PARAMETER ESTIMATION FROM GROUPED DATA WITH APPLICATIONS TO META-ANALYSIS

By YEN-HONG KUO

A Dissertation submitted to the

School of Public Health

University of Medicine and Dentistry of New Jersey

and the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

UMDNJ-School of Public Health

Awarded jointly by these institutions and

Written under the direction of

Professor Dirk F. Moore, PhD

And Approved by

_____

_____

_____

_____

Piscataway/New Brunswick, New Jersey

May, 2010

ABSTRACT OF THE DISSERTATION

PARAMETER ESTIMATION FROM GROUPED DATA WITH APPLICATIONS TO META-ANALYSIS

By YEN-HONG KUO

Dissertation Director:

Professor Dirk F. Moore, PhD

Categorizing a continuous variable is easy for communication and statistical analysis in public health and medical research. However, categorization loses information, reduces statistical power, and biases the estimate of a dose-response association while reducing its efficiency. Further, it jeopardizes the validity and efficiency of a meta-analysis because of the single cutoff point and/or inconsistent cutoff points in the included studies.

In order to appropriately summarize the estimates from each study in a meta-analysis with comparable categories or dose-response association, a new approach on re-estimating the underlying distribution of a categorized covariate by using the published information is the first step.

This dissertation research proposes two types of approaches to estimate the underlying distribution. The first approach is linear model approach. When the underlying distribution follows a normal distribution, a linear model can be constructed by using the mean, standard deviation, and cutoff points with their cumulative probabilities in each study. The parameters can be estimated via the weighted mixed-effect linear regression model. When the underlying distribution follows a gamma distribution, a linear model is derived by applying a property of the incomplete gamma distribution. The parameters can be estimated by using a numerical iteration algorithm.

The second approach is a goodness-of-fit approach. When the parameters of the underlying distribution cannot be linearized, based on the cutoff points and their cumulative probabilities in each study, the parameter estimates minimize the distance between the expected and observed values. We also applied this approach to estimate the parameters of a categorized zero-inflated distribution: the proportion of excess zero and the continuous variables.

In addition, we discuss the impacts from categorization on the relative efficiency of estimating the parameters and the dose-response association, and the validity of the dose-response association by maximum likelihood approach via the multinomial distribution and simulation studies.

In summary, the main contribution from this dissertation is that our approaches use published data to convert from the disadvantage of inconsistent cutoff points in many studies into useful information and to improve meta-analysis. We also generalize the approaches of evaluating the impacts from categorizing a continuous variable.

ACKNOWLEDGEMENT

I like to express my deeply-felt thanks to my thesis advisor, Dr. Dirk F. Moore, for his thoughtful guidance and warm encouragement. I also thank the members of my thesis committee: Dr. John M. Davis, whose discussions ensure the clinical applicability; Dr. Yong Lin, whose comments strengthen the theoretical contents; and Dr. Shou-En Lu, whose suggestions reduced ambiguity.

I like to express my gratitude to Dr. Weichung Joe Shih for his encouragement and supports.

I want to express my sincere thanks to my friends who provide supports during my dissertation research. This is an impossible task to name all of you here but you are always on my mind.

I would like to extend my deepest gratitude to my family for their continuous supports and encouragement: my parents, my mother-in-law, my sisters and brother, my brothers-in-law and sisters-in-law, my nieces and nephew, and my nieces-in-law and nephews-in-law.

Last, but not the least, I would like to acknowledge my daughter Allie and my son Evan. Their love and understanding support me to overcome many challenges. I want to express my deepest gratitude and very special thanks to my wife Shuling for everything. She is always there for me. Without her, this thesis is impossible.

DEDICATION

To my father

Ching-Tsai Kuo

and

To my mother

Li-Ying Kuo-Lin, *in memoriam*.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Categorizing a continuous variable is frequently used in epidemiological and medical studies. The main reasons for grouping data are to make it easier to perform statistical analyses and to improve clarity for interpretation and communication. However, categorization causes loss of information, statistical power, and efficiency.

