]$, the probability of correct classification vs. $\Pr[Y > c|\bar{D}]$, the probability of a false positive, for all possible thresholds c, in the unit square from $(0, 0)$ to $(1, 1)$. The ROC curve provides a visible means of assessing the accuracy of the diagnostic test to make decisions about the threshold c, as determined by sensitivity and specificity for the specific application.

When interpreting the ROC curve there are two extreme cases. The first case is when the ROC curve is a diagonal line from $(0, 0)$ through $(1, 1)$. This plot indicates that Y is a useless classifier and is equivalent to flipping a coin. In this case the distribution of Y for the diseased group is identical to the distribution for the non-diseased group. The second case is when the ROC curve passes through the vertex $(0, 1)$ illustrating that Y is a perfect classifier and that the kernel of the density function for the diseased group is disjoint from the kernel of the density function for the non-diseased group.

A commonly used summary statistic for the ROC curve is the area under the ROC curve (AUC). The AUC is the probability that a randomly chosen subject is classified into the correct group and is given by

$$\Pr[Y^D > Y^{\bar{D}}].$$

The AUC for a classifier Y ranges from 0.5 to 1, where $AUC = 0.5$ in the useless classifier (diagonal line) and $AUC = 1$ in the perfect classifier. Figure 1.1 illustrates the ability of the ROC curve and AUC to discriminate between two different densities. Since, the AUC is a measure of discrepancy between two density functions, it can be used to determine the effectiveness of an active treatment when compared with a control group in a clinical trial setting. This follows since the Mann-Whitney rank sum U statistic provides a nonparametric estimate of the AUC (Bamber (1975))

Dodd and Pepe (2003) proposed a method of using logistic regression to model the AUC. The model allows for the inclusion of covariates when estimating the AUC. The next section defines the Mann-Whitney U statistic and the semi-parametric AUC

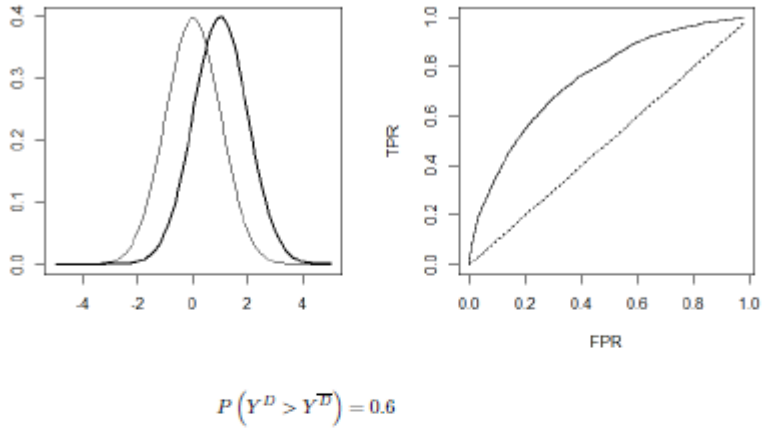

Figure 1.1: The relationship between the ROC curve and AUC

regression model.

1.3 Semi-parametric AUC Regression Model

1.3.1 Mann-Whitney rank sum U Statistic

The Mann-Whitney U statistic was developed in 1947 (Mann and Whitney (1947)) and provides a nonparametric approach for comparing the location parameters for two independent samples. Let $x_1, ..., x_n$ and $y_1, ..., y_m$ be independent random samples from $F_x(.)$ and $F_y(.)$, respectively. Combine the two data sets and let

$s_1, ..., s_n$ denote the ranks of $x_1, ..., x_n$. The Wilcoxon rank-sum statistic T is

$$T = \sum_{i=1}^{n} s_i,$$

and the Mann-Whitney U statistic is

$$U = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} I\left(x_i > y_j\right)}{nm} \tag{1.1}$$

where

$$I\left(x_i > y_j\right) = \begin{cases} 1 & x_i > y_j \\ \frac{1}{2} & x_i = y_j \\ 0 & x_i < y_j \end{cases}.$$

The relationship between T and U is

$$nmU = T - \frac{1}{2} n\left(n+1\right).$$

It should be noted that one is unable to adjust for covariates using either of these statistics. Van Elteren (1960) proposed a method for combining Mann-Whitney statistics across discrete block or strata variables provided these variables do not have a multiplicative effect with the treatment effects.

Dodd and Pepe (2003) provided a semi-parametric AUC regression model for discrete covariates.

1.3.2 AUC Regression Model

Let Y_i^D and $Y_j^{\overline{D}}$ denote the i^{th} response in the diseased (treatment) group $(i = 1, ..., N_D)$ and the j^{th} response variable in the non-diseased (control) population $(j = 1, ..., N_{\overline{D}})$, respectively. Suppose one wished to determine whether or not Y could be used as an effective classifier. In which case, one could test

$$H_0 : \Pr[Y_i^D > Y_j^{\overline{D}}] = 0.5 \quad \text{versus} \quad H_1 : \Pr[Y_i^D > Y_j^{\overline{D}}] > 0.5$$

or

$$H_0 : \text{AUC} = .5 \quad \text{versus} \quad H_1 : \text{AUC} > .5.$$

It has been shown that the Mann-Whitney statistic given in (1.1) is a non-parametric estimate of AUC given by

$$AUC = \frac{\sum_{i=1}^{N_D} \sum_{j=1}^{N_{\overline{D}}} I(Y_i^D > Y_j^{\overline{D}})}{N_D N_{\overline{D}}}. \tag{1.2}$$

Suppose that one has a covariate X, then the covariate-specific AUC can be expressed as

$$AUC_{ij} = \Pr[Y_i^D > Y_j^{\overline{D}} | X_i, X_j].$$

In many clinical settings one is interested in Y's ability to classify at a specific covariate X, in which case, the AUC of interest becomes

$$\Pr[Y_i^D > Y_j^{\overline{D}} | X_i = X_j = X]. \tag{1.3}$$

Dodd and Pepe (2003) proposed the use of logistic regression for the AUC with the generalized linear model, $g(AUC) = \mathbf{X}^{\mathbf{T}}\beta$, where g is a link function. The estimates of parameters are solutions to the score equations

$$\sum_{i=1}^{N_D} \sum_{j=1}^{N_{\overline{D}}} \frac{(I_{ij} - \mathrm{AUC}_{ij})}{var(I_{ij})} \frac{\partial AUC_{ij}}{\partial \beta},$$

The solutions can be found with commonly used GLM software, such as, SAS/Proc Genmod (SAS (2008a)) or SAS/ Proc Logistic (SAS (2008b)).

It should be noted that the binary variables, I_{ij}, in the score equations are correlated. In which case, the solutions to these estimating equations are correct but resulting standard errors as produced by the standard GLM software are incorrect. Dodd and Pepe used the bootstrap to obtain the standard errors. Zhang (2008) proposed a modification of the method provided by DeLong et al. (1988) to derive the variance of the nonparametric AUC where the variance of the regression parameters are calculated using the delta method (Zhang et al. (2011)).

1.4 Outline of the Dissertation

In this dissertation, we investigate semi-parametric regression models for the AUC to determine the effect of covariates on the treatment effect. We examine a

procedure for testing $k > 2$ ordered treatment arms while adjusting for covariates. A new multiple comparison method based on this model is also presented.

In chapter 2, we investigate two alternate analytical methods for estimating the standard errors. The methods are compared with the bootstrap method suggested by Dodd and Pepe (2003) and the analytical method by Zhang (2008). Both methods are computationally feasible and simple to implement. They are based on procedures given in Fligner and Policello (1981) and Birnbaum and Klose (1957). In chapter 3, we introduce the Jonckheere-Terpstra statistic (Jonckheere (1954) and Terpstra (1952)) and apply it to the AUC regression framework. In Chapter 4, we introduce a new method for multiple comparisons based on the model defined in Chapter 3. This method is compared to two nonparametric multiple comparison methods. Chapter 5 contains concluding remarks and possible directions for future research.

CHAPTER TWO

Estimating Standard Errors of AUC Regression

This chapter presents three analytical methods for computing the standard errors for the regression parameters in AUC regression.

2.1 Introduction

Dodd and Pepe (2003) presented a semi-parametric logistic regression model for the area under the receiver operating characteristics curve (AUC). Bamber (1975) demonstrated that the Mann-Whitney statistic is an unbiased estimate of AUC $= P\left(Y^D > Y^{\bar{D}}\right)$. Dodd and Pepe (2003) used this result to develop an AUC regression model that accommodates discrete and continuous covariates. However, since the binary responses used in the Mann-Whitney statistics are correlated the standard errors produced by software packages are incorrect. Zhang (2008) and Zhang et al. (2011) presented a method given by DeLong et al. (1988) to estimate the variance of the AUC, from which the delta method was used to estimate the standard errors for the regression parameters. In this chapter, we present two alternative analytical methods for estimating the variance of the AUC. The first method is based upon a modification of a method given in Fligner and Policello (1981). The second method considers an approach suggested by Birnbaum and Klose (1957) for computing upper and lower bound on the variance of the Mann-Whitney statistic. Both methods can be used to estimate the standard errors of the logistic regression parameters. The three analytical methods are compared with the bootstrap method using a simulation study and a problem arising from a real data application.

The AUC regression model by Dodd and Pepe (2003) and proposed bootstrapping method for finding the standard errors is briefly described in Section 2.2. The DeLong method (DeLong et al. (1988)), the Fligner method (Fligner and Policello

(1981)), and the Birnbaum method (Birnbaum and Klose (1957)) for variance estimation of the AUC are described in Section 2.3. The method for computing parameter estimates and standard errors as given by Zhang (2008) and Zhang et al. (2011) is presented in Section 2.3.5. The methods are compared using a simulation study in Section 2.4. The results for the three methods as applied to a real data example are given in Section 2.5. Section 2.6 contains a brief discussion of the results and conclusions.

2.2 Semi-parametric AUC Regression Model

Let Y_i^D denote the i^{th} response in the diseased (or treatment) group and $Y_j^{\bar{D}}$ the j^{th} response variable in the non-diseased (or control) group for $i = 1, ..., N_D$ and $j = 1, ..., N_{\bar{D}}$. Suppose that one wants to estimate the AUC as a linear function of covariate X, then the covariate-specific AUC is

$$AUC_{ij}(X_i, X_j) = \Pr\left(Y_i^D > Y_j^{\bar{D}} | X_i, X_j\right). \tag{2.1}$$

It is rare that researchers are interested in estimating the AUC at different levels of a covariate, so the AUC at a specified covariate level is

$$AUC_{ij}(X) = \Pr\left(Y_i^D > Y_j^{\bar{D}} | X_i = X_j = X\right). \tag{2.2}$$

Dodd and Pepe (2003) applied this model to the Generalized Linear Model (GLM) framework. This approach allowed them to model the AUC and adjust the treatment effects for covariates. Their model can be written as

$$g\left(AUC_{ij}\right) = \mathbf{X_{ij}^T}\beta \tag{2.3}$$

where g is a monotone increasing link function, $\mathbf{X_{ij}}$ is a vector function of the covariates X_i and X_j, and β is a vector of the model parameters of interest. Dodd and Pepe proposed using the logistic or probit link function since

$$E\left(I\left(Y_i^D > Y_j^{\bar{D}}\right) | Z_i = z_i, Z_j = z_j\right) = AUC_{ij}$$

In the case where there are no covariates, the Mann-Whitney ranked sum U statistic given by,

$$\widehat{AUC} = \frac{\sum_{i=1}^{N_D} \sum_{j=1}^{N_{\bar{D}}} I\left(Y_i^D > Y_j^{\bar{D}}\right)}{N_D N_{\bar{D}}}$$

is an unbiased estimator of the AUC. Dodd and Pepe (2003) noted the parameters can be estimated by the usual score equations given by

$$\sum_{i=1}^{N_D} \sum_{j}^{N_{\overline{D}}} \frac{(I_{ij} - AUC_{ij})}{var\left(I_{ij}\right)} \frac{\partial AUC_{ij}}{\partial \beta}.$$

where $I_{iJ} = I\left(Y_i^D > Y_j^{\bar{D}}\right)$ is a Bernoulli random variable with probability of sucess given by $P(Y^D > Y^{\bar{D}})$. Solutions to the score equations can be found using statistical software such as the SAS/ Proc Logistic (SAS (2008b)). However, the model given above is based on the Mann-Whitney U statistic, which is just a sum of dependent Bernoullli random variables. In which case, the usual standard errors of the estimates are incorrect.

2.3 Estimation of Standard Errors

In this section, we discuss the four methods for finding the standard errors of the AUC regression model.

