Notes on COPYRIGHT:

MAXIMALLY SELECTED TEST STATISTICS:

METHODOLOGY AND APPLICATION

by

YUHUI MA

A dissertation submitted to the

Graduate School – New Brunswick Rutgers, The State University of New Jersey

and

The Graduate School of Public Health

University of Medicine and Dentistry of New Jersey

In partial fulfillment of the requirements

For the degree of Doctor of Public Health Graduate Program in Biostatistics

Written under the direction of Dr. Pamela Ohman Strickland

And approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

JANUARY 2012

ABSTRACT OF THE DISSERTATION

Maximally selected test statistics: Methodology and Application

By YUHUI MA

Dissertation Director:

Dr. Pamela A. Ohman Strickland

In clinical or public health research studies, an investigator often assumes that some continuous predictive variable X allows classifying study population into a risk and a normal group with respect to a response variable Y. The aim of these research efforts is to transform a continuous variable into a binary variable by identifying a threshold or cutpoint in the predictor to distinguish different groups with high or low probabilities of favorable outcomes. Several methods including maximally selected chi-square statistics, maximally selected rank statistics and Koziol's exact finite sample distribution approach to search for the optimal cut point have been reviewed and compared in Chapter 1. Since utilizing the maximally selected rank statistic to analyze semi-continuous predictors has not been discussed in the literatures, this dissertation provides the comparison of the null distribution, power curve, precision of cut point estimation between semi-continuous and continuous predictive variables via simulation. In Chapter 2, we confirmed the critical values to reject the null hypotheses are lower in semi-continuous predictors compare to continuous predictors. In Chapter 3, we show the power of maximally selected rank statistic from the semi-continuous predictor is stochastically larger than that from the continuous predictor. In Chapter 4, we found besides the sample size and effect size, the location of the true cut-point also affects the precision of the cut-point estimates. Compared to the continuous predictor, the semi-continuous predictor has higher percentage of correct cut-point estimates. The null distributions for semi-continuous predictor simulated in Chapter 1 are then applied to the study of "lead exposure, HPA dysfunction, blood pressure and hypertension risk" (Fiedler 2010) in Chapter 6 to determine the cut point in blood lead level that triggers increased stress. This application focused on the multivariate relationship between predictor variables and response variable, which were not discussed in the literature. After adjusted by other confounder

variables through the regression residuals, a significant cut-point of 2 μg/dL in blood lead level is identified. Since the use of the regression residual of the response variable violates the independence assumption of this maximally selected rank statistics, this dissertation also demonstrated the robustness of this assumption in chapter 5.

ACKNOWLEDGEMENTS

I would never have been able to finish my dissertation without the guidance of my committee members.

I would like to express my deepest gratitude to my advisor, Dr. Pamela A. Ohman Strickland, for her excellent guidance and patience. I would like to thank Dr. Shih, who provides me with an exceptional research atmosphere. I would also like to thank Dr. Lu and Dr. Xie for guiding my research and providing valued suggestions and comments. Special thanks go to Dr. Fiedler, who was willing to provide me the data and motivated the research topic.

DEDICATION

I dedicate this dissertation to my wonderful family. Particularly to my two precious sons, Wenhao and Wentao, who are the joys of our lives. Thank you both for the understanding and patience when Mommy is always busy at work. And to my husband, who is always there for me in good days and bad days.

I must also thank my loving mother and my terrific in-laws who have helped so much with baby-sitting and have given me their fullest support.

Finally, I dedicate this work in memory of my father, who believed in diligence, science and the pursuit of academic excellence. Daddy, you have given me so much, thanks for your faith in me, and for teaching me that I should never surrender.

We made it…

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: LITERATURE REVIEW AND METHODOLOGY IMPLEMENTATION

1.1 Abstract

In clinical and public health research studies, the prognostic factors or predictive variables are often measured on a continuous or semi-continuous scale. In certain circumstances, cutpoints can be defined for these prognostic factors to delineate the study population into normal and at risk groups with high or low probabilities of favourable outcomes with respect to a response variable Y. An optimal cutpoint search method is introduced in this chapter. Methods from the statistical and medical literature for calculating appropriate p-values of these optimal cutpoints, such as large sample approximations and small sample exact methods of maximally selected Chi-square statistics and maximally selected rank statistics, are discussed and compared in this chapter.

1.2 Introduction

In clinical or public health studies, an investigator often assumes that some prognostic factor X allows for a classification of population into a risk and a normal group with respect to a response variable Y. The aim of the research effort is to find a threshold or cutpoint in the prognostic factor to distinguish different groups with high or low probabilities of favourable outcomes. From medical research point of view, a cutpoint may be preferred for (1) offering a simple risk classification into "high" verse "low" (Schulgen 1994), (2) establishing eligible criteria for prospective studies (Mazumdar 2000), (3) setting diagnostic criteria for disease and assisting in making treatment recommendations (Mazumdar 2000), and (4) imposing an assumed biological threshold.

For a quantitative prognostic factor, the straightforward and popular method to find a cutpoint is to test each observed value in a systematic manner and select the cutoff value which maximizes the measure of difference in response between the two groups of subjects. In other words, the cutpoint defining "normal" and "risk" groups is the one with minimal p-value relating the prognostic factor to outcome. The cutpoint so chosen is often termed "optimal", but this description without any p-value correction is

inadvisable because of the well-known problem of multiple testing. Altman has demonstrated in a simulation that when this approach is used, the probability of obtaining a significant result (at α=0.05 level) from the logrank test, when there is no actual relationship between the variables, i.e. the type I error, is inflated to 0.4 (Altman 1994).

Several methods for calculating appropriate p-values and unbiased effect measures with optimal cutpoints have been documented in the statistical and medical literature, and will be discussed throughout this chapter. Theoretical considerations behind these methods allow a correction of the minimal p-value for the multiple testing, leading to a true false-positive rate of 5%.

In the case when the outcome variable Y is binary and the predictive variable X is continuous, a cutpoint might be examined by searching in the range of X and then performing an association test for the obtained 2×2 contingency table using the chi-square statistics. Miller and Siegmund show that the distribution of the maximally selected chi-square statistic, i.e. the maximal chi-square statistic over all possible cutpoints, converges to a normalized Brownian bridge under the null-hypothesis of no association between X and Y, which is different from the known chi-square distribution (Miller and Siegmund 1982). The distribution of this maximally selected Chi-square statistic in the small sample case is examined by Halpern in a simulation study (Halpern 1982), while Koziol derives the exact distribution of maximally selected Chi-squared statistic using a combinatorial approach (Koziol 1991). In the case when the outcome variable Y is binary and the predictive variable X is in nominal or ordinal scales, Boulesteix has developed the exact distribution to handle the case of optimally selected splits (Boulesteix 2006a; Boulesteix 2006b).

In the case when the outcome is ordered, quantitative or censored variable, Lausen and Schumacher developed the asymptotic null distribution of maximally selected rank statistic. This asymptotic null distribution of maximally selected rank statistic is then compared with Monte Carlo simulation results by using continuous predictive variable X (Lausen 1992; Lausen 2004).

This chapter is organized as follows. In section 1.3, we introduce the graphic diagnosis for the appropriateness of using a cutpoint model. In section 1.4, we introduce the

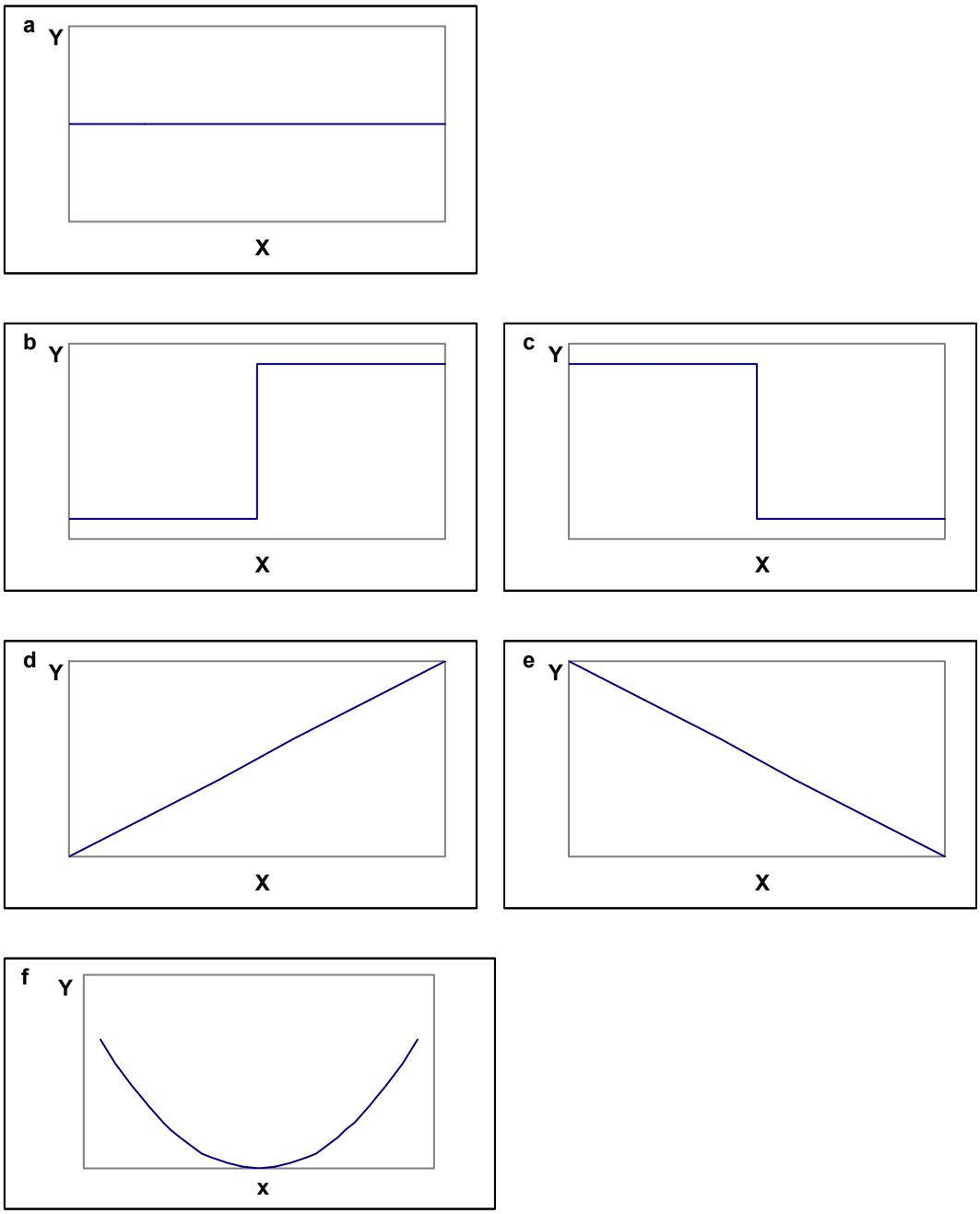
maximally selected Chi-square statistics used for binary outcome variable, which includes the asymptotic approximation, Koziol's exact method for small samples with continuous predictor as well as Boulesteix's exact method for ordinal predictor. In section 1.5, we introduce the maximally selected rank statistics used for continuous and censored response variables, which includes the asymptotic approximation and simulation results for finite samples. All of these maximally selected statistics are discussed under the null hypothesis that predictor and response variables are stochastically independent. To illustrate the use of these maximally selected statistics methods, in section 1.6 we describe some studies that have utilized these cut point finding techniques as examples.

1.3 Graphic diagnosis

In the absence of any a priori clinical information regarding the prognostic relationship between a covariate and outcome, the appropriateness of a cutpoint model is often determined empirically with graphical and numerical results.

Boulesteix (2007) and Mazumdar (2000) depicted some extreme examples in figure 1.1.1, in which X is the predictor variable and could be continuous or ordinal; Y is the outcome variable and could be binary or continuous. The mean of Y is plotted against X, therefore when Y is binary the mean of Y is the proportion at each value of X. An approximately horizontal graph of type (a) indicates poor association between X and Y. Types (b) and (c) correspond to ideal situation to use a single cutpoint model, since the underlying relationship between the outcome and prognostic variable is a step function. Types (d) and (e) correspond to strong monotonic associations in which no apparent cutpoint dividing X into high and low outcome risk groups. Type (f) indicates non-monotonic association, a single cutpoint model can not be used to divide X into two distinct (high and low) outcome risk groups.

Figure 1.1.1 Plot of mean of outcome variable against predictor variable



1.4 Maximally selected chi-square statistics

1.4.1 Asymptotic Approximation of Maximally Selected Chi-Square Statistics

Suppose we have a binary outcome Y (Y = 0, 1) and a continuous predictor X. When searching for the point in X that best separates the two groups, this predicator variable X will be dichotomized. And the following 2 x 2 table is used to calculate the χ^2 statistic.

Y	$X \le x$	$X > x$	
1	a	b	n_1
0	c	d	n_2
	a+c	b+d	N

For a given *x*, the square root of the chi-square statistic can be written as

$$(\chi^2)^{\frac{1}{2}} = \frac{\left|\hat{F}_1(x) - \hat{F}_2(x)\right|}{\left[\hat{F}(x)\{1-\hat{F}(x)\}(\frac{1}{n_1}+\frac{1}{n_2})\right]^{\frac{1}{2}}} \qquad (1)$$

Where $\hat{F}_1, \hat{F}_2$ *and* $\hat{F}$ are estimates of the empirical distribution of X for population 1 (subjects with Y=1), population 2 (subjects with Y=0) and the common population under $H_0 : F_1 = F_2 = F$, respectively with forms of

$$\hat{F}_1(x) = pr(X_1 \le x) = \frac{a}{a+b},$$
$$\hat{F}_2(x) = pr(X_2 \le x) = \frac{c}{c+d},$$
$$\hat{F}(x) = pr(X \le x) = \frac{a+c}{N},$$

Miller and Siegmund (1982) show that the maximally selected statistic of $(\chi^2)^{\frac{1}{2}}$ converges to a normalized Brownian bridge of $\frac{|W_0(t)|}{\{t(1-t)\}^{\frac{1}{2}}}$ under the null-hypothesis of no

association between X and Y. Where $W_0(t)$ is a tie-down Wiener process (Brownian bridge) on [0, 1] with $t=F(x)$. In order to have a considerable number of observations in each of the separated groups, the search of x, the cut point, will be over the interval of $[F^{-1}(\varepsilon_1), F^{-1}(\varepsilon_2)]$ in order statistic, where ε_1, ε_2 are the percentiles with range from (0, 1) and $\varepsilon_1 < \varepsilon_2$. Typically $\varepsilon_1 = 0.1$ and $\varepsilon_2 = 0.9$. $F(x)$ is the empirical distribution function on total sample size of N. Use M (ε_1, ε_2) to indicate the maximally selected $(\chi^2)^{\frac{1}{2}}$ statistics over the interval of $F^{-1}(\varepsilon_1)$ to $F^{-1}(\varepsilon_2)$, its asymptotic approximation is listed as following.

$$pr[M(\varepsilon_1, \varepsilon_2) \geq d] = \frac{4\varphi(d)}{d} + \varphi(d)(d - \frac{1}{d}) \log(\frac{\tau_2}{\tau_1}) + o\{d^{-1}\varphi(d)\} \text{ for } 0 < \varepsilon_1 < \varepsilon_2 < 1,$$

Where $\tau_j = \varepsilon_j / (1 - \varepsilon_j)$ and $\varphi(d)$ is the standard normal density $(2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2} d^2)$.

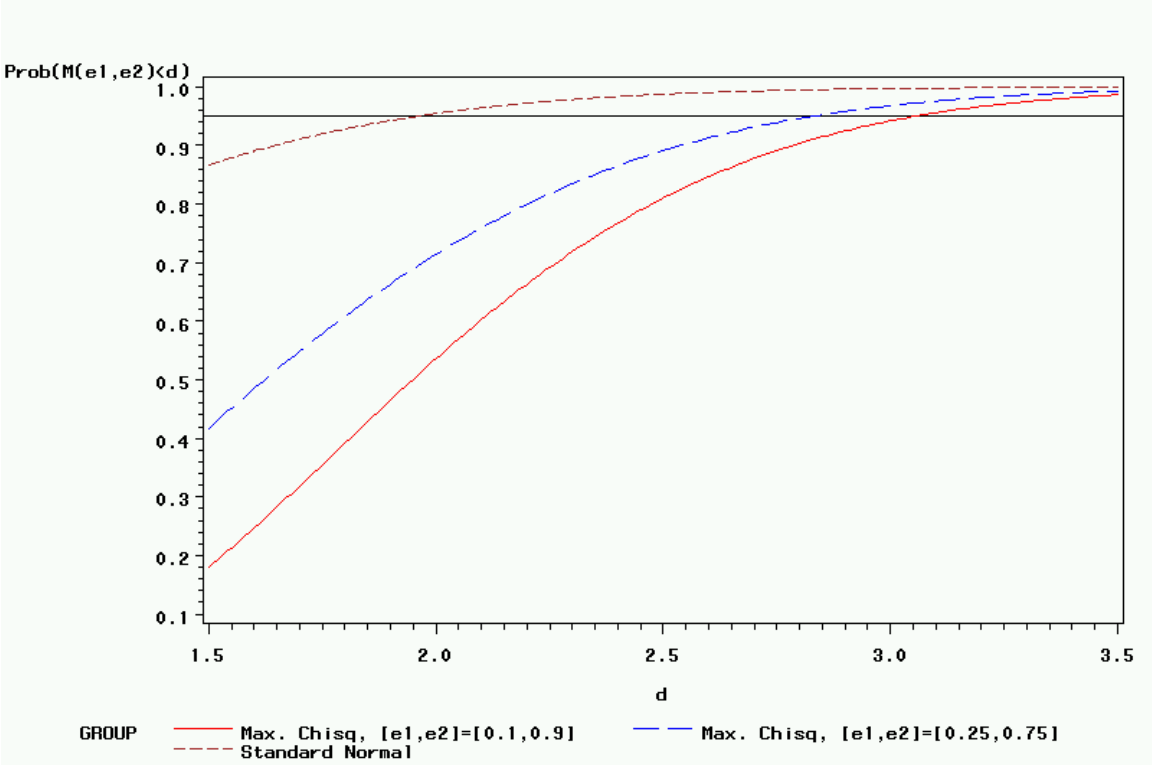
Since the $(\chi^2)^{\frac{1}{2}}$ follows standard normal distribution on [0, +∞), table 1.1.1 and figure 1.1.2 show the comparison between M (ε_1, ε_2) and standard normal density. Table 1.1.1 gives the critical constants of M (ε_1, ε_2) for α=.10, .05, .01, ε_1=.25, .10 and ε_2=1- ε_1. These critical constants should be compared with 1.64, 1.96 and 2.59, respectively.

According table 1.1.1 and figure 1.1.2, by using the optimal search of the M (ε_1, ε_2), the critical values increased in order to maintain the same significance level as those on the ordinary Chi square table. The most commonly used critical value of 1.96 corresponds to a type I error rate around 0.5 in the asymptotic distribution of M (ε_1, ε_2) when the search is within the central 80% of the combined data.

Table 1.1.1 Upper α quartile points of the M (ε_1, ε_2), where the maximum is over $F^{-1}(\varepsilon_1)$ to $F^{-1}(\varepsilon_2)$, compare to the standard normal density.

α	[ε_1, ε_2]		Standard Normal Density
	[0.25, 0.75]	[0.10, 0.9]	
0.10	2.54	2.78	1.64
0.05	2.83	3.05	1.96
0.01	3.40	3.59	2.58

Figure 1.1.2 Asymptotic cumulative probability of M (ε1, ε2) compare to the standard normal density under H0.



Note: The maximum is over $F^{-1}(\varepsilon_1)$ to $F^{-1}(\varepsilon_2)$

1.4.2 Null Distributions of Maximally Selected Chi-Square Statistics for Small Sample Size

Halpern (1982) studied the finite-sample distribution of this maximally selected chi-square statistic via simulation. Koziol (1991) derived the exact distribution of maximally selected χ^2 statistics given the sample size in each response group as *n* and *m* using Durbin's combinatorial approach (Durbin, 1971 and1973).

Consider the notation in the following 2×2 table as an example.

Group	$X \le X_i$	$X > X_i$	Total
Y=1	n_i		n
Y=2	i- n_i		m
Total	i		n+m

Let X_{11}, X_{12}, ..., X_{1n} denote the observations from the first group with empirical distribution function F_n; X_{21}, X_{22}, ..., X_{2m} the observations from the second group with empirical distribution function G_m; and $X_1 \le X_2 \le ... \le X_{n+m}$ the ordered observations from the combined sample. For the x in the interval $[X_i, X_{i+1})$,

$$F_n(x) - G_m(x) = \frac{n_i}{n} - \frac{i - n_i}{m} = \frac{n+m}{m}\left(\frac{n_i}{n} - \frac{i}{n+m}\right),$$

where n_i is the number of X's that are less than or equal to X_i.

Let $F_{nm}(i) = n_i / n$ and $H_{nm}(i) = i/(n+m)$, $i = 1, \ldots, n+m$. Then from the equation (1) in section 1.4.1, the exact maximally selected $(\chi^2)^{\frac{1}{2}}$ statistic for testing the equality of the underlying continuous distributions F and G of the first and second groups, respectively, may be written as

$$A_{nm}^{+} = \frac{n+m}{m} \max_{i=1,\ldots,n+m-1} \left(F_{nm}(i) - H_{nm}(i)\right) \Bigg/ \left[H_{nm}(i)[1 - H_{nm}(i)]\left(\frac{1}{n} + \frac{1}{m}\right)\right]^{1/2}$$

and

$$A_{nm}^{-}=\frac{n+m}{m}\max_{i=1,\ldots,n+m-1}\left(H_{nm}(i)-F_{nm}(i)\right)\Big/\left[H_{nm}(i)[1-H_{nm}(i)]\left(\frac{1}{n}+\frac{1}{m}\right)\right]^{1/2}$$

Using $A_{nm}=\max\left(A_{nm}^{+},A_{nm}^{-}\right)$ to indicate the exact maximally selected $(\chi^2)^{\frac{1}{2}}$ statistic, the finite-sample null distribution of A_{nm} can be directly determined by using the combinatorial approach of Durbin. Let d > 0 be arbitrary, and consider the graph of $F_{nm}(i)$ as a function of i. if $A_{nm}^{+}\le d$, then all points (i, $F_{nm}(i)$) must lie below or on the curve of

$$y=\frac{md}{n+m}\left[\frac{i}{n+m}\left(1-\frac{i}{n+m}\right)\left(\frac{1}{n}+\frac{1}{m}\right)\right]^{1/2}+\frac{i}{n+m}$$

Similarly, if $A_{nm}^{-}\le d$ then all points (i, $F_{nm}(i)$) must lie above or on the curve of

$$y=\frac{-md}{n+m}\left[\frac{i}{n+m}\left(1-\frac{i}{n+m}\right)\left(\frac{1}{n}+\frac{1}{m}\right)\right]^{1/2}+\frac{i}{n+m}$$

Let $b_j=\max\left\{i:\frac{j}{n}>\frac{md}{n+m}\left[\frac{i}{n+m}\left(1-\frac{i}{n+m}\right)\left(\frac{1}{n}+\frac{1}{m}\right)\right]^{1/2}+\frac{i}{n+m}\right\},1\le j\le n,$

$$c_j=\min\left\{i:\frac{j}{n}<\frac{-md}{n+m}\left[\frac{i}{n+m}\left(1-\frac{i}{n+m}\right)\left(\frac{1}{n}+\frac{1}{m}\right)\right]^{1/2}+\frac{i}{n+m}\right\},0\le j\le n-1,$$

The graph of $F_{nm}(i)$ will cross the upper boundary or the lower boundary or both if and only if it passes through at least one of the points $(b_j,j/n),(c_j,j/n)$. Let these points satisfying this criterion be labeled $B_1,B_2,\ldots,B_q$ in the order of increasing i, with the convention that if there are two such points for the same i they are labeled in increasing order of j. Then $A_{nm}>d$ if and only if the graph of $F_{nm}(i)$ passes through one or more of $B_1,B_2,\ldots,B_q$. Based on Durbin's methods, Koziol (1991) derived recursive formulas for computing this probability as given below.

$$\Pr\left[A_{nm}>d\right]=\binom{n+m}{n}^{-1}\sum_{r=1}^{q}\binom{n+m-i_r}{n-j_r}b_r$$

Where $b_s = \begin{pmatrix} i_s \\ j_s \end{pmatrix} - \sum_{r=1}^{s-1} \begin{pmatrix} i_s - i_r \\ j_s - j_r \end{pmatrix} b_r, s = 2, \ldots, q, \quad b_1 = \begin{pmatrix} i_1 \\ j_1 \end{pmatrix}$

The results from Koziol's exact finite sample distribution approach agree well with Halpern's simulated values. Table 1.1.2 gives the comparison on the critical values obtained from the simulated finite sample distribution (Halpern 1982), the Koziol's exact finite sample distribution approach and from asymptotic approximation. Figure 1.1.3 gives the comparison on the critical values obtained from the exact maximally selected Chi-square statistics (Koziol 1991) and those from asymptotic approximation (Miller and Siegmund 1982) as well as the standard normal distribution.

When sample size is 50, the 90th and 95th quartile points of the A_{nm} differ by about 7 – 8% from the Miller and Siegmund's asymptotic values. When sample size is 100, the 90th and 95th quartile points of the A_{nm} differ by about 4 – 5% from the Miller and Siegmund's asymptotic values. The asymptotic approximation (Miller and Siegmund 1982) is comparable with those from exact sample distribution at sample size of 200.

Table 1.1.2 Upper α quartile points of A_{nm} under H_0 from Halpern's finite sample simulation results, Koziol's exact finite sample distribution approach and large sample asymptotic approximation.

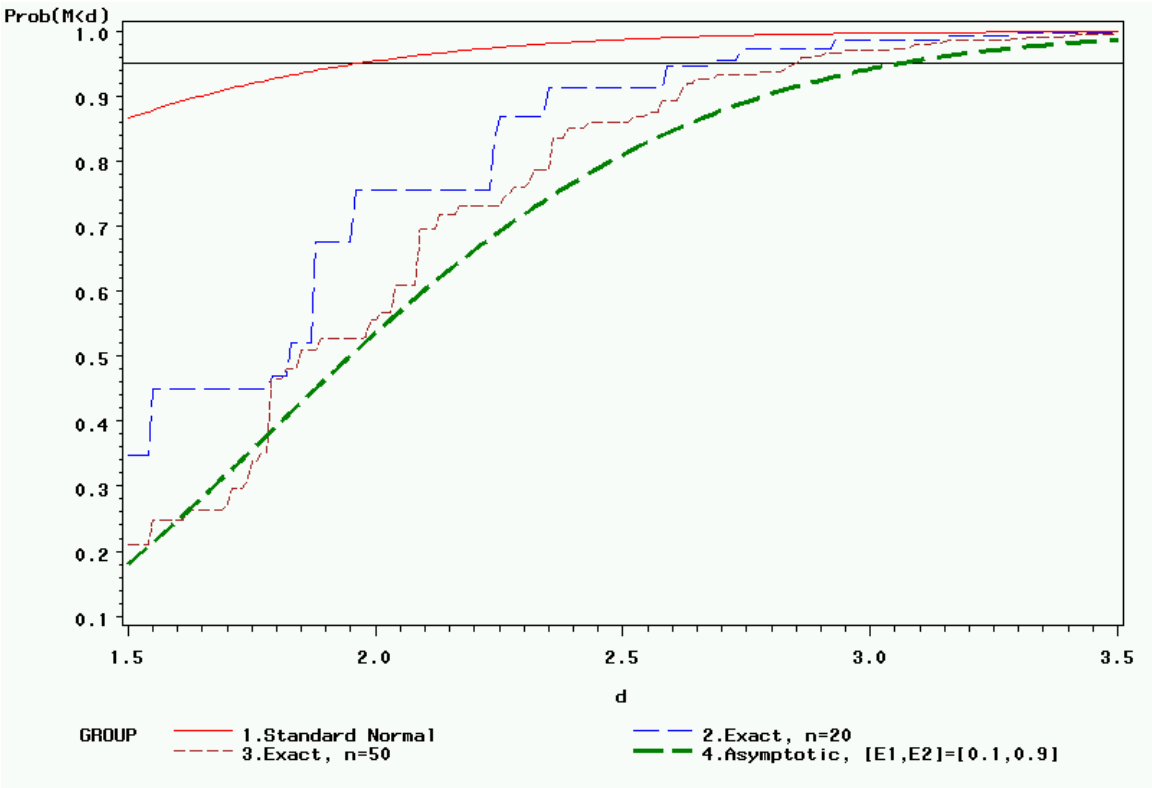
N	α=0.10			α=0.05			α=0.01		
	Halpern†		Koziol‡	Halpern†		Koziol‡	Halpern†		Koziol‡
	ε=0.25	ε=0.1		ε=0.25	ε=0.1		ε=0.25	ε=0.1	
20	2.247	2.343	2.345	2.343	2.583	2.684	2.927	2.927	3.147
50	2.358	2.612	2.612	2.687	2.853	2.850	3.150	3.150	3.312
100	2.431	2.700	2.701	2.800	3.030	2.949	3.491	3.564	3.430
200	2.520	2.711		2.807	2.944		3.225	3.557	
∞*	2.54	2.78		2.83	3.05		3.40	3.59	

†: obtained after taking square root on Halpern's simulation results table 1b;

*: obtained from Miller and Siegmund (1982)'s asymptotic approximation.

‡: due to the large sample size, the A_{nm} is not able to be calculated for N = 200 from Koziol's exact finite sample distribution approach.

Figure 1.1.3 Koziol's exact probability of A_{nm} compare with Miller and Siegmund's asymptotic probability of M (ε1, ε2).



Note: M is Anm from Koziol's exact statistics and M (ε1, ε2) from Miller and Siegmund's asymptotic statistics.

1.4.3 Null Distributions of Maximally Selected Chi-Square Statistics for ordinal predictor variables

Boulesteix (2006) presented an exact method to compute the maximally selected Chi-Square statistics for non-continuous predictor variables of at least ordinal measurement scale (which include e.g. classical ordinal or discrete variables). More specifically, some instances of discrete variables, which are essentially continuous variables measured in a discrete form in practice, include the height of a newborn baby given in centimeters or the blood lead level measured in integer scale. For such variables, if there are K distinct values and N subjects, we generally have K<N if N is large enough, whereas continuous variables may be assumed to take N distinct values in the sample. The exact method proposed by Boulesteix is similar to Koziol's (1991) exact method, which uses Durbin's combinatorial as an approach but takes into account the possibility of multiple samples with same value of X.

Considering the following 2 x 2 contingency table for a given sample $(x_i, y_i)_{i=1,\ldots,N,}$ let $a_1<\ldots<a_K$ denote the different values taken by X.