When a variable of interest is only available in the categorized form and this variable might be a mixture of a continuous distribution and excess zeroes, it may not be possible to know the underlying information about this mixture distribution. Therefore, if methods for modeling this continuous variable as well as the true zero are available, they can provide more useful information from the estimated parameters than from what was originally presented.

The worst impact from categorizing a continuous variable would be an incorrect estimate from a multiple regression analysis, a multi-center clinical trial, or a meta-analysis.

Meta-analysis has been widely used to summarize results from similar studies. Meta-analysis is a quantitative method for summarizing a large number of studies, and in some cases it may improve the accuracy of estimation by using a larger sample size than each individual study (Deeks et al, 2009). As a consequence, its application in medicine has strengthened the practice of evidence-based medicine. However, if the impacts from categorizing data cannot be handled properly, the results from a meta-analysis could be misleading and potentially even jeopardize patient safety.

We start by summarizing the methods used for categorizing continuous variables in Section 1.1. In Section 1.2, three published meta-analyses and one pooled analysis will be used to illustrate the methods of categorizing data. From those four examples, some statistical issues and clinical concerns will be described in Section 1.3. These concerns are the motivations for this dissertation research. In Section 1.4, the methods used for estimating the underlying distribution based on categorized data will be introduced, and the limitations of using those methods on the examples which are described in Section 1.2 will be discussed.

1.1 Methods for Categorizing a Continuous Variable

Many different methods for categorizing a continuous variable have been used in the literature. Based on the rationale of choosing cutoff points, they may be classified into the following three types.

1) Evidence-based

   This type of categorization is based on the existing cutoff points, such as: the cutoff points used in published studies, the cutoff points in the user's manual of a product, or cutoff points defined by professional organizations or governmental agencies.

   For example, body mass index (BMI) has been commonly used to classify weight status. These BMI groups are then used to evaluate the association with some diseases or conditions of interest. Flegal and colleagues (2007) assessed the cause-specific excess death associated with being underweight, overweight and obese. They used the BMI criteria provided by the National Institute of Health

and World Health Organization to classify weight status. The classifications for BMI from NIH (NIH, 1998) are: Underweight (BMI < 18.5 kg/m^2); Normal weight (BMI: 18.5-24.9 kg/m^2); Overweight (BMI: 25.0-29.9 kg/m^2); Obesity (Class 1) (BMI: 30-34.9 kg/m^2); Obesity (Class 2) (BMI: 35.0-39.9 kg/m^2); Extreme obesity (Class 3)(BMI: ≥40 kg/m^2).

Similar definitions can also be found from the World Health Organization (WHO, 1999). They are: Underweight (BMI < 18.50 kg/m^2); Normal range (BMI: 18.50-24.99 kg/m^2); Overweight (BMI: ≥25.00 kg/m^2); Preobese (BMI: 25.00-29.99 kg/m^2); Obese class I (BMI: 30.00-34.99 kg/m^2); Obese class II (BMI: 35.00-39.99 kg/m^2); Obese class III (BMI: ≥40.00 kg/m^2).

2) Interval-based

This type of categorization is based on pre-defined intervals by using integers as the cutoff point. Usually the intervals have equal length. For example, Flum and colleagues (2002) assessed the association between age and the chance of misdiagnosed appendicitis. They grouped age of study participants into the following age groups: 0-4, 5-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79 and ≥80.

3) Statistics-based

This type of categorization is data-driven. Quartiles or other quantiles are frequently used to get equal number of subjects in each group. The other popular method is the minimum P-value method (Altman et al, 1994). This classification is based on the cutoff point for which the smallest P-value can be achieved after testing a range of possible cutoff points.

For example, Bartali and colleagues (2008) assessed the association between serum micronutrient concentrations and decline in physical function among older persons. They used the quartiles of vitamin E concentration (μg/ml) of all study participants to group them into four groups: 1st quartile (<1.1 μg/ml), 2nd quartile (1.1-1.3 μg/ml), 3rd quartile (1.3-1.5 μg/ml) and 4th quartile (≥1.5 μg/ml).

1.2 Examples of Categorized Variable in the Literature

In the following sub-sections, four examples will be used to demonstrate the types of categorization in epidemiological and medical studies.