2.3.1 Bootstrap Method

Zhang (2008)) summarized their bootstrap method in the following steps:

(1) Stratify the range of the covariate variable as S strata. With a discrete covariate, each level of the covariate is a stratum. If the covariate is continuous, the values should be discretized into stratum making sure to have enough data in each strata.

(2) For a discrete covariate, generate all of the 0 or 1 indicator data within each stratum $s = 1, ..., S$ comparing the diseased and non-diseased group, $I\left(Y_{is}^D > Y_{js}^{\bar{D}}\right)$. The model is then $g\left(AUC_{ij}\right) = \beta_0 + \beta_s$.

(3) For a continuous covariate, an additional parameter is included in the model in order to fit the from comparing two responses from different covariate

values, such as $I\left(Y_{is}^{D} > Y_{js}^{\bar{D}}\right)$ where the covariate values are different but from the same stratum s. The model for this case is $g\left(AUC_{ij}\right) = \beta_0 + \beta_1 \mathbf{X_i} + \beta_{\mathbf{2}}\left(\mathbf{X_i} - \mathbf{X_j}\right)$.

(4) Use the logistic regression model to fit the data with covariates to obtain parameter estiamtes.

(5) Bootstrap the original data within each stratum to compute the parameter standard errors.

The bootstrapping procedure is computationally intensive and its use is highly restrictive if the covariates are continuous. In the remainder of this section we propose three analytical methods for estimating the standard errors for the regression parameters. The three methods are given as 1) DeLong (DeLong et al. (1988)), 2) Fligner (Fligner and Policello (1981)), and 3) Birnbaum (Birnbaum and Klose (1957)).

2.3.2 DeLong Method

DeLong et al. (1988) proposed an analytical method for estimating the variance of the AUC. The Mann Whitney rank sum U is the probability that an observation from a diseased (or treatment) group will be greater than or equal to an observation from a control (or placebo) group. Equivalence to the nonparametric AUC was first shown by Bamber (1975) where an estimate of AUC is given by

$$\widehat{AUC} = \sum_{i=1}^{N_D} \sum_{j=1}^{N_{\bar{D}}} I_{ij} \tag{2.4}$$

where

$$I_{ij} = \begin{cases} 1 & x_i > y_j \\ \frac{1}{2} & x_i = y_j \\ 0 & x_i < y_j \end{cases}.$$

Let

$$V_i^D = \frac{1}{N_{\overline{D}}} \sum_{j=1}^{N_{\overline{D}}} I_{ij}, \quad V_j^{\overline{D}} = \frac{1}{N_D} \sum_{i=1}^{N_D} I_{ij}$$

for $i = 1, ..., N_D$ and $j = 1, ..., N_{\bar{D}}$ where V_i^D is the relative rank of the i^{th} response of the diseased group in the non-diseased group and $V_j^{\bar{D}}$ is the relative rank of a non-diseased observation in the diseased group.

The DeLong estimate of the variance for the nonparametric AUC is

$$\text{Var}\left(\widehat{AUC}\right) = \frac{\text{Var}\left(V^D\right)}{N_D} + \frac{\text{Var}\left(V^{\overline{D}}\right)}{N_{\overline{D}}}. \tag{2.5}$$

2.3.3 Fligner Method

Fligner and Policello (1981) proposed a modification of the general class of nonparametric rank tests for the two-sample location parameter case. These included the Mann-Whitney Wilcoxon test. The method allows one to test the equality of two populations with fewer restriction on the shape of the null populations while preserving power as in the original tests. Fligner's procedure is based upon the use of placement values.

Define $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(m)}$ and $Y_{(1)} \leq Y_{(2)} \leq ... \leq Y_{(n)}$ to be the ordered X and Y samples, where X and Y are the two populations of interest. Let $Q_i, i = 1, ..., m$ be the rank of the i^{th} smallest X observation in the combined sample and R_j be the rank of of the j^{th} smallest Y observation in the combined sample. The placement value $P_i = Q_i - i$ is the number of Y observations less than X_i and $S_j = R_j - j$ is the placement value for Y_j. Note, that the Mann-Whitney statistic can be expressed in terms of the placement values where $U = \sum_{i=1}^{m} P_i$.

The Fligner estimate of the variance for the nonparametric AUC is

$$\text{Var}\left(\widehat{AUC}\right) = \frac{\sum \left(S_j - \bar{S}\right)^2 + \sum \left(P_i - \bar{P}\right)^2 + \bar{P}\bar{S}}{(nm)^2} \tag{2.6}$$

where $\bar{S} = \sum_{j=1}^{n} S_j/n$ and $\bar{P} = \sum_{i=1}^{m} P_i/m$ are the mean of the placement values.

2.3.4 Modified Birnbaum Method

Birnbaum and Klose (1957) derived a sharp upper bound and sharp lower bound for the variance of the Mann Whitney statistic. Our initial investigation of these bounds found that they were somewhat conservative. As a result, we propose to use

the midpoint between the upper and lower bound as a conservative estimate of the variance of the AUC. Let p denote the AUC, $p = \Pr(X > Y)$ and n, m the respective samples sizes as in the previous section. The upper bound for the variance of the Mann-Whitney statistics, U, as derived by Van Dantzig (1951).

$$\sigma^2(U) \leq mnp(1-p)\max(m,n). \tag{2.7}$$

The lower bound as derived by Birnbaum and Klose (1957) is

$$\sigma^2(U) \geq \begin{cases} \mu\nu\left[\mu r(1-r) - \frac{(\mu-1)^2}{12(\nu-1)}\right] & \frac{\mu-1}{\nu-1} \leq 2r \\ \mu\nu\left[\frac{4}{3}r\sqrt{2(\mu-1)(\nu-1)r} - (\mu+\nu-2)r^2 + r(1-r)\right] & \frac{\mu-1}{\nu-1} \leq 2r \end{cases} \tag{2.8}$$

where $\mu = \min(m,n)$, $\nu = \max(m,n)$, and $r = \min(p, 1-p)$. A simulation study in section 2.4 will compare all four methods of estimation.

2.3.5 Delta Method for the AUC Regression

In this section, the methods proposed by Zhang (2008) and Zhang et al. (2011) for estimating the standard errors in the logistic regression model are described. This method can be used for each of the three analytic methods presented in the previous section. The logistic regression model is

$$g\left(AUC_i\right) = \alpha + \beta x_i$$

with link function g. Suppose that the covariate of interest X_i, is a binary covariate such as gender. When $x = 0$, the model gives

$$\hat{\alpha} = g(\widehat{AUC} \mid x = 0).$$

When $x = 1$, the model results in

$$\hat{\alpha} + \hat{\beta} = g(\widehat{AUC} \mid x = 1)$$

An estimate of the β parameter is

$$\hat{\beta} = g(\widehat{AUC} \mid x = 1) - g(\widehat{AUC} \mid x = 0)$$

If we assume independence among the levels of the covariate, the variances of the parameters can be calculated as

$$var(\hat{\alpha}) = var(g(\widehat{AUC} \mid x = 0)),$$

and

$$var(\hat{\beta}) = var(g(\widehat{AUC} \mid x = 0)) + var(g(\widehat{AUC} \mid x = 1)).$$

By use of the delta method when g is the logit function, we have

$$var(logit(\widehat{AUC_i})) = \frac{var(\widehat{AUC_i})}{\widehat{AUC_i}^2(1 - \widehat{AUC_i}^2)}.$$

2.4 Simulation Study

A simulation study was designed to compare the three analytical methods [DeLong, Fligner, and Birnbaum midpoint] to the bootstrap method proposed by Dodd and Pepe. The simulation cases were adapted from an approach given by Dodd and Pepe (2003). The models have a covariate consisting of three strata levels with sample size n for each group and level. Data were generated such that $Y_i^{\bar{D}} = -log\,(u_1) + \delta_{1i}$ and $Y_i^D = -log\,(u_2) + \delta_0 + (\delta_{1i} + \delta_{2i})$ where $u_1 \sim$ exponential(1) and $u_2 \sim$ exponential(1). The parameters in the model can be derived using

$$AUC_i = F\,(\delta_0 + \delta_{2i})$$

where $F(x) = (1 + e^{-x})^{-1}$ is the cdf of a logistic random variable Balakrishnan and Nevzorov (2003).

For $i = 1$ and $\delta_{2i} = 0$, one has

$$AUC_1 = F\,(\beta_0)$$

and when $i = 2, .., S$

$$AUC_i = F\,(\beta_0 + \beta_{i-1})\,,$$

where $\beta_0 = \delta_0$ and $\beta_j = \delta_{2(j+1)}$ where $j = 1, ..., S-1$. Table 2.1 and Table 2.2 contain the results for sample sizes $n = 30$ and $n = 100$ in each level of each group when

Table 2.1: Parameter estimates for the four methods (n = 30)

		Bootstrap			DeLong		
	True			Coverage			Coverage
Parameter	Value	Est.	S.E.	95% CI	Est.	S.E.	95% CI
β_0	0.15	0.16	0.32	0.963	0.14	0.31	0.959
β_1	0.50	0.50	0.46	0.958	0.53	0.45	0.947
β_2	1.00	1.02	0.49	0.971	1.03	0.47	0.955

		Fligner			Birnbaum		
	True			Coverage			Coverage
Parameter	Value	Est.	S.E.	95% CI	Est.	S.E.	95% CI
β_0	0.15	0.16	0.31	0.962	0.16	0.34	0.968
β_1	0.50	0.50	0.45	0.968	0.51	0.49	0.976
β_2	1.00	1.02	0.48	0.950	1.02	0.51	0.957

Table 2.2: Parameter estimates for the four methods (n = 100)

		Bootstrap			DeLong		
	True			Coverage			Coverage
Parameter	Value	Est.	S.E.	95% CI	Est.	S.E.	95% CI
β_0	0.15	0.15	0.17	0.956	0.15	0.17	0.952
β_1	0.50	0.51	0.24	0.952	0.51	0.24	0.960
β_2	1.00	1.00	0.25	0.957	1.01	0.25	0.951

		Fligner			Birnbaum		
	True			Coverage			Coverage
Parameter	Value	Est.	S.E.	95% CI	Est.	S.E.	95% CI
β_0	0.15	0.15	0.17	0.957	0.15	0.18	0.964
β_1	0.50	0.51	0.24	0.953	0.51	0.26	0.966
β_2	1.00	1.01	0.25	0.950	1.01	0.27	0.958

$\delta_0 = 0.15$, $\delta_{1i} = 0$, $\delta_{22} = 0.5$, and $\delta_{23} = 1$. Results represent 1000 realizations of the model and bootstrap samples of size 200 for each population.

From both tables we can see that the Fligner method is nearly identical to the DeLong method. The Birbaum method tends to come in just slightly above the values obtained from the other methods. Based on this simulation, the Fligner

method perform exceptionally well at estimating the standard errors. The Birnbaum method clearly overestimates the standard errors, as we expected. In general we saw coverage well above the other methods. The Fligner and DeLong methods had very similar coverage close to the expected 95%. We would consider both methods a viable replacement for the bootstrap.

2.5 Clinical Trial Example

All four methods were applied to data from a clinical trial investigating the efficacy a drug for urinary incontinence in North American women using a placebo control. This is the same example considered by Zhang (2008). The primary efficacy measure is the percent change in number of episodes per week from baseline to the final visit. The severity of the problem at baseline is used to define 4 strata and can take values from 1, indicating mild, to 4, indicating severe. A covariate of interest indicates whether a patient has had hormone replacement therapy (horm = yes(1) or no(0)). We also investigated the interaction between disease severity and this covariate. Since the data is non-normal and highly skewed in each group and level, it is ideal for our nonparametric method. Overall we are interested in analyzing the joint predictive and prognostic effects of the disease severity and hormone replacement therapy.

The hypothesis of interest is

$$H_0 : P\left(Y^T > Y^P\right) = 0.5$$

versus

$$H_1 : P\left(Y^T > Y^P\right) > 0.5,$$

where Y^T is the relative percent reduction in incontinence episodes from baseline to endpoint for the treatment group and Y^P is likewise for the the placebo group. Note that we only considered the bootstrap, DeLong, and Fligner methods for the estimates of the variance of the AUC. The AUC logistic regression model of interest is

$$\begin{aligned}\text{logit}(\text{AUC}(\text{horm,z})) = \beta_0 &+ \beta_1 I(\text{horm} = 0) + \beta_2 I(\text{z} = 1) \\ &+ \beta_3 I(\text{z} = 2) + \beta_4 I(\text{z} = 3) + \beta_5 I(\text{z} = 1 \ \& \ \text{horm} = 0) \\ &+ \beta_6 I(\text{z} = 2 \ \& \ \text{horm} = 0) + \beta_7 I(\text{z} = 3 \ \& \ \text{horm} = 0)\end{aligned}$$

The tables below summarize the results, with Table 2.3 containing a comparison of the parameter estimates for the four methods and Table 2.4 containing a comparison of the estimates of the AUC with confidence intervals for three methods, excluding the Birnbaum method. We see that disease severity level 3 is significantly different from the other levels for overall treatment effects. Also those patients with disease severity level 3 that have had hormone replacement therapy has a significant effect. While the bootstrap method results in insignificance for those patients in disease severity level 4 and no hormone replacement therapy The DeLong and Fligner methods find significance as well.