Group	$X \le a_k$	$X > a_k$	Total
Y=1	$n_{1,\le a_k}$	$n_{1,>a_k}$	N_1
Y=2	$n_{2,\le a_k}$	$n_{2,>a_k}$	N_2
Total	$n_{\cdot,\le a_k} = \sum_{j=1}^{k} m_j$	$n_{\cdot,>a_k} = \sum_{j=k+1}^{K} m_j$	N

Use A_k to indicate the maximally selected $(\chi^2)^{\frac{1}{2}}$ statistic obtained from the above table, A_k can be formulated as $A_k = \max\left(A_k^+, A_k^-\right)$,

where $A_k^+ = \max_{(i=1,\ldots,\sum_{j=1}^{K-1} m_j)} \dfrac{\frac{N}{N_1}\left(\frac{n_{2,\le a_k}}{N_2} - \frac{n_{\cdot,\le a_k}}{N}\right)}{\sqrt{\frac{n_{\cdot,\le a_k}}{N}\left(1-\frac{n_{\cdot,\le a_k}}{N}\right)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}$

and $A_k^- = \max_{(i=1,\ldots,\sum_{j=1}^{K-1} m_j)} \dfrac{\dfrac{N}{N_1}\left(\dfrac{n_{\cdot,\le a_k}}{N} - \dfrac{n_{2,\le a_k}}{N_2}\right)}{\sqrt{\dfrac{n_{\cdot,\le a_k}}{N}\left(1 - \dfrac{n_{\cdot,\le a_k}}{N}\right)\left(\dfrac{1}{N_1} + \dfrac{1}{N_2}\right)}}$

Similar to Koziol's method, let d > 0 be arbitrary. In order to have $A_k \le d$ if and only if all the points with coordinates ($n_{\cdot,\le a_k}$, $n_{2,\le a_k}$) for k=1, ..., K-1 lie on or above the curve of

$$lower_d(x) = \frac{N_2 x}{N} - \frac{N_1 N_2 d}{N}\sqrt{\frac{x}{N}\left(1 - \frac{x}{N}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

and below or on the curve of

$$upper_d(x) = \frac{N_2 x}{N} + \frac{N_1 N_2 d}{N}\sqrt{\frac{x}{N}\left(1 - \frac{x}{N}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

A sufficient and necessary condition for $A_k \le d$ is that the graph $(i, N_2(i))$ does not pass through any point of integer coordinates (i, j) with $i = n_{\cdot,\le a_k}$ and $upper_d(i) < j \le \min(N_2, i)$ or $\max(0, i - N_1) \le j < lower_d(i)$. Use $B_1, \ldots, B_q$ to denote these points of $(i_1, j_1), \ldots, (i_q, j_q)$, where $B_1, \ldots, B_q$ are labeled in the order of increasing i and increasing j within each i. The same as Koziol's method, by using Durbin's combinatorial approach, the probability of $A_k \le d$ can be calculated using the following equation.

$$\Pr[A_k > d] = \binom{N}{N_2}^{-1} \sum_{s=1}^{q} \binom{N - i_s}{N_2 - j_s} b_s$$

Where $b_s = \binom{i_s}{j_s} - \sum_{r=1}^{s-1} \binom{i_s - i_r}{j_s - j_r} b_r, s = 2, \ldots, q, \quad b_1 = \binom{i_1}{j_1}$

According to Boulesteix (2006), Koziol's approach is inappropriate for measuring the association between a binary variable Y and a non-continuous variable X where the number of distinct values of X is substantially less than the sample size.

Figure 1.1.4a and 1.1.4b show the distribution of the exact maximally selected $(\chi^2)^{\frac{1}{2}}$ statistic for small samples. When the predictor variable is continuous, the Koziol's exact method is applied. When the predictor variable is at least ordinal, the Boulesteix's exact method is applied. In figure 1.1.4a the sample size is 20 and 4 distinct levels are used for the at least ordinary predictor. In figure 1.1.4b the sample size is 50. The numbers of 5 and 10 distinct levels are used for the at least ordinary predictor.

According figures 1.1.4a and 1.1.4b, when the predictor variable is discrete the quartile points of the exact maximally selected $(\chi^2)^{\frac{1}{2}}$ statistic are stochastically less than those from the continuous predictor variable given the same sample size. When the predictor variable is in discrete scale, the quartile points of the A_k statistic increase when the number of distinct levels, i.e. number of potential cut points, increases.

Figure 1.1.4a Comparison of the exact probability of maximally selected $(\chi^2)^{\frac{1}{2}}$ statistics between continuous predictor variable and at least ordinal predictor variable with sample size of 20.

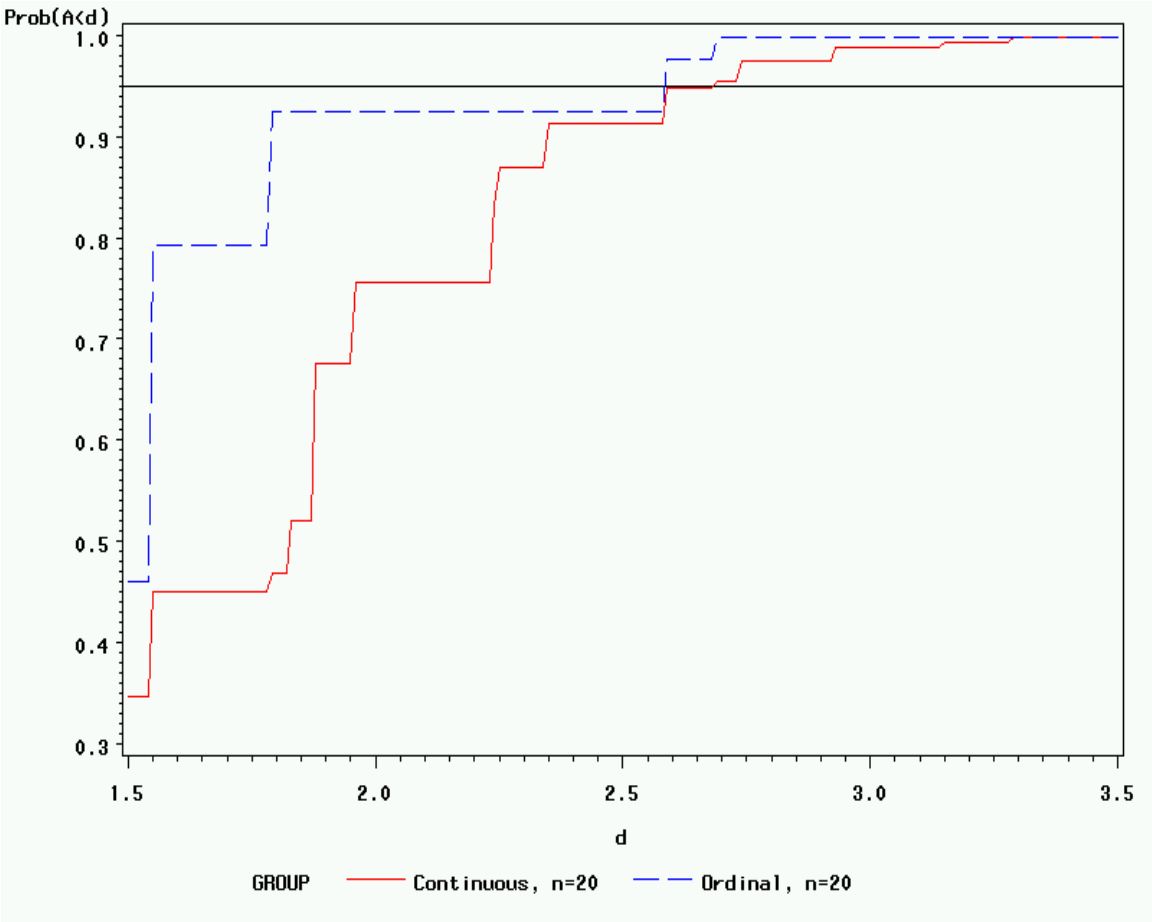
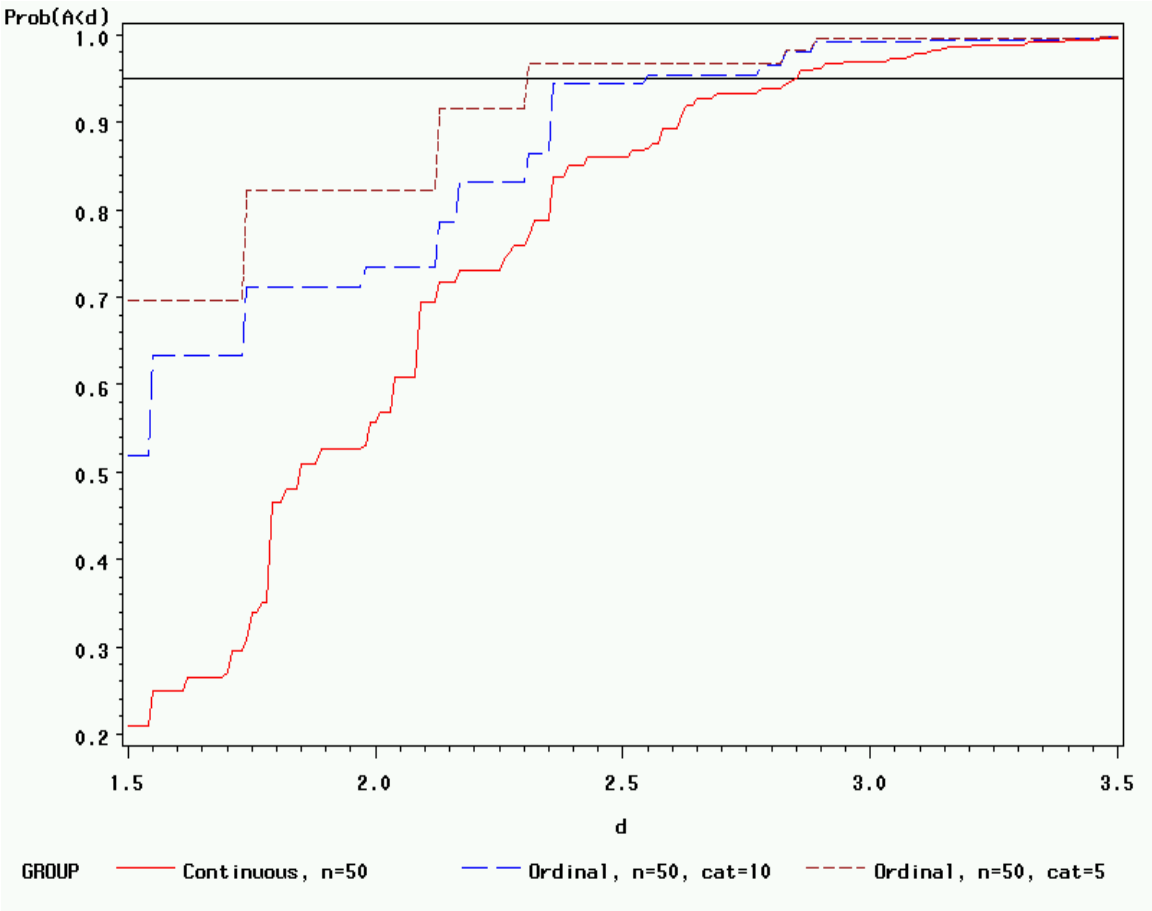


Figure 1.1.4b Comparison of the exact cumulative probability of maximally selected $(\chi^2)^{\frac{1}{2}}$ statistics between continuous predictor and at least ordinal predictor variable with different number of distinct levels given sample size of 50.



Note: A is A_{nm} used in the Koziol's exact method and A_k used in the Boulesteix's exact method.

1.5 Maximally selected rank statistics for Continuous Response Variables

1.5.1 General form of the test statistics

When the response variable is not binary, Lausen and Schumacher (1992) developed the asymptotic approximation for the maximally selected rank statistic, which extended the area of application to continuous and censored response variables. When there exists a cutpoint μ in X, it determines two groups of observations; the groups are defined by all individuals whose X values are either below (or equal to) or above a certain cutpoint. Since the cutpoint is unknown, the estimation of and testing the significance about the cutpoint would be of interest. Analysing the significance of a cutpoint, the null hypothesis H is that the event $X \le \mu$ has no influence on the distribution of Y for all μ:

$$H_0: \quad \Pr(Y \le y \mid X \le \mu) = \Pr(Y \le y \mid X > \mu) \text{ for all } y,\ \mu \in \Re$$

A simple linear rank statistic $S_{n\mu}$ is introduced by Lausen and Schumacher (1992) as

$$S_{n\mu} = \sum_{\{X_i \le \mu\}} a_n(R_{in}),$$

Where $R_{1n}, \cdots, R_{in}, \cdots, R_{nn}$ denote the ranks of $Y_1, \cdots, Y_i, \cdots, Y_n$ and $a_n(R_{in})$ denote the scores, i.e., in the case of tied or censored observations, $a_n(R_{in})$ denotes the mid-scores or the scores given by the log-rank statistic or when set the scores equal to the ranks this $S_{n\mu}$, i.e. $a_n(R_{in}) = R_{in}$, defines the Wilcoxon two-sample rank statistic.

Under the null hypothesis the following conditional expectation and the conditional variance of $S_{n\mu}$ can be written as

$$E(S_{n\mu}) = nF_{nX}(\mu)\bar{a}_n$$

and $\text{var}(S_{n\mu}) = A_n^2 nF_{nX}(\mu)(1 - F_{nX}(\mu))$,

where $A_n^2 = [1/(n-1)]\sum_{i=1}^{n}(a_{in} - \bar{a}_n)^2$, with $a_{in} = a_n(i)$, $\bar{a}_n = (1/n)\sum_{i=1}^{n} a_{in}$, and $F_{nX}(\mu) = (1/n)\sum_{i=1}^{n} I_{\{X_i \le \mu\}}$ is the empirical distribution function of X, in which *I* denotes an indicator function; i.e., I=1 for $\{X_i \le \mu\}$ and I=0 otherwise.

With the above expectation and variance of $S_{n\mu}$ Lausen and Schumacher (1982) developed a standardized rank test statistic $T_{n\mu}$ as in below

$$T_{n\mu} = \frac{S_{n\mu} - E(S_{n\mu})}{(\mathrm{var}(S_{n\mu}))^{1/2}} = \frac{1}{A_n[nF_{nX}(\mu)(1-F_{nX}(\mu))]^{1/2}} \sum_{i=1}^{n} (I_{|X_i \le \mu|} a_n(R_{in}) - F_{nX}(\mu)\bar{a}_n). \quad (2)$$

In order to have a reasonable amount of data in both groups, the hypothetical cutpoint μ is restricted to an interval and the sample quartiles are used for the interval bounds; i.e., $\mu \in [F_{nX}^{-1}(\varepsilon_1), F_{nX}^{-1}(\varepsilon_2)]$, where $0 < \varepsilon_1 < \varepsilon_2 < 1$ and $F_{nX}^{-1}(t) = \min\{x : F_{nx}(x) \ge t\}$.

The maximally selected rank statistic $M_n(\varepsilon_1, \varepsilon_2)$ is of interest and defined by

$$M_n(\varepsilon_1, \varepsilon_2) = \max_{\mu \in [x_1, x_2]} |T_{n\mu}|,$$

where $x_1 = F_{nx}^{-1}(\varepsilon_1), x_2 = F_{nx}^{-1}(\varepsilon_2)$, *and* $0 < \varepsilon_1 < \varepsilon_2 < 1$.

1.5.2 A special case in Wilcoxon two-sample rank statistic

Use the special case of Wilcoxon rank-sum test as an example the equation (2) of standardized rank test statistic in section 1.5.1 can be simplified as

$T_{n\mu} = \dfrac{S_{n\mu} - E(S_{n\mu})}{\sqrt{\mathrm{var}(S_{n\mu})}}$, which is the large sample approximation on Wilcoxon rank-sum test. The $S_{n\mu}$ is the sum of the ranks on Y for the group with X values below or equal to a certain cut point of μ.

For the Wilcoxon rank-sum test without ties, use $n_{\le\mu}, n_{>\mu}$ to denote the number of observations with X values below (or equal to) or above a certain cut point of μ

respectively, and use *n* to denote the total sample size, the $E(S_{n\mu})$ and $\text{var}(S_{n\mu})$ can be calculated using the following equations.

$$E(S_{n\mu}) = \frac{n_{\leq u}(n+1)}{2}$$

$$\text{var}(S_{n\mu}) = \frac{n_{X\leq\mu} n_{X>\mu}(n+1)}{12}$$

1.5.3 The use of Log rank statistic

When the outcome interest is in survival time the equation (2) of standardized rank test statistic in section 1.5.1 can be written as log rank statistic. In which the $S_{n\mu}$ is the sum of the events at each time point for group with values in predictor variable below the hypothetical cutpoint.

Let X be the risk factor of interest measured as a continuous variable and T be the outcome variable of survival time. The population is divided into two groups based on the cutpoint. Let $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ be the ordered observed event times of the outcome variable T. Let O_i be the number of events at time $t_{(i)}$, N_i be the number of subjects at risk prior to time $t_{(i)}$ and O_i^- and N_i^- be the number of events at time $t_{(i)}$ in group with values in predictor variable below the hypothetical cutpoint. Then the $S_{n\mu}$, $E(S_{n\mu})$ and $\text{var}(S_{n\mu})$ can be calculated using the following equations.

$$S_{n\mu} = \sum_{i=1}^{k} O_i^-$$

$$E(S_{n\mu}) = \sum_{i=1}^{k} O_i \frac{N_i^-}{N_i}$$

$$\text{var}(S_{n\mu}) = \sum_{i=1}^{k} \frac{O_i(\frac{N_i^-}{N_i})(1-\frac{N_i^-}{N_i})(N_i - O_i)}{N_i - 1}$$

.

1.5.4 Asymptotic Approximation

As mentioned in section 1.5.1, since the cut point is unknown, the estimation of and testing the significance about the cut point are of interest. Searching for the μ that corresponds to the maximally selected rank statistic $M_n(\varepsilon_1, \varepsilon_2)$ defined in section 1.5.1 would be the straightforward approach to estimate the cut point. Lausen and Schumacher showed the asymptotic distribution for the maximally selected rank statistic $M_n(\varepsilon_1, \varepsilon_2)$ is the same as the asymptotic distribution of the square root of the maximally selected χ^2 statistic derived by Miller and Siegmund (1982) and has the following approximation for the distribution

$$pr[\max_{\mu\in[x_1,x_2]}\left|T_{n\mu}\right| \geq b] = \frac{4 = \varphi(b)}{b} + \varphi(b)(b - \frac{1}{b})\log(\frac{\varepsilon_2(1-\varepsilon_1)}{(1-\varepsilon_2)\varepsilon_1}) + o(\frac{\varphi(b)}{b}),$$

for $0 < \varepsilon_1 < \varepsilon_2 < 1$ and $b \to \infty$, where $\varphi(b)$ denotes the standard normal density.

1.5.5 Simulation results for finite samples

In order to gain some insight into the finite null distribution of the maximally selected rank statistics, Lausen and Schumacher also conducted a Monte Carlo simulation study with different sample sizes. The critical values of the 95% quartile in the simulated distributions of maximally selected Wilcoxon rank statistic and maximally selected log-rank statistic under the null hypothesis that X and Y are independent are listed in table 1.1.3 and table 1.1.4 respectively.

Both table 1.1.3 and table 1.1.4 suggest that the asymptotic distribution has larger critical values than finite sample distribution, which indicates the use of asymptotic critical values will result in a conservative test, fairly to reject at greater than a 0.05 level.

Table 1.1.3 Simulated and approximated 95% quartiles of the distributions of the maximally selected Wilcoxon rank statistics under the null hypothesis

n	$(\varepsilon_1, \varepsilon_2)$			
	(.1, .9)	(.25, .75)	(.4, .6)	(.4, .9)
10	2.39	2.39	2.19	2.35
20	2.59	2.50	2.39	2.47
30	2.70	2.60	2.41	2.58
50	2.78	2.64	2.43	2.66
100	2.90	2.73	2.46	2.78
200	2.93	2.77	2.50	2.82
∞	3.05	2.83	2.56	2.88

Table 1.1.4 Simulated and approximated 95% quartiles of the distributions of the maximally selected log-rank statistics under the null hypothesis

n	$(\varepsilon_1, \varepsilon_2)$			
	(.1, .9)	(.25, .75)	(.4, .6)	(.4, .9)
10	2.55	2.30	2.04	2.29
20	2.70	2.45	2.26	2.50
30	2.76	2.52	2.30	2.62
50	2.82	2.57	2.36	2.67
100	2.91	2.65	2.41	2.74
200	2.94	2.72	2.49	2.81
∞	3.05	2.83	2.56	2.88

1.6 Application Studies from the literature

In the analysis involving data from clinical or epidemiological studies, significant attention is given to continuous variables such as blood pressure, certain biomarkers etc., but the predictive importance of such variables can not be established easily. Transforming a continuous variable into a categorical variable usually binary, makes the

model more interpretable. In this section, two examples of the cut point determination in clinical treatment procedures are given which implies the choice of a cutpoint to convert a continuous covariate to a binary covariate is often based on biological knowledge about the particular risk factor or on the results already published in other studies.

1.6.1 Treatment for Unresponsive Lymphoma (Mazumdar, 2000)

For patients with advance-stage or poor-prognosis malignant lymphoma that have not responded to conventional-dose chemotherapy, high-dose (HD) chemotherapy is recommended (Haas, 1994). But since high-dose chemotherapy is not only toxic to patient's cancer cells but also toxic to patient's healthy blood cells, the HD chemotherapy alone can lead to poor recovery and great risk of infections. The use of HD chemotherapy followed by re-infusion of the peripheral blood stem cells (PBSC), which collected from other donors or the patients themselves before the HD chemotherapy, is effective in treating relapsed non-Hodgkin's lymphoma. A high complete response rate is seen and a significant fraction of patients appear to be cured by this approach. There has shown an inverse correlation between the quantity of the PBSC re-infused and the speed of patient recovery. Or in another word, the greater the quantity of the PBSC re-infused, the faster the patient recovers from the toxicity of the treatment (Moskowitz, 1998). The number of PBSC collected from patients or other donors can in most cases be increased by the administration of a stem cell growth factor, called Granulocyte Colony-Stimulating Factor (G-CSF) (Valbonesi 1996). In clinical practice, the threshold dose of PBSC necessary to ensure patients would have a rapid recovery of their HD chemotherapy blood counts would be of interest. In the lymphoma literature, cutpoints defining threshold doses PBSC needed range from 1.2 million to 5.0 million. And the most frequently reputed cutpoint is 2.5 million cells (Mazumdar, 2000). But Mazumdar point out that all of these cutpoints found were simply by examining scatter plots for threshold effects.

The outcome of interest from the lymphoma literatures is the time for patients to recover 20,000 platelets in days by using two categories. One is in less than 14 days and another is in greater or equal to 14 days, namely 'successful' recovery and 'slow' recovery. This 14-day threshold is based not only on mortality but also on hospital costs and quality of

life consideration. By using this binary outcome variable, Mazumdar analyzed 55 patients who were treated with an HD chemotherapy followed by re-infusion of PBSC regimen at Memorial Sloan-Kettering Cancer Center between 1994 and 1997. The data shows positive correlation between the quantity of PBSC re-infused and the proportion of patients who recovered to 20,000 platelets in less than 14 days. By using the uncorrected Chi-square statistics, Mazumdar observed at 2.0 million PBSCs there exists a maximum chi-squared and minimum p-value of 8.73 and 0.003, respectively. Mazumdar reassessed the significance of this cut point using the asymptotic approximation of the maximally selected chi-square statics with searching interval at the central 80%, i.e. set ε=0.1. The p-value changed to 0.068, which remains of marginal significance (Mazumdar 2000). However since this study is with finite sample size of 55 patients, the Koziol's exact statistics would be preferred to adjust the p-value. Without the detailed data, I would use the critical values of $2.85^2 = 8.12$ or $2.949^2 = 8.696$ for sample size of 50 and 100 respectively (see table 1.1.2 in section 1.4.2 of this chapter), which assumes the cut point is located in the median position of the data, to compare with the maximum chi-squared statistics. Then we would conclude the quantity of 2.0 million PBSCs is a significant cut point. Based on the combined evidences from other lymphoma literatures, this 2.0 million PBSCs re-infusion appears to be a reasonable cutpoint for this type of patient to achieve a successful recovery from HD chemotherapy. Although there is evidence of inverse correlation between the quantity of the PBSC re-infused and the speed of patient recovery. During the practice, clinicians would like to have a recommended quantity of the PBSC collected and re-infused to ensure a rapid recovery. Rather than using the scatter plots, the maximally selected chi-square statistics method can be used as a tool to claim the statistic significance of the identified cutpoint which separates the patient population into two groups with high or low probabilities of favourable outcomes.

1.6.2 Treatment for Seminoma (Pub 1996)

Puc et al published a study in Journal of Clinical Oncology in 1996, identified a cut point in the size of residual tumour mass to relate with the poor prognosis for post

chemotherapy patients with advanced seminoma and claimed to use this cut point to justify the post chemotherapy for these patients.

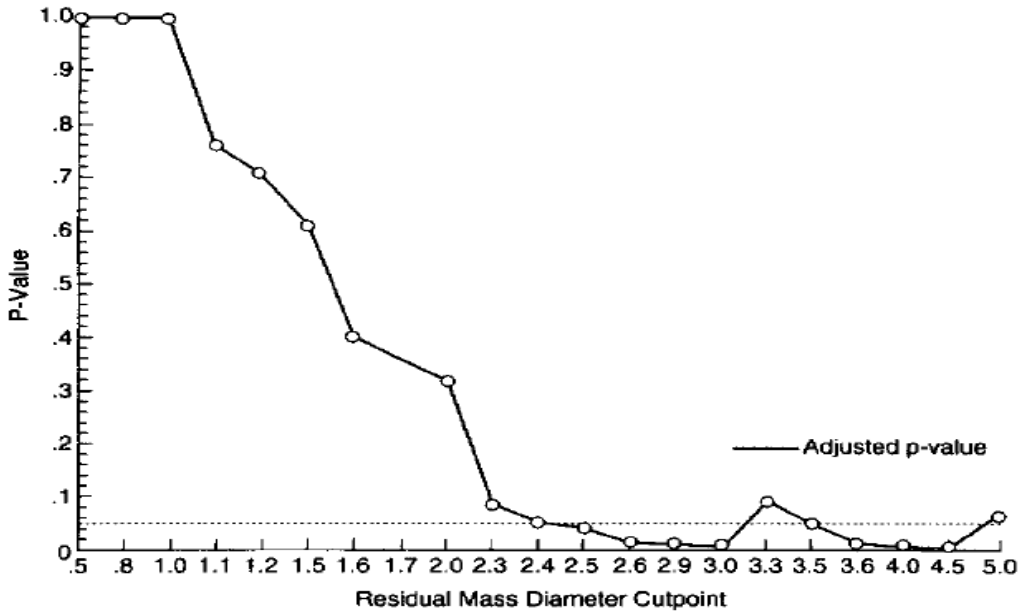
The majority of patients with seminoma who are treated initially with chemotherapy are found to have a residual tumour mass. The guidelines for management of these post chemotherapy patients are controversial, and may involve surgery, radiotherapy or close observation. Efforts to reduce treatment related morbidity include the avoidance of post chemotherapy surgery. The size of the residual mass on a post-chemotherapy computer tomography (CT) scan, measured as the largest diameter, is reported in the literature to predict poor prognosis. Pub et al analyzed the retrospective data from 104 advanced seminoma patients treated with various chemotherapy regiments at Memorial Sloan-Kettering Cancer Center from 1979 to 1992 The outcome studied was whether a patient had a site failure or non-site failure during the post chemotherapy follow up period, which is a binary variable. Site failure was defined as either the presence of cancerous tumour found at post chemotherapy surgery or clinical relapse at the assessed site. The site non-failure was defined as no cancerous tumour found at post chemotherapy surgery or no clinical relapse during the follow up period. The median follow-up time is 47 months with range of 5 to 153 months. Out of these 104 patients studied, 10 had site failures and 94 had no site failures. Various prognostic factors have been studied, including primary disease site, chemotherapy regimen, pre-treatment serum tumour markers, residual disease site, residual tumour mass size etc. to relate with the outcome of site failure. Out of them the residual tumour mass size is the only significant variable predictive of outcome. The p-value is determined as 0.0316 from log-rank statistic by analyzing the failure-free survival time and size of the tumour mass size.

The unadjusted chi-square statistic is calculated based on the binary outcome of site failure and binary predictor of residual tumour mass size dichotomized using every available data point in the range. The results show when the cut point is set at 3 cm, the analysis yields the highest chi-square and lowest p-value of 14.1 and 0.006 respectively. The author then states by using 3 cm as the cut point the adjusted lowest p-value is 0.0316, which is significant. Furthermore the author recommended that for patients with a residual tumour mass less than 3 cm, the post chemotherapy can be managed only by close observation and additional intervention is not indicated. Although in this literature

published by Pub et al, the author did not specify which method was used to adjust the p-value, by using the asymptotic approximation of the maximally selected chi square statistics, and with the searching interval set at the central 80% (ε=0.1), the square root of the highest chi-square value of 14.1 leads to a p-value of 0.032, which matches with the 0.0316 in this article.

The data set used in this article has relatively large sample size with 104 patients. The predictive variable of the residual tumour size is measured at semi-continuous level (see figure 1.1.5), in which the number of the potential cut points, i.e. the number of distinct values in X, is less than the sample size. The most appropriate method to adjust the p-value in this case would be the Boulesteix's (2006) method described in section 1.4.3 of this chapter. But because the large sample asymptotic approximation provides the most conservative result, the change of p-value adjusting method won't alter the final conclusion of declaring the 3 cm in the size of residual tumour mass as the significant cut point. Also figure 1.1.5 shows there are multiple data points with adjusted p-values below 0.05, but the data point with maximum of the Chi-square statistics is the one that best separates the patient population.

Figure 1.1.5 Comparison of cutpoint for residual mass size according to adjusted p value based on relationship between site failure and site non-failure.



Note: This figure is cited from Figure 1 in Puc et al. 1996

CHAPTER 2: THE EFFECT OF DISCRETENESS OF VALUES FOR THE PREDICTOR ON THE NULL DISTRIBUTION FOR MAXIMALLY SELECTED RANK STATISTICS

2.1 Abstract

In Lausen's paper (1992), only the possibility of a continuous predictor is discussed for the maximally selected rank statistics. When a predictor is semi-continuous or ordinal, the number of cutpoints available, which is the number of distinct values in X, is substantially less than those from a continuous predictor. Consequently for a given cut point the alpha value is inflated less than when the predictor variable is continuous. In this chapter, the Monte Carlo simulation is used to compare the null distributions of maximally selected rank statistics from predictors with continuous and semi-continuous measurements.

2.2 Introduction

In Lausen's paper (1992), a rank statistic is given and the asymptotic distribution of the maximally selected rank statistics is developed. This asymptotic result is then compared with the Monte Carlo simulation results with different sample sizes. However all of the simulations only considered continuous predictor variables. When the predictor is semi-continuous or ordinal, compared to the continuous one, it would require a less stronger level of evidence to be observed in order for a cutpoint to be deemed "significant", as to compensate for less number of inferences being made. In this chapter, the following two aims will be achieved via Monte Carlo simulation.

Aim 1: To identify the level of discreteness in a semi-continuous predictor at which investigators need to use alternative null distributions to the approximate distribution for large samples (n>=200).

Aim 2: To identify the level of discreteness in a semi-continuous predictor at which investigators need to use alternative null distributions to the approximate or exact distributions for small samples (n=20, 30, 50 or 100).