1.2.1 Dichotomized Quantitative Covariate

Ferrandina and colleagues (1997) performed a meta-analysis of the association between the dichotomized cathepsin-D level and the disease-free survival in node-negative breast cancer patients (Ferrandina, 1997). The survival time is the outcome of interest. There are 11 clinical studies which met the selection criteria. Cutoff points were chosen by the researchers of each study to classify the cathepsin-D status as either positive or negative. From the 12 published articles, 11 different cutoff points ranging from 20 to 78 pmol/mg protein were reported in this meta-analysis from 10 studies. These cutoff points were decided by using two statistics-based methods: median or minimum P-value method. They are summarized from the original article and shown in Table 1.2.1 on the following page.

Even though the authors acknowledged the issue of inconsistent cutoff points, they still used the dichotomized covariate to summarize the effect from cathepsin-D status by using the method developed by Peto (1987).

This article has been used as an example to demonstrate the unreliability of results due to the inconsistent cutoff points (Altman, 2001). However, no corresponding solution was proposed.

Table 1.2.1 Cutoff points summarized from studies included in the meta-analysis by Ferrandina et al, 1997

Reference	No. of patients according to cathepsin-D content		Cut-ff (pmol mg^{-1} protein)	Positivity (%)
	Low	High		
Isola et al (1993)	167	95	NA	36
Janicke et al (1993)	64	33	50	34
Kandalafi et al (1993)	84	51	NA	37.7
Kute et al (1992)	45	93	39	28*
Namer et al (1991)	132	114	35	46
Pujol et al (1993)	38	26	20	40
Seshadri et al (1994)	117	237	25	67
Ravdin et al (1994)	467	460	54	50
Spyratos et al (1989)	39	29	45	42.6
	57	11	70	16
Tandon et al (1990)	135	64	75	32
Thorpe et al (1989)	93	26	78	22
Thorpe et al (1989)	24	57	24	70

*: This is an error in the original article. The correct value should be “67”.

1.2.2 Categorized Quantitative Covariate

Kamat and colleagues (2009) evaluated the association between body mass index (BMI) and the risk of Barrett's esophagus (BE) by performing a meta-analysis (Kamat et al, 2009). The included studies categorized BMI into different weight status to assess the association with BE by using relative risk or odds ratio. The number of groups and their corresponding cutoff points from all of the included studies are summarized in Table 1.2.2 on the next page.

Even though the classifications provided by the World Health Organization are commonly used (WHO, 1999), a wide variation of numbers of groups (2 to 5) as well as different cutoff points co-exist in the studies included in their meta-analysis. Therefore, when two meta-analyses were performed by using different cutoff points (BMI of 30 or 25 kg/m^2), the authors needed to exclude one or two studies to accommodate this inconsistency. The exclusion resulted in losing information from relevant studies.

Table 1.2.2 Summary of body mass index categories used in the meta-analysis of BMI and Barrett's esophagus by Kamat (2009).

First Author (Year)	BMI (kg/m^2) category
Gerson (2002)	≤25 > 25
Bu (2006)	Quartile I (<22) - low Quartile II (22–24.9) Quartile III (25–29.9) Quartile IV (>30) - high
Ronkainen (2005)	<30 ≥30 obesity
Corley (2006)	<30 >30
Johansson (2007)	Low tertile (</23.6) Middle tertile (23.6-26.6) High tertile (>/26.6)
Corley (2007)	<25.0 25.0–27.4 27.5–29.9 30.0–34.9 >35.0
Gerson (2007)	Underweight (BMI < 18.5) Normal (BMI 18.4–24.9) Overweight (BMI 25–29.9) Obese (BMI > 30)
Stein (2005)	<25 (normal, reference) 25–30 (overweight) >30 (obese)
Veugelers (2006)	Underweight (BMI < 20) normal weight (BMI of ≥20 and < 25) overweight (BMI of ≥25 and < 30) obesity (BMI ≥30)
Edelstein (2007)	normal weight (BMI <25) overweight (BMI 25–29.99) obese (BMI ≥30)
El-Serag (2005)	<25 25–30 >30
Smith (2005)	<18.5, ‘‘underweight’’ 18.5-24.9, ‘‘normal’’ 25-29.9, ‘‘overweight’’ ≥30, ‘‘obese’’