2.6 Discussion

In this chapter, we investigated three analytical methods for computing the parameter estimates and standard errors for the semi-parametric AUC regression model, as proposed by Dodd and Pepe (2003), with discrete covariates. All three were compared to the bootstrapping method and were found to be much less computationally intense to calculate and easier to implement.

The DeLong method used the relative rank of the observations, the Fligner method made use of the relationship between the placement values and the AUC, and our modified Birnbaum method used the midpoint of the sharp upper bound and sharp lower bound of the AUC. We were able to calculate the standard errors of the parameters and the variance of the AUC using the delta method based on the development done by Zhang (2008) and (Zhang et al. (2011). Simulation studies showed that for small sample sizes, the DeLong and Fligner methods are comparable to the bootstrap method. We expected the midpoint method to be considerably more conservative then the other methods presented in this paper. Note that we

Table 2.3: Parameter Estimates for the Four Methods

			Bootstrap		DeLong	
Parameters	Level	Estimate	SD	95% CI	SD	95% CI
Intercept		-0.51	0.30	(-1.11, 0.08)	0.31	(-1.11, 0.09)
z	1	0.65	1.92	(-3.12, 4.41)	0.97	(-1.25, 2.54)
z	2	0.76	0.40	(-0.02, 1.55)	0.43	(-0.09, 1.62)
z	3	-1.46	0.72	(-2.87, -0.05)	0.63	(-2.69, -0.23)
Horm	0	-0.13	0.45	(-1.01, 0.75)	0.47	(-1.04, 0.79)
z*Horm	1 and 0	0.16	1.17	(-2.14, 2.46)	1.38	(-2.54, 2.86)
z*Horm	2 and 0	-0.43	0.66	(-1.72, 0.86)	0.78	(-1.96, 1.10)
z*Horm	3 and 0	1.71	0.90	(-0.05, 3.47)	0.91	(-0.07, 3.49)

		Fligner		Birnbaum	
Parameters	Level	SD	95% CI	SD	95% CI
Intercept		0.30	(-1.11 0.08)	0.32	(-1.14, 0.13)
z	1	0.75	(-0.85, 2.15)	0.77	(-0.86, 2.16)
z	2	0.42	(-0.08, 1.60)	0.46	(-0.14, 1.66)
z	3	0.62	(-2.7, -0.22)	0.70	(-2.83, -0.09)
Horm	0	0.45	(-1.01, 0.75)	0.49	(-1.09, 0.83)
z*Horm	1 and 0	1.17	(-2.14, 2.46)	1.21	(-2.21, 2.53)
z*Horm	2 and 0	0.75	(-1.93, 1.07)	0.83	(-2.06, 1.20)
z*Horm	3 and 0	0.90	(-0.05, 3.47)	0.99	(-0.23, 3.65)

Table 2.4: Estimates of AUC with the Four methods

z	Horm	AUC Estimate	Bootstrap 95% CI	DeLong 95% CI	Fligner 95% CI
1	0	0.458	(0.068, 0.848)	(0.138, 0.778)	(0.11, 0.81)
	1	0.467	(0.019, 0.914)	(0.140, 0.793)	(0.127,0.807)
2	0	0.576	(0.420, 0.732)	(0.412, 0.739)	(0.418, 0.734)
	1	0.437	(0.289, 0.585)	(0.300, 0.575)	(0.289, 0.585)
3	0	0.596	(0.435, 0.757)	(0.428, 0.764)	(0.432, 0.76)
	1	0.878	*(0.763, 0.993)*	*(0.751, 1.000)*	*(0.76, 0.996)*
4	0	0.654	(0.497, 0.811)	*(0.514, 0.794)*	*(0.502, 0.806)*
	1	0.625	(0.484, 0.766)	(0.492, 0.759)	(0.487, 0.763)

originally investigated using the maximum as an estimate of the variance, however we found this to be overall too conservative. In general a 95% confidence interval of

the AUC, using only the sharp upper bound as the variance, covered the entire range of all possible values of the AUC. We found the midpoint to be a more reasonable estimate, although still somewhat overestimated. Any of these methods are simple to implement but we found the Fligner and DeLong to be the most successful methods, as they are both easy to implement, as compared to the bootstrap, and accurate in estimation, as compared to our Birnbaum midpoint method.

CHAPTER THREE

AUC Regression for $k > 2$ Ordered Treatment Effects

In this chapter, we extend the AUC regression model, with and without discrete covariates, to the situation where there are $k > 2$ treatment levels as the alternative hypothesis. This assumption allows one to test the null hypothesis with the Jonckheere Trend statistic (Jonckheere (1954) and Terpstra (1952)). A simulation study is used to investigate the properties and performance of the new method. A real world example is investigated to further illustrate the proposed method.

3.1 Jonckheere-Terpstra Statistic

Suppose that one observes data from $k > 2$ populations. Let U_{uv}, $u < v = 2, 3..., k$ denote the Mann-Whitney statistic for the u^{th} and v^{th} samples given by

$$U_{uv} = \sum_{s=1}^{n_u} \sum_{t=1}^{n_v} I\left(X_{vt} > X_{us}\right).$$

Suppose that one is interested in testing

$$H_0 : \theta_1 = \theta_2 = ... = \theta_k$$

vs.

$$H_1 : \theta_1 \leq \theta_2 \leq \cdots \leq \theta_k$$

with at least one strict inequality. Terpstra (1952) and Jonckheere (1954) working independently considered this problem and derived what is now called the Jonckheere-Terpstra statistic,

$$V_2 = \sum_{u<v}^{k} \sum U_{uv}.$$

Computationally, V_2 is the sum of $k(k-1)/2$ Mann-Whitney statistics.

It has been shown that for ordered alternatives, the Jonckheere-Terpstra statistic is preferable to tests of more general class difference alternatives (such as the

Kruskal-Wallis). V_2 has been shown to have an asymptotically normal distribution under the null hypothesis where

$$Z = \frac{V_2 - E_0(V_2)}{\sqrt{Var_0(V_2)}} \sim N(0,1)$$

when H_0 is true where

$$E_0(V_2) = \frac{N^2 - \sum_{j=1}^{k} n_j^2}{4}$$

and

$$Var_0(V_2) = \frac{N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)}{72}.$$

Puri (1965) derived the asymptotic efficiency and asymptotic power of the test. Odeh (1971) originally provided tables of exact probabilities and critical values for nominal values of $\alpha = 0.5, 0.2, 0.1, 0.05, 0.025, 0.01,$ and 0.005 for $k = 3$ groups and small sample sizes (from $n = 2$ through 8). In addition he derives a recurrence formula for computing the exact distribution. The asymptotic distribution and power is also considered by Bartholomew (1961).

Randles and Wolfe (1991) presented an alternate method of calculating the Jonckheere-Terpstra statistic as follows: Let

$$U_s^* = \sum_{i=1}^{s-1} U_{is}$$

where U_s^* is the Mann-Whitney statistic between the first $(s-1)$ groups and the s^{th} group for $s = 2, ..., k$. Odeh (1971) showed that $U_2^*, ..., U_k^*$ are independent and the Jonckheere-Terpstra statistic can be written as the sum of $k-1$ Mann-Whitney statistics,

$$V_2 = \sum_{s=2}^{k} U_s^*.$$

To illustrate this method, let $k = 3$, in which case we compute $k-1 = 2$ Mann-Whitney statistics. The first is U_2^* for $H_{01} : \alpha_1 = \alpha_2$ and U_3^* for $H_{02} : (\alpha_1, \alpha_2) = \alpha_3$. In this alternative approach, we have exchanged an increase in book-keeping (combining groups) with a significant decrease in the number of computed Mann-Whitney statistics when k is moderate to large.

Since the JT statistic is the sum of $(k-1)$ Mann-Whitney statistics, we can use the approach given by Zhang (2008) for adding covariates when considering the equality of the k treatment groups versus the ordered alternative hypotheses. A simulation study given in the next section presents a method for incorporating covariates in the JT setting.

3.2 Simulation Study without Covariates

We want to investigate the performance of the Jonckheere-Terpstra statistic in a model with $k = 3$ different treatment levels. Suppose we have the model

$$E(y_{ij}) = \mu + \alpha_i$$

where the treatment effect of interest is α_i for $i = 1, 2, 3$. The cases of interest are:

(1) $\alpha_1 < \alpha_2 < \alpha_3$

(2) $\alpha_1 < \alpha_2 = \alpha_3$

(3) $\alpha_1 = \alpha_2 < \alpha_3$

(4) $\alpha_1 = \alpha_2 = \alpha_3$

Note that we expect situation 2 may not perform as well as the other cases. Recall that to calculate the Jonckheere-Terpstra statistic we are grouping the first and second groups together to compare to the third group, which would not be appropriate in the second situation. We will still provide the results of the situation to see how the Jonckheere-Terpstra statistic performs. For these four cases, we consider two simulations from different distributions.

(1) We sample from a normal distribution with mean 0 and $\sigma^2 = 4, 16$. Each treatment group is of size $n = 30$. Figures 3.1, 3.2, and 3.3 contain the results for the Jonckheere-Terpstra statistic, the corresponding Z score, and p-value when $\sigma^2 = 4$. Figures 3.4, 3.5 and, 3.6 contain the similar results when $\sigma^2 = 16$.

(2) In this case $Y = -log\,(u_1)$ where $u_1 \sim$ exponential(1). Each group has sample size $n = 30$. Figures 3.7, 3.8, and 3.9 contain the results for the JT statistic, Z score, and p-value.

The horizontal dotted lines on the figures represent the rejection region for Type I error, type 1 error = 0.05. The results are based on 1000 realizations of the model. In Figures 3.1 through 3.6 we see that the Jonckheere-Terpstra statistic performs exceptionally well for normally distributed data. As anticipated, we reject the global null hypothesis for the first, second, and third situations and fail to reject for the fourth situation where the null is true.

When using the non-normal data, Figures 3.7, 3.8, and 3.9 indicate that the JT statistic also performs well. We anticipate that one would reject the null hypotheses in the first three cases and fail to reject in the fourth case. We observe an increase in the values for the p-value of the JT statistic. This is especially in case 2, although we are still correctly rejecting the null hypothesis whenever there is at least one strict inequality in the treatment effects. Overall, the Jonckheere-Terpstra statistic performs well when using non-normal data.