2.3 Method

The null distribution of maximally selected rank statistic from continuous predictor and that from the semi-continuous or ordinal predictor will be compared via Monte Carlo simulation. For the semi-continuous or ordinal predictor, two different levels of discreetness which are with 10 ordinal levels and 15 ordinal levels are explored.

The random numbers are generated for both continuous predictor and semi-continuous predictor. For continuous quantitative predictor, the uniform distribution on (0, 1) was used. The random numbers were generated by using SAS function of RANUNI. For semi-continuous predictor, the random numbers were also generated from uniform distribution on (0, 1). For the case of having 10 ordinal scales in X, the generated random numbers are multiplied by 10. For the case of having 15 ordinal scales in X, the generated random numbers are multiplied by 15. The obtained values will then be rounded to the smallest integers that are greater than or equal to the random numbers. Since the null distribution of the maximally selected rank statistic is nonparametric, it is invariant for arbitrary distribution of X and Y (Lausen, 1992). The response variable will be set to standard normal, *N (0, 1)*, using the SAS function of RANNOR under the null hypothesis that there is no relationship between X and Y. The sample size, n, will cover 20, 30, 50, 100 and 200 the interval of (ε_1, ε_2) will cover (0.25, 0.75), (0.4, 0.6) and (0.1, 0.9). The Lausen's rank statistic described in section 1.5 of chapter 1 are derived with the scores, $a_n(R_{in})$, set equal to the ranks. 10,000 Monte Carlo repetitions are generated on the calculation of maximally selected rank statistics.

The null hypothesis will be rejected if the maximally selected rank statistic is greater than a critical value. Hence the upper quartiles of the simulated null distribution are of interest. The null distributions of the maximally selected rank statistic that generated by using continuous predictor and semi-continuous predictor will be compared through tables and plots.

2.4 Results and Discussion

When the predictor variable X and the response variable Y are independent under the null hypothesis, table 2.2.1, figure 2.2.1a and figure 2.2.1b show the cumulative distribution of maximally selected rank statistics for both continuous predictor and semi-continuous predictor with different sample sizes. The searching interval is set at the central 50% with ε of 0.25 for figure 2.2.1a and at the central 80% with ε of 0.10 for figure 2.2.1b.

The results suggest that the critical values from the distribution on continuous predictor are stochastically larger than those on the semi-continuous predictor given the same sample size (figure 2.2.1a and 2.2.1b). Compared to the semi-continuous predictor with 10 discrete levels, the critical values obtained from the continuous predictor are roughly increased by 10% at 0.05 α level. Also as expected, the figure 2.2.2 shows that the critical values increase when the discrete levels increased from 10 to 15. Corresponds to aim 1 in section 2.2, when the sample size is at 200, using the semi-continuous predictor can result in a much smaller critical value compared to continuous predictor. When the searching interval is set at the central 80%, the critical values are 2.94, 2.74 and 2.59 for predictors measured in continuous, semi-continuous with 15 levels and semi-continuous with 10 levels respectively. Similar results can be seen in small sized samples. Based on these simulated results, when the discrete level is at 15 or below the alternative null distribution should be recommended compared to the null distribution obtained from continuous predictor..

Figure 2.2.1a shows that when the sample size is increased from 50 to 100 the critical values are increased substantially for both continuous and semi-continuous predictors when the searching interval is set at $(\varepsilon_1, \varepsilon_2) = (.25, .75)$. For example, when the α is set at 0.05 level, the critical values are increased from 2.64 to 2.73 for continuous predictor and from 2.43 to 2.47 for semi-continuous predictor with 10 discrete levels.

However the figure 2.2.3 shows that at the searching interval of central 50% the critical values are very close to each other when to compare the sample size of 100 to the sample size of 200, given the predictor variables are measured in the same scales, i.e. both continuous or both semi-continuous with the same discrete levels.

Further more when set the searching interval at central 80%, the quartile curves have little difference for sample size of 50, 100 and 200 (figure 2.2.4) from semi-continues predictor with 10 discrete levels. However when the discrete level is set at 15 the quartile curves do show difference from sample size of 50, 100 and 200 (figure 2.2.5). This indicates there might exist a threshold ratio between the number of potential cut points available and the sample size. When this ratio gets smaller the quartile points may converge to constants.

As expected, figure 2.2.6 to 2.2.9 reveals the critical values are impacted by ε. Figure 2.2.6 and 2.2.7 plot the simulated and approximated upper quartiles of the maximally selected rank statistics from continuous predictor, where the finite samples are with size of 30, 100 and 200. When the searching intervals are increased from the central 20% to 50% or from the central 50% to 80% the critical values can be increased by 5 to 10%. The same pattern can be been seen in semi-continuous predictor with 10 or 15 discrete levels (figure 2.2.8 and 2.2.9).

Figure 2.2.1a Simulated and approximated upper α quartile of the maximally selected rank statistics, where the maximum is over the interval of $(\varepsilon_1, \varepsilon_2) = (.25, .75)$, for continuous predictor and semi-continuous predictor.

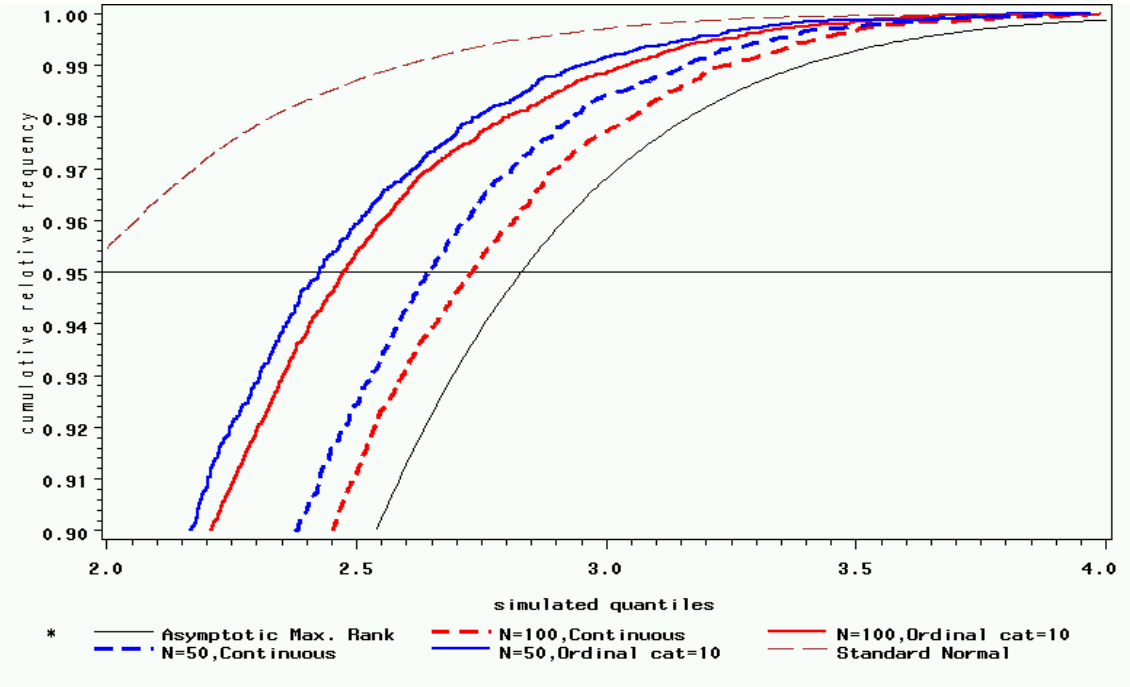


Figure 2.2.1b Simulated and approximated upper α quartile of the maximally selected rank statistics, where the maximum is over the interval of $(\varepsilon_1, \varepsilon_2) = (.10, .90)$, for continuous predictor and semi-continuous predictor.

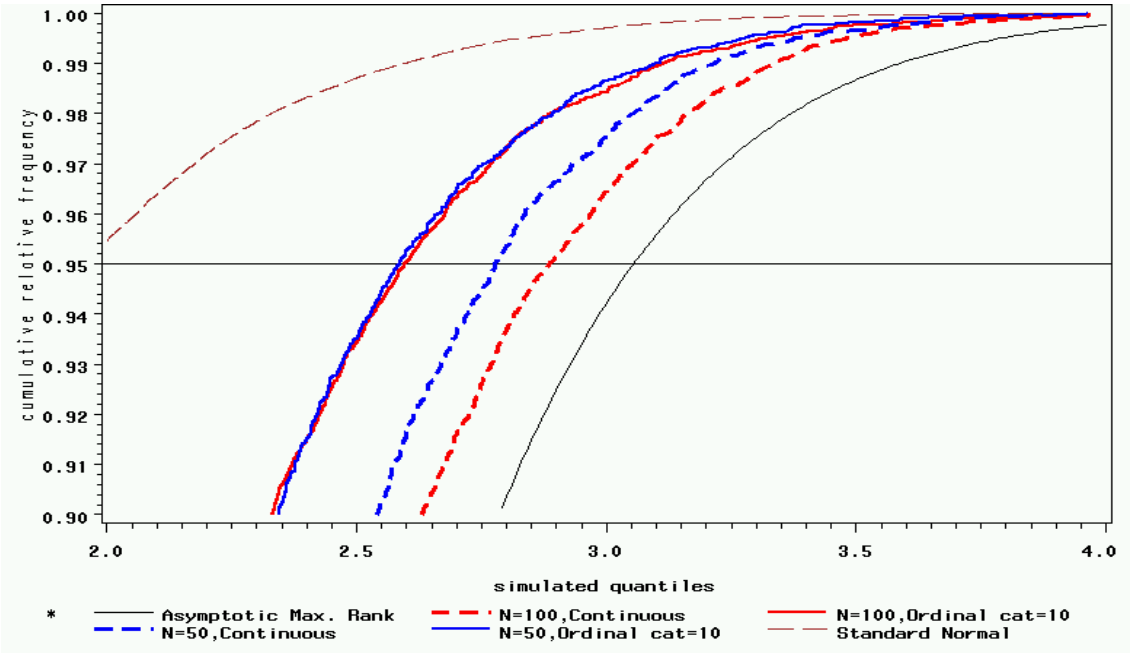


Figure 2.2.2 Simulated upper α quartile of the maximally selected rank statistics, where the maximum is over the interval of (ε_1, ε_2) = (.25, .75), for semi-continuous predictor with 10 and 15 discrete scales.

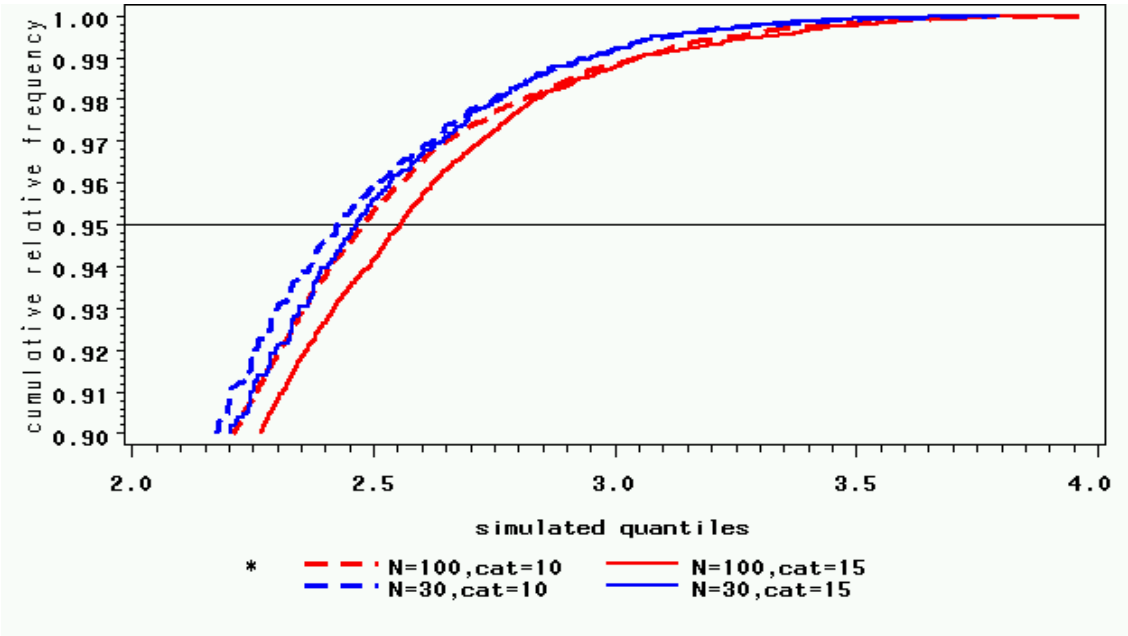


Figure 2.2.3 Simulated and approximated upper α quartile of the maximally selected rank statistics, where the maximum is over the interval of (ε_1, ε_2)=(.25, .75), n=100, 200

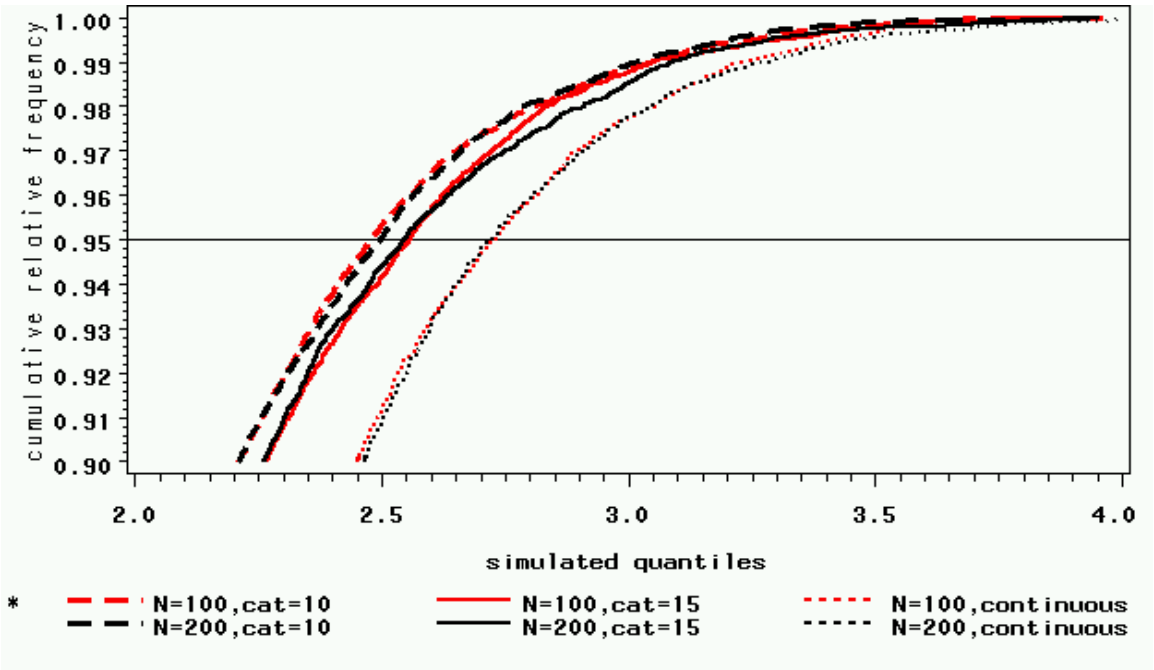


Figure 2.2.4 Simulated upper α quartile of the maximally selected rank statistics from finite sample size of 20, 30, 50, 100 and 200 from semi-continuous predictor with 10 discrete levels, where the maximum is over the intervals of central 80% (ε =0.1).

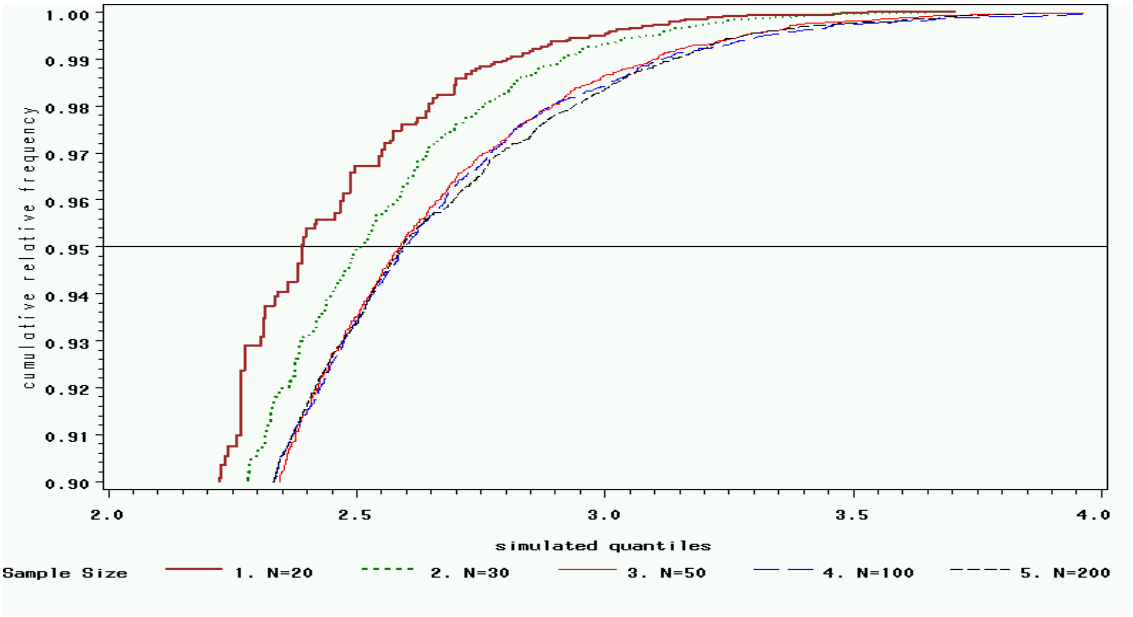


Figure 2.2.5 Simulated upper α quartile of the maximally selected rank statistics from finite sample size of 20, 30, 50, 100 and 200 from semi-continuous predictor with 15 discrete levels, where the maximum is over the intervals of central 80% (ε =0.1).

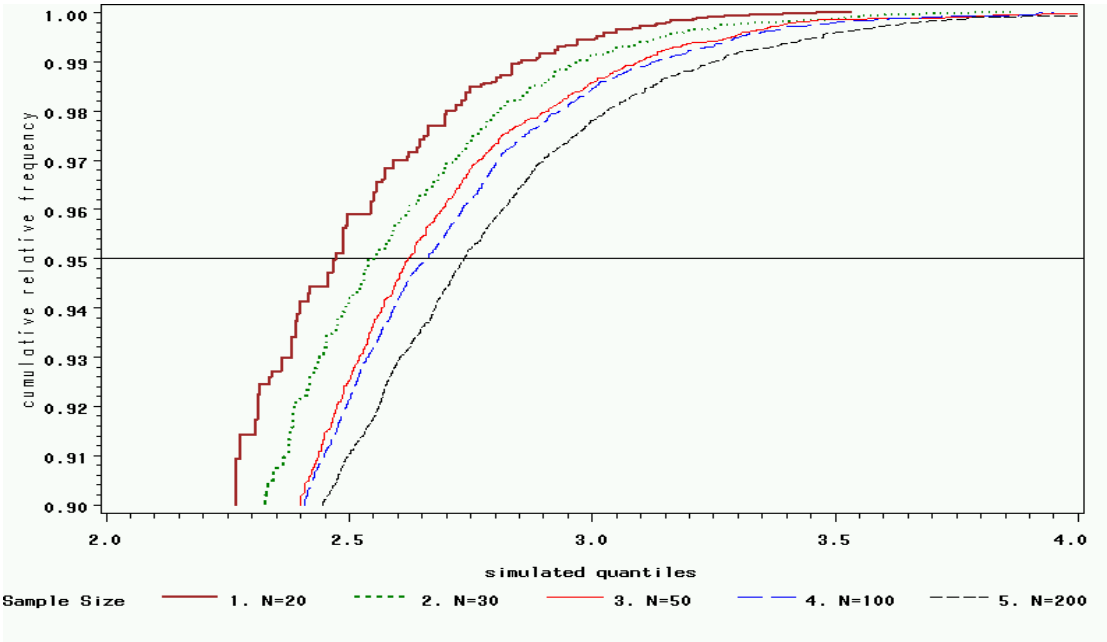


Figure 2.2.6 Simulated upper α quartile of the maximally selected rank statistics from finite sample size of 30 and 100 with continuous predictor, where the maximum is over the intervals of central 80% (ε =0.1), 50% (ε =0.25) and 20% (ε =0.4).

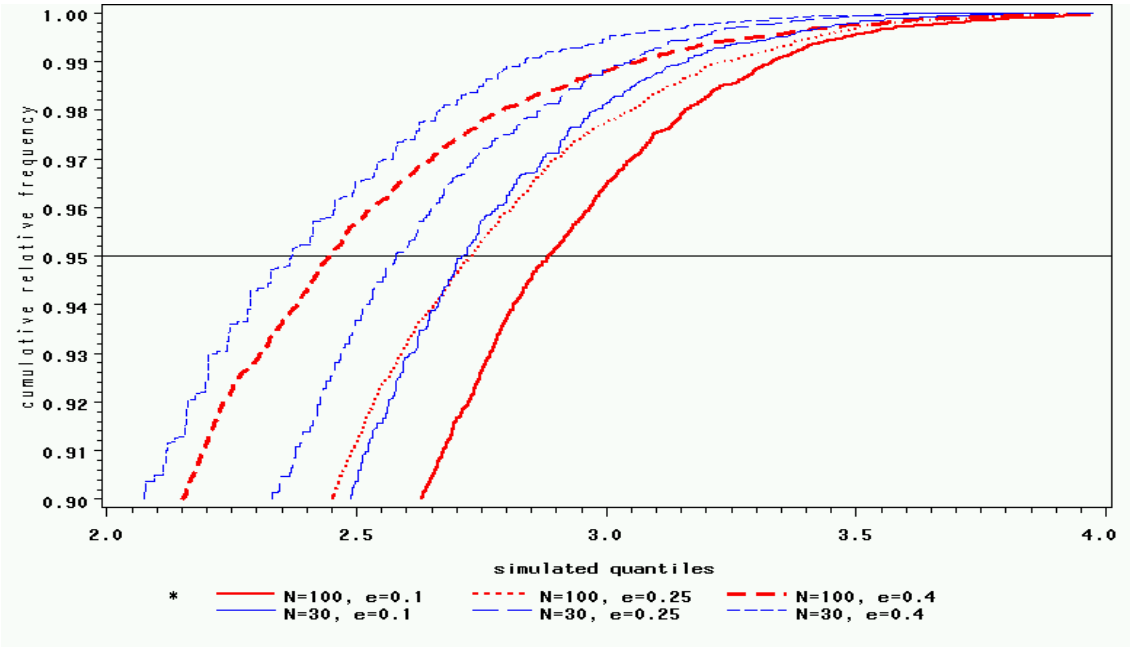


Figure 2.2.7 Simulated upper α quartile of the maximally selected rank statistics from finite sample size of 200 and large sample asymptotic approximation of the maximally selected rank statistics with continuous predictor, where the maximum is over the intervals of central 80% (ε =0.1), 50% (ε =0.25) and 20% (ε =0.4).

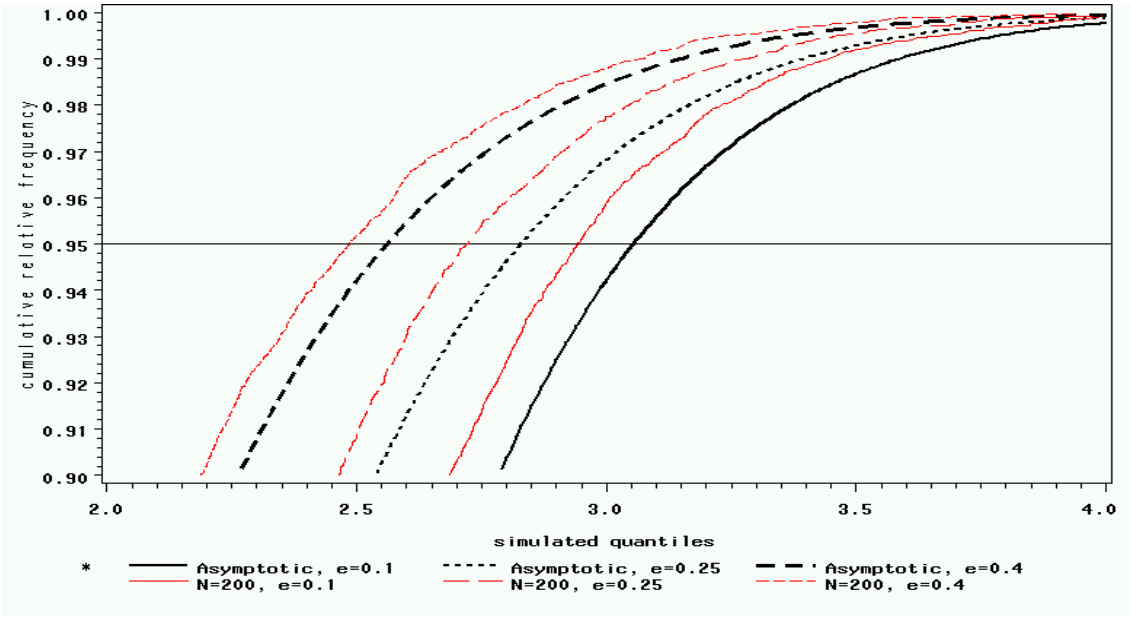


Figure 2.2.8 Simulated upper α quartile of the maximally selected rank statistics from semi-continuous predictor with 10 discrete levels, where the maximum is over the intervals of central 80% (ε =0.1), 50% (ε =0.25) and 20% (ε =0.4) and n=50, 200.

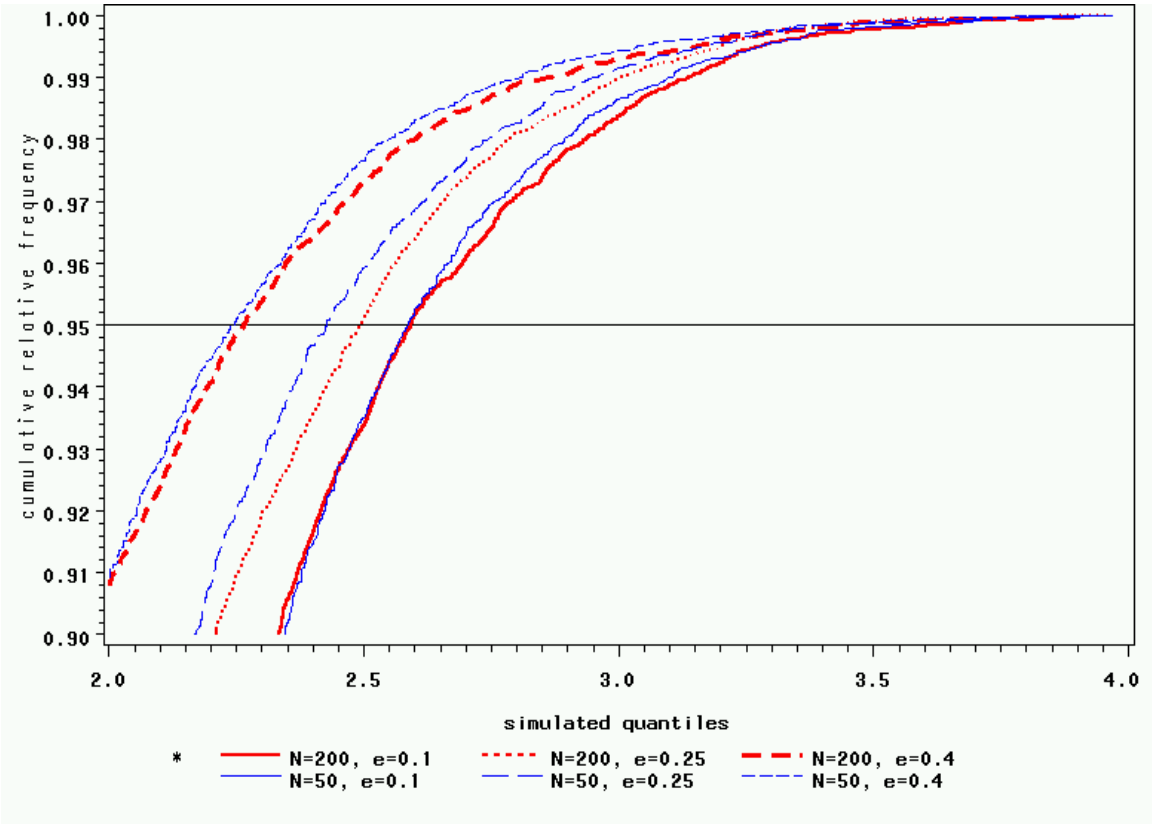


Figure 2.2.9 Simulated upper α quartile of the maximally selected rank statistics from semi-continuous predictor with 15 discrete levels, where the maximum is over the intervals of central 80% (ε =0.1), 50% (ε =0.25) and 20% (ε =0.4) and n=50, 200.

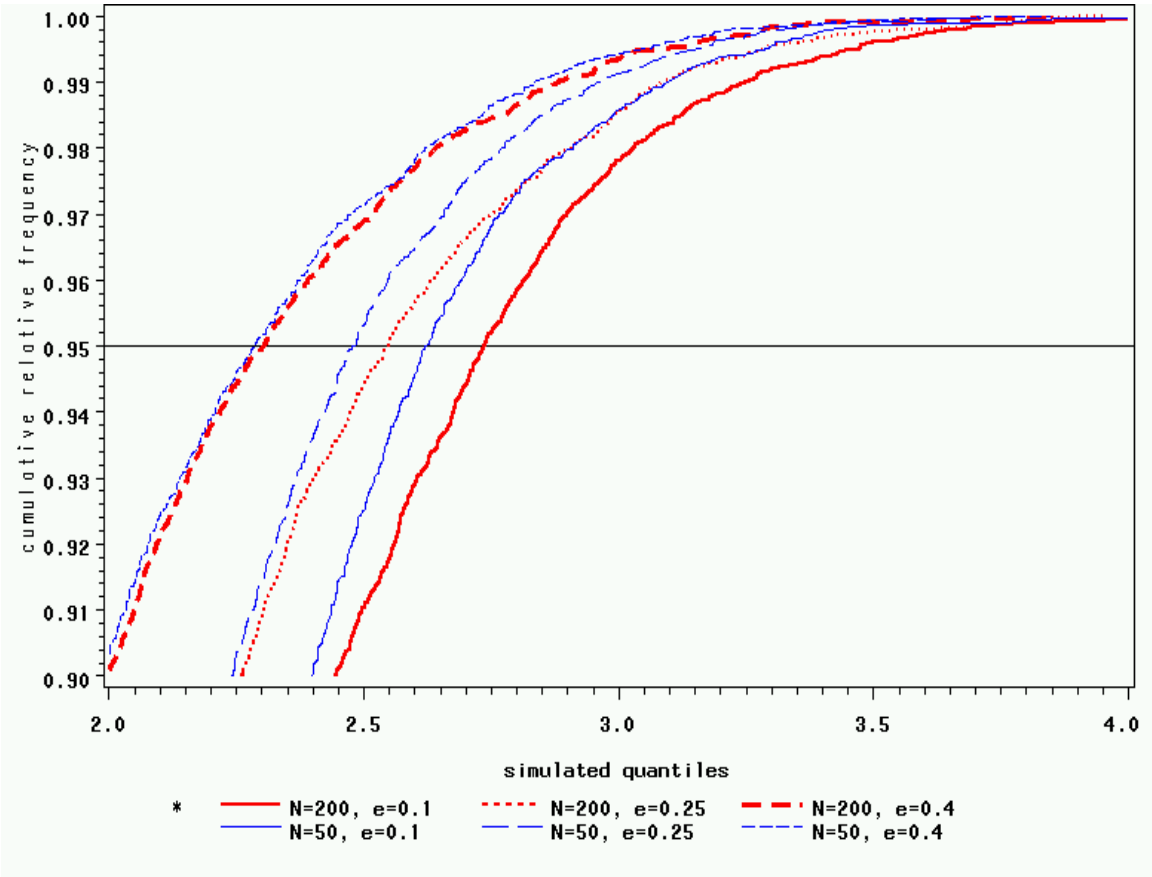


Table 2.2.1 Simulated and approximated upper α quartile of the maximally selected rank statistics, where the maximum is over the interval of (ε1, ε2) = (.25, .75), for continuous predictor and semi-continuous predictor with 10 discrete scales.