1.2.3 Categorized Quantitative Covariate Containing Excess Zeroes

Morton and colleagues performed a pooled analysis of the association between alcohol consumption and the non-Hodgkin lymphoma risk (Morton et al, 2005). A pooled analysis is one type of meta-analysis in which the individual patient data are combined to perform the analysis. To categorize the patients who consumed alcohol, the published articles of the included case-control studies used various numbers of cutoff points ranging from 0 to 5 (Table 1.2.3 on next page). The cutoff points used for categorizing the patients were decided by using either a statistics-based (tertile, quartile) or an interval-based method.

Most of the studies used the non-drinker group as the reference group to report the relative risk via the odds ratio. However, one study (Chang et al, 2004) combined the non-drinker and the low level of alcohol consumption subjects together as the reference group. One study (Willett et al, 2004) used a low level group instead of the non-drinker group as the reference group.

Currently there is no statistical method available to estimate the parameters of a continuous variable if it is reported as a mixture of a zero exposure group and several categories for other groups. One publication uses the Expectation Maximization (EM) algorithm to estimate the mixture of a doubly truncated log-normal distribution (McLachlan and Jones, 1988). Also there is no method to convert the association from the dichotomized status of drinker vs. non-drinker to the association based on the level of alcohol consumption. Therefore, without the data from individual patients, these published articles cannot be included in meta-analysis with currently available methodology.

Table 1.2.3 Cutoff points used in the case-control studies on the alcohol consumption and non-Hodgkin lymphoma risk

Author (Year)	Level	Cutoff Points	
		Female	**Male**
Holly et al (1999)	None* Low Medium High	0 <=2.2 >2.2-<5.8 >=5.8 (drinks/week)	0 <=5.5 >5.5-<13.6 >=13.6 (drinks/week)
Morton et al (2003)	Never Ever	Never Ever	*No male in the study*
Chang et al (2004)	Q1: 0.00-0.21** Q2: >0.21-0.78 Q3: >0.78-1.80 Q4: >1.80	Q1: 0.00-0.21 Q2: >0.21-0.78 Q3: >0.78-1.80 Q4: >1.80 (drinks/day)	Q1: 0.00-0.21 Q2: >0.21-0.78 Q3: >0.78-1.80 Q4: >1.80 (drinks/day)
Willett et al (2004)	Never >0-1 >1-2 >2-4 >4-6 >6 (units/day)	*(adjusted for gender)*	*(adjusted for gender)*
Tavani et al (2001)	None <3 3-6 >=7 (drinks/day)	*(combined)*	*(combined)*

*: The underlined level represents the reference group.

**:"Qn" represents the n^{th} quartile.

Another meta-analysis example involves assessing the association between tea consumption and endometrial cancer by Bandera and colleagues (Bandera et al, 2007). A summary of some relevant studies is shown in Table 1.2.4. From the included studies, some of the subjects did not consume tea. Therefore, when the subjects were mixed with the subjects who consumed small amount as the reference group, the estimated odds ratio impacts the estimate in meta-analysis.

Table 1.2.4 Summary of studies which studied the association between tea consumption and endometrial cancer

Author (Year)	**No. of Group**	**No. of Group containing 0**	**Level or Cutoff Points**	**Notes**
Levi et al (1993)	2	N/A	Low Intermediate	Did not specify the cutoff point in the article.
Zheng et al (1996)	4	1	Never/Monthly Weekly 1 cup/day ≥2 cups/day	Used 1 cup = 237 ml.
Goodman et al (1997)	4	2	0 0-34 34-237 >237	Used quartiles. Therefore, the "0-34" group might be a mixture of 0 and low dose. The unit of cutoff points is gram.
Jain et al (2000)	4	1	0 0-250 250-500 >500	The unit of cutoff points is gram.

1.3 Issues identified from the Examples

When the results from meta-analysis have been used as evidence for practicing evidence-based medicine, the examples in the previous section raised some statistical issues and clinical concerns. They will be discussed in the following subsections.

1.3.1 Categorizing Quantitative Covariates

As we have seen, a quantitative variable is frequently grouped as a categorical variable in the medical and