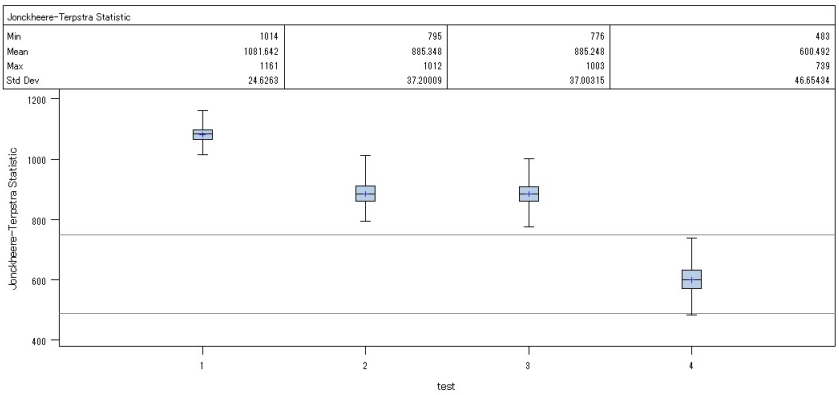

Figure 3.1: The Jonckheere-Terpstra Statistic for the four situations where $n_i = 30$ and $\sigma^2 = 4$

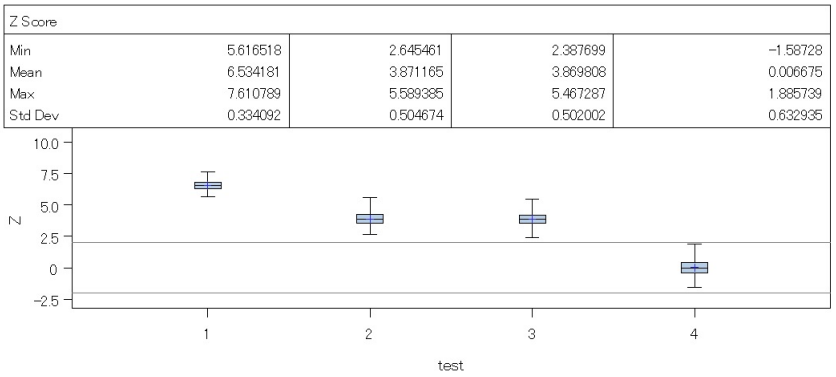

Figure 3.2: The Z score

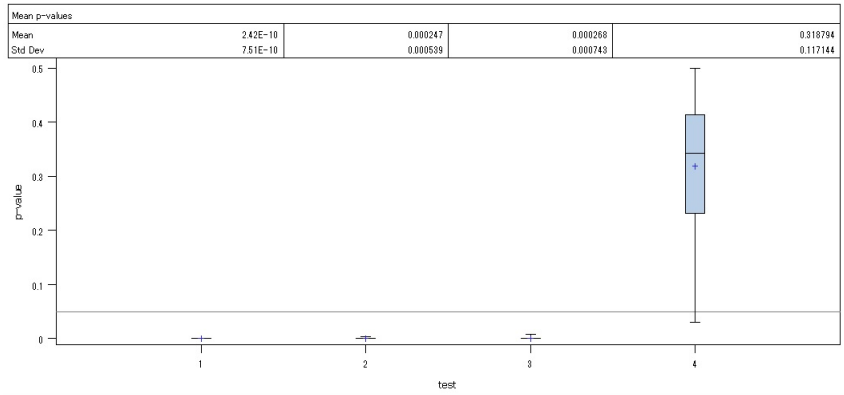

Figure 3.3: The p-value

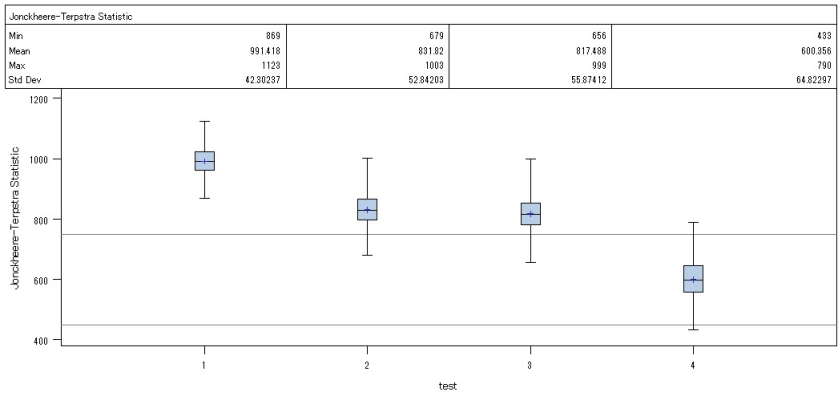

Figure 3.4: The Jonckheere-Terpstra Statistic for the four situations where $n_i = 30$ and $\sigma^2 = 16$

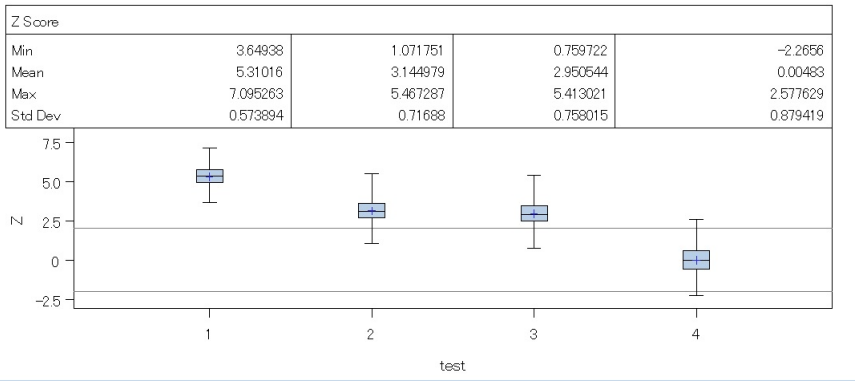

Figure 3.5: The Z score

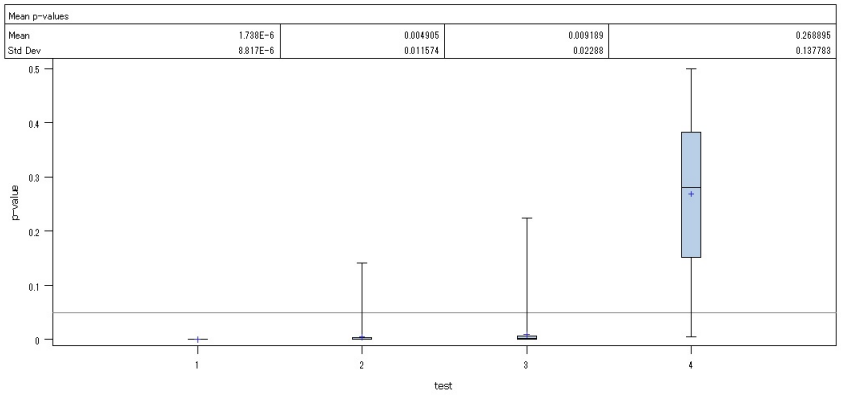

Figure 3.6: The p-value

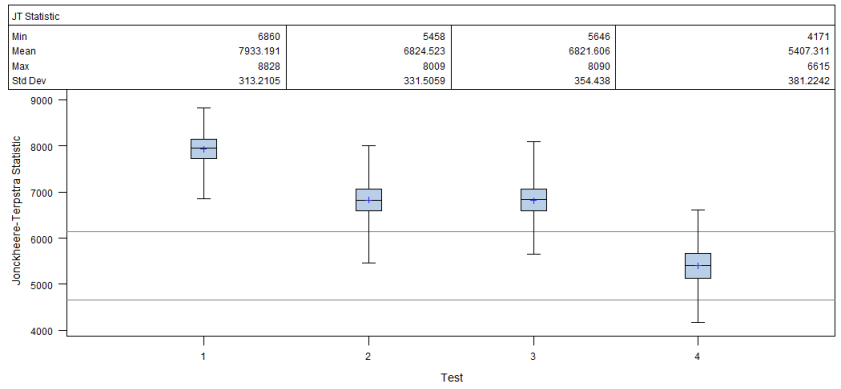

Figure 3.7: The Jonckheere-Terpstra Statistic for the four cases with non-normal data

3.3 The JT Statistic with Discrete Covariates

Dodd and Pepe (2003) and Zhang (2008) presented procedures for including discrete covariates in the general two populations case. This involved estimating the

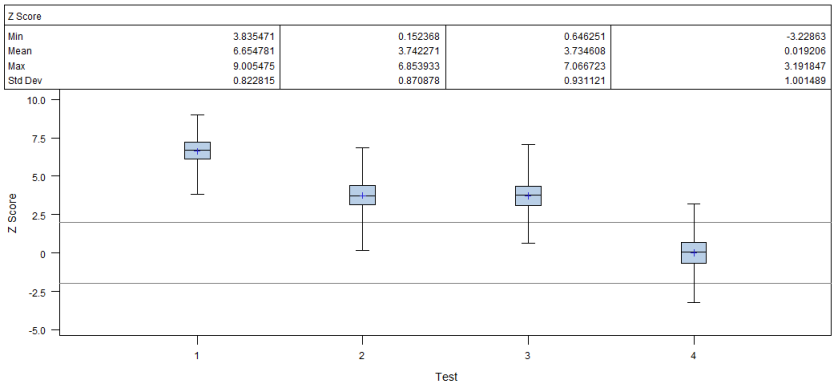

Figure 3.8: The Z-Scores

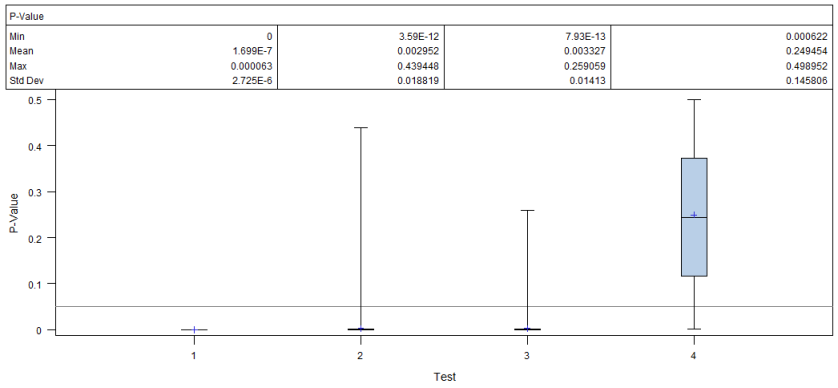

Figure 3.9: The p-values

correct standard errors by the bootstrap method or an alternative analytical method (such as DeLong et al. (1988)). In order to add covariates to the Jonckheere-Terpstra model framework we must address a few unique hurdles. For a discrete covariate, X the model is given by

$$y_{ijk} = \alpha_i + \beta_j x_{ij} + \epsilon_{ijk}$$

where α_i is the treatment effect for $i = 1, .., k$, x_{ij} is the covariate vector and β_j is the unknown parameter for the j^{th} covariate. Recall that the JT statistic is appropriate when an ordered alternative hypothesis holds. Thus we must assume that this ordering holds for each level of the covariates. In which case, we assume that one of the following situations are true:

(1) $\alpha_i + \beta_j$ has the same ordering as the treatment effect α_i for each value of the covariate.

(2) β_j is constant for all j.

These assumptions are highly restrictive, but if the covariate β violates the ordered assumption of the model, the Jonckheere-Terpstra statistic will no longer appropriate.

Adapting the AUC logistic regression model approach to the problem considered by the Jonckheere-Terpstra statistic is simple to implement. Suppose that the covariate, X is binary. If $x_{ij} = 0$

$$y_{ijk} = \alpha_i + \epsilon_{ijk}$$

the basic Jonckheere-Terpstra trend test considered in a previous section. If $x_{ij} = 1$ then

$$y_{ijk} = \alpha_i + \beta + \epsilon_{ijk}.$$

In this model the medians are simply shifted over β_j units. Thus, the Jonckheere-Terpstra statistic can be calculated by separating the data into groups based on covariate values. For each level of the covariate $(k-1)$ Mann-Whitney statistics are used to compute the JT statistic. Under the above assumption, the Jonckheere-Terpstra statistic should come to the same conclusions at each level of the covariate. The estimates of the parameter can be obtained using the methods described in Zhang (2008) and Dodd and Pepe (2003) where the standard errors are computed using the analytical method given in DeLong et al. (1988) and Zhang (2008). The next section presents a simulation study to investigate the performance of the statistic when including covariates.

3.4 Simulation Study

The purpose of this simulation study is to evaluate the performance of the Jonckheere-Terpstra statistic when applied to AUC regression with discrete covariates. Consider a model with $k = 4$ different treatment effects: a placebo and three

ordered treatment levels. Cases were generated for models having a categorical covariate with three levels, $x_{ijk} = I_{(0,1,2)}(j)$, and sample size n for each group and level. We assume that β_j has the same ordering as α_i. Data were generated such that $Y_i^P = -log\,(u_1) + \alpha_1 + \beta_j x_{ijk}$ and $Y_i^T = -log\,(u_i) + \alpha_i + \beta_j x_{ijk}$ where $i = 2, 3, 4$. We considered the Jonckheere-Terpstra statistic for the following five cases

(1) $\alpha_1 < \alpha_2 < \alpha_3 < \alpha_4$

(2) $\alpha_1 < \alpha_2 < \alpha_3 = \alpha_4$

(3) $\alpha_1 = \alpha_2 < \alpha_3 < \alpha_4$

(4) $\alpha_1 = \alpha_2 = \alpha_3 < \alpha_4$

(5) $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$

The results for the Jonckheere-Terpstra statistic in all 5 cases are given in Figures 3.10, 3.11, and 3.12. The horizontal lines represent the rejection region for $\alpha = 0.05$. The results are based on 500 realizations of the model. Table 3.1 gives the average summary estimates from the box plots. Note that parameter estimates of the treatment effects and covariate are given only for case 1 in Table 3.2.