$(\varepsilon_1, \varepsilon_{2)}$	N	α=0.10			α=0.05			α=0.01		
		Cont. Pred.[1]	S15 Pred[3]	S10 Pred[2]	Cont. Pred.[1]	S15 Pred[3]	S10 Pred[2]	Cont. Pred.[1]	S15 Pred[3]	S10 Pred[2]
(.10, .90)	20	2.39	2.27	2.23	2.59	2.47	2.39	3.01	2.85	2.80
	30	2.49	2.33	2.28	2.71	2.55	2.51	3.16	2.97	2.92
	50	2.54	2.40	2.34	2.78	2.62	2.59	3.21	3.10	3.10
	100	2.63	2.41	2.33	2.89	2.66	2.60	3.33	3.13	3.12
	200	2.68	2.44	2.33	2.94	2.74	2.59	3.44	3.26	3.14
	∞*	2.79			3.05			3.59		
(.25, .75)	20	2.27	2.23	2.14	2.49	2.42	2.39	2.97	2.89	2.83
	30	2.33	2.20	2.17	2.58	2.47	2.43	3.05	2.95	2.95
	50	2.38	2.24	2.17	2.64	2.48	2.43	3.16	2.96	2.95
	100	2.45	2.27	2.21	2.73	2.56	2.47	3.24	3.05	3.04
	200	2.46	2.26	2.21	2.72	2.55	2.50	3.28	3.09	3.01
	∞*	2.54			2.83			3.40		
(.40, .60)	20	2.01	2.00	1.94	2.32	2.26	2.19	2.80	2.74	2.73
	30	2.07	1.99	1.95	2.37	2.28	2.20	2.84	2.77	2.77
	50	2.12	1.98	1.95	2.39	2.28	2.25	2.99	2.83	2.81
	100	2.15	1.99	1.95	2.44	2.31	2.25	3.07	2.89	2.84
	200	2.19	2.00	1.96	2.49	2.30	2.26	3.07	2.88	2.87
	∞*	2.26			2.56			3.15		

*: From asymptotic estimation.
1: Continuous predictor.
2: Semi-continuous predictor with 10 discrete levels
3: Semi-continuous predictor with 15 discrete levels

CHAPTER 3: THE EFFECT OF ASSESSING PREDICTORS ON DIFFERENT SCALES ON THE POWER OF A STUDY WHEN USING A MAXIMALLY SELECTED RANK STATISTIC IN HYPOTHESIS TESTING

3.1 Abstract

The power curves of the maximally selected rank statistics from continuous predictor and that from semi-continuous predictor with different level of discreteness are compared at the same cutpoint and the same sample size via Monte Carlo simulation by using their respective critical values obtained from chapter 2. The results show the power of maximally selected rank statistic from the semi-continuous predictor is stochastically larger than that from the continuous predictor. The simulation results also show that besides the sample size and effect size, the location of the true cut point also affects the power of test using the maximally selected rank statistic. The power reaches its highest when the true cut point is at the 50^{th} percentile of the predictor variable and the power decreases when the location of the true cut point moves to the lower or upper quartiles.

3.2 Introduction

When planning a study, investigators may be able to choose between measuring a subject characteristic using continuous scale and using discrete semi-continuous or ordinal scale. For example, in pregnancy and birth data, in order to access the relationship between the weight of the baby at birth and the type of delivery (natural or cesarean), the baby weight at birth maybe measured at precision of 1 g or 100 g or by 3 categories of <2500 g, 2500 – 4000 g or >4000 g. In this case the cut point of birth weight in semi-continuous scale may be more clinical meaningful and have larger power to identify an association. In this chapter, the following aim 1 will be examined using Monte Carlo simulation.

Aim 1: Under the appropriate null distributions, assess whether using a discrete semi-continuous or ordinal predictor rather than a purely continuous predictor will produce more power when using a maximally selected rank statistic to test for significance.

Often, when investigators study the relationship between some continuous response and predictor variables, appropriate diagnostics are not considered. Instead the linearity is the most common assumption used in the data analyses. The objective of the following aim

2 is to compare the power using simple linear regression to that using the maximally selected rank statistic when detecting associations when the effect of the predictor on the response is a simple step function.

Aim 2: Determine whether simple linear regression is robust to misspecification of the effect of predictor, when the effect of predictor on the response is a step function.

The following aim 3 is to study if the maximally selected rank statistics can detect association given the underlying relationship is linear.

Aim3: Compare the power between linear regression method and maximally selected rand statistics when the true relationship between predictor and response is linear.

3.3 Methods

3.3.1 Power comparison of maximally selected rank statistics between continuous predictors and discrete semi-continuous predictor

At the same cutpoint and the same sample size the power curves of the maximally selected rank statistics from continuous predictor and that from semi-continuous predictor with different level of discreteness will be compared via Monte Carlo simulation. The estimated power of this maximally selected rank statistic will be plotted against different effect sizes, θ. The following model is used.

$$Y = \mathrm{I}_{X>\mu}\theta + Z,$$

where $\begin{cases} I_{X>\mu} = 1 \ \ if \ \ X > \mu \\ I_{X>\mu} = 0 \ \ if \ \ X \le \mu \end{cases}$ and $Z \sim N(0,1)$

In order to compare the power curve from the continuous predictor and that from semi-continuous variable with 10 ordinal levels, the predictor variable X will be generated from uniform distribution of [0, 1]. For the semi-continuous predictor variable, instead of multiplying by 10 and rounded to the smallest integer as described in section 2.3 of chapter 2, the random number will be rounded to the smallest tenth decimal point in order to have the same cut point as the continuous predictor. The true cut point is set at 0.2 and 0.5. At each condition, Z is generated from standard normal distribution and Y is then computed for two alternatives. The effect size, θ, is set from 0 to 2.5 by every 0.1

increment. The sample size n will cover 20, 50, 100 and 200. The interval (ε_1, ε_2) of (.10, .90) will be used.

The maximally selected rank statistics will be developed for 10,000 Monte Carlo repetitions. The corresponding 95^{th} percentiles obtained under the null distribution in section 2.4 of chapter 2 will be used as the critical values. Out of these 10,000 Monte Carlo samples, the simulated power will be calculated based on the proportion of samples with the maximally selected rank statistics greater than or equal to their corresponding critical values. These simulated powers will be then compared between continuous and semi-continuous predictor variables via tables and plots.

3.3.2 Robustness of simple linear regression when model is mis-specified.

When the effect of predictor on the response is a step function, i.e. when there exists a cutpoint in the predictor variable, the power curve from the maximally selected rank statistics and that from the linear regression method will be compared via Monte Carlo simulation. The obtained simulated powers will then be plotted against different effect sizes. The same data with continuous predictor variable generated in section 3.3.1 of this chapter will be used. The simple linear regression model will be applied to these data. Again 10,000 Monte Carlo repetitions will be produced and the simulated power for simple linear regression is calculated as the proportion of the regression analyses with p-values < 0.05. These simulated powers will be compared between the maximally selected rank statistic method in section 3.3.1 of this chapter and the linear regression method for continuous predictor variable via tables and plots.

3.3.3 Apply the maximally selected rank statistics to linear association

The following model is used to simulate data with linear association between the predictor and response variable.

$Y = \beta X + Z$, where X~U[0,1] and $Z \sim N(0,1)$

In which X is continuous predictor generated from uniform distribution of [0, 1] and Z is the random error from standard normal distribution. If a step function is mis-specified to the above simulated data then $\beta=2$ corresponds to the effect size of 1 standard deviation between the two groups with the assumed cut point at 0.5 in X. Corresponds to the effect

size θ set from 0 to 2.5 by 0.1 increment in section 3.3.2 above, the regression slop will be set from 0 to 5 by 0.2 increment. The sample size n will cover 20, 50 and 100. Under each condition, 1,000 Monte Carlo repetitions will be produced. Both linear regression and maximally selected rank statistics with searching interval of central 90% will be conducted. The simulated power for simple linear regression is calculated as the proportion of the regression analyses with p-values < 0.05. The simulated power for maximally selected rank statistics will be calculated as the proportion with maximum rank statistics greater than the corresponding 95^{th} percentiles obtained under the null distribution in section 2.4 of chapter 2.

3.4 Results and Discussion

3.4.1 Power comparison of maximally selected rank statistics between continuous predictors and discrete semi-continuous predictor

Table 3.3.1 presents the simulated power of the maximally selected rank statistics for continuous predictor and semi-continuous predictor with 10 discrete levels at different sample sizes, different effect sizes and different locations of the true cut points. Since the predictor variables are generated from the uniform distribution of [0, 1], the true cut points are set at 0.2 and 0.5, which corresponds to the 20^{th} and 50^{th} percentiles of the simulated data. The searching interval in this chapter is set at (0.10, 0.90). When the sample size is 50, the critical values over the searching interval of $(\varepsilon_1, \varepsilon_2) = (.10, .90)$ and at α of 0.05 level are 2.59 and 2.78 for semi-continuous with 10 discrete scales and continuous predictors respectively (Table 2.2.1 in chapter 2).

Figure 3.3.1 presents the power curve of the maximally selected rank statistics for continuous predictor and semi-continuous predictor at sample size of 50 and 100. Both table 3.3.1 and figure 3.3.1 demonstrate that at the same effect size and with the same sample size the power of maximally selected rank statistic from the semi-continuous predictor is stochastically larger than that from the continuous predictor. One intuitive reason for this is that it has been demonstrated in chapter 2 the critical values used to reject the null hypotheses are always smaller in semi-continuous predictor than in the continuous predictor. In addition, Figure 3.3.1 and table 3.3.2 also show that the power increases along with the total sample size as well as the effect size.

When the effect size is 0, at which there is no association between the predictor and response variable, the power of detecting the association at the pre-specified 20^{th} or 50^{th} percentiles is very close to 0. Under the same situation if the location of the cut point is not pre-specified, instead the searching of the maximum of the rank statistics is performed over the central 80% of the data the power should be 5%.

Besides the sample size and effect size, the location of the true cut point also affects the power of this maximally selected rank statistics. When the effect size and the sample size are at the same, figure 3.3.2 presents the power of the maximally selected rank statistics when the true location of the cut point is from 10^{th} percentile to the 90^{th} percentile of the continuous predictor by every 10% increment. This figure shows that the power reaches to its highest when the true cut point is at the 50^{th} percentile of the predictor variable and the power decreases when the location of the true cut point moves to the lower or upper quartiles.

Table 3.3.1 Simulated power of the maximally selected rank statistics between continuous predictor and semi-continuous predictor with 10 discrete scales at different cut point locations, where the maximum is over the interval of (ε_1, ε_2) = (.10, .90).

N	Effect size (θ)	μ=0.2		μ=0.5	
		Continuous Predictor	S10 Predictor[1]	Continuous Predictor	S10 Predictor[1]
20	0	0.5	0.7	0.5	0.6
	0.5	1.7	2.7	4.8	8.1
	1.0	9.3	15.5	22.2	31.6
	1.5	27.9	38.2	57.9	70.3
	1.7	34.8	47.3	71.0	80.7
	2.0	48	57.5	86.6	92.3
	2.5	62.8	70.0	97.4	98.8
50	0	0.4	0.6	0.3	0.7
	0.5	6.6	9	12.2	18.5
	1.0	38.4	47.8	67.5	75.6
	1.1	46.0	57.0	77.4	83.8
	1.2	57.1	66.1	87.6	90.1
	1.4	73.1	81.5	96.9	96.8
100	0	0.5	0.6	0.4	1.1
	0.5	15.4	24.6	31.2	38.6
	0.7	40.8	48.6	66.2	75.7
	0.8	51.6	64.7	82.2	87.3
	0.9	66.1	78.4	90.4	95.6
	1.0	78.9	85.8	97.0	98.9
200	0	0.4	1.4	0.4	0.9
	0.5	42	54.6	66.6	79.1
	0.6	61.3	76.3	84.9	94.7
	0.7	80.9	88.6	97.0	99.0
	0.8	89.4	95.4	99.3	99.6
	0.9	97.0	98.6	99.8	100.0
	1.0	99.5	99.7	100.0	100.0

1: Semi-continuous predictor with 10 discrete levels.

Figure 3.3.1 Power comparisons for maximally selected rank statistics between continuous and semi-continuous predictor with 10 discrete scales at α = 0.05 and search interval of (ε_1, ε_2) = (.10, .90) when the location of the true cut point is at 50th percentile.

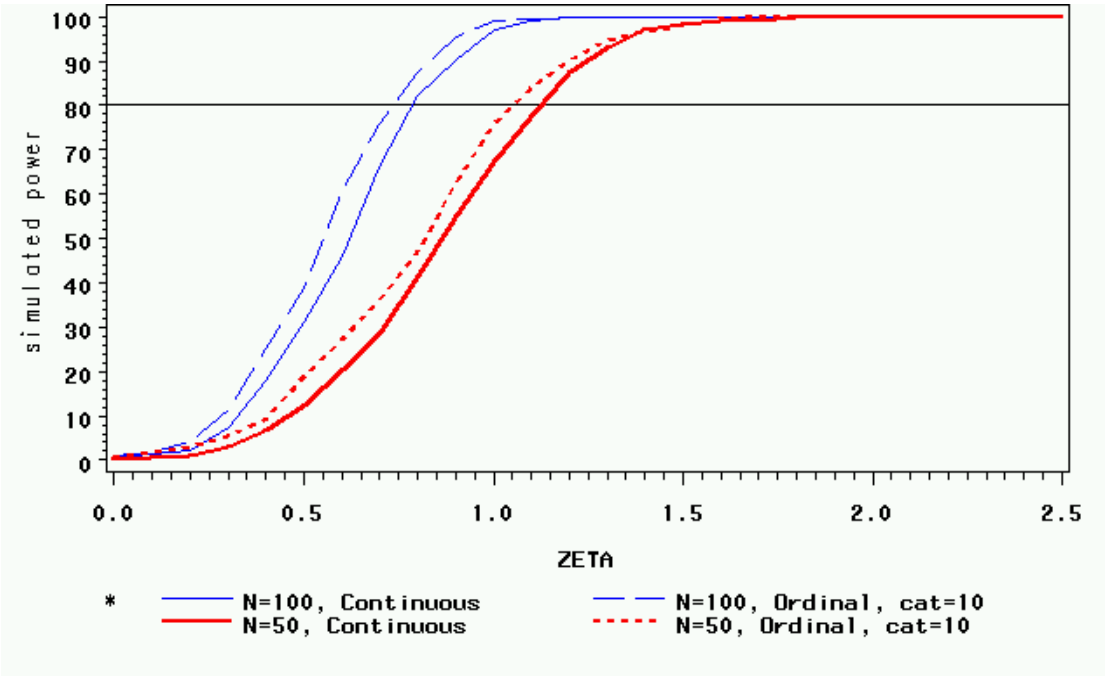
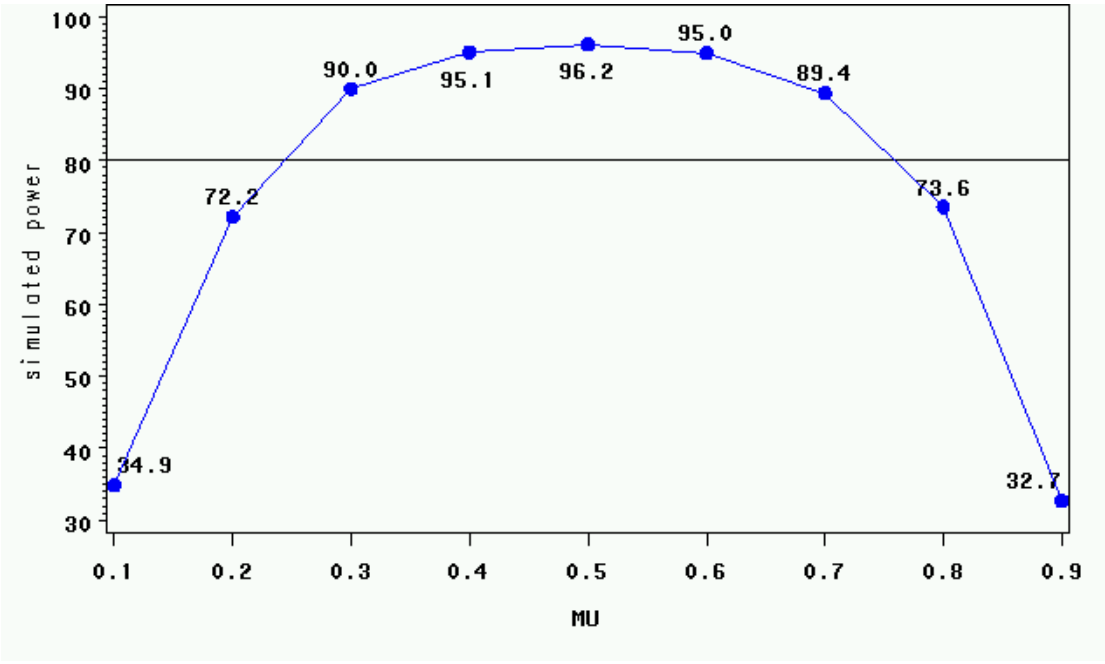


Figure 3.3.2 Power of the maximally selected rank statistics when the true cut-point is at different location. Using continuous X, n=50, (ε_1, ε_2)=(.10, .90), θ = 1.4 at α = 0.05.



3.4.2 Robustness of simple linear regression when model is mis-specified.

Table 3.3.2 presents the simulated power of the maximally selected rank statistics for continuous predictor and that of simple regression analysis at different sample sizes, different effect sizes and different locations of the true cut points.

Figure 3.3.3 shows the power curve comparison between the maximally selected rank statistics and simple linear regression when the sample size is set at 100 and the true cut point is set at the 20^{th} and 50^{th} percentiles. Both table 3.3.2 and figure 3.3.3 show that when the effect size is small, the simple linear regression has larger power to detect the association between predictor and response variables. When the effect size reaches to certain level, these two methods have equal power to detect the association. However since these two methods will result in two different interpretations for the underling association between predictor and outcome, appropriate diagnosis through scatter plot should be done in order to apply the correct analysis method.

Figure 3.3.3 Comparison of the simulated power curve of the maximally selected rank statistics for continuous predictor and that of regression analysis with sample size of 100.

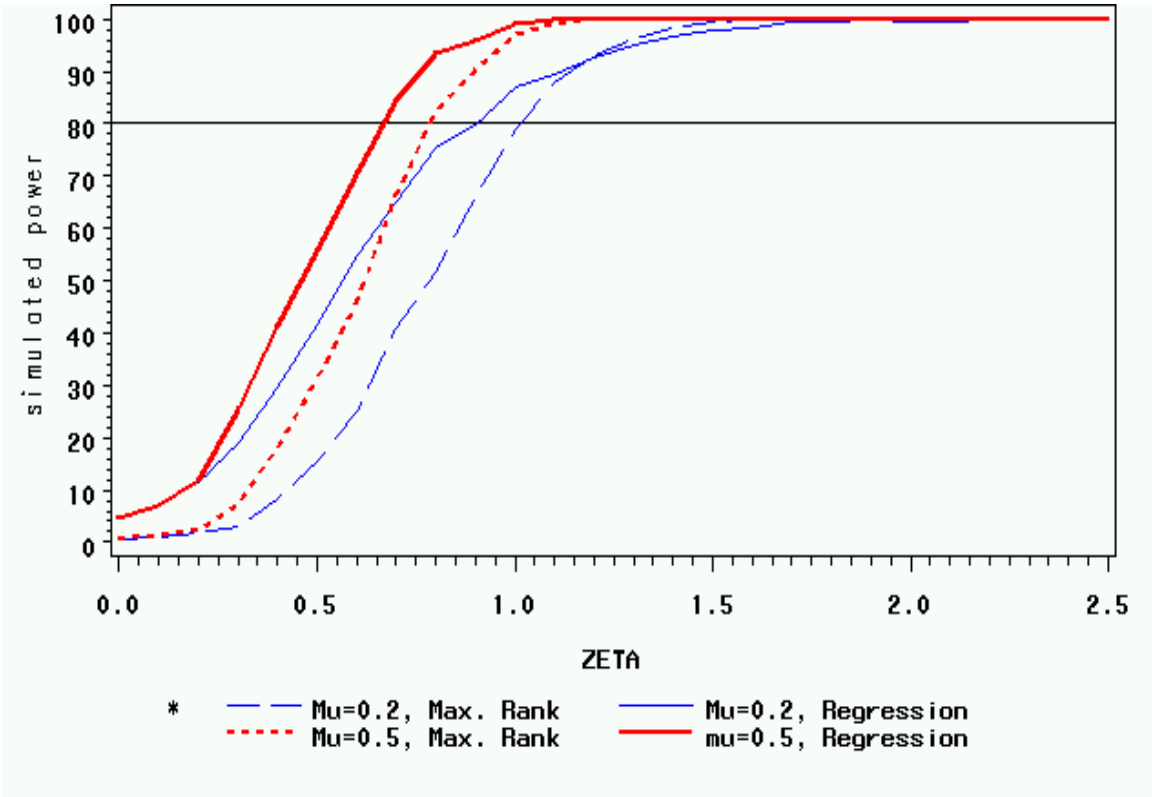


Table 3.3.2 Simulated power of linear regression and maximally selected rank statistics, when the effect of the continuous predictor on the response is a step function.

N	Effect size (θ)	μ=0.2		μ=0.5	
		Max. Rank [1]	Linear Reg. [2]	Max. Rank [1]	Linear Reg. [2]
20	0	0.5	2.3	0.5	2.1
	0.5	1.7	4.5	4.8	14.5
	1.0	9.3	19.4	22.2	43.1
	1.5	27.9	33.1	57.9	71.8
	1.7	34.8	40.5	71.0	81.5
	2.0	48	51.6	86.6	92.0
	2.5	62.8	62.7	97.4	97.4
50	0	0.4	5.1	0.3	4.1
	0.5	6.6	15.2	12.2	32
	1.0	38.4	47.5	67.5	82.8
	1.1	46.0	50.4	77.4	90.2
	1.2	57.1	55.3	87.6	94.1
	1.4	73.1	72.3	96.9	99.0
100	0	0.5	3.1	0.4	1.5
	0.5	15.4	41.7	31.2	55.8
	0.7	40.8	65.05	66.2	84.4
	0.8	51.6	75.25	82.2	93.3
	0.9	66.1	79.95	90.4	95.8
	1.0	78.9	87.0	97.0	99.1
200	0	0.4	1.9	0.4	2.8
	0.5	42	49.6	66.6	85.1
	0.6	61.3	63.4	84.9	94.7
	0.7	80.9	76.8	97.0	98.6
	0.8	89.4	84.7	99.3	99.8
	0.9	97.0	91.8	99.8	89.9
	1.0	99.5	96.0	100.0	100

1. Maximally selected rank statistics, where the maximum is over the interval of $(\varepsilon_1, \varepsilon_2) = (.10, .90)$.
2. Simple linear regression model.

3.4.3 Apply the maximally selected rank statistics to linear association

Table 3.3.3 presents the simulated power of the maximally selected rank statistics for continuous predictor and that of simple regression analysis at different sample sizes, different effect sizes given the underlying relationship between the predictor and response is linear.

Figure 3.3.4 shows the power curve comparison between the maximally selected rank statistics and simple linear regression for sample size of 20 and 50 given the underlying relationship between the predictor and response is linear.

Both table 3.3.3 and figure 3.3.4 show that the regression method has larger power to detect the association compare to maximally selected rank statistics. When the effect size is large enough, such as the slop equals 4 and 2.8 for sample size of 20 and 50 respectively, these two methods have equal power to detect the association. However if the underlying association between the predictor and outcome is linear, the cut point searching model should not be used because in such monotonic association no cutpoint dividing population into high and low outcome risk groups is apparent.

Based on the results presented in sections 3.4.2 and 3.4.3, given a study data, different analysis model would result in different interpretation on the underlying association. In the absence of any a priori information regarding the prognostic relationship between a covariate and outcome, exploratory plots are necessary to reveal if there exists an obvious thresholds that suggest potential cutpoints, or provide a range of values in which the search for a cutpoint should be performed.

Figure 3.3.4 Comparison of the simulated power curve of the maximally selected rank statistics and that of simple regression analysis for linear association between continuous predictor and response.

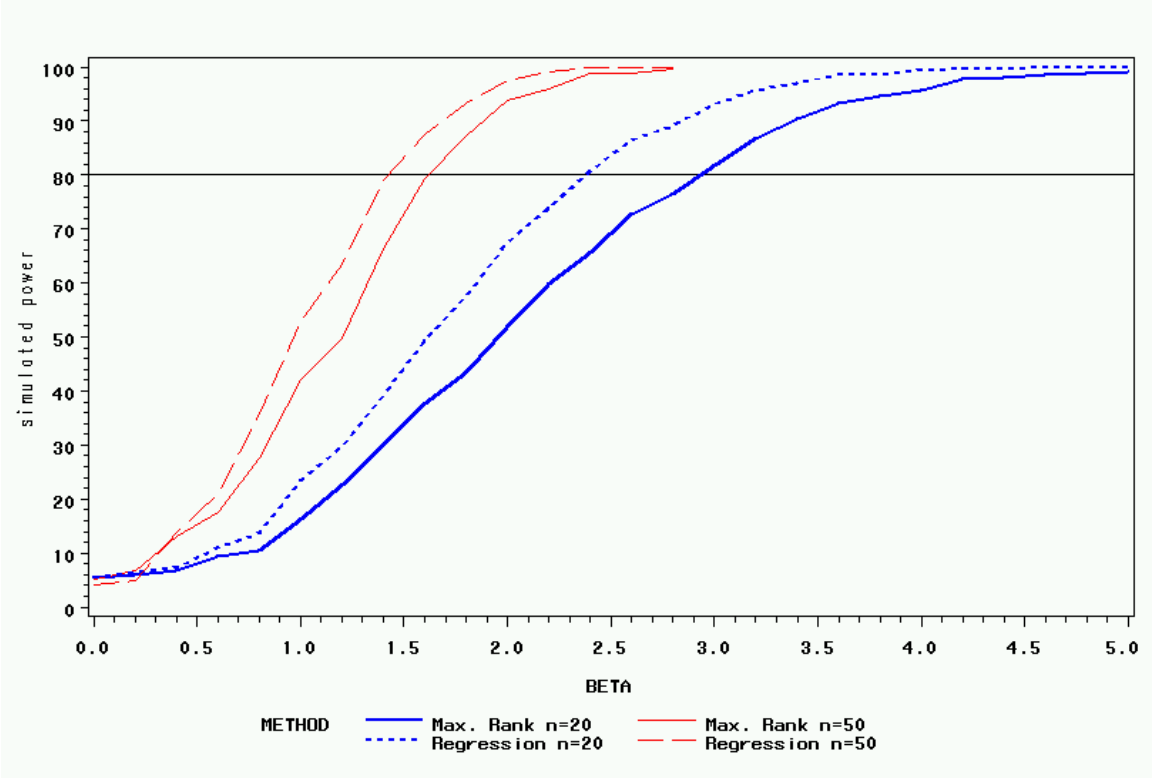


Table 3.3.3 Simulated power of linear regression model and maximally selected rank statistics, when the underlying relationship between the continuous predictor and response variable is linear.

N	Effect size (θ)	Regression slop (β)	Max. Rank [1]	Linear Reg. [2]
20	0	0	5.2	6.1
	0.5	1	15	22.6
	1	2	50.5	66.3
	1.2	2.4	64	80.2
	2	4	95.1	99.1
	2.5	5	99.1	100
50	0	0	5.3	4
	0.5	1	42.2	53
	0.8	1.6	79.2	87.5
	1	2	93.8	97.6
	1.4	2.8	99.7	100

CHAPTER 4: THE PRECISION OF CUT-POINT ESTIMATE OBTAINED USING THE MAXIMALLY SELECTED RANK STATISTIC

4.1 Abstract

In this chapter, the use of bootstrap method to create confidence interval for the estimated cut-point is explored. The results show that the location of the true cut-point in addition to the sample size and effect size can substantially affect the precision of the cut-point estimates. Compared to the continuous predictor, the semi-continuous predictor has higher percentage of correct cut-point estimates.

4.2 Introduction

When a significant effect is identified from a particular study using a cutpoint analysis, creating a confidence interval for the identified cutpoint would be of interest. A bootstrap simulation would provide a simple approach for creating such a confidence interval.

In order to explore the performance of these intervals, Monte Carlo simulation is used to create 95% confidence intervals for the cut-point obtained from a simulated sample using the maximally selected rank statistics. Different sample sizes and effect sizes as well as different locations for the true cutpoint are considerred.

The Monte Carlo simulation is also used to determine the sample size and effect size necessary to obtain a precise estimate of the cut-point. The 95% confidence interval of the cut-point estimation for continuous predictor and semi-continuous predictor with 15 ordinal scales are studied.

4.3 Methods

For continuous predictor, the uniform distribution of [0, 1] is used to generate the random numbers. For semi-continuous predictor with 15 ordinal scales, the random number generated from the uniform distribution of [0, 1] will be multiplied by 15 and then rounded to the smallest integers that are greater than or equal to the random numbers. The same model presented in section 3.3.1 of chapter 3 will be used to generate the response random variable of Y, which follows the standard normal distribution.

$$Y = \mathrm{I}_{X>\mu}\theta + Z,$$

where $\begin{cases} I_{X>\mu} = 1 \ \ if \ \ X > \mu \\ I_{X>\mu} = 0 \ \ if \ \ X \le \mu \end{cases}$ and $Z \sim N(0,1)$

The effect size, θ, is set at 1, 2 and 5. The true cut points, μ, are set at 10^{th}, 25^{th} and 50^{th} percentiles of the predictor. The sample size n will equal either 50 or 100. The searching interval (ε_1, ε_2) of (.10, .90) is used. In order to compare the precision of cut point estimation, under the same sample size, the data set contains the same random numbers of X and Z are used to generate the simulation data sets with different effect sizes and with cut points at different locations. For the simulated sample generated under each circumstance, 10,000 bootstrap samples were constructed and the cutpoints corresponding to the maximally selected rank statistics will be estimated for each of these 10,000 bootstrap samples.

The 95% confidence intervals of these estimated cut points will then be derived from the percentiles of the simulated cumulative frequencies. The true cut points will also be used to compare with the estimated cut points. Because there is possibility that the true cut point of μ is not selected into the bootstrap samples, in such case the largest X less than μ will be identified as the true cut point. The percentage of correct estimations will be calculated and discussed.