We use the method given in Zhang (2008) to obtain estimates of the covariates and individual intercepts for the treatment effects. We can see from the figures that the Jonckheere-Terpstra statistic with discrete covariates performs exceptionally well on skewed data. The first four cases have at least one strict inequality and each have a significant Jonckheere-Terpstra statistic resulting in the rejection of the null hypothesis. For cases 2,3, and 4 we observe increased p-values, but the same overall significance results. As the number of strict inequalities decreases, our JT statistic and Z scores become smaller and closer to the rejection line. The final case, with all equalities, is almost entirely contained between the two lines and correctly fails to reject the null. As expected, the Jonckheere-Terpstra statistic for each covariate is approximately the same, since the data is simply shifted β_j units. Overall, the model

performs exceptionally well whenever the restrictive assumptions hold and the model is appropriate.

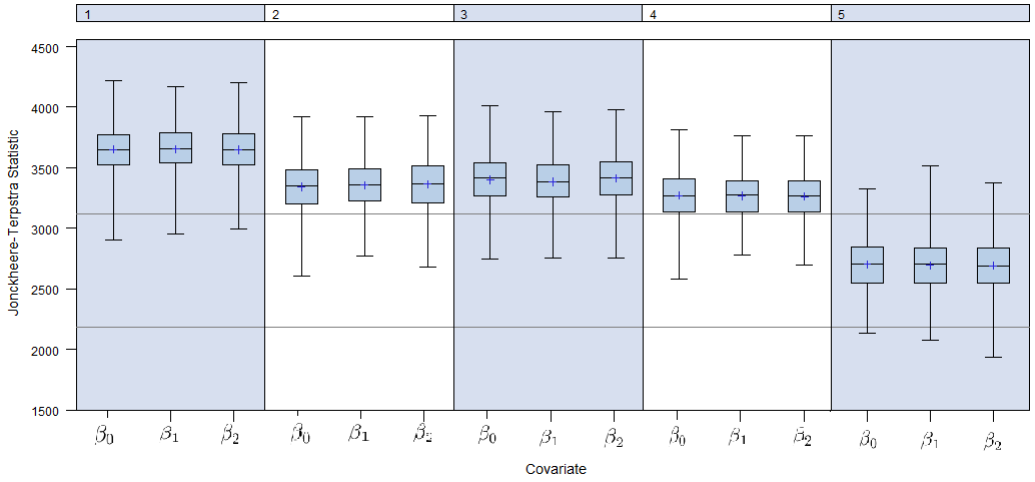

Figure 3.10: The Jonckheere-Terpstra Statistic for the five cases by three levels of the covariate β_j (n=30)

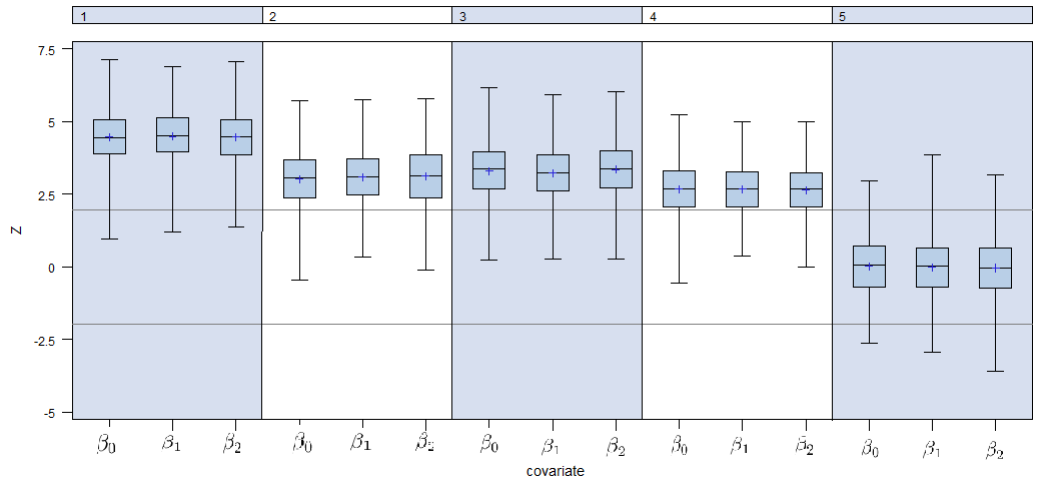

Figure 3.11: The Z score corresponding to the Jonckheere-Terpstra statistic for the five cases by three levels of β_j

3.5 Clinical Trial Example

The method is applied to data from the cognitive rehabilitation study where the purpose is to determine the efficacy of two treatment programs for preventing cognitive impairment and functional decline in hospital patients 70 years of age or older.

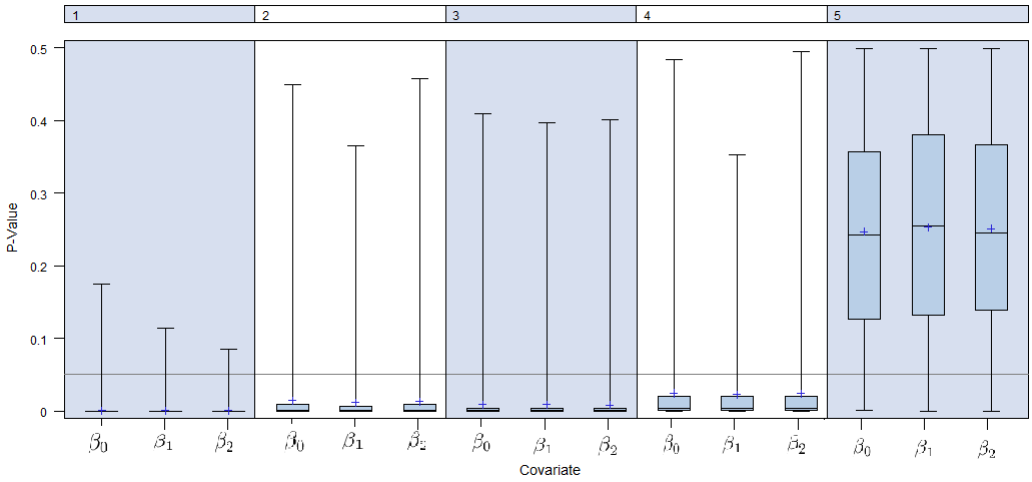

Figure 3.12: The p-values corresponding to the Jonckheere-Terpsta statistic for the five cases by three levels of β_j

Table 3.1: Average Jonckheere-Terpstra Statistic, Z score, and p-value estimates for the five cases

	Parameter	JT	Z Score	p-value
Case 1	β_0	3651	4.46	0.0005
	β_1	3655	4.47	0.0007
	β_2	3650	4.45	0.0006
Case 2	β_0	3344	3.02	0.015
	β_1	3358	3.09	0.012
	β_2	3364	3.11	0.013
Case 3	β_0	3402	3.29	0.008
	β_1	3385	3.21	0.009
	β_2	3412	3.34	0.008
Case 4	β_0	3271	2.68	0.025
	β_1	3269	2.67	0.022
	β_2	3263	2.64	0.024
Case 5	β_0	2703	0.01	0.24
	β_1	2696	−0.01	0.25
	β_2	2692	−0.03	0.25

A significant percentage of older patients in hospitals, experience delirium episodes. This study was designed to help increase quality of life and reduce adverse events for the elderly patient population. The Hospital Elder Life Program (HELP) was de-

Table 3.2: Parameter Estimates for Case 1

Parameter	True Value	Est	DeLong S. E.	Coverage 95% CI
α_1	0.55	0.58	0.32	0.950
α_2	0.90	0.91	0.34	0.948
α_3	1.20	1.24	0.36	0.952
β_1	0.50	0.48	0.47	0.948
β_2	1.00	1.03	0.51	0.952

signed for delirium prevention and maintaining or improving cognitive and functional ability. Researchers are investigating if including interactive gaming technology, the Wii^{TM} Nintendo system, can provide additional improvement in cognitive, functional, and quality outcomes of patients.

The response variable is the percent change in a patient's score on a functional test, the Patient Activities of Daily Living (ADL), from baseline (time of enrollment in the study) to the end of the study (discharge from hospital care). The ADL score ranges in value from 0 to 20, where a higher score indicates greater independence. Functional abilities such as grooming, dressing, feeding, etc. are evaluated in the patients. The null hypothesis is

$$H_0 : \alpha_{\text{No Treatment}} = \alpha_{\text{HELP}} = \alpha_{\text{HELP+Wii}}$$

vs.

$$H_0 : \alpha_{\text{No Treatment}} \geq \alpha_{\text{HELP}} \geq \alpha_{\text{HELP+Wii}},$$

with at least one strict inequality where the "no treatment" group was randomly selected from historical (pre-HELP) data. A patient's gender is a covariate (women(1) or men (0)) of interest since we expect women to have a better response than men. We assume that the covariate is constant for each treatment group. Table 3.3 contains the mean (and standard deviation) of ADL at baseline and discharge for each of the patient arms and the average percent ADL change per patient. The % change in

ADL is calculated by

$$\% \text{ ADL change } = \frac{\text{ADL Score at Baseline} - \text{ADL Score at Discharge}}{\text{ADL Score at Baseline}} \times 100$$

Table 3.4 gives the sample sizes by gender for each of the study arms. The majority of patients see some function decline while in the hospital and the goal of this study is to limit this. Table 3.3 illustrates that the patients in the study tend to have minimal decline or even greater functional ability compared to the retrospective group. Since we would like to see a higher patient ADL at discharge, a smaller or negative % ADL change is more desirable.

Table 3.3: Average ADL Scores

Level	No Treatment	HELP	HELP + Wii
Baseline	13.71 (6.16)	12.99 (5.42)	12.61 (5.27)
Discharge	13.27 (5.20)	12.85 (5.35)	13.21 (5.82)
Average % ADL Change	10.62	8.06	1.24

Table 3.4: Sample sizes by Gender

Level	No Treatment	HELP	HELP + Wii
Female	80	67	68
Male	70	48	52
Total	150	115	120

The results given in Table 3.5, indicate significant difference in the treatment effects for the three groups. The Jonckheere-Terpstra statistic is nearly identical when adjusting for gender. The results demonstrate that the new rehabilitation program involving the Wii is effective in reducing delirium episodes and increasing functional status in older patients when compared with the historical control group of having no treatment. In the next chapter we introduce a method for making multiple treatment comparisons, thereby, enabling one to determine if there are any differences among the two therapeutic treatment groups.

Table 3.5: Average Jonckheere-Terpstra Statistic, Z score, and p-value estimates by covariate

Parameter	JT	Z Score	p-value
JT_{β_0}	4166	4.56	0.0004
JT_{β_1}	4167	4.56	0.0002

3.6 Discussion and Conclusions

In this chapter we presented a method for applying the Jonckheere Trend test to a semi-parametric regression model with discrete covariates. Dodd and Pepe (2003) introduced AUC regression by exploiting the relationship between the Mann-Whitney U statistic and the AUC. Since the Jonckheere-Terpstra statistic can be shown to be the sum of independent Mann-Whitney statistics, we were able to exploit this relationship and extend the semi-parametric regression model to the case of having $k > 2$ ordered treatment effects when one needs to adjust the model for discrete covariates. The Jonckheere-Terpstra statistic is more powerful when used in cases when treatment effects are ordered under the alternatives hypothesis. Since, this ordered relationship must be maintained when introducing covariates, we placed a restriction on the covariates such that one of the following conditions is true: 1) the covariate is constant for all treatment levels, or 2) the covariate maintains the same order as found in the unadjusted treatment effects.

The simulation study showed that the small-sample performance of the method performs exceptionally well at detecting the specific differences in the treatment groups. The results using the covariates performed equally well, but this was as expected since these covariates are highly restrictive for the Jonckheere Trend test.

CHAPTER FOUR

Multiple Comparison Methods for the Jonckheere Trend Tests

In the previous chapter, we extended the method suggested by Zhang (2008) and Zhang et al. (2011) to the problem of testing for treatment differences under the ordered alternative hypotheses as considered by the Jonckheere trend test. The Jonckheere-Terpstra statistic is used to determine if there is an overall global treatment difference among the k treatment levels. In this chapter, we present a new multiple comparison procedure when the Jonckheere trend test is appropriate. A simulation study is performed to compare the proposed method with two existing nonparametric multiple comparison procedures (Shirley (1977) and Nashimoto and Wright (2007)) that are used under the ordered alternative assumption. A real data example is also presented.