4.4 Results and Discussion

In this chapter, the confidence interval of the estimated cutpoint is explored using bootstrap simulation approach. We note that the null hypothesis should be test first before constructing the confidence interval. The confidence interval should not be constructed unless the null hypothesis is rejected, i.e. we declare there is a cutpoint that significantly affects the response.

Table 4.4.1 presents the results of confidence interval estimation for the cut point for both continuous predictor and semi-continuous predictor with different sample sizes, effect sizes and locations of the true cut points based on the bootstrap simulation.

Figure 4.4.1a and 4.4.1b provide the box plots to cover the 95% confidence intervals of the cut point estimates at sample size of 100 for continuous and semi-continuous predictors respectively. In these plots the true cut point locations are set at 10^{th} 25^{th} and 50^{th} percentile and the effect sizes are set at 1, 2 and 5. Similarly Figure 4.4.2a and 4.4.2b plots the 95% confidence intervals at sample size of 50 for continuous and semi-continuous predictors respectively.

In the box plots, the lower and upper bounds of the boxes represent the lower and upper limits of the 95% confidence intervals. The whiskers represent the minimum and maximum of the cut point estimates. The dots represent the true cut points. And the lines within the boxes represent the median estimates.

As expected, for both continuous predictor and semi-continuous predictor, larger effect size or larger sample size results in a narrower confidence interval, which indicates the larger the effect size or the larger the sample size the more precise the cut point estimate.

From the simulated data, when using a continuous predictor and sample size of 100 as an example, the true cutpoints μ at 10^{th}, 25^{th} and 50^{th} percentile of the continuous predictor, are 0.097, 0.217 and 0.451 respectively. Since when performing the bootstrap simulation, these true cutpoints may have chances of not being selected into the bootstrap samples. In such cases, the true cut points from those particular bootstrap samples would be the largest values below the 10^{th}, 25^{th} or 50^{th} percentile in the original data, i.e. the true cut point would be the largest value below 0.097, 0.217 or 0.451 when the true cut point location is set at 10^{th}, 25^{th} or 50^{th} percentile.

The results also show that besides the effect size, the location of the true cut point affects the cut point estimation precision substantially. With the location of the true cut point moves towards to the 50% percentile, the range of the 95% confidence interval narrows. For example, consider the sample size of 50 and effect size at 5, indicating the difference between the two separated groups is 5 standard deviations of underling distribution. The 95% confidence interval is (0.084, 0.901) when the true cutpoint location is at the 10^{th} percentile whereas the confidence interval is (0.524, 0.569) when the true cutpoint location is at 50^{th} percentile. This low precision of the cut point estimates is caused not only by the imbalance of the sample size between the two groups defining by above and

below the cut point but also the searching interval (ε_1, ε_2) of (.10, .90), which may not cover the true cut point for all bootstrap samples. The result for semi-continues predictor with the true cutpoint located at 10^{th} percentile sees these same problems. These simulation results show the precision of the CI estimation for the cut point depends strongly on its location.

We noticed that for continuous predictor when the sample size is at 100 and the true cut point is at 50^{th} percentile, all cutpoints estimated from the bootstrap samples are below the true cut point of 0.451. This is due to the distribution of the simulated data plotted in figure 4.4.3. The data points that above and close to the 50^{th} percentile are with relatively high values. They will not be selected as the best separation data points in the bootstrap samples based on their rank statistics.

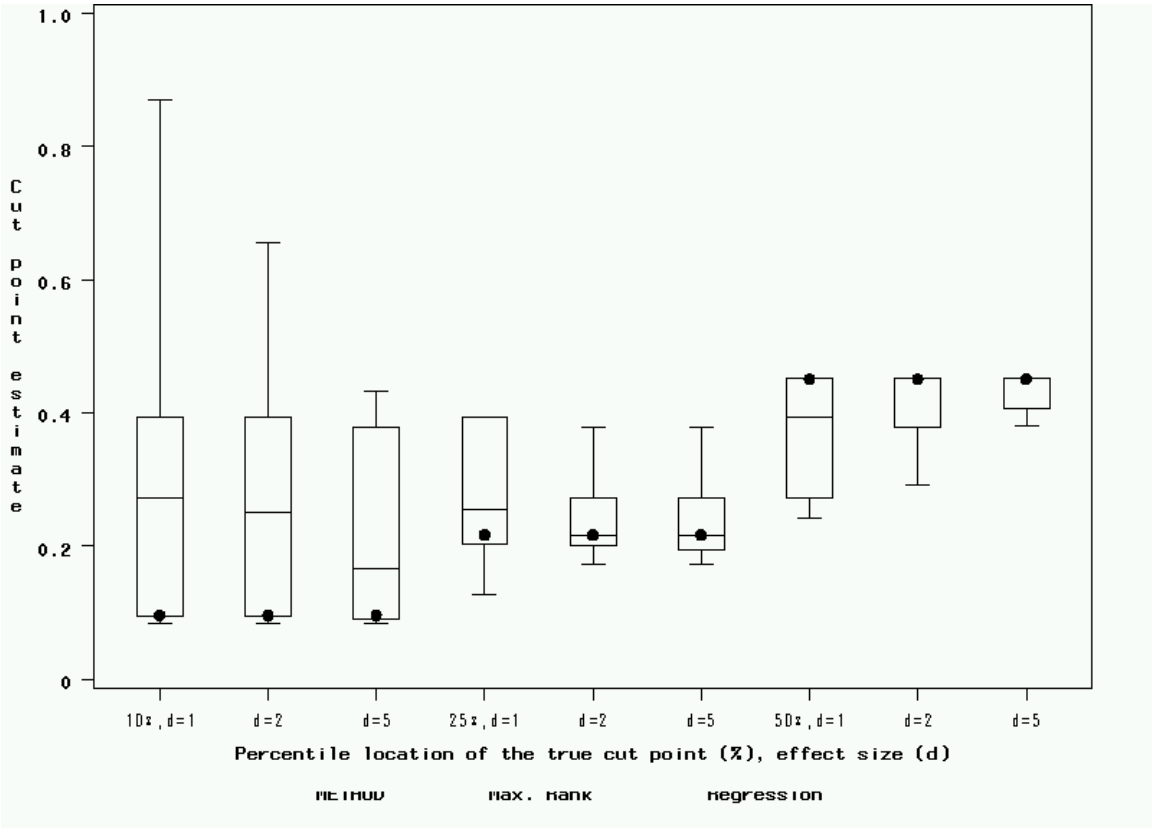
All of the results presented in this chapter are based on the simulation from one random sample generated for each of the particular circumstance depicted in section 4.3 of this chapter. Therefore the bootstrap results depend strongly on the studied samples. In order to study the precision of the cutpoint estimates, the ideal approach would be to generate 10,000 Monte Carlo random samples under each condition proposed in section 4.3, and 10,000 bootstrap simulations in order to calculate CIs for each one of the 10,000 Monte Carlo samples. Covreage average CI width etc could then be studied under the defined conditions. However to implement this approach is beyond the computation ability. The results stated in this chapter are merely an example to show how this approach would work.

A better approach to evaluate the precision of the cutpoint estimates either via simulation or via asymptotic methods would be the interest of the future works.

Table 4.4.1 95% CI of the estimated cut point from the bootstrap simulation

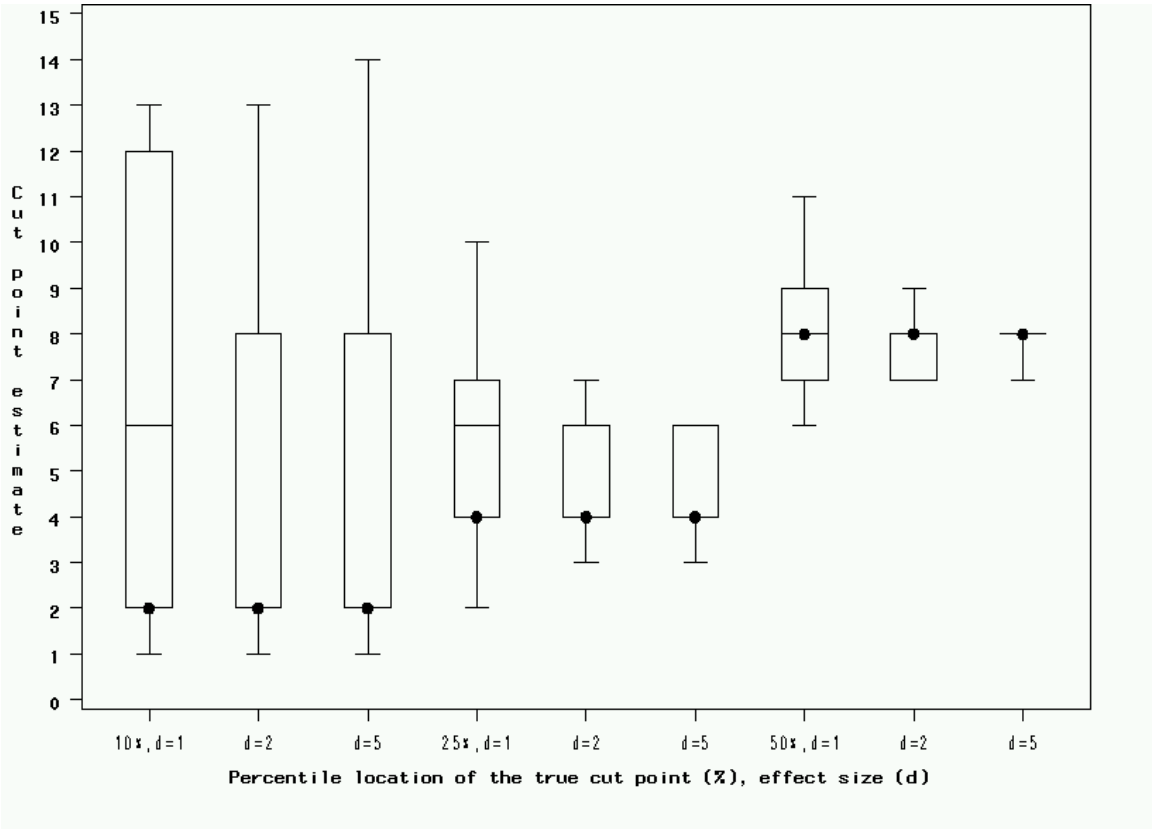
N	θ	Location of the true cut point (Percentile)	μ	95% CI for Continuous predictor	μ	95% CI for Semi-continuous predictor with 15 levels
100	1	10%	0.097	0.095, 0.394	2	2, 12
	2			0.095, 0.394		2, 8
	5			0.091, 0.378		2, 8
	1	25%	0.217	0.203, 0.394	4	4, 7
	2			0.200, 0.272		4, 6
	5			0.194, 0.272		4, 6
	1	50%	0.451	0.272, 0.451	8	7, 9
	2			0.378, 0.451		7, 8
	5			0.406, 0.451		8, 8
50	1	10%	0.1325	0.084, 0.901	2	2, 14
	2			0.084, 0.901		2, 14
	5			0.084, 0.901		2, 12
	1	25%	0.217	0.133, 0.594	4	2, 13
	2			0.154, 0.524		3, 7
	5			0.180, 0.405		4, 5
	1	50%	0.5694	0.405, 0.569	8	2, 14
	2			0.524, 0.569		7, 9
	5			0.524, 0.569		8, 8

Figure 4.4.1a 95% confidence intervals of the estimated cut points when the predictor is continuous and sample size is 100



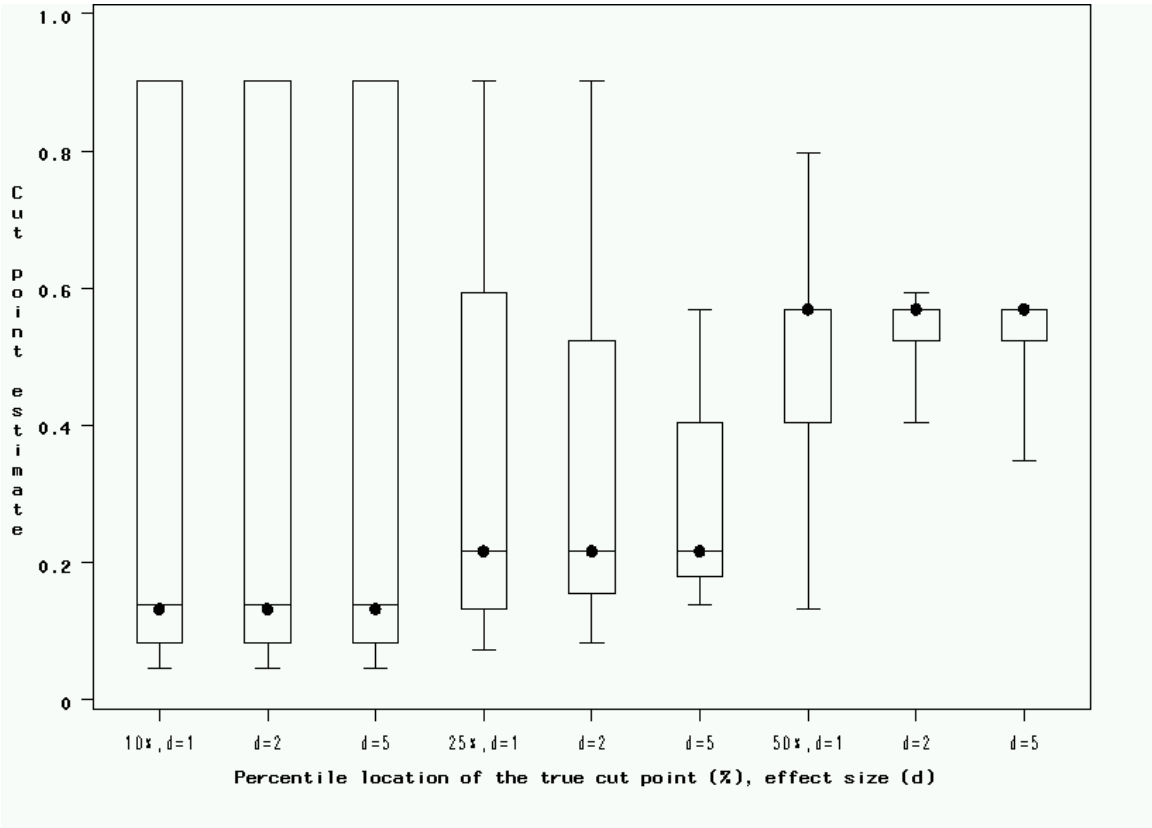
Note: The true cut point location is set at 10th, 25th and 50th percentile, and effect size is at 1, 2 and 5. The lower and upper limits of the 95% confidence intervals are plotted as the lower and upper bounds of the boxes. The whiskers represent the minimum and maximum of the cut point estimates from the bootstrap samples. The dots represent the true cut points and the lines within the boxes represent median cut point estimates.

Figure 4.4.1b 95% confidence intervals of the estimated cut points when the predictor is semi-continuous with 15 levels and sample size is 100.



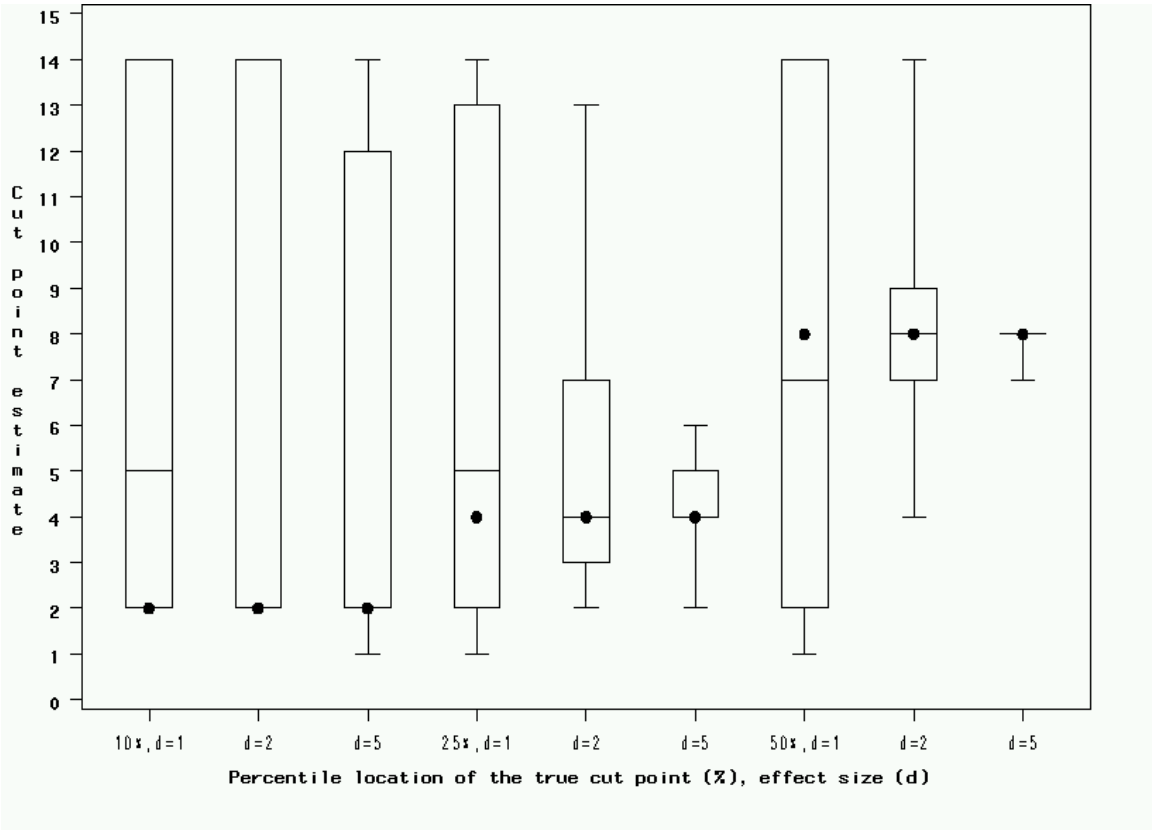
Note: The true cut point location is set at 10th, 25th and 50th percentile, and effect size is at 1, 2 and 5. The lower and upper limits of the 95% confidence intervals are plotted as the lower and upper bounds of the boxes. The whiskers represent the minimum and maximum of the cut point estimates from the bootstrap samples. The dots represent the true cut points and the lines within the boxes represent median cut point estimates.

Figure 4.4.2a 95% confidence intervals of the estimated cut points when the predictor is continuous and sample size is 50



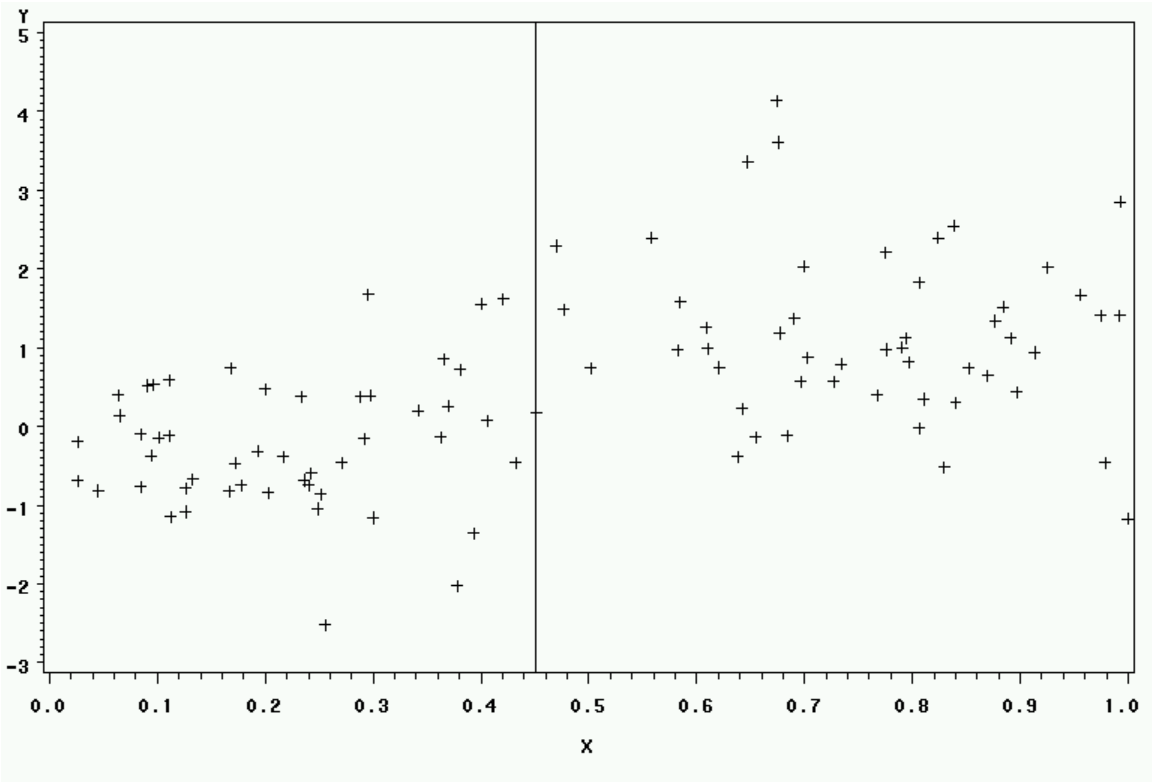
Note: The true cut point location is set at 10th, 25th and 50th percentile, and effect size is at 1, 2 and 5. The lower and upper limits of the 95% confidence intervals are plotted as the lower and upper bounds of the boxes. The whiskers represent the minimum and maximum of the cut point estimates from the bootstrap samples. The dots represent the true cut points and the lines within the boxes represent median cut point estimates.

Figure 4.4.2b 95% confidence intervals of the estimated cut points when the predictor is semi-continuous with 15 levels and sample size is 50.



Note: The true cut point location is set at 10th, 25th and 50th percentile, and effect size is at 1, 2 and 5. The lower and upper limits of the 95% confidence intervals are plotted as the lower and upper bounds of the boxes. The whiskers represent the minimum and maximum of the cut point estimates from the bootstrap samples. The dots represent the true cut points and the lines within the boxes represent median cut point estimates.

Figure 4.4.3 Simulated data* with continuous predictor for bootstrap sampling.



*: This data is simulated based on the model described in section 4.3 with sample size of 100, effect size of 1 and true cut point at 50^{th} percentile

CHAPTER 5: EFFECT OF USING RESIDUALS AFTER COVARIATE ADJUSTMENT

5.1 Abstract

In clinical or public health studies, multiple predictor variables are often collected. This chapter discusses the application of the maximally selected rank statistics when the response variable is not only associated with the predictor of interest which is used to find a cut point but also associated with other covariates. By using the regression residuals of the response variable from the other covariates, the violation in the independent observations assumption is a major concern. This chapter demonstrates the robustness of this independence assumption through Monte Carlo simulations.

5.2 Introduction

The method of maximally selected rank statistic that was discussed in the previous chapters is focused on the association between the response variable and only one predictor variable. In clinical or public health studies, the response variable is often not dependent on only one predictor variable. Usually multiple predictor variables are collected in these studies. When the response variable is associated with more than one predictor variable, instead of using the response variable itself the most intuitive method to find the cut point from the predictor of interest would be to use the regression residuals obtained from a multivariate regression model. However since the regression residuals are not independent samples in that they are correlated with each other, it violates the independent observation assumption of the maximally selected rank statistic. The robustness of this independence assumption will be examined through out this chapter.

5.3 Method

In order to compare the cut point estimation behaviour between the independent response variable and non-independent response variable, the following two types of data are simulated.

1. One response variable and three predictor variables follow multivariate normal distribution by using the variance-covariance structure obtained from the lead exposure study data (chapter 6).

2. Use the same data sets generated in 1. Keep only the blood lead level variable as the predictor and generate the independent response variable from standard normal distribution

5.3.1 More than one predictor variable

The lead exposure data that is discussed in chapter 6 is used as the source to generate the simulated data. The covariates of age, solvent exposure index and blood lead level identified in chapter 6 are used. The change from baseline at minute -15 in ACTH level are used as the response variable in that at this time the stress test has not started yet, the difference between the ACTH measured at this time point and at minute 0 would not be very small and should not be associated with blood Pb level, so that the effect size that we will add in the steps below are not affected. The variance-covariance matrix as well as the means for each of these four variables are calculated and then used as the parameters to generate the multivariate normal distribution. The sample size is set at 100 for the simulated data. The blood lead levels are then assigned with the actual values from the lead exposure study data based upon their corresponding quartiles. For example, in the lead exposure study, 22.86% of subjects are with blood lead level of 2 so in the simulated the data subjects with blood lead levels below 23 percentile are assigned as a value of 2, so on and so forth. Based on this approach, 10,000 Monte Carlo samples are constructed. The true cut points, μ, are again set at 25th and 50th percentiles of the blood lead level variable in each simulated sample. The same model used in section 3.3.1 chapter 3 will be used to re-derive the response variable, which is to add the effect size θ of 1 and 2 standard deviations when the associated blood lead level is above the true cut point.

$$Y = \mathrm{I}_{X>\mu}\theta + Z,$$

where $\begin{cases} I_{X>\mu} = 1 \ \ if \ \ X > \mu \\ I_{X>\mu} = 0 \ \ if \ \ X \le \mu \end{cases}$ and $Z \sim N(0,1)$

For each Monte Carlo sample, the regression residuals are obtained from the regression model with the change from baseline in ACTH as response variable, age and solvent exposure index as covariates. The method of maximally selected rank statistic is then applied to these regression residuals for the cut point estimate in blood lead levels. The percentage of samples with the estimated cut points matched with true ones is determined. In addition obtained maximum of the rank statistics will be compare with the critical value of 2.66 from table 1.1 in chapter 1, in which the discrete level is 15 and sample size is 100. The percentage of samples that reject the null hypothesis at 0.05 significance level will be reported.

5.3.2 One predictor variable using lead exposure study data

The predictor variable of blood lead level in the 10,000 simulated data sets from the above section 5.3.1 is used in this section. The true cut points, μ, are again set at 25^{th} and 50^{th} percentiles of this blood lead level variable in each Monte Carlo sample. The response variable is first generated from the standard normal distribution and then added with the effect size of 1 and 2 standard deviations when the associated blood lead level is above the true cut point. Compare to section 5.3.1, the response variable in the data generated in this section depend only on the predictor variable of blood lead level. The cutpoints corresponding to the maximally selected rank statistics will be estimated for each of the 10,000 Monte Carlo samples. The percentage of the samples with the estimated cut points matched with true ones is determined. In addition obtained maximum of the rank statistics will be compare with the critical value of 2.66 from table 1.1 in chapter 1, in which the discrete level is 15 and sample size is 100. The percentage of samples that reject the null hypothesis at 0.05 significance level will be reported.

5.4 Results and Discussion

Table 5.5.1 shows the percentage of the cut points that are correctly estimated in each of the two situations described in section 5.3 above. When the true cut point is set at 25^{th} percentile and the effect size is set at 1 standard deviation, by using the regression residual as the response variable, 82.5% Monte Carlo samples have their cut point estimates matched with the true cut points. The numbers are close by comparing it to

89.6% which are obtained by using the simulated blood lead level as the only predictor. Table 5.5.1 also presents the percentage for other situations such as the effect size of 2 standard deviations and the true cut point locations set at 25^{th} and 50^{th} percentile etc. The numbers are all close to each other among the three simulated conditions which indicate that although the independent observations assumption is violated by using the regression residual as the response, it would not affect the accuracy of the cut point estimates using the maximally selected rank statistic method.

Out of each Monte Carlo sample, the Maximum of the rank statistic that corresponds to the estimated cutpoint are also used to compare with the critical value of 2.66 at α=0.05 level for semi-continuous predictor with 15 levels and (ε_1, ε_2)=(0.1, 0.9) from table 2.2.1 in chapter 2. The percentage of the simulated Monte Carlo samples that reject the null hypothesis at α=0.05 for both simulation conditions described in section 5.3 above are presented in table 5.5.1 as well. For theta at 2 standard deviation level almost all samples rejected the null hypothesis. For theta at 1 standard deviation level, the numbers are very close to each other and both above 95% among the two simulated conditions. This result is consistent with table 3.3.1 in chapter 3, in which the power is 98.9 with sample size of 100 and effect size of 1 standard deviation. These results from table 5.5.1 again indicate the robustness of the independence assumption.

Table 5.5.1 The effect of the independence assumption on the accuracy of cut point estimates using maximally selected rank statistics

θ	Quartile	Estimate = True cut (%)		Percentage of samples reject the H_0	
		Multivariate [1]	Indep. Blood_Pb Sim [2]	Multivariate [1]	Indep. Blood_Pb Sim [2]
1 STD	25^{th}	82.5	89.6	98.9	95.1
	50^{th}	98.9	99.8	99.6	99.7
2 STD	25^{th}	82.1	84.5	99.9	96.9
	50^{th}	92.7	99.5	100	100

Note: [1] Data obtained from section 3.1. [2] Data obtained from section 3.2.

CHAPTER 6: APPLICATION OF THE MAXIMALLY SELECTED RANK STATISTIC TO BLOOD LEAD EXPOSURE STUDY

6.1 Abstract

In this chapter the method of maximally selected rank statistics for semi-continuous predictor is applied to a lead exposure study conducted by University of Medicine and Dentistry of New Jersey - Robert Wood Johnson Medical School. A significant cut point of 2 μg/dL in the blood Pb level has been identified as relate to the marginal association with the increase of the ACTH at the time point subjects demonstrated most elevated ACTH levels during the reactivity phase (p-value = 0.0310). The significance level of this identified cut point increased after adjustment for confounder variables of age and solvent exposure index (p-value = 0.0093). Since this study contains both reactivity and recovery phases and the response variable of ACTH is measured at multiple time points, based on this nature of the study design the method of baseline adjusted AUC is used to represent the overall ACTH during the entire study. No significant linear association is found between the blood level and the baseline adjusted AUC in the multivariate regression analysis. And there is no significant cut point identified from the blood lead level in relation to the baseline adjusted AUC.

6.2 Study Background of lead exposure, HPA dysfunction, blood pressure on hypertension risk

6.2.1 Study Rationale

Lead (Pb) continues to pose health risks for those sectors of the population in which exposure is highest, particularly individuals of lower socioeconomic status and workers in the construction trades. In addition to notorious effects on cognitive function, chronic Pb exposure is associated with hypertension and cardiovascular disease. Lab experimental studies reveal that chronic Pb exposure permanently alters corticosterone levels and stress reactivity in rats. If similar alterations in the hypothalamic pituitary adrenal axis (HPA) are demonstrated among humans chronically exposed to Pb, then such alterations may contribute over time to disease vulnerability (Fiedler 2010).

6.2.2 Blood Lead Levels in United States

National Health and Nutrition Examination Surveys (NHANES) is the only survey providing national data on lead exposure. This NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States. It is part of the programs in Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation. The survey examines a nationally representative sample of about 5,000 persons each year. These persons are located in counties across the country, 15 of which are visited each year.

The blood lead level (BLL) varies by age, gender, racial/ethnic groups and income status. Overall, during the 1999—2002 survey period, the geometric mean of blood lead level in the studied population is 1.6 μg/dL (Table 6.2.1) (CDC, 2005). The survey from 1999 — 2002 indicated continuous decrease in BLLs for all age groups and racial/ethnic populations from the 1991 — 1994 survey.