4.1 Introduction

Multiple comparison methods are regularly used in many areas of data analysis. A common goal of clinical trials is to compare active treatment groups with a placebo group in order to determine the efficacy of the treatments as compared with a control. The investigation does not usually end here as the investigators may be interested in how and which treatments differ from the placebo or in some cases from each other. For example, some treatment may work better for younger patients, for women than versus men, or for patients with a lower body mass index. Questions of this type require models that allow for covariance structure. Another question of interest involves multiple comparison procedures. For example in dosage studies, researchers are interested in determining the minimum effective dose of a treatment by comparing many levels. In examples of this type, it is not enough to know that there are

treatment differences. One needs to determine the lowest level dose for when a treatment is effective.

Multiple comparison methods have been widely studied and there are many different types of procedures where the choice of which one uses depends upon a variety of factors and desired outcomes. However, each share a common attribute, that as the number of comparisons increase there is an increase in the probability of detecting treatment differences due to chance alone. Thereby, leading to incorrect conclusions. Westfall et al. (1999) and Hsu (1996) provide extensive discussion on different multiple comparison methods and their properties. In this chapter, we present a multiple comparison method using the Jonckheere-Terpstra statistic as applied in the AUC regression setting in order to test for treatment differences in the case where the alternative hypothesis has assumed ordered treatment effects. A brief discussion of two existing multiple comparison procedures, Shirley (1977) and Nashimoto and Wright (2007), is given in Section 4.2. The proposed multiple comparison method is presented in Section 4.3. Section 4.4 provides a simulation study comparing our method with the existing procedures. Section 4.5 uses an example given in Shirley (1977) to compare the three methods. Conclusions and a final discussion are given in Section 4.6.

4.2 Multiple Comparison Methods

The problem of interest is to determine differences in location parameters for $k > 2$ treatment groups using a non-normal response variable, Y when the researchers believe that the treatment effects are ordered under the alternative hypothesis. Shirley (1977) and Nashimoto and Wright (2007) proposed nonparametric methods that have been used for this situation. A brief discussion of their procedures is included below.

4.2.1 The Shirley Method

Shirley (1977) considered the problem of determining differences in treatment groups that are created by increasing dosage levels of an active compound as compared

with a zero dose control group. The test is a nonparametric version of a parametric test given by Williams (1971).

Suppose there are k active treatment levels and a zero-dose control group. Williams (1971) proposed a procedure based upon the maximum likelihood estimates of the location parameters, M_i, subject to the constraint that $M_1 \leq M_2 \leq ... \leq M_k$. The statistic is

$$t_k = \frac{\hat{M}_k - X_0}{(S^2/r_k + S^2/c)^{-1/2}}$$

where S^2 is an estimate of the residual variance, $c = r_0$ is the number of observations in the control group and X_0 is the control group sample mean. Williams (1971) provided tables for the critical points of the t_k statistic.

Shirley (1977) developed a nonparametric version of the Williams test by analyzing the observed ranks instead of the actual data. The results were based on the Wald-Wolfowitz limit theorem (Wald and Wolfowitz (1944)), where the vector $\bar{R} = \left(\bar{R}_0, \bar{R}_1, ...\bar{R}_k\right)$ has a limiting multivariate normal distribution and $\bar{R}_i$ is the mean rank of group i. The Shirley multiple comparison test is as follows. For equal group sizes, let

$$t = C_{N,k} \left[\underset{1\leq u\leq k}{\text{Max}} \sum_{j=u}^{k} \bar{R}_j(k-u+1) - R_0\right] \tag{4.1}$$

where $C_{N,k} = [(k+1)(N+1)/6]^{1/2}$ and N is the total sample size. t is shown to be approximately distributed as Williams' t_k with $\nu = \infty$. If the samples sizes are unequal or there are a considerable number of ties in the data, the statistic becomes

$$t = C_{N,k} \left[\underset{1\leq u\leq k}{\text{Max}} \left(\sum_{j=u}^{k} r_j\bar{R}_j / \sum_{j=u}^{k} r_j\right) - \bar{R}_0\right] \tag{4.2}$$

where $C_{N,k} = [(N(N+1)/12)(1/r_k + 1/C)]^{1/2}$. The Shirley multiple comparison test compare each treatment level to the zero-dose control group using either equation (4.1) or (4.2) and the critical points given by Williams (1971).

4.2.2 The Nashimoto and Wright Method

Hayter (1990) proposed a one-sided Studentized-range test for an ordered alternative hypothesis. Nashimoto and Wright (2007) extended Hayter (1990) to a

rank-based multiple comparison procedure (NPM). Assume the k populations are identical except for possibly different locations and, M_i, then one has a significant median inequality $M_i < M_j$ for any $1 \leq i < j \leq k$ if

$$\underset{i \leq m \leq m' \leq j}{\text{Max}} \left(\frac{\bar{R}_{m'} - \bar{R}_m}{\sigma/\sqrt{n}} \right) \geq h_{\alpha,k,\nu} \tag{4.3}$$

where $\bar{R}_i$ is the average rank for the i^{th} group and $\sigma_a = \sqrt{N(N+1)/12}$.

The critical value, $h_{\alpha,k,\nu}$, is determined by Hayter (1990) so that the familywise error rate is α. The degrees of freedom is $\nu = \infty$ when using equation (4.3). Hayter (1990) provides the critical value, $h_{\alpha,k,\nu}$ for $k = 3, 4, \ldots, 9$ with a range of degrees of freedom, ν, between 5 and ∞ and $\alpha = 0.10, 0.05$.

4.3 A New Multiple Comparison Method

In this section, we present a new procedure for multiple comparisons of $k > 2$ treatment medians when the alternative hypothesis is ordered. The method is based upon AUC regression and the Jonckheere-Terpstra statistic. The Jonkcheere trend test assumes

$$H_0 : \theta_1 = \theta_2 = \theta_3 = \ldots = \theta_k \tag{4.4}$$

versus

$$H_1 : \theta_1 \leq \theta_2 \leq \ldots \leq \theta_k$$

with at least one strict inequality. Note: the assumed direction of the inequality above is without loss of generality. Suppose that one can reject the global hypothesis (4.4) at the α-level, the problem of interest is to determine where the strict inequalities hold while preserving the familywise error at α. Randles and Wolfe (1991) describe a procedure originally proposed by Odeh (1971) as an alternate method of calculating the Jonckheere-Terpstra statistic. That is, let

$$U_s^* = \sum_{i=1}^{s-1} U_{is}$$

for $s = 2, ..., k$. $U_2^*, U_3^*, ..., U_k^*$ are independent Mann-Whitney statistics where U_s^* is the Mann-Whitney statistic for comparing the s^{th} group with a group formed by

combining the first $(s-1)$ groups. It should be pointed out that $U_s^* \geq 0$ whenever the alternative hypothesis in (4.4) holds. The alternative form for the Jonckheere-Terpstra statistic becomes

$$V_2 = \sum_{s=2}^{k} U_s^*.$$

The statistics $U_2^*, U_3^*, \ldots, U_k^*$ can be used to define a multiple comparison procedure. Suppose that one rejects H_0 at the $\alpha-$level where

$$P\left(W \geq V_2 \mid H_0 \text{ is true}\right) = p \leq \alpha$$

in which case there is at least one strict inequality among the medians for the k groups. Our objective is to find its location

(1) Compute

$$P\left(W \geq U_s^* \mid H_0 \text{ is true}\right) = p_s.$$

(2) Let s_1 be the smallest value such that $p_s \leq \alpha$. In which case group s_1 is the first group for which a strict inequality holds when testing (4.4). Since we knew that there was at least one strict inequality under the alternative hypothesis, we are assured of having the above satisfied at least once when the global null hypothesis (4.4) is rejected. Are there are any additional cases with strict inequality? If $s_1 < k$ then continue to the next step, otherwise the procedure has identified the single strict inequality as being between groups $(k-1)$ and k.

(3) Test the new hypothesis

$$H_0 : \theta_{s_1} = \theta_{s_1+1} = \theta_{s_1+2} = \ldots = \theta_k \tag{4.5}$$

versus

$$H_1 : \theta_{s_1} \leq \theta_{s_1+1} \leq \ldots \leq \theta_k$$

at the $\alpha/2$-level. Repeat the above steps with the new hypothesis (4.5) to identify the index $s_2 > s_1$ as at the smallest index satisfying $p_s \leq \alpha/2$. Note

one must recompute U_s^* when testing (4.5) since we are no longer interested in the first $s_1 - 1$ groups.

(4) Repeat until one can no longer reject the new null hypothesis. Note: each time we repeat the above step and form a new null hypothesis (new level), the original level α becomes α/m at the m^{th} level.

In the next section we investigate this procedure with the two existing procedures using a simulation study.

4.4 Simulation

We conduct a simulation study to compare the Shirley (1977) method and the Nashimoto and Wright (2007) test with the proposed multiple comparison test. The model of interest is given by

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

where $\epsilon_{ijk} = -\log(u)$ and $u \sim$ exponential(1). The treatment of interest has $k = 4$ treatment groups and data are generated from models with the logit link. We illustrate the model using a treatment effect, α_i and a blocking factor β_j where, $i = 1, 2, 3, 4$, and $j = 1, 2,$, and $\beta_1 \leq \beta_2$. Each treatment group has $n = 30$ and results represent 1000 realizations of the model.

To investigate the proposed multiple comparison test we considered seven situations with $k = 4$ treatments. The three multiple comparison methods were performed for each situation. The cases of interest are

(1) $\alpha_1 < \alpha_2 < \alpha_3 < \alpha_4$

(2) $\alpha_1 < \alpha_2 < \alpha_3 = \alpha_4$

(3) $\alpha_1 = \alpha_2 < \alpha_3 < \alpha_4$

(4) $\alpha_1 < \alpha_2 = \alpha_3 < \alpha_4$

(5) $\alpha_1 = \alpha_2 = \alpha_3 < \alpha_4$

(6) $\alpha_1 < \alpha_2 = \alpha_3 = \alpha_4$

(7) $\alpha_1 = \alpha_2 < \alpha_3 = \alpha_4$

Tables 4.1 and 4.2 contain the results for the seven cases for the three multiple comparison procedures. The tables indicate the proportion of the times the null hypothesis was rejected for each case. The average for the p_s are included for the new procedure. Since, Shirley's method is designed for comparing each of the treatment levels to a zero treatment method, this method only compares α_1 with α_2, α_3, and α_4. Since, the NPM method by Nashimoto and Wright (2007) performs all pairwise comparisons, we have included the results for each comparison.

The simulated results show that the new method performed extremely well. It outperforms the NPM method, which is shown to be very conservative. Nashimoto and Wright's method is also shown to be somewhat inconsistent. The Shirley method performed well when comparing the active treatments to zero dose responses, however, when comparing group 1 vs group 2 it is consistently inferior to the new method. As expected this method easily rejects the comparison of group 1 with group 4, since there is a large difference between the medians for these two groups. There are cases when the proposed method has frequent occurrences of rejecting incorrectly by finding a difference when there is no difference. For example in Case 5 the test is incorrectly rejects the first two comparisons at a rate of almost 10%. Overall, however, the new method performed extremely well and consistently detects the appropriate treatment differences.

In the next section we compare the three comparison methods using an example found in Shirley (1977).

4.5 Example

Shirley (1977) presented an example using the reaction times of mice to various stimuli. The reaction times for each group is highly skewed. The original data had many ties that were obscuring the results of the different nonparametric procedures.

Table 4.1: Results for Cases 1 - 4

Case 1: $\alpha_1 < \alpha_2 < \alpha_3 < \alpha_4$					
	Proposed JT			Shirley	Nashimoto
Comparison	% Rejected	p-value	Comparison	% Rejected	% Rejected
1 vs 2	0.932	0.013	1 vs 2	0.786	0.457
2 vs 3	0.909	0.011	2 vs 3	-	0.588
3 vs 4	0.842	0.012	1 vs 3	1.000	0.998
			1 vs 4	1.000	1.000
			3 vs 4	-	0.440
			2 vs 4	-	1.000
Case 2: $\alpha_1 < \alpha_2 < \alpha_3 = \alpha_4$					
1 vs 2	0.942	0.012	1 vs 2	0.859	0.563
2 vs 3	0.894	0.012	2 vs 3	-	0.711
3 vs 4	0.037	0.249	1 vs 3	1.000	1.000
			1 vs 4	1.000	1.000
			3 vs 4	-	0.001
			2 vs 4	-	0.720
Case 3: $\alpha_1 = \alpha_2 < \alpha_3 < \alpha_4$					
1 vs 2	0.105	0.251	1 vs 2	0.031	0.002
(1,2) vs 3	0.979	0.004	2 vs 3	-	0.689
3 vs 4	0.829	0.014	1 vs 3	0.886	0.581
			1 vs 4	1.000	1.000
			3 vs 4	-	0.001
			2 vs 4	-	1.000
Case 4: $\alpha_1 < \alpha_2 = \alpha_3 < \alpha_4$					
1 vs 2	0.939	0.012	1 vs 2	0.908	0.681
2 vs 3	0.037	0.251	2 vs 3	-	0.001
(2,3) vs 4	0.967	0.004	1 vs 3	0.875	0.679
			1 vs 4	1.000	1.000
			3 vs 4	-	0.703
			2 vs 4	-	0.678

Therefore, we slightly modified the data as to eliminate many of the ties and give the data a more ordered structure. The modified data and their respective ranks are given in Table 4.3. The hypothesis of interest is

$$H_0 : \theta_0 = \theta_1 = \theta_2 = \theta_3 \tag{4.6}$$

vs.

$$H_1 : \theta_0 \leq \theta_1 \leq \theta_2 \leq \theta_3$$

Table 4.2: Results for Cases 5 - 7

	Proposed JT			Shirley	Nashimoto
Comparison	% Rejected	p-value	Comparison	% Rejected	% Rejected
Case 5: $\alpha_1 = \alpha_2 = \alpha_3 < \alpha_4$					
1 vs 2	0.088	0.256	1 vs 2	0.041	0.006
(1,2) vs 3	0.086	0.246	2 vs 3	-	0.005
(1,2,3) vs 4	1.000	0.001	1 vs 3	0.026	0.002
			1 vs 4	0.975	0.920
			3 vs 4	-	0.001
			2 vs 4	-	0.914
Case 6: $\alpha_1 < \alpha_2 = \alpha_3 = \alpha_4$					
1 vs 2	0.948	0.012	1 vs 2	0.928	0.773
2 vs 3	0.046	0.248	2 vs 3	-	0.006
(2,3) vs 4	0.048	0.255	1 vs 3	0.921	0.773
			1 vs 4	0.912	0.769
			3 vs 4	-	0.005
			2 vs 4	-	0.005
Case 7: $\alpha_1 = \alpha_2 < \alpha_3 = \alpha_4$					
1 vs 2	0.098	0.249	1 vs 2	0.038	0.009
(1,2) vs 3	0.982	0.005	2 vs 3	-	0.785
3 vs 4	0.045	0.248	1 vs 3	0.904	0.748
			1 vs 4	0.912	0.776
			3 vs 4	-	0.003
			2 vs 4	-	0.786

The Shirley (1977) method requires the calculation of $\bar{t}_k$ for each treatment level. Table 4.4 below gives the value of the test statistic as well as the critical point given in Williams (1971), and whether each treatment level is considered significantly different from the control (level 0). Nashimoto and Wright (2007) consider all pairwise comparisons, and thus ends up being more conservative then both Shirley and the new method. Table 4.5 contains the results for the NPM test with the modified Shirley data. The critical value given by Hayter (1990) is $h_{\alpha,k,\infty} = 3.295$. The NPM procedure reaches the same conclusion as given with the Shirley test. Each treatment level differs from the control, however, we are unable to find differences in reaction times among the remaining three groups.

Table 4.3: Modified reaction times in seconds of mice

Group 0		Group 1		Group 2		Group 3	
Reaction Time	Rank	Reaction Time	Rank	Reaction Time	Rank	Reaction Time	Rank
2.35	8	2.80	12	9.80	39	7.00	26
3.00	13	2.27	6	3.24	18	9.90	40
3.10	15	3.80	22	5.80	24	9.46	38
2.10	2	9.40	36	7.80	28	8.80	33
2.20	3	8.40	31	2.60	10	8.85	34
2.21	4	3.15	16	2.30	7	3.45	21
2.22	5	3.20	17	6.20	25	9.00	35
2.79	11	4.40	23	9.42	37	8.48	32
2.00	1	3.25	19	7.82	29	2.40	9
3.05	14	7.40	27	3.40	20	7.89	30
Means	7.6		20.9		23.7		29.8

Table 4.4: Test statistics for Shirley's method

Level	Test Statistic	1% Critical Value	Significance
$\bar{t}_3$	4.246	2.49	Yes
$\bar{t}_2$	3.079	2.48	Yes
$\bar{t}_1$	2.544	2.43	Yes

Table 4.5: Test Statistics for Nashimoto and Wright's Method, NPM

Comparison	Test Statistic	Significance
0 vs 1	3.598	Yes
1 vs 2	0.757	No
2 vs 3	1.650	No
0 vs 2	4.355	Yes
1 vs 3	2.407	No
0 vs 3	6.005	Yes

The results on reaction times in mice for the four groups using the proposed method are as follows. The Jonckheere-Terpstra statistic for the global hypothesis

(4.6) is

$$P(W \geq 487) = P(Z \geq 4.513) = 0.000004 < \alpha = 0.05$$

which is highly significant, hence, we conclude there is some difference among the medians for the four treatment levels. In computing the p-value for U_s^*, we observe that

$$P\left(W \geq U_1^*\right) = P(W \geq 92) = P(Z \geq 3.175) = 0.00075.$$

Indicating that there is a significant difference between the group 1 and the control. The new hypothesis becomes

$$H_0 : \theta_1 = \theta_2 = \theta_3 \tag{4.7}$$

vs

$$H_1 : \theta_1 \leq \theta_2 \leq \theta_3$$

Excluding the control group, we find the new p-value for the Jonckheere-Terpstra statistic is

$$P(W \geq 208) = 0.0137 < 0.025.$$

Indicating that there is a significant treatment differences in the medians for the remaining groups. We recomputed U_s^* for $s = 1, 2$ and found

$$P\left(W \geq U_2^*\right) = P(W \geq 61) = P(Z \geq 0.832) = 0.203 \tag{4.8}$$

and

$$P\left(W \geq U_3^*\right) = P(W \geq 147) = P(Z \geq 2.07) = 0.019 < 0.025. \tag{4.9}$$

Thus, we determine that the next significant difference comes between groups 2 and 3. Since the p-value in equation (4.8) is too large, whereas, the p-value in equation (4.9) is less than $\alpha/2 = 0.025$. In which case, we conclude, with a family-wise error of < 0.05, using the new procedure that

$$\theta_0 < \theta_1 = \theta_2 < \theta_3. \tag{4.10}$$

As with the other two procedures, this method concludes that groups 1-3 are significantly different from group 0. Whereas, the new method was the only procedure to find a significant differences among groups 1, 2, and 3 at the 0.05 level.

4.6 Discussion

If several treatment medians are to be compared with one another and the medians satisfy an ordered assumption, then a test procedure can be chosen to have good power properties under this ordered alternative. The method proposed in this paper is an intuitive and simple multiple comparison procedure based on the Jonckheere-Terpstra statistic. The method utilizes the ordered alternative assumption to allow for less comparisons to be performed and yet still draw significant conclusions about differences in treatment groups.

The simulation study showed the proposed method performed exceptionally well with non-normal data. It correctly detected strict inequalities in every location and outperformed two nonparametric multiple comparison procedures found in the literature.

CHAPTER FIVE

Conclusions and Future Work

In a clinical trial with multiple treatment arms, the evaluation of the active treatments is of the utmost importance to investigators. Ascertaining the most effective drug or the most preferential treatment allows researchers to make better decisions for patient outcomes. Our approach extends the case with an active treatment and a control to multiple treatment arms. This approach allows one to model the main effects of multiple arms as functions of a discrete covariate. In addition, the proposed multiple comparison test accurately handles the multiplicity issue and tests for treatment differences. Chapters 3 and 4 demonstrated that the approach is promising for applications in this area.

Since the AUC regression model with the Jonckheere-Terpstra statistic with covariates is a unique and complex situation, much research remains to be done. In chapter 3, we discussed the method for estimating covariate effects with an ordered alternative and multiple treatment effects. To include a covariate involved making many significant assumptions. We did not address the situation where there are interactions between two or more covariates or if the covariates are continuous. It is possible more extensive assumptions would have to be made, but the potential problem of interest is, can the proposed AUC regression model handle an interaction variable?

A second area of interest is in the multiple comparisons problem. Our proposed multiple comparison test was only tested with the no covariate situation with an ordered alternative. More research would have to be done to include covariate effects in this procedure.

APPENDIX

APPENDIX A

SAS Programs

A.1 SAS Code for Simulation Study using Fligner and Birnbaum's Methods

This program compared the Fligner and the modified Birnbaum Methods to the bootstrap and the DeLong method for estimating the standard errors of the parameters of the binary AUC regression model with logit link in Chapter 2. The following SAS file is the main file which calls the macros for the Fligner, the DeLong, and the Birnbaum procedure.

```
%macro comp_auc ( dataset= , num= ,sample_size= ,level=, link=, int=, beta=);
/************************************
  use the alternative analytical methods
************************************/

/* The Fligner method to compute the standard error */
%place(data=&dataset, y=x, group=therapy, size=&sample_size,
level=&level, int=&int,beta=&beta,ret=pla); run;
proc append base = presult data=pla force; run;
/* The DeLong method to compute the standard error */
%delong(dataset= &dataset,  level=&level, size=&sample_size,
ret= del, int=&int , beta=&beta);
proc append base= dresult data=del force; run;
/* The Birnbaum method to compute the standard error */
%max(data=&dataset, y=x, group=therapy, size=&sample_size,
level=&level, int=&int, beta=&beta, ret=mmm);
proc append base = mresult data=mmm force;
run;

%mend comp_auc;
```

```
%macro sim_auc( sim_time=, level= ,int=, beta=,  samplesize=, bnum= ,link=logit) ;
****************************************;
sim_time: the number of realizations to be generated
level:  the level number of the discrete covariate
int:  the intercept of the AUC model
beta: the parameters  for the covariate level effect.
(the number of the parameters = level -1 .
 For example, when level=3,  beta = 0.5 1 . )
samplesize: the sample size in each covariate level of each group
bnum: the number of resamples for each realization in the boostrap
procedure
link: the link to be used for the GLM (default is logit)
****************************************;
  proc datasets library=WORK nolist;
           delete  dresult;
                   delete  presult;
                   delete  mresult;
   run;

/******** The true AUC *************/

  /*read user input to macro variables*/
   %let beta0=0;
   %let D = 1;
   %let beta&D = %nrbquote(%scan(%bquote(&beta), &D, %str( )));
   %do %while(%nrbquote(&&beta&D) ~= );
     %let D = %eval(&D + 1);
     %let beta&D = %nrbquote(%scan(%bquote(&beta), &D, %str( )));
   %end;

        /* Compute the true AUC values according to the user input parameters */
  data auc_true (keep= variable true);
        do cov=0 to &level-1;
                variable='auc'||put(cov,1.);
```

```
                beta= symget('beta'||left(cov));
                true= 1/(1+exp(-beta-&int));
                 output;
         end;
 run;

 data int_true;
variable='Intercept';
 true=&int;
run;
data cov_true;
  do cov=1 to &level-1;
        variable='cov'||put(cov,1.);
        true= input(symget('beta'||left(cov)), 4.);
        output;
  end;
run;

data true (drop= cov);
length variable $ 9;
set  int_true cov_true auc_true;
run;

proc sort data=true; by variable; run;

/*****************************************************
                              Simulate the data
*****************************************************/
%do num =1 %to &sim_time;
data simdata ;
  do cov= 0 to &level-1;
     do n= 1 to &samplesize;
          beta= symget('beta'||left(cov));
          treat= -log(ranexp(0))+beta+&int;
```

```
          placebo= -log(ranexp(0));
        output;
      end;
  end;
run;

data treat  ( keep=cov therapy x);
set simdata; x=treat; therapy= 'T';  run;
data placeb  ( keep=cov therapy x);
set simdata; x=placebo;  therapy='P';  run;
data simd;
set treat placeb;run;

proc sort data=simd; by cov therapy x;run;

%comp_auc( dataset= simd, num=&bnum ,sample_size=&samplesize,
level=&level, link=&link, int=&int, beta =&beta);
 %end;
%mend;

/* Input the parameters and call the main macro */
%sim_auc ( sim_time=1, level=3,int=0.15 ,  beta=0.5 1,
 samplesize=100, bnum=200,link= logit);
```