The elevated BLLs were defined as BLLs $\geq$10 μg/dL for all ages. From 1999 — 2002 survey, the overall prevalence of elevated BLLs for the U.S. population was 0.7%, a decrease of 68% from 2.2% in the 1991--1994 survey. Eliminating blood lead levels (BLLs) $\geq$10 μg/dL in children is one of the national health objectives for 2010 (US Department of Health and Human Services, 2010).

Table 6.2.1 Geometric means (GMs) of blood lead levels (measured as µg/dL), by race/ethnicity, sex and age group – National Health and Nutrition Examination Survey (NHANES), United States, 1999—2002. (CDC 2005)

Sex/Age (yrs)	No. in sample	NHANES 1999–2002* GM (95% confidence interval) All racial/ethnic groups	White, non-Hispanic	Black, non-Hispanic	Mexican American
Both sexes					
≥1	16,825	1.6 (1.5–1.6)	1.5 (1.5–1.6)†	1.8 (1.7–1.9)¶	1.6 (1.6–1.7)
1–5	*1,610*	*1.9 (1.8–2.1)*	*1.8 (1.6–2.0)†*	*2.8 (2.5–3.1)§¶*	*1.9 (1.8–2.0)†*
6–19	*6,283*	*1.1 (1.1–1.2)*	*1.1 (1.0–1.1)†§*	*1.5 (1.4–1.6)§¶*	*1.3 (1.2–1.4)†¶*
20–59	*5,876*	*1.5 (1.5–1.6)*	*1.5 (1.4–1.5)†§*	*1.7 (1.6–1.8)¶*	*1.8 (1.6–1.9)¶*
≥60	*3,056*	*2.2 (2.1–2.3)*	*2.2 (2.1–2.3)†*	*2.7 (2.5–2.8)§¶*	*2.1 (1.9–2.3)†*
Male					
≥1	8,202	1.9 (1.8–2.0)	1.9 (1.8–1.9)†§	2.1 (2.0–2.3)¶	2.0 (1.9–2.2)¶
1–5	*846*	*1.9 (1.8–2.1)*	*1.8 (1.6–2.0)†*	*2.8 (2.5–3.1)§¶*	*2.0 (1.8–2.1)†*
6–19	*3,158*	*1.3 (1.3–1.4)*	*1.2 (1.1–1.3)†§*	*1.7 (1.5–1.8)¶*	*1.5 (1.4–1.6)¶*
20–59	*2,689*	*2.0 (1.9–2.0)*	*1.9 (1.8–2.0)§*	*2.1 (2.0–2.3)*	*2.3 (2.2–2.5)¶*
≥60	*1,509*	*2.7 (2.6–2.8)*	*2.6 (2.5–2.7)†*	*3.4 (3.1–3.6)§¶*	*2.6 (2.3–2.8)†*
Female					
≥1	8,623	1.3 (1.3–1.3)	1.3 (1.2–1.3)†	1.5 (1.4–1.6)§¶	1.3 (1.2–1.4)†
1–5	*764*	*1.9 (1.8–2.1)*	*1.8 (1.5–2.1)†*	*2.8 (2.5–3.2)§¶*	*1.8 (1.7–2.0)†*
6–19	*3,125*	*1.0 (0.9–1.0)*	*0.9 (0.8–1.0)†§*	*1.3 (1.2–1.5)§¶*	*1.1 (1.0–1.2)†¶*
20–59	*3,187*	*1.2 (1.2–1.2)*	*1.2 (1.1–1.2)†*	*1.4 (1.3–1.5)¶*	*1.3 (1.2–1.4)*
≥60	*1,547*	*1.9 (1.8–2.0)*	*1.9 (1.8–2.0)†*	*2.3 (2.1–2.4)§¶*	*1.8 (1.6–2.0)†*

*Differences in GMs between NHANES 1999–2002 and NHANES 1991–1994 are all significant at p<0.05.
†Significantly different from non-Hispanic blacks at p<0.05, with Bonferroni adjustment.
§Significantly different from Mexican Americans at p<0.05, with Bonferroni adjustment.
¶Significantly different from non-Hispanic whites at p<0.05, with Bonferroni adjustment.

6.2.3 Blood Lead Exposure and Adrenocortical Responses to Acute Stress study in the literature.

Gump et al (2008) conducted a study of low-level prenatal and postnatal blood lead exposure and adrenocortical response to acute stress in children, i.e. which is aim to study the HPA response to acute stress as a function of Pb exposure.

In this study, 169 children from an ongoing longitudinal study of the effects of environmental toxicants on development (Stewart et al. 2000) were enrolled. The child's response to an acute laboratory stressor was assessed within 2 weeks of attaining 9.5 years of age. To measure adrenocortical reactivity, the response variable of cortisol level was collected from the saliva specimens. The predictor variables include prenatal blood level and postnatal blood lead level measurements. The prenatal blood levels were measured from children's cord blood specimens at the time of delivery. And the postnatal blood lead data was collected from the children at an average (± SD) age of 2.62 ± 1.20 years through the children's pediatricians and county public health agencies.

The prenatal blood Pb levels are ranged from <1.0 to 4.4 μg/dL and the postnatal blood Pb levels are ranged from 1.5 to 13.10 μg/dL with only six children having blood lead concentrations >10 μg/dL.

This study divided subjects into 4 quartile groups by using their prenatal and postnatal Pb exposure data and performed an ANCOVA analysis with other confounder variables. The quartile ranges in prenatal blood Pb data are ≤1.0 μg/dL, 1.1 – 1.4 μg/dL, 1.5 – 2.1 μg/dL and 2.1 – 4.4 μg/dL and the quartile ranges in postnatal blood Pb data are 1.5 – 2.8 μg/dL, 2.9 – 4.1 μg/dL, 4.2 – 5.4 μg/dL and 5.5 – 13.1 μg/dL.

The study results show both prenatal and postnatal lead exposure had significant positive effect on cortisol responses to acute stress at 9.5 years of age with increasing Pb level associated with an increasing cortisol response. The p-values are <0.001 and <0.005 for prenatal and postnatal lead exposure respectively. Especially according to the figure 2 in this article, the first quartile groups in both prenatal and postnatal lead exposure groups show substantial lower salivary cortisol change from baseline to the end of the stress test. The effects that found in the study were significant in a population with low levels of Pb exposures in which the blood Pb levels are well below the 10 μg/dL defined by the CDC as elevated in young children (CDC, 1991).

6.2.4 Lead Exposure Study Design and Procedure

In this chapter the data collected from the lead exposure study will be analyzed to address if there exists a threshold in blood lead level in relate to the increment of stress, which is measured as ACTH through the stress test.

This lead exposure study was conducted by University of Medicine and Dentistry of New Jersey - Robert Wood Johnson Medical School under the principal investigator of Nancy Fiedler with the grant number of 5R21ES15135. The detailed study design and procedures can be found in the grant report (Fiedler 2010).

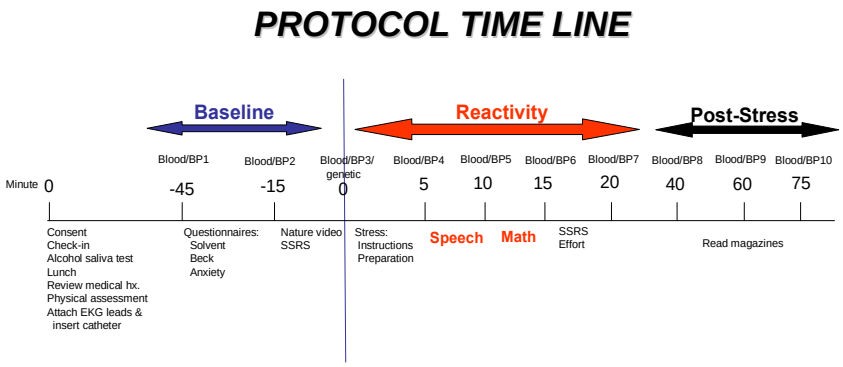
This study recruited voluntary subjects including painters, dry wall/tapers and carpenters who met the inclusion/exclusion criteria to complete the experimental stress procedure.

During the stress procedure (See Figure 6.2.1 for time line), all subjects were tested in the afternoon to maximize stress responsively and to control for the effects of circadian

rhythm. On the day of test, after signed the informed consent, all subjects received a check-in by research nurse in a separate private room to ascertain that the subject met the inclusion criteria. The EKG electrodes were attached by the nurse, and the subject was seated in a comfortable chair behind a table that contained a cassette recorder, blood pressure machine, and computer. The nurse placed an indwelling, catheter in the antecubital vein of the non-dominant arm from which blood Pb and blood samples for glucocorticoid and genetic polymorphisms were drawn. Immediately after placement of the catheter, blood was drawn for the standard chemistry panel and blood lead. Blood pressure was monitored with the BpTRU during the experimental session. Subjects were asked to sit quietly for 60 minutes to acclimate to the room and equipment. During the first 45 minute period, the nurse assisted the subject in completing the DS14, Spielberger Anxiety Scale, Beck Depression Inventory, and solvent questionnaire. After 45 minutes, blood pressure was taken and blood drawn. For the next 15 minutes, the subject viewed a nature video after which a blood sample was drawn and blood pressure taken. Instructions for the speech and math tasks were given and the subject was allowed a 10 minute preparation period. Then three person audience entered the room and was seated before the table with a video camera aimed directly at the subject for the purpose of recording the subject's responses. The Trier Social Stress Test (i.e., public speaking task and math tests) was administered without an intervening rest period between the speech and math tests to maximize the acute stress level. Blood pressure and a blood samples were taken immediately after completion of each stressor and the SSRS and scales assessing perception of the task (e.g., competence, difficulty) and effort put forth were completed. Blood pressure and blood samples were then collected at 15, 30, and 60 minutes post-stressor while the subject sat quietly reading pre-selected magazines provided by the research technician. After the one hour rest period, the catheter was removed, the EKG leads and blood pressure cuff detached and the subjects was given a light snack and underwent debriefing to clarify the purpose of the stressor and alleviate any remaining stress regarding their performance.

The blood lead was assessed by Quest laboratories at a part of routine chemistries. Plasma ACTH concentrations were measured using DSL Active Cortisol EIA kit (DSL-10-200).

Figure 6.2.1 Protocol Timeline



6.3 Application of the Maximally Selected Rank Statistics

6.3.1 Analysis objective and endpoints

The purpose of the lead exposure study is to test in adult workers the effect of chronic Pb on HPA axis function, including its ability to alter responsivity to stress changes. It is reasonable to assume there might exist a threshold in the blood Pb level which allows for a classification of population into a risk and a normal group with respect to stress response. The analysis objective in this chapter is to test if this assumption is true. The following endpoints will be examined.

1. To find the cut point in the blood Pb level in the marginal association to the change from baseline in ACTH at time point of 15 minutes.
2. To find the cut point in the blood Pb level in relate to the change from baseline in ACTH after adjusted by covariates of age and solvent exposure index or age alone at time point of 15 minutes.
3. To find the cut point in the blood Pb level in relate to the overall ACTH level during the entire study period via baseline adjusted AUC with and without adjusted by covariates.

The conclusion would be drawn based on the marginal association at time point of 15 minutes in that the stress level would reach to its highest at this time point based on the study design. The reason to choose age and solvent exposure index as the covariates is that based on the analyses results from the original grant report these two covariates were identified through the model selection.

6.3.2 Study Data Description on demographics and baseline characteristics

Seventy two (72) subjects who enrolled into this study have available data. The predictor of interest is blood lead level. The response variable is adrenocorticotropic hormone (ACTH), which is secreted from the pituitary in response to corticotropin-releasing hormone from the hypothalamus. Corticotropin-releasing hormone is secreted in response to many types of stress.

Out of these 72 enrolled subjects, 48 are painters and 24 are carpenters or drywall tapers. There 70 male subjects and 2 female subjects. The average age is at 46.3 years. The subject demographics are presented on table 6.3.1.

Table 6.3.1 Demographics

Demographic Characteristics	Category/Statistics	Total (N=72)
Occupation	Painter	48
	Carpenters/Dry wall taper	24
Gender	Female	2
	Male	70
Race	Black	6
	Hispanic	15
	Other	6
	White	44
Age (Years)	N	72
	Mean	46.3
	Std	6.99
	Median	45.5
	Min	30.0
	Max	60.0

In addition to demographics, other baseline covariates such as blood lead level, baseline ACTH and solvent exposure index etc. are also collected in this study. Their summary

statistics are presented on table 6.3.2. The mean blood lead level in this study is 6.1 μg/dL with the standard deviation of 5.12 and range from 2 to 31 μg/dL. According to the study grant report (Fiedler 2010), two confounder variables of age and solvent exposure index have been identified when to assess the lead exposure effect on ACTH level. The model building techniques used in this study report (Fiedler 2010) to assess whether the effects of lead levels were confounded by additional covariates are described in the following. At first the spearman correlation between all of the baseline characteristic variables and the predictor of interest, the blood lead level, were imputed. Those covariates that were significantly correlated with the blood lead level were identified. These were then added to the regression models looking at the effect of blood lead levels in order to assess whether the effects of lead changed either in magnitude or significance and whether the additional covariates added significantly to the prediction of the response. Table 6.3.2 also lists the spearmen correlation coefficients as well as the associated p-values between these baseline covariates and blood lead level.

Although the Negative Affect and the Social Inhibition correlated with blood lead level, these two measures were not included in the model in that these could be a health outcome of lead exposure and therefore, result in an over fitting of the model leading to increased chances of Type II error.

Since the predictor variable of blood Pb levels are measured at semi-continuous level, there are fourteen (14) discrete levels. Their frequencies are displayed on table 6.3.3.

6.3.3 Study Data Description on the response variable of ACTH

There are 3 pre-baseline ACTH measurements at minute -45, -15 and 0 before the Trier Social Stress Test was administered. The measurements taken at time 0 were treated as baseline. There are 7 post-baseline ACTH measurements, out of which 4 measurements were taken during the reactivity phase at minute 5 after the instruction, at minute 10 after the speech task, at minute 15 after the math test and at minute 20 after the perception of the task assessment on competence, difficulty etc., and 3 measurements were taken during the post-stress recovery phase at minute 40, 60 and 75.

Table 6.3.2 Baseline Characteristics

Baseline Characteristics	N	Mean	Std	Median	Min	Max	Corr [1]	p-value
Blood Lead (ug/dL)	72	6.1	5.12	4.0	2.0	31.0	1.00	-
Solvent Exposure Index	72	3.6	7.48	0.6	0.0	33.4	0.57	<.0001
Negative Affect	72	5.8	4.90	4.0	0.0	19.0	0.37	0.0015
Social Inhibition	72	8.6	6.04	7.5	0.0	22.0	0.33	0.0056
Age	72	46.3	6.99	45.5	30.0	60.0	-0.22	0.0619
Cortisol at Baseline	72	11.9	5.80	10.3	3.3	31.1	-0.22	0.0627
ACTH at Baseline	72	22.2	12.62	19.1	1.5	62.5	0.22	0.0658
State Depression	72	5.4	5.36	4.0	0.0	32.0	0.19	0.1221
Lifetime Nicotine (pack years)	72	11.7	15.96	2.7	0.0	77.5	0.17	0.1585
Years Worked	72	20.0	7.15	20.0	9.0	35.0	-0.15	0.2045
State Anxiety	72	30.8	9.32	28.0	20.0	51.0	-0.02	0.8989

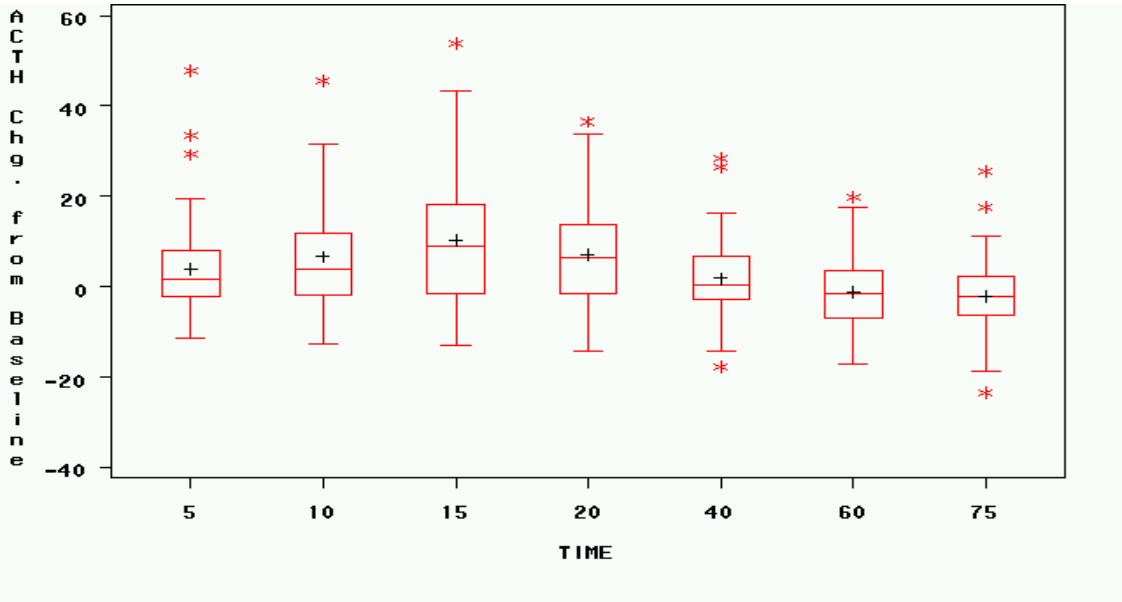
[1]: Spearman correlation

Table 6.3.3 Frequency of blood lead levels.

Blood Lead Level (μg/dL)	Frequency	Cumulative Frequency
2	16	16
3	11	27
4	9	36
5	5	41
6	9	50
7	3	53
8	1	54
9	3	57
12	3	60
13	2	62
14	4	66
15	1	67
17	2	69
31	1	70

Figure 6.3.2 show the box plots of change from baseline in ACTH measurements at each time point over all subjects. The mean ACTH increased from the baseline at each stressor section and reached peak at minute 15 during the reactivity phase. The mean ACTH then started to decrease during the post-stress phase and recovered to the baseline level at minute 60 and 75.

Figure 6.3.2 Box plots of ACTH change from baseline by time points



6.3.4 Analyses results from the original grant report

In the study grant report (Fielder 2010), the final multivariate regression model included blood lead level, baseline ACTH, age and solvent exposure index as the covariates and used the ACTH at minutes 5, 10, 15, 20, 40, 60 and 75 as response variables. The analyses results show ACTH, which is an indicator of stress, was significantly increased in response to and recovery from stressor among the groups more highly exposed to lead. Specifically, it revealed a significant effect of blood lead on ACTH when comparing subjects at the 25^{th} vs. the 75^{th} percentile after adjustment for baseline and covariates of age and solvent exposure index. The p-values are 0.014 and 0.0063 for unadjusted and adjusted models respectively. The detailed analyses results are presented on table 6.3.4.

Table 6.3.4 Blood lead effects unadjusted and adjusted for age and solvent exposure index. Both unadjusted and adjusted effects are adjusted for baseline value of outcome.

Outcome	Time (min)	Blood lead effect (95% CI) p-value	
		Unadjusted	Adjusted
ACTH	5	0.5 (-9.4, 11.6) .92	0.7 (-10.7, 13.6) .91
	10	12.4 (-0.8, 27.4) .067	14.2 (-1.2, 32.0) .072
	15	21.1 (4.1, 40.8) .014	27.6 (7.4, 51.5) .0063
	20	13.1 (-0.7, 28.9) .064	17.7 (1.2, 36.8) .035
	40	13.2 (1.2, 26.5) .030	16.5 (2.4, 32.6) .021
	60	15.6 (3.3, 29.4) .012	17.7 (3.2, 34.3) .016
	75	-3.7 (-14.5, 8.5) .53	-4.2 (-16.6, 10.1) .54

Note 1: Estimates (95% confidence intervals) represent the % increase/decrease in response at the specified time for those at the 75th (7 μg/dL) versus the 25th percentile (3 μg/dL) of blood lead levels. P-values are transcribed directly from the regression model results.

Note 2: This table is cited from the Fiedler's 2010 study grant report

6.3.5 Methods

6.3.5.1 Marginal association between blood Pb level and ACTH

Corresponds to the first research endpoint in section 6.3.1, which is to find the cut point in the blood Pb level in the marginal association to the change from baseline in ACTH at time point of 15 minutes. The maximally selected rank statistics for semi-continuous predictor described in chapter 1 will be applied to this lead exposure data.

Since the distribution of blood Pb level is right skewed, there are 16 out 70 subjects are at the lowest level of 2 μg/dL. When the searching interval is set at the central 80%, the rank statistic will be calculated in relate to the change from baseline in ACTH score at time point of 15 minute for every blood Pb level ranged from 2 to 13 (μg/dL). The maximum of these calculated rank statistics will be identified and compared with the critical value obtained from the null distribution with sample size of 72 and semi-continuous predictor with 14 discrete levels using the method described in chapter 1. The

associated p-value will be reported. The existence of the cut point in blood Pb level in relate to the stress response will be claimed based on this p-value.

6.3.5.2 Adjusted by covariates

In this study grant report, besides the blood lead level, which is the predictor of interest, age, solvent exposure index and baseline level of the ACTH were also included into the final multivariate regression model. The baseline level of the ACTH was included in order to account for the effect of regression to the mean. The age and solvent exposure index were included as confounders. This study grant report revealed a significant effect of blood lead when comparing subjects at the 25^{th} vs. the 75^{th} percentile after adjustment for baseline and covariates. Based on these results we understood age and solvent exposure index have an effect on the outcome variable of ACTH, and we also know these two variables are correlated with the blood Pb exposure (Table 6.3.2), we would like to study if the association between the ACTH and the cut point in blood Pb level, which is the predictor we are interested in, will change after adjusting for these two covariates of age and solvent exposure index or for age alone.

Again the change from baseline in ACTH at minute 15 will be used as the primary response variable. A multivariate regression model with the predictors of age and solvent exposure index or with the predictor of age alone will be conducted. The resulting residual in change from baseline ACTH will then be obtained. The maximally selected rank statistics will be applied to search for the cutpoint in blood lead level using these regression residuals at time point of 15 minute as response variables. Similar to section 6.3.5.1, the maximum rank statistic will be identified and compared with the critical values obtained from the null distribution with sample size of 72 and 14 discrete scales using the method described in chapter 1. The associated p-value will be reported. Any changes in the significance level of the identified cut point compare to the marginal association will be discussed.

6.3.5.3 Analysis on repeated measurements

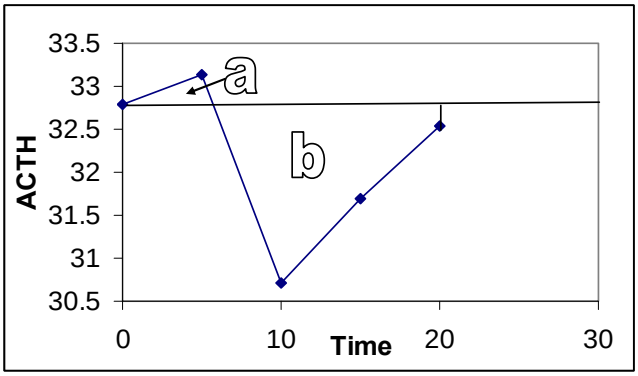
Based on the design, this study consists reactivity and recovery phases. And the response variable of ACTH was collected at multiple time points throughout these two different phases. This provides another research interest, which would be if there exists a cut point

in the blood Pb level to classify the population into two groups and the ACTH score would be significantly increased in response to and recovery from stressor among the groups more highly exposed to lead.

For the analysis on overall ACTH during the entire study duration, the baseline adjusted AUC (BAUC) will be calculated and used as the response variable. The baseline adjusted AUC will be calculated as the area above the baseline level under the curve of ACTH levels subtracted by the area over the curve of ACTH levels below the baseline level. Use figure 6.3.3 as an example, the baseline adjusted AUC is calculated as area a – area b. By using this baseline adjusted AUC, subjects who had higher ACTH score during the reactivity phase or had slower recovery would have higher BAUC value. Subjects who had low ACTH score in the reactivity phase or had rapid recovery after the stress test would have lower BAUC value.

The multivariate regression model with the predictors of baseline level ACTH, age, solvent exposure and blood lead level will be first conducted to explore if there is any significant relationship between the overall ACTH level and blood lead level. Then the same maximally selected rank analyses describe in section 6.3.5.1 and 6.3.5.2 above will be applied to identify if there is any cutpoint in blood lead level associated with this overall stress response as indicated by BAU with and without adjusting by covariates.

Figure 6.3.3 Baseline adjusted AUC.



6.4 Results

6.4.1 Cutpoint in marginal association

Figure 6.4.1a and 6.4.1b presents the scatter plots of ACTH and its change from baseline at time point of 15 minutes versus blood Pb level, which provide intuitive pictures on the marginal association between these two variables.

Since the blood Pb level is measured in semi-continuous level, the descriptive statistics of the change from baseline in ACTH sores at each of the blood Pb levels by every time point are presented in table 6.4.1 and figure 6.4.2.

The maximally selected rank statistics at time point of 15 minute is obtained at the blood lead level of 2 μg/dL with the value of 2.80. The associated p-value is 0.0310 according to the simulated null distribution with sample size of 100 and 15 discrete levels in the predictor.

From table 6.4.1, at minute 15 subject with blood Pb level of 2 μg/dL had small change in ACTH level from baseline with the mean of 0.73. However, the change in ACTH from baseline for subjects with blood Pb level of 3 μg/dL or higher is much bigger. The mean change values are 21.06 and 10.96 for subjects with blood Pb level of 3 μg/dL and its above respectively. Figure 6.4.2 also suggests that at minute 15 the ACTH increases along with the blood lead level. These descriptive summary results further confirmed the significant cut point of 2 μg/dL claimed from the maximally selected rank statistics analysis.

According figure 6.4.2, it seems the change from baseline of ACTH increases along with the blood lead level at minute 10, 15 and 20. The change from baseline of ACTH seems the same at all blood lead levels at all other time points especially at minute 5, 60 and 75. This can be explained by the study design. Minute 10 is the time point that subjects finished the public speech task and is in the middle of the Trier Social Stress Test (TSST). Minute 15 is the time point that subjects continued to finish the math test and is the time point subjects received the maximum stress in this study. Minute 20 is the time point subjects finished the scales assessing perception of the task after subjects finished

the TSST. We observed the trend of the stress response in relate to the blood Pb level at these 3 time points. And the trend reaches to the significant level at its peak time point.

Figure 6.4.1a ACTH at minute 15 vs. blood lead level.

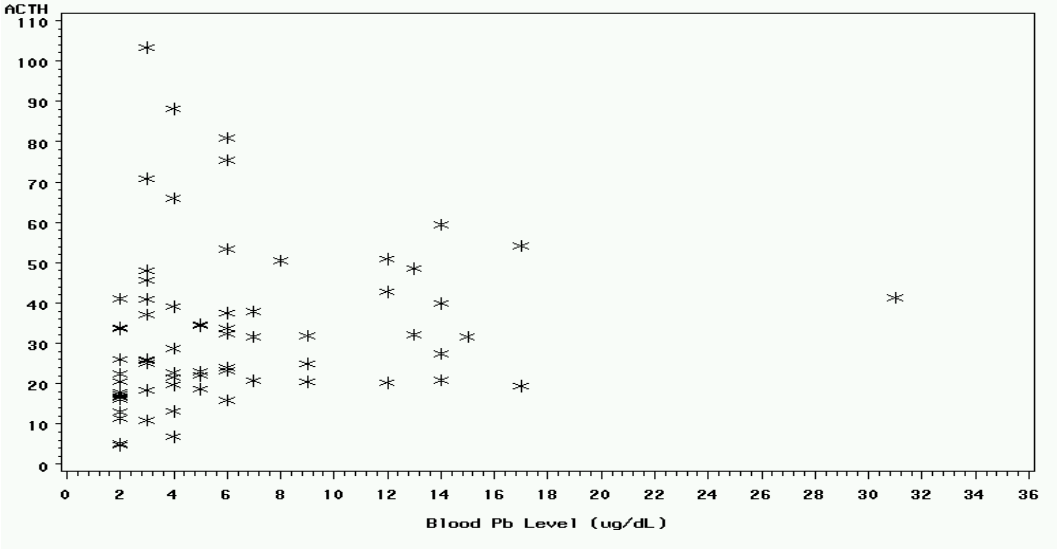


Figure 6.4.1b Change from baseline in ACTH at minute 15 vs. blood lead level.

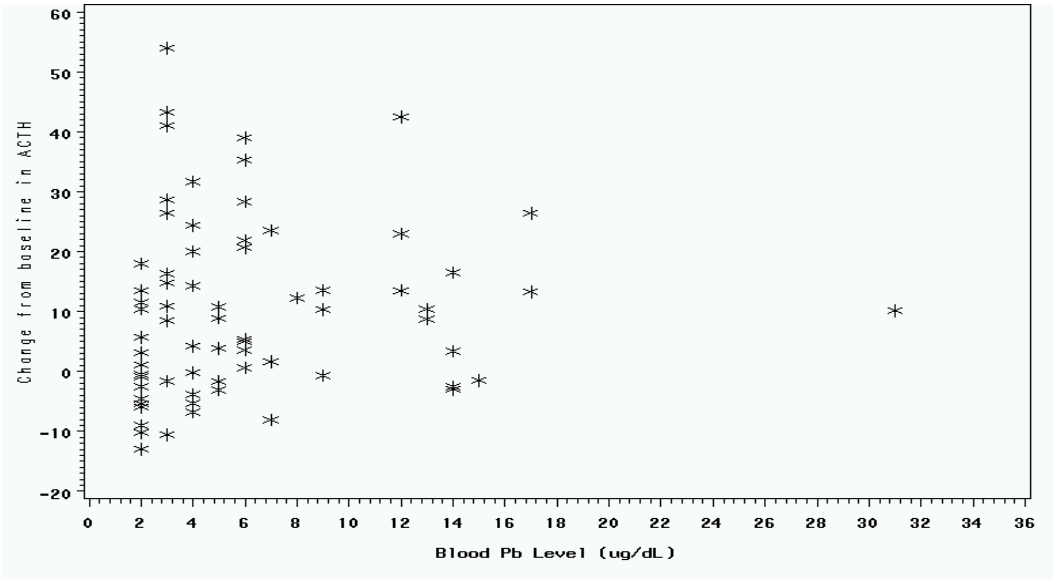


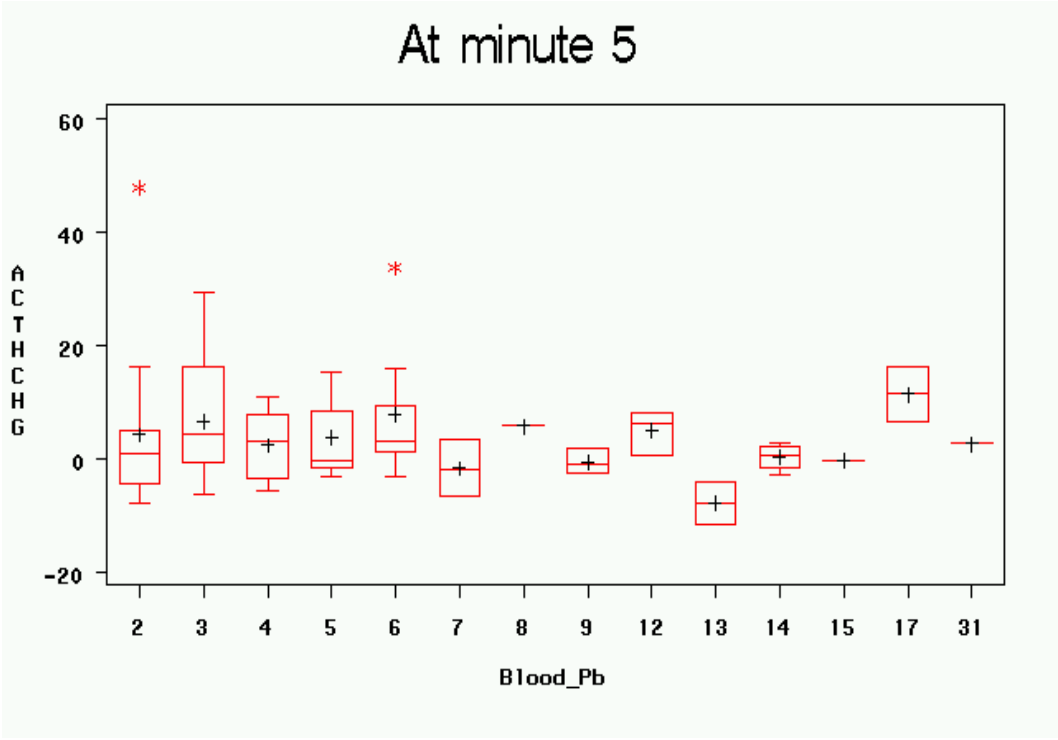
Table 6.4.1 Descriptive Statistics of ACTH change from baseline by time points and blood Pb levels.

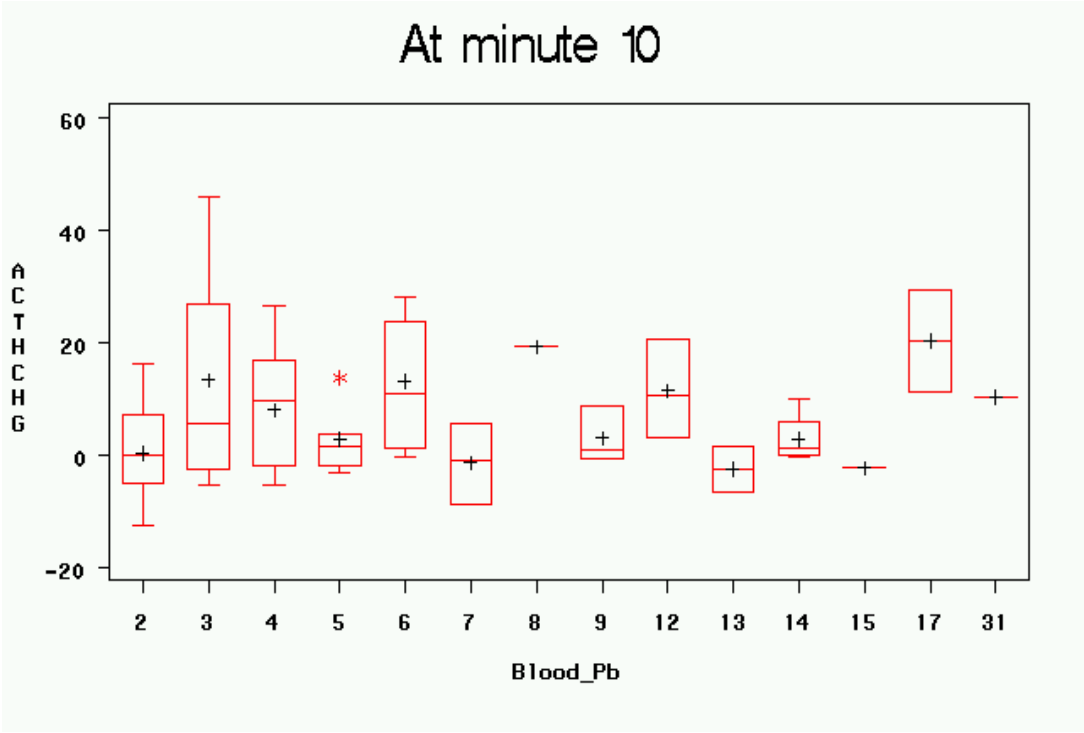
		ACTH Change from Baseline at						
Blood Pb (ug/dL)	**Statistics**	**Min. 5**	**Min. 10**	**Min. 15**	**Min. 20**	**Min. 40**	**Min. 60**	**Min. 75**
Overall	N	70	70	70	70	69	69	69
	MEAN	3.86	6.63	10.21	7.00	1.78	-1.18	-2.11
	STD	9.77	11.65	14.67	11.48	8.64	8.37	8.49
	MEDIAN	1.60	3.75	8.83	6.32	0.47	-1.49	-2.12
	MIN	-11.61	-12.66	-12.96	-14.24	-17.67	-17.34	-23.57
	MAX	47.81	45.74	53.94	36.57	28.44	19.82	25.51
2	N	16	16	16	16	16	16	16
	MEAN	4.17	0.26	0.73	1.36	-0.37	-2.91	0.92
	STD	13.59	8.37	9.04	7.92	11.26	7.74	10.52
	MEDIAN	0.80	-0.10	-0.62	-0.42	-1.66	-6.17	0.55
	MIN	-8.00	-12.66	-12.96	-9.87	-14.20	-15.87	-18.24
	MAX	47.81	16.25	18.01	15.20	28.44	9.13	25.51
3	N	11	11	11	11	10	10	10
	MEAN	6.49	13.25	21.06	14.66	5.09	0.00	-1.29
	STD	10.86	16.42	19.73	14.64	7.84	11.84	9.72
	MEDIAN	4.19	5.55	16.35	14.81	3.19	0.50	-1.15
	MIN	-6.34	-5.52	-10.57	-10.31	-6.43	-17.34	-23.57
	MAX	29.26	45.74	53.94	33.63	16.39	19.82	11.05
4	N	9	9	9	9	9	9	9
	MEAN	2.30	7.91	8.73	3.81	0.03	-1.79	-2.51
	STD	6.33	11.22	14.25	9.13	4.88	6.58	8.41
	MEDIAN	3.04	9.72	4.27	6.18	-0.38	-3.63	-0.90
	MIN	-5.70	-5.37	-6.80	-7.11	-7.20	-10.08	-18.71
	MAX	10.71	26.38	31.71	15.41	9.68	9.97	7.78
5	N	5	5	5	5	5	5	5
	MEAN	3.74	2.67	3.77	3.50	-1.12	-6.90	-5.47
	STD	7.87	6.80	6.18	5.27	4.49	4.66	5.30
	MEDIAN	-0.37	1.47	3.90	1.39	-0.83	-8.19	-6.23
	MIN	-3.12	-3.38	-3.07	-1.44	-7.13	-12.19	-11.13
	MAX	15.29	13.78	10.79	11.26	5.15	-0.08	1.41

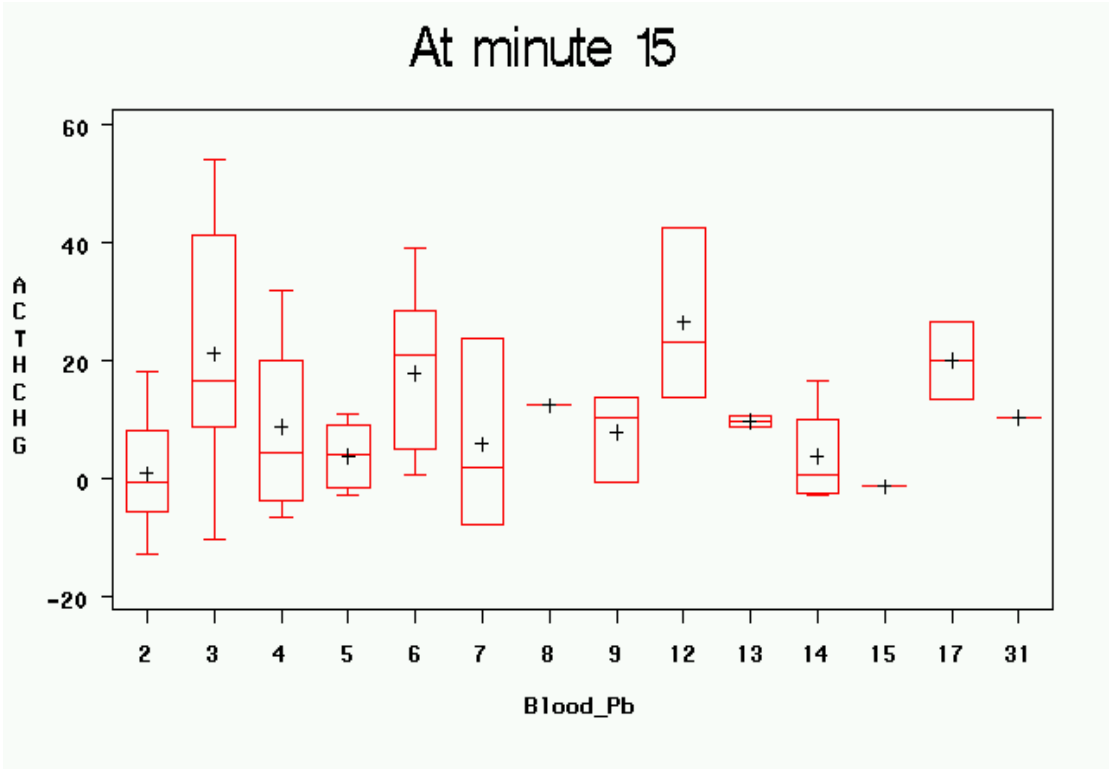
		ACTH Change from Baseline at						
Blood Pb (ug/dL)	**Statistics**	**Min. 5**	**Min. 10**	**Min. 15**	**Min. 20**	**Min. 40**	**Min. 60**	**Min. 75**
6	N	9	9	9	9	9	9	9
	MEAN	7.81	13.08	17.77	13.17	3.33	0.76	-4.63
	STD	11.28	11.81	14.60	10.31	4.90	8.31	8.20
	MEDIA	3.05	10.73	20.74	13.70	2.74	0.44	-4.38
	MIN	-3.39	-0.59	0.65	-4.32	-2.30	-12.74	-17.56
	MAX	33.60	28.06	38.99	27.89	11.81	17.48	10.95
7	N	3	3	3	3	3	3	3
	MEAN	-1.77	-1.48	5.73	3.12	-1.50	-4.65	-4.99
	STD	5.06	7.14	16.20	17.93	10.90	7.96	3.47
	MEDIA	-2.12	-1.08	1.68	-0.32	-0.16	-3.19	-3.99
	MIN	-6.64	-8.82	-8.06	-12.83	-13.00	-13.24	-8.85
	MAX	3.46	5.45	23.57	22.52	8.67	2.47	-2.12
8	N	1	1	1	1	1	1	1
	MEAN	5.69	19.20	12.32	9.11	-4.60	-7.38	-9.89
	STD	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MEDIA	5.69	19.20	12.32	9.11	-4.60	-7.38	-9.89
	MIN	5.69	19.20	12.32	9.11	-4.60	-7.38	-9.89
	MAX	5.69	19.20	12.32	9.11	-4.60	-7.38	-9.89
9	N	3	3	3	3	3	3	3
	MEAN	-0.78	2.93	7.75	3.72	1.75	2.64	1.55
	STD	2.23	4.87	7.49	9.10	6.26	6.41	2.24
	MEDIA	-1.19	0.91	10.32	4.26	5.05	6.24	0.73
	MIN	-2.78	-0.60	-0.68	-5.64	-5.47	-4.76	-0.15
	MAX	1.62	8.48	13.63	12.54	5.67	6.45	4.09
12	N	3	3	3	3	3	3	3
	MEAN	4.88	11.34	26.37	22.29	14.61	5.67	-0.21
	STD	4.04	8.74	14.79	13.66	10.66	9.54	4.54
	MEDIA	6.19	10.63	23.03	20.94	11.37	4.19	2.26
	MIN	0.35	2.98	13.54	9.36	5.94	-3.05	-5.45
	MAX	8.11	20.42	42.54	36.57	26.51	15.86	2.56

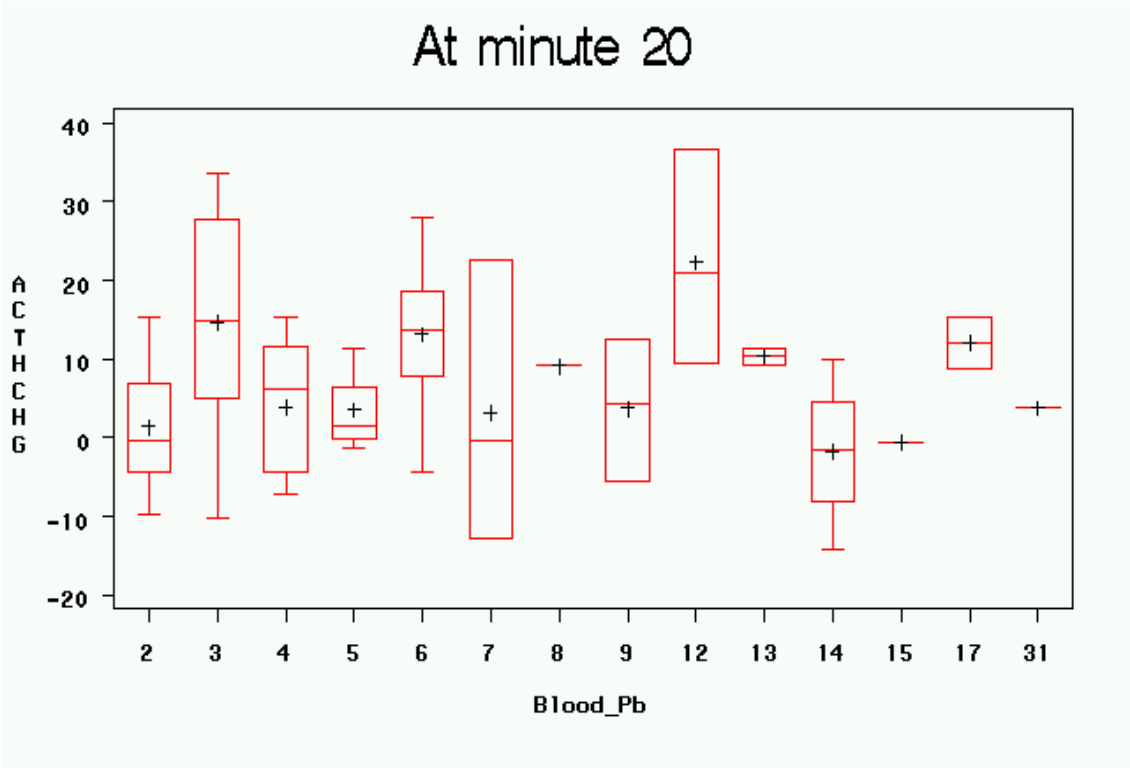
Blood Pb (ug/dL)	Statistics	ACTH Change from Baseline at Min. 5	Min. 10	Min. 15	Min. 20	Min. 40	Min. 60	Min. 75
12	N	3	3	3	3	3	3	3
	MEAN	4.88	11.34	26.37	22.29	14.61	5.67	-0.21
	STD	4.04	8.74	14.79	13.66	10.66	9.54	4.54
	MEDIAN	6.19	10.63	23.03	20.94	11.37	4.19	2.26
	MIN	0.35	2.98	13.54	9.36	5.94	-3.05	-5.45
	MAX	8.11	20.42	42.54	36.57	26.51	15.86	2.56
13	N	2	2	2	2	2	2	2
	MEAN	-7.89	-2.64	9.61	10.32	9.81	14.39	2.79
	STD	5.26	5.79	1.21	1.59	4.38	0.33	11.91
	MEDIAN	-7.89	-2.64	9.61	10.32	9.81	14.39	2.79
	MIN	-11.61	-6.73	8.75	9.20	6.71	14.16	-5.64
	MAX	-4.18	1.45	10.46	11.44	12.91	14.63	11.21
14	N	4	4	4	4	4	4	4
	MEAN	0.24	2.84	3.60	-1.84	-5.13	-5.46	-7.34
	STD	2.55	4.73	9.09	9.91	9.91	5.30	5.83
	MEDIAN	0.65	1.00	0.43	-1.56	-4.67	-4.83	-5.02
	MIN	-3.09	-0.41	-3.00	-14.24	-17.67	-12.43	-15.96
	MAX	2.75	9.76	16.53	10.00	6.51	0.22	-3.38
15	N	1	1	1	1	1	1	1
	MEAN	-0.34	-2.42	-1.44	-0.59	-1.32	1.90	10.13
	STD	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MEDIAN	-0.34	-2.42	-1.44	-0.59	-1.32	1.90	10.13
	MIN	-0.34	-2.42	-1.44	-0.59	-1.32	1.90	10.13
	MAX	-0.34	-2.42	-1.44	-0.59	-1.32	1.90	10.13
17	N	2	2	2	2	2	2	2
	MEAN	11.32	20.19	19.85	12.00	8.82	1.66	-7.91
	STD	6.91	12.67	9.25	4.71	3.48	1.09	8.41
	MEDIAN	11.32	20.19	19.85	12.00	8.82	1.66	-7.91
	MIN	6.44	11.23	13.31	8.67	6.36	0.89	-13.86
	MAX	16.21	29.15	26.39	15.33	11.28	2.43	-1.96
31	N	1	1	1	1	1	1	1
	MEAN	2.56	10.16	10.20	3.77	-2.54	-6.92	-5.41
	STD	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MEDIAN	2.56	10.16	10.20	3.77	-2.54	-6.92	-5.41
	MIN	2.56	10.16	10.20	3.77	-2.54	-6.92	-5.41
	MAX	2.56	10.16	10.20	3.77	-2.54	-6.92	-5.41

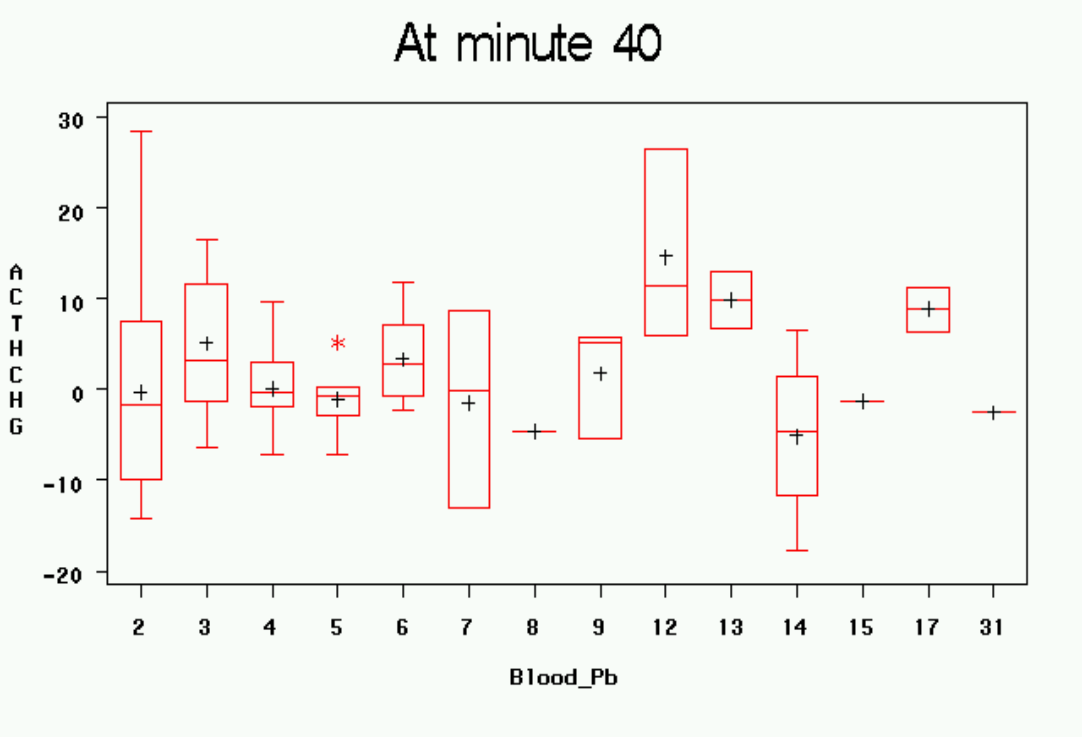
Figure 6.4.2 Change from baseline in ACTH at each blood Pb level by time points.

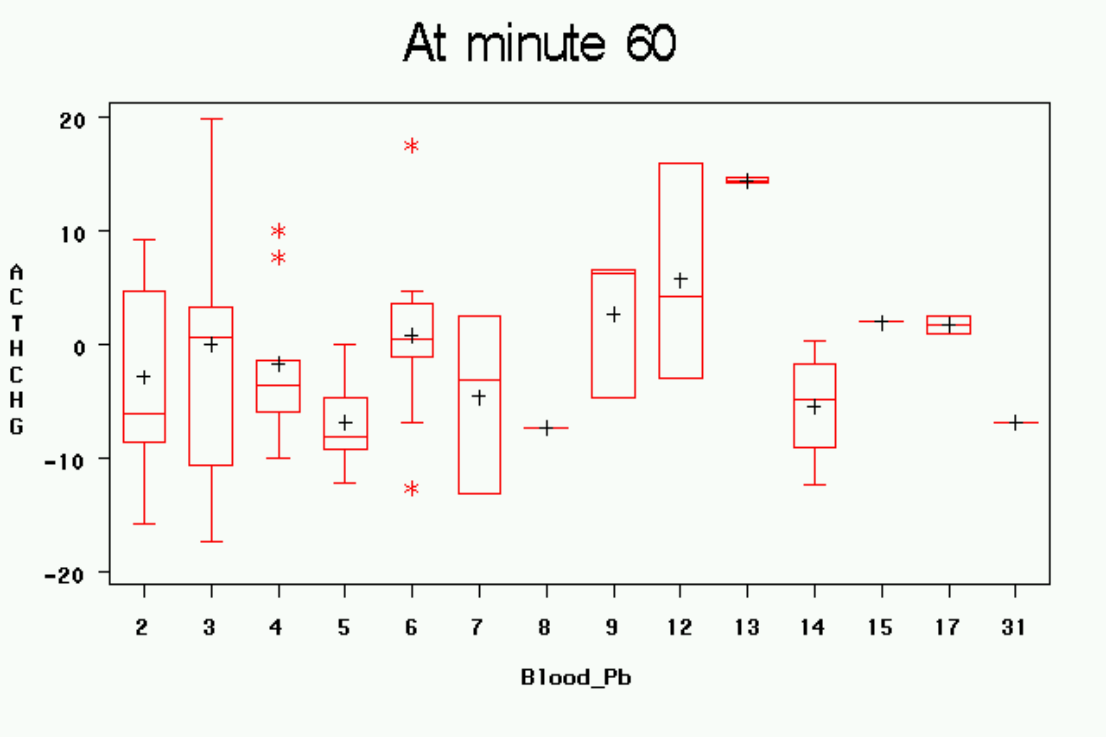


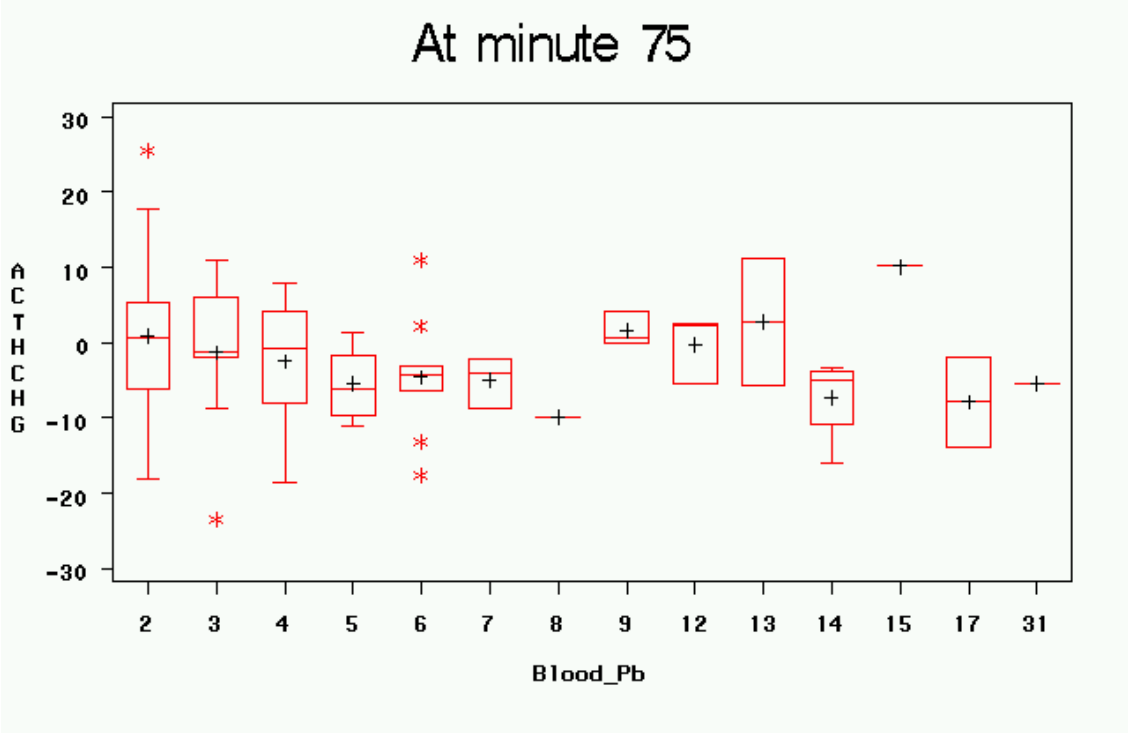



At minute 15
ACTHCHG
60
40
20
0
-20
2
3
4
5
6
7
8
9
12
13
14
15
17
31
Blood_Pb


At minute 20
ACTHCHG
40
30
20
10
0
-10
-20
2
3
4
5
6
7
8
9
12
13
14
15
17
31
Blood_Pb


At minute 40
ACTHCHG
30
20
10
0
-10
-20
2
3
4
5
6
7
8
9
12
13
14
15
17
31
Blood_Pb


At minute 60
ACTHCHG
20
10
0
-10
-20
2
3
4
5
6
7
8
9
12
13
14
15
17
31
Blood_Pb



6.4.2 Adjusted by covariates

After adjusted for the confounder variables of age and solvent exposure index, figure 6.4.3 plots the regression residuals of change from baseline in ACTH versus blood lead level at time point of 15 minute.

Table 6.4.2 shows the cut points and associated p-values in blood lead level in relation to age and solvent exposure index adjusted ACTH change from baseline. The cut point of 2 µg/dL in blood lead level is significantly identified at minute 15. Compared to the marginal association, the significance increased after adjusted by the covariates of age and solvent exposure with the maximum of rank statistic changed from 2.80 to 3.15 and p-value changed from 0.0346 to 0.0102 The detailed results are presented on table 6.4.2.

Figure 6.4.4 shows the box-plots of regression residuals of change from baseline in ACTH at each of the blood lead level by time point. The reference lines are drawn at the mean of the residuals at blood lead level of 2 µg/dL. This figure reveals that the ACTH level increased for subjects with blood level above 2 µg/dL at minute 15 and 20. For the rest of the time point, there seems to be no association between ACTH and blood lead

level. Table 6.4.2 also presents the results by adjusting for covariate of age only. The result show little change with the maximum of the rank statistic of 3.14 and p-value of 0.0106, which means age is the most confounded variable in the relationship between blood lead exposure and ACTH increase.

Figure 6.4.3 Plot of regression residual of ACTH change from baseline after adjusted by age and solvent exposure index at time point of 15 min vs. blood lead level.