1.1.1 SAS Macros for Alternative Procedures

The following SAS file include the macros for the Fligner and modified Birbaum procedure, which is called by the main SAS file above. The SAS macro for the DeLong procedure (and bootstrap) is given by Zhang (2008)

```
/* Birnbaum procedure for the standard errors in AUC regression */
%macro max(data=, y=, group=, size=, level=, int=, beta=, ret=);

proc datasets library=WORK nolist;
delete max_auc;
```

```
run;
ods listing close;
%do j=0 %to &level-1;
data mv;
set &data;
if cov=&j;
run;

proc sort data=mv;
by &group &y;
run;

proc iml;
use mv;
read all ;
rep=x;
x=J(&size,1);
   do i=1 to &size;
        x[i]=rep[i];
   end;
y=J(&size,1);
   do i=1 to &size;
        y[i] = rep[i+&size];
   end;
/* Compute the U statistic */
u0=0;
do i=1 to &size;
   do j=1 to &size;
                if (x[j]<y[i]) then u0=u0+1;
          else if (x[j]=y[i]) then u0=u0+1/2;
        end;
end;

u0=u0/(&size*&size);
```

```
/* Find the upper and lower bound as defined by Birnbaum */
max= sqrt((&size*&size*u0*(1-u0)*&size)/(&size**2*&size**2));
q = 2*u0;
r= min(u0, (1-u0));
if 1 < q then min = (&size*&size*(&size*r*(1-r)
                  -(((&size-1)**2)/(12*(&size-1)))))/( &size**2*&size**2);
if 1 > q then min = (&size*&size*((4/3)*r*sqrt(2*(&size-1)*(&size-1)*r)
                  -(&size+&size-2)*r**2+r*(1-r)))/(&size**2*&size**2);
min1 = sqrt(min);
sd=(min1+max)/2;
out=&j||u0||sd;
varname='cov'||'estimate'||'sd';
create max1 from out[colname=varname];
append from out;
quit;
data max1;
set max1;
variable ='auc'||put(cov,1.);
run;
proc append base= max_auc data=max1 force; run;
%end;
/************************************************************
Combining the delta methods and the Birnbaum method to
 estimate the beta
************************************************************/
data maximum;
set max_auc;
g_auc=log(estimate/(1-estimate));
sd_g=sd/(estimate*(1-estimate));
var_g=sd_g**2;
run;

data parm (keep=g_auc);
set maximum;
```

```
run;
data var_parm(keep=var_g);
set maximum;
run;
proc iml;
use parm;
read all into est;
int= est[1,1];
cov=J(&level-1,1);
     do i= 1 to &level-1;
          cov[i]= est[i+1,1]-int;
     end;
out=int;
    do i=1 to &level-1;
          out=out||cov[i];
    end;
varname='Intercept';
    do i=1 to &level-1;
          varname=varname||cats( 'cov', char(i));
    end;
create parm1  from out[colname= varname];
append from out;
quit;

proc transpose data=parm1 out =parm;
run;

data parm (keep=variable estimate);
set parm;
variable=_NAME_;
estimate=col1;
run;

proc iml;
```

```
use var_parm;
read all into est;
int= est[1,1];
cov=J(&level-1,1);
     do i= 1 to &level-1;
          cov[i]= est[i+1,1]+int;
     end;
out=int;
     do i=1 to &level-1;
          out=out||cov[i];
     end;
varname='Intercept';
     do i=1 to &level-1;
          varname=varname||cats( 'cov', char(i));
     end;
create var_parm1  from out[colname= varname];
append from out;
quit;

proc transpose data=var_parm1  out =var_parm;
run;
data var_parm (keep=variable  sd);
set var_parm;
variable=_NAME_;
var=col1;
sd= sqrt(var);
run;
data max_p;
merge parm var_parm;
by variable;
run;
data max_all;
set max_p max_auc;
run;
```

```
proc sort data=max_all;
by variable;
run;
data &ret;
merge max_all true;
by variable;
ciu=estimate+sd*1.96;
cil=estimate-sd*1.96;
if (ciu>true) and (cil<true) then hit=1; else hit=0;
drop cov;
run;
%mend max;

/* Run the macro to calculate the Fligner method of standard errors */
%macro place(data=, y=, group=, size=, level=, int=, beta=, ret=);
proc datasets library=WORK nolist;
delete place_auc;
run;
ods listing close;
        %do j=0 %to &level-1;
data pv;
set &data;
if cov=&j;
run;
proc sort data=pv;
by &group &y;
run;
proc rank data=pv ties=mean out=rankr;
var &y;
ranks rank;
run;

 proc iml;
 use rankr;
```

```
read all;
rep = rank;
x=J(&size,1);
do i=1 to &size;
x[i]=rep[i];
end;

p=J(&size,1);
q=1;
do i=1 to &size;
p[i]=x[i]-q;
q=q+1;
end;

meanp=J(&size,1)-1;

diffp=J(&size,1);

do i =1 to &size;
meanp=meanp+p[i];
end;

do i=1 to &size;
meanp[i]=meanp[i]/&size;
end;

do i =1 to &size;
diffp[i]=(p[i]-meanp[i])**2;
end;

out = p||diffp;
varname='p'||'diffp';
create place from out[colname=varname];
append from out;
```

```
 quit;
 run;

 proc iml;
  use rankr;
 read all;
 rep =rank;
  y=J(&size,1);
 do i=1 to &size;
 y[i]=rep[i+&size];
 end;
s=J(&size,1);
q=1;
 do j=1 to &size;
 s[j]=y[j]-q;
 q=q+1;
 end;
 means=J(&size,1)-1;
  diffs=J(&size,1);
   do i=1 to &size;
 means=means+s[i];
 end;
 do i=1 to &size;
means[i]=means[i]/&size;
end;
do i=1 to &size;
diffs[i]=(s[i]-means[i])**2;
end;
out = s||diffs;
varname='s'||'diffs';
create place2 from out[colname=varname];
append from out;
quit;
run;
```

```
/*Find the average ranks and sample sizes */
  proc means data=place noprint;
  var p;
  output out=m1 mean=mean1;
  run;
  proc means data=place2 noprint;
  var s;
  output out=m2 mean=mean2;
  run;
 proc means data=place noprint;
 var diffp;
 output out= total sum=sum ;
 run;
 proc means data=place2 noprint;
 var diffs;
 output out= total2 sum=sum;
 run;
 data total;
 set total;
 group = 1;
 run;
 data total2;
 set total2;
 group = 2;
 run;
 data tot;
 set total total2;
 run;
 proc means data=tot noprint;
 var sum;
 output out=num sum=sum;
 run;
 data var;
 merge M1 M2 num;
```

```
 cov=&j;
 keep mean1 mean2 sum cov;
 run;

    proc iml;
           use pv;
read all into data;
rep= data[,2];

x= J(&size,1);
do i= 1 to &size;
        x[i]=rep[i];
end;
y=J(&size,1);
do i= 1 to &size;
    y[i]=rep[i+&size];
end;

/* compute u stat */
u0=0;
do i=1 to &size;
   do j=1 to &size;
               if (x[j]<y[i]) then u0=u0+1;
          else if (x[j]=y[i]) then u0=u0+1/2;
        end;
end;

u0=u0/(&size*&size);
out = &j||u0;
varname = 'cov'||'estimate';
create var1 from out[colname=varname];
append from out;
quit;
run;
```

```
 data var;
 merge var var1;
 var=(sum+mean1*mean2)/(&size**2*&size**2);
 std=sqrt(var);
variable='auc'||put(cov,1.);
 drop sum;
drop mean1;
drop mean2;
drop var;
drop cov;
 run;
 proc append base=place_auc data=var force;
 run;
 %end;

data placement;
set place_auc;
g_auc=log(estimate/(1-estimate));
std_g=std/(estimate*(1-estimate));
var_g=std_g**2;
run;

  data parm (keep= g_auc);
set placement;
run;

data var_parm(keep= var_g);
set placement;
run;
proc iml;
use parm;
read all into est;
int= est[1,1];
cov=J(&level-1,1);
```

```
      do i= 1 to &level-1;
            cov[i]= est[i+1,1]-int;
      end;
out=int;
     do i=1 to &level-1;
            out=out||cov[i];
     end;
varname='Intercept';
     do i=1 to &level-1;
            varname=varname||cats( 'cov', char(i));
     end;
create parm1  from out[colname= varname];
append from out;
quit;

proc transpose data=parm1 out =parm;
run;

data parm (keep=variable estimate);
set parm;
variable=_NAME_;
estimate=col1;
run;

proc iml;
use var_parm;
read all into est;
int= est[1,1];
cov=J(&level-1,1);
      do i= 1 to &level-1;
            cov[i]= est[i+1,1]+int;
      end;
out=int;
      do i=1 to &level-1;
```

```
            out=out||cov[i];
        end;
varname='Intercept';
        do i=1 to &level-1;
            varname=varname||cats('cov', char(i));
        end;
create var_parm1  from out[colname= varname];
append from out;
quit;

proc transpose data=var_parm1  out =var_parm;
run;

data var_parm (keep=variable  std);
set var_parm;
variable=_NAME_;
var=col1;
std= sqrt(var);
run;
data place_p;
merge parm var_parm;
by variable;
run;
data place_all;
set place_p place_auc;
run;
proc sort data=place_all;
by variable;
run;
data &ret;
merge place_all true;
by variable;
ciu=estimate+std*1.96;
cil=estimate-std*1.96;
```

```
if (ciu>true) and (cil<true) then hit=1; else hit=0;
run;
 %mend place;
```

BIBLIOGRAPHY

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

Bamber, D. (1975), "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology*, 12, 387–415.

Bartholomew, D. J. (1961), "Ordered tests in the analysis of variance," *Biometrika*, 48, 325–332.

Birnbaum, Z. and Klose, O. (1957), "Bounds for the Variance of the Mann-Whitney Statistic," *The Annals of Mathematical Statistics*, 28, 933–945.

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.

Fligner, M. and Policello, G. (1981), "Robust Rank Procedures for the Behrens-Fisher Problem," *Journal of the American Statistical Association*, 76, 162–168.

Hayter, A. (1990), "A one-sided studentized range test for testing against a simple ordered alternative," *Journal of the American Statistical Association*, 85, 778–785.

Hsu, J. (1996), *Multiple Comparisons: Theory and Methods*, London:Chapman and Hall.

Jonckheere, A. R. (1954), "A distribution-free k-sample test against ordered alternatives," *Biometrika*, 41, 133–145.

Kruskal, W. H. and Wallis, W. A. (1952), "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, 47, 583–621.

Lehmann, E. (1975), *Nonparametrics: Statistical Methods based on Ranks*, Holden-Day, San Francisco.

Mann, H. and Whitney, D. (1947), "On a test of whether one of two random variables is stochasitcally larger than the other," *Annals of Mathematical Statistics*, 18, 50–60.

Nashimoto, K. and Wright, F. (2007), "Nonparametric multiple-comparison methods for simply ordered medians," *Computational Statistics and Data Analysis*, 51, 5068–5076.

Odeh, R. E. (1971), "On Jonckheere's k-sample test against ordered alternatives," *Technometrics*, 13, 912–918.

Puri, M. L. (1965), "Some distribution-free k-sample rank tests of homogeneity against ordered alternatives," *Comm. Pure Applied Mathematics XVIII*, 51–63.

Randles, R. and Wolfe, D. (1991), *Introduction to the Theory of Nonparametric Statistics*, Malabar, Florida.

SAS (2008a), "SAS – The GENMOD Procedure," .

— (2008b), "SAS – The LOGISTIC Procedure," .

Shirley, E. (1977), "A Non-Parametric Equivalent of Williams' Test for Contrasting Increasing Dose Levels of a Treatment," *Biometrics*, 33, 386–389.

Terpstra, T. (1952), "The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking," *Indagationes Mathematicae*, 14, 327–333.

Van Dantzig, D. (1951), "On the consistency and the power of Wilcoxon's two-sample test," *Koninklijke Nederlandse Akademie Van Wetenschappen, Proceedings*, 54, 1–9.

Van Elteren, P. (1960), "On the combination of independent two sample tests of Wilcoxon," *Bulletin of the International Statistical Institute*, 37, 351–361.

Wald, A. and Wolfowitz, J. (1944), "Statistical tests based on permutations of the observations," *Annals of Mathematical Statistics*, 15, 358–372.

Westfall, P., Tobias, R., Rom, D., W. D., and Y., H. (1999), *Multiple Comparisons and Multiple Tests Using SAS*, Cary, NC: SAS Institute Inc.

Williams, D. (1971), "A Test for Differences between Treatment Means when Several Dose Levels are Compared with a Zero Dose Control," *Biometrics*, 27, 103–117.

Zhang, L. (2008), "Semiparametric AUC Regression for Testing Treatment Effect in Clinical Trial," Ph.D. thesis, Baylor University.

Zhang, L., Zhao, Y., and Tubbs, J. D. (2011), "Inference for semiparametric AUC regression models with discrete covariates," *Journal of Data Science*, 10.