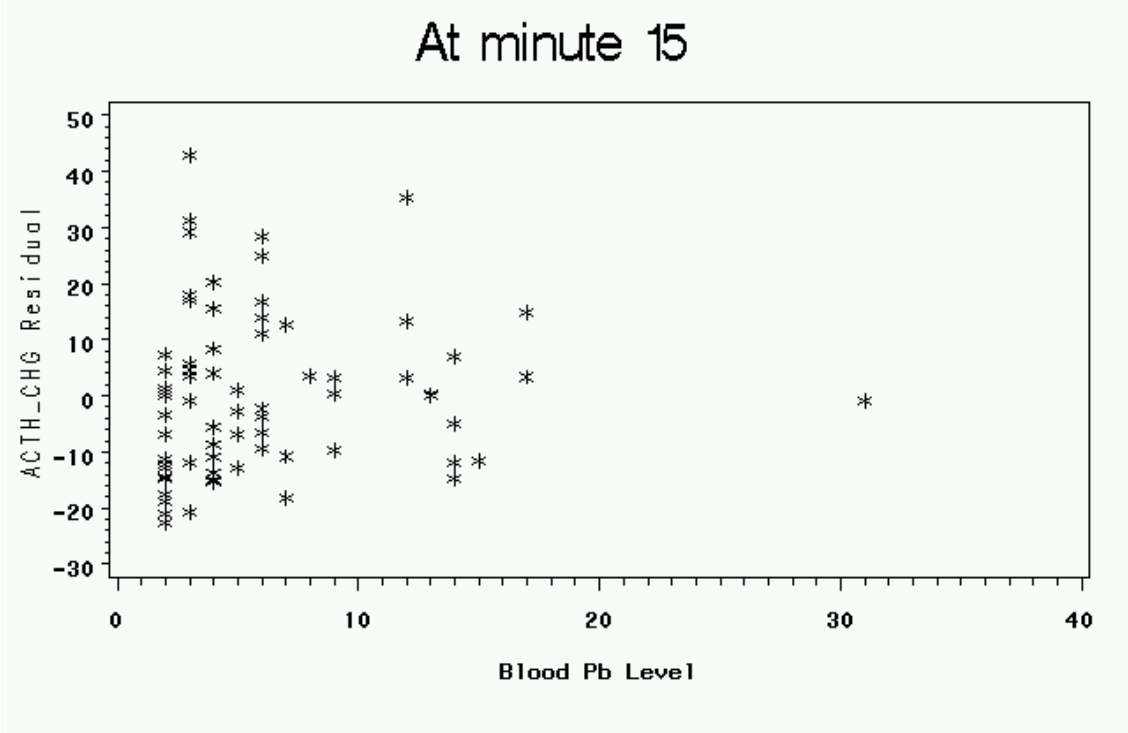
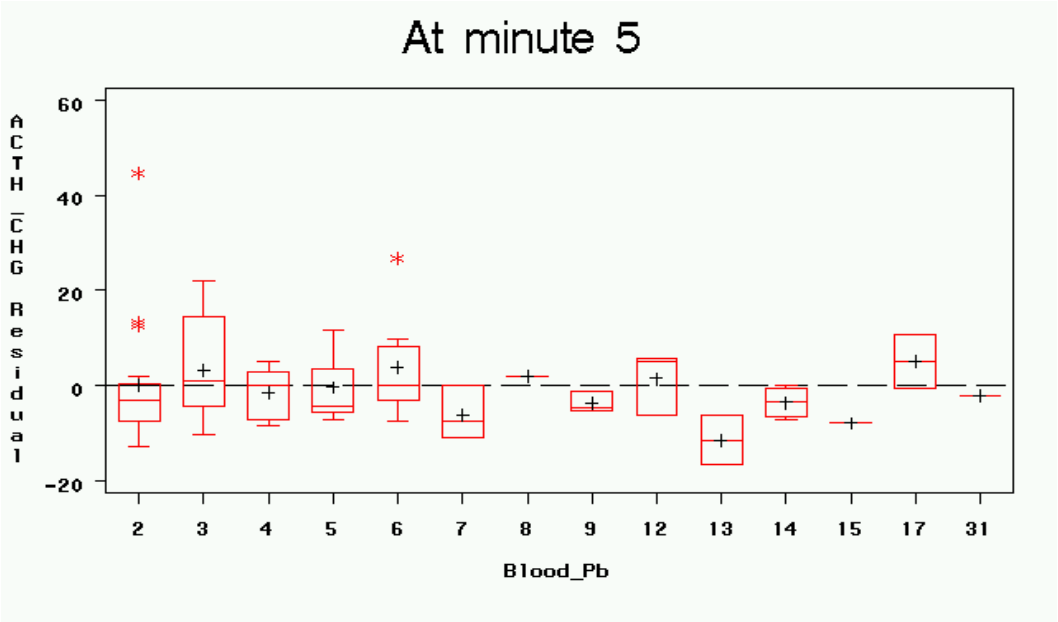
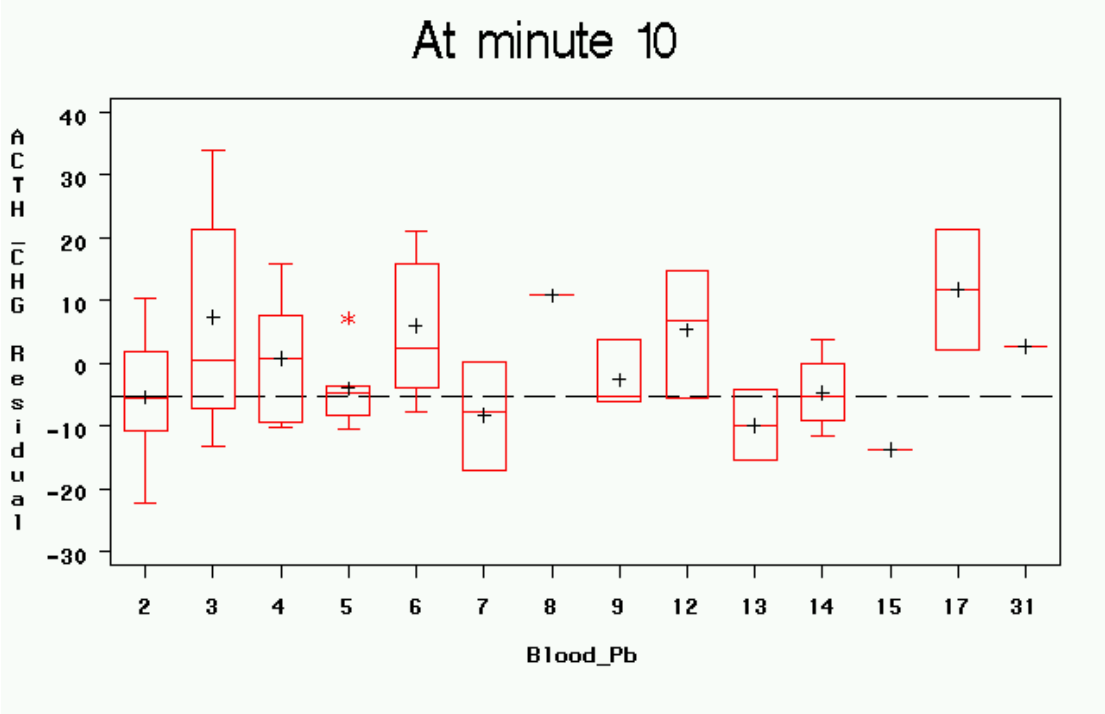
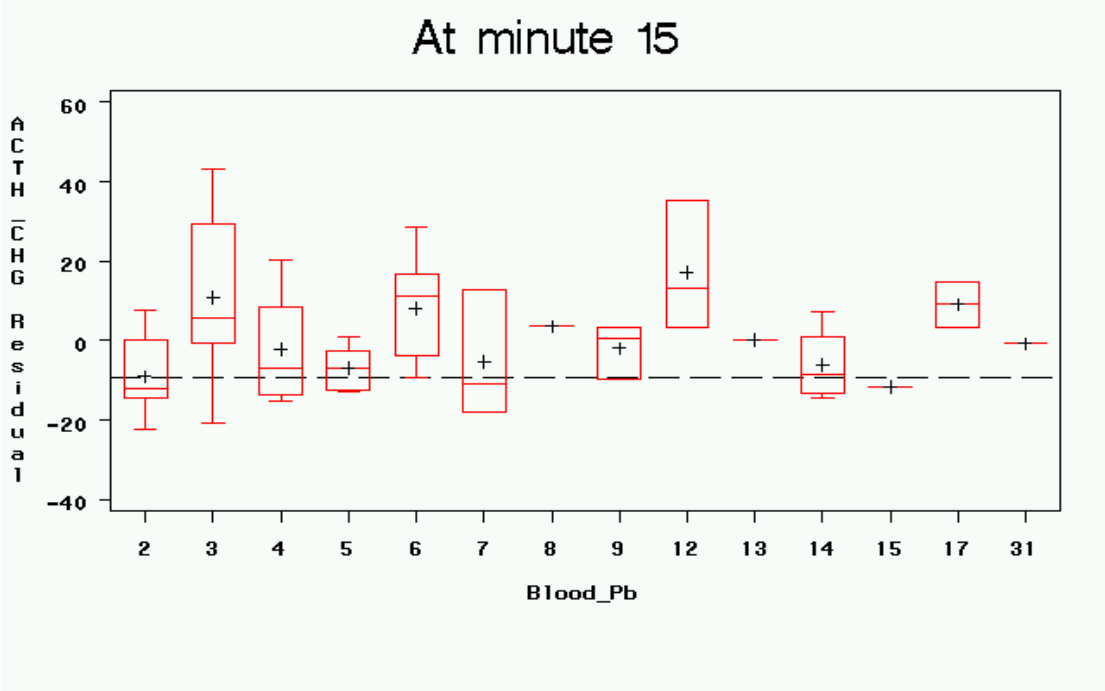
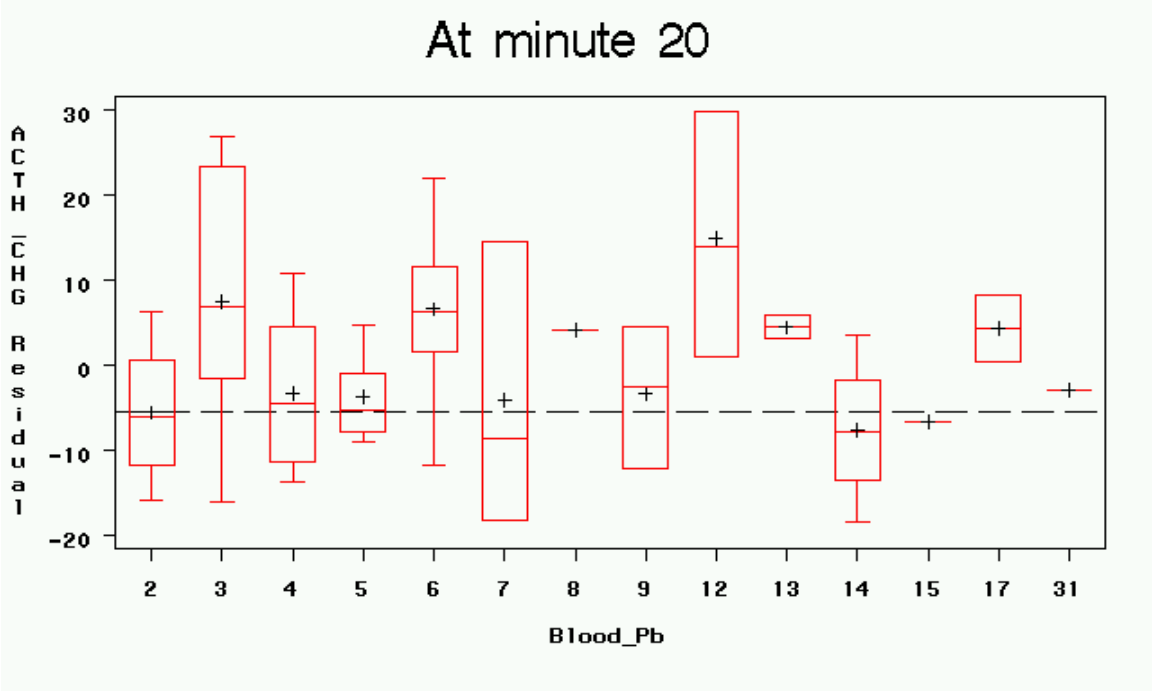


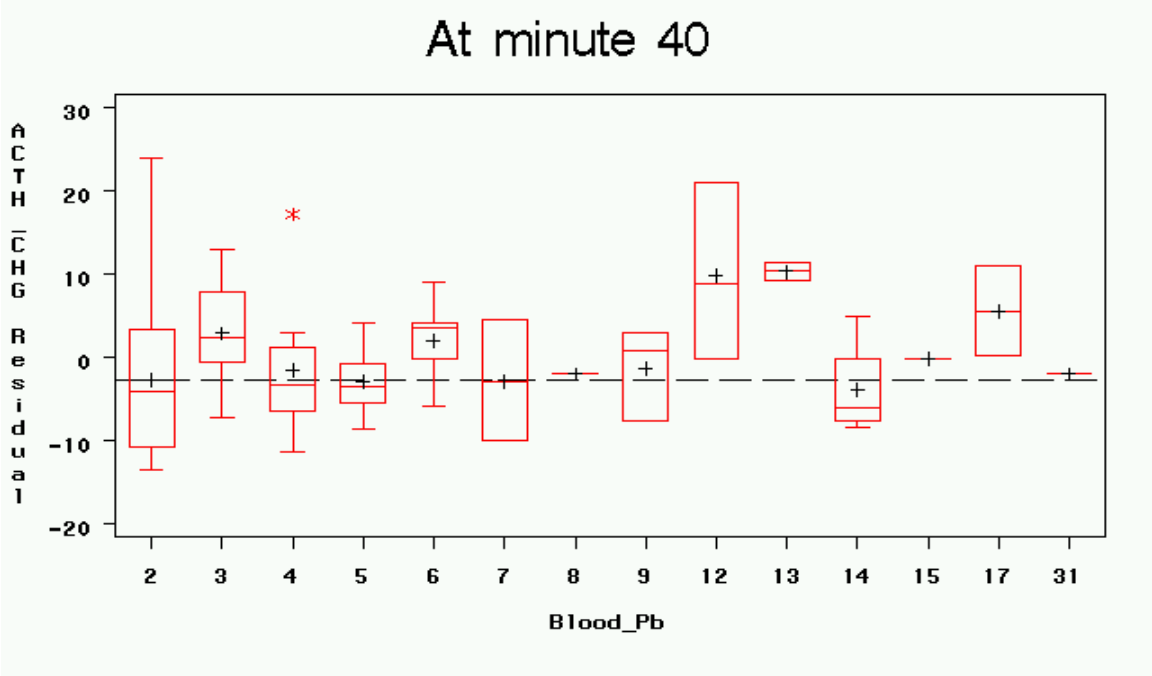
Figure 6.4.4 Box plots of regression residuals of ACTH change from baseline after adjusted by age and solvent exposure index at each blood Pb level by time points.

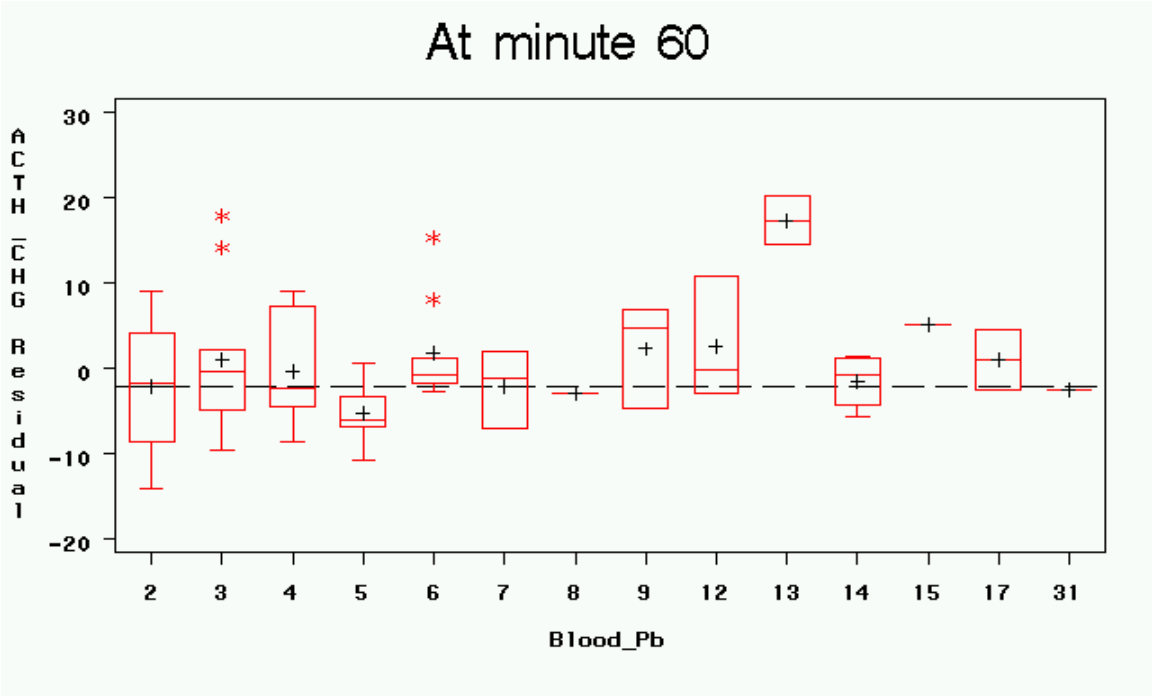


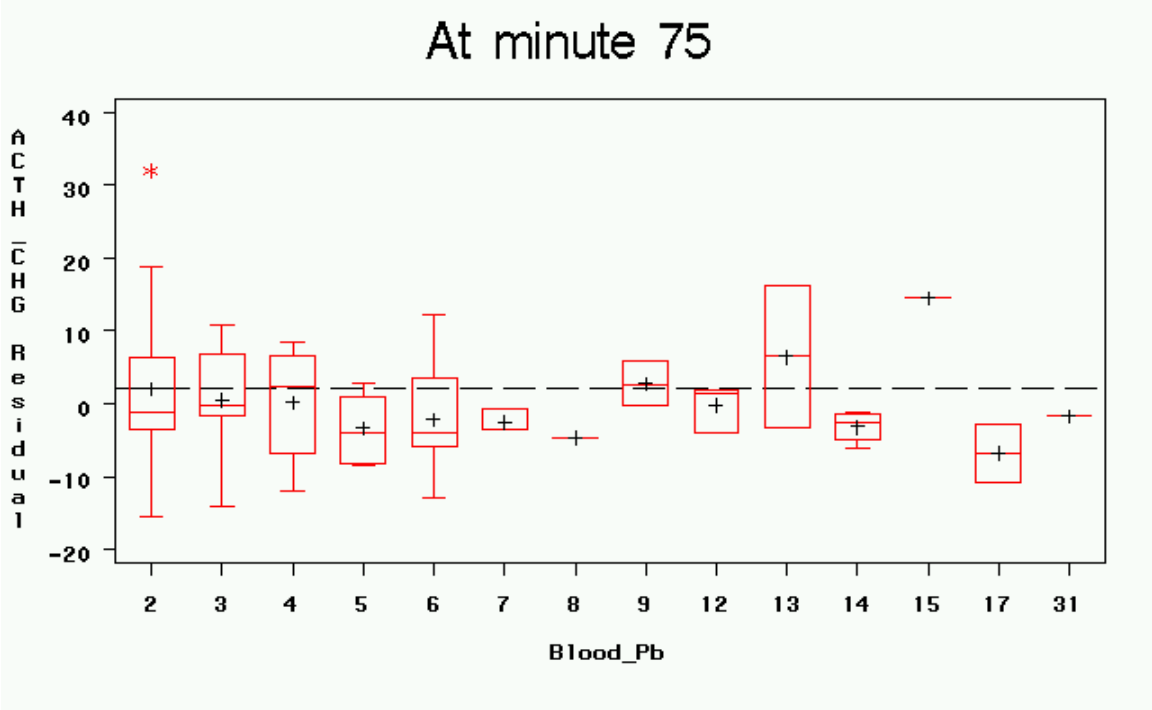



At minute 15
ACTH_CHG Residual
60
40
20
0
-20
-40
2
3
4
5
6
7
8
9
12
13
14
15
17
31
Blood_Pb


At minute 20
ACTH_CHG Residual
30
20
10
0
-10
-20
2
3
4
5
6
7
8
9
12
13
14
15
17
31
Blood_Pb


At minute 40
ACTH_CHG Residual
30
20
10
0
-10
-20
2
3
4
5
6
7
8
9
12
13
14
15
17
31
Blood_Pb


At minute 60
ACTH_CHG Residual
30
20
10
0
-10
-20
2
3
4
5
6
7
8
9
12
13
14
15
17
31
Blood_Pb



6.4.3 Baseline adjusted AUC

Figure 6.4.5 displays the box plot of the baseline adjusted AUC against the blood lead level. No apparent pattern can be seen through this plot.

The regression results also suggest no association between the baseline adjusted AUC and blood lead level. The p-values from the regression models with and without adjustment for covariates of age and solvent exposure index are 0.6520 and 0.7242 respectively.

Table 6.4.2 confirms the regression results by showing there is no significant cut point identified based on the unadjusted and covariates adjusted BAAUC with p-values of 0.3882 and 0.5066 respectively.

Although the ACTH scores are measured at different time point on the same subject, it can not be analyzed by using the repeated measurement method in that the underlying relationship between the ACTH score and time is not linear. The ACTH scores are expected to increase during the reactive phase and then decrease during the recovery phase. The baseline adjusted AUC would be appropriate to summarize the ACTH score over time for this type of study design.

Figure 6.4.5 Box plot of baseline adjusted AUC vs. blood lead level

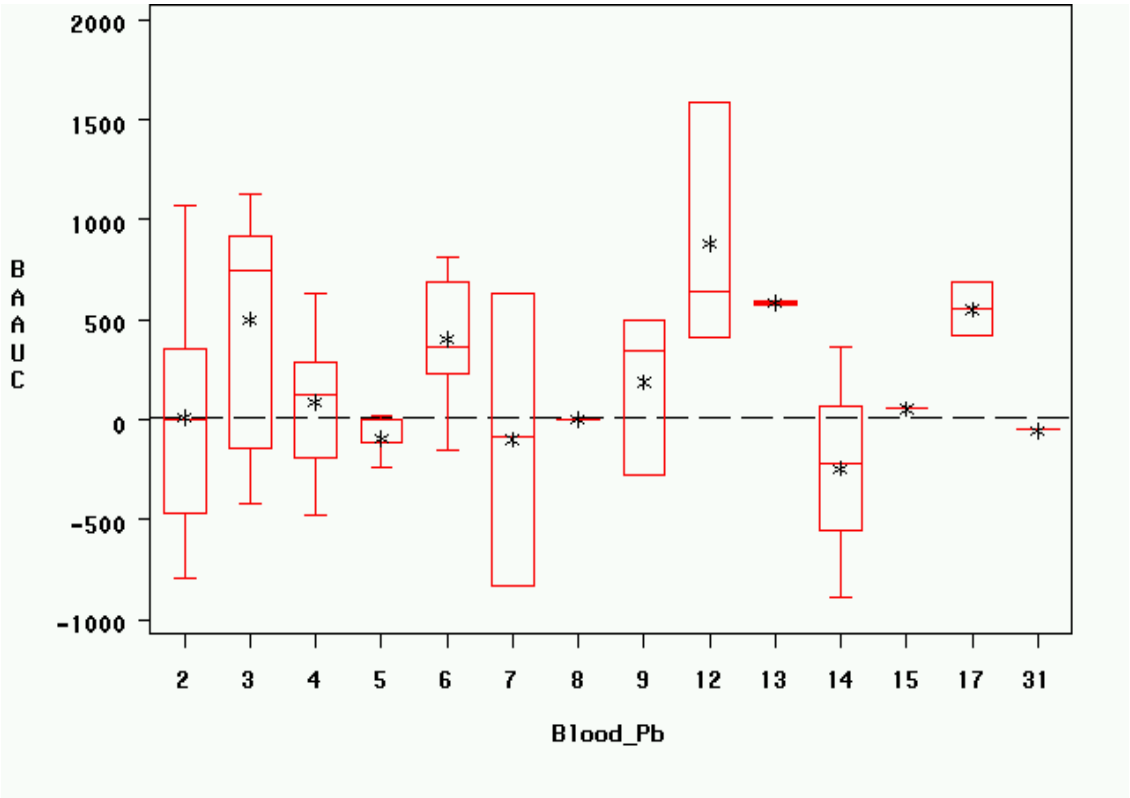


Table 6.4.2 Cut point of blood lead level in related to change from baseline ACTH scores at minute 15 and overall.

Methods	Change from baseline in ACTH at min. 15			BAAUC [1]		
	BLL (μg/dL)	Maximum Rank Stat.	p-value	BLL (μg/dL)	Maximum Rank Stat.	p-value
Unadjusted	2	2.80	0.0346	5	1.75	0.3882
Adjusted [2]	2	3.15	0.0102	2	1.58	0.5066
Adjusted [3]	2	3.14	0.0106	5	1.77	0.3750

1. Baseline adjusted AUC
2. Adjusted by age and solvent exposure index.
3. Adjusted by age only.

6.5 Discussion

This study enrolled subjects who were identified because of the expectation of high level lead exposure and control subjects who were unlikely to be exposed to high level lead.

The predictor variable of blood lead level collected from this study therefore has wide range from 2 to 17ug/dL with an outlier of 31ug/dL. This hence provides a large range to search for the cut point of the blood lead level.

In addition, the analyses results from the grant study report show that by using linear regression method, no association was detected between ACTH and blood lead level, however when comparing the highest quartile group to the lowest quartile group in blood lead level the ACTH was significantly increased in response to the stressor, which suggests the potential of cut point existence. The sample size of this study 72 and there are 14 discrete levels in the predictor variable of blood lead. We used the null distribution generated from sample size of 72 and 14 discrete levels in the predictor variable using the method described in chapter 1, which provides an exact result for the conclusion of a cut point.

With a continuous outcome variable, the cut point search from a continuous predictor has been thoroughly discussed in the literatures. The effect of discreteness of values for the predictor on the cut point search has not been studied. The predictor of blood lead level collected in this study is measured in a discrete scale with multiple observations at the same value, which triggers this dissertation topic on how cut point search behaves differently between the continuous predictor and semi-continuous predictor.

The ACTH is released from pituitary after stimulated by the corticotrophin releasing hormone, which is released from hypothalamic-pituitary-adrenal (HPA) axis after triggered by psychological stress. The details between the HPA axis and stress response can be found in the study grant report. The grant report also mentioned the HPA axis dysfunction has been associated with disease states including depression, metabolic disease, obesity and hypertension. This lead exposure study reveals a cut point of 2 μg/dL in blood lead level of adults in related to a significant increase of ACTH in response to stressor. Followed by a stress challenge, subjects with blood lead level greater than 2 μg/dL responded in higher ACTH increase compare to subjects with blood lead level of 2 μg/dL. By using the cut point selection method, we transformed the semi-continuous variable of blood lead level into a binary variable. The selected cut point allows a classification of the population into two distinct groups at which it maximizes

the measure of difference between the groups. Ideally the cutpoint is suggested by theories of biological functioning behind the relationship between the blood lead level and HPA dysfunction, but this information is not available. Instead, we explored the study sample and to find the empirical cutpoint based on the observed data of outcome and prognostic variables to differentiate between high and low risk groups. This obtained cut point of 2 μg/dL is the best separation point in terms of with maximum difference in ACTH response between groups based on this study sample. And the difference between the two groups is statistically significant after adjustment of multiple testing. In addition to the statistical significance, the choice of a cutpoint to convert a continuous covariate to a binary covariate is also rely on biological knowledge about the particular risk factor or on the results already published in other studies. Compare to the population blood lead level, this cutpoint identified in this study is relatively low compare to the 10 μg/dL by CDC as elevated for public health purposes ((US Department of Health and Human Services, 2010). However this result is consistent to study results from Gump et al 2008, in which the same significant positive association is found between the blood Pb level and cortisol responses to acute stress for children with low blood lead level at 9.5 years of age by comparing the lowest quartile group with range of 1.5 – 2.8 μg/dL to the combined three upper quartile groups with range of 2.9 – 13.1 μg/dL.

CHAPTER 7: DISCUSSION AND FUTURE WORKDS

7.1 Discussion

In the analysis involving data from clinical or epidemiological studies, significant attention is given to continuous or semi-continuous variables such as blood pressure, age etc., but the predictive importance of such variables can not be established easily. It is common in practice to transform a continuous prognostic variable into a binary variable for clinical use. Dichotomizing the predictor variable may result in a loss of information, but is often necessary for practical decision-making. For instance, this is done in order to set up practical criteria as mentioned in the PBSCT example in section 1.6.1, or to guide clinicians and patients in their choice of therapy as mentioned in the advanced seminoma example in section 1.6.2. An optimal cut point searching method in which cutpoints are systematically tested and identified as the one with minimum p-value is introduced in this dissertation. The methods for adjustment of the minimum p-value that accounts for having taken multiple, but not independent, looks at the data are thoroughly discussed and studied.

Ideally a cutpoint is suggested by theories of biological functioning, but this information is rarely available. Instead, observed data on the outcome and prognostic variable often are obtainable from experimental samples and can be explored in order to find empirically a cutpoint which appears to differentiate between high and low risk groups. Some experimental samples may show multiple data points with adjusted p-values below 0.05, the conclusion draw based only on the one with minimum p-value may not represent the general population. If these data points are shown close to each other as a cluster, in order to apply the obtained cutpoint to the population the more appropriate approach would be to group these data and report the range. If these data points are separated, it may be due to an outlier or their might be multiple cutpoints. In this case a though look on the scatter plot is needed.

The cutpoint model discussed in this dissertation is for the existence of single cut point. It can not be used to the variable such as blood pressure, which may have two cutpoints that with values too high and too low are both associated with increased risk.

7.2 Future works

In this dissertation, we reviewed several methods including maximally selected chi-square statistics, maximally selected rank statistics to search for the optimal cut point. Since utilizing the maximally selected rank statistic to analyze semi-continuous predictors has not been discussed in the literatures. This dissertation provides the comparison of the null distribution, power curve, precision of cut point estimation between semi-continuous and continuous predictive variables through simulation. It is not yet clear how to derive the exact methods to compute the maximally selected ranks statistics for finite samples and for samples with predictor variables measured in ordinal or semi-continuous levels. Also in chapter 4, we met computational difficulties to construct the averaged confidence interval for estimated cutpoints. Therefore both the exact method and a better way to create the confidence interval via simulation would be of interest in the future works.

Although we used the baseline adjusted AUC to summarize the multiple measurements of the response variable due to the nature of the lead exposure study design, how to apply the maximally selected statistics to analyze the conventional repeated measurement studies would be of another interest in the future works.

REFERENCE

Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute* 1994; 86: 829-835

Boulesteix, A. L., 2006. Maximally selected chi-square statistics for ordinal variables. *Biometrical Journal* 48, 451-462;

CDC, 1991. Preventing Lead Poisoning in Young Children Atlanta, GA Centers for Disease Control and Prevention.

CDC. Update: blood lead levels --- United States, 1999—2002. MMWR May 27, 2005 / 54 (20); 513-516. Available at http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5420a5.htm

Durbin, J. 1971 Boundary-crossing probabilities for the Brownian motion and Poisson processes and techniques for computing the power of the Kolmogorov-Smirnov test. *Journal of Applied Probability* 8, 431-453.

Durbin, J. 1973. Distribution theory for test based on the sample distribution function. *Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics*.

Fiedler, Nancy. Grant Number: 5R21ES15135; Final Progress Report Lead Exposure, HPA Dysfunction, Blood Pressure: Hypertension Risk

Gump BB, Stewart P, Reihman J, Lonky E, Darvil T, Parsons PJ, Granger DA. Low-level prenatal and postnatal blood lead exposure and adrenocortical responses to acute stress in children. Environmental Health Perspective 2008 Feb; 116(2): 249-255.

Halpern, J., 1982. Maximally selected chi-square statistics for small samples. *Biometrics* 38, 1017-1023.

Hass, R., Mohle, R., Fruhauf, S., Goldschmidt, H., Witt, B., Flentje, M., Wannenmacher, M. and Hunstein, W. Patient characteristics associated with successful mobilizing and autograpting of peripheral blood progenitor cells in malignant lymphoma, Blood, 83, (12), 3787-3794 (1994)

Koziol, J.A., 1991. On maximally selected chi-square statistics. *Biometrics* 47, 1557 – 1561.

Lausen, B., hothorn, T., Bretz, F., and Schumacher M. 2004 Assessment of optimal selected prognostic factors. *Biometrical Journal* 46. 364-374

Lausen, B. and Schumacher, M. (1992). Maximally selected rank statistics. *Biometrics* 48, 73-85.

Mazumdar M, Glassman JR. Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Statistics in Medicine* 2000; 19: 113-132.

Miller, R. and Siegmund, D. (1982). Maximally selected Chi-square statistics. *Biometrics* 48, 1011-1016.

Moskowitz, C., Glassman, J., Wuest, D., Maslak, P., Reich, L., Gucciardo, A., Coady-Lyons, N., Zelenetz, A. and Nimer, S. Factors affecting mobilization of peripheral blood progenitor cells in patients with lymphoma, Clinical Cancer Research, 4, 311-316 (1998).

Puc, H., Heelan, R., Mazumdar, M., Herr, H., Scheinfeld, J., Vlamis, V., Bajorin, D., Bosl, G., Mencel, P. and Motzer, R. Management of residual mass in advanced

seminoma: results and recommendations from the Memorial Sloan-Kettering Cancer Center, Journal of Clinical Oncology, 14, 454-460 (1996)

Schulgen G, Lausen B, Olsen J, and Schumacher M. Outcome-oriented cutpoints in analysis of quantitative exposures. *American Journal of Epidemiology* Vol.140, No.2 172-184.

Stewart P, Pagano J, Sargent D, Darvill T, Lonky E, Reihman J. Effects of Great Lakes fish consumption on brain PCB pattern concentration and progressive-ratio performance. *Environmental Research* 2000 (81): 18-32

US Department of Health and Human Services. Healthy people 2010 (conference ed, in 2 vols). Washington, DC: US Department of Health and Human Services; 2000. Available at http://www.health.gov/healthypeople.

Valbonesi M, Pollicardo N, Carlier P, Florio G, Ruzzenenti MR, Pungolino E, Benvenuto F, Figari Q. PBSC collection from G-CSF primed donors. Transfusion Science 1998 Dec; 17(4): 619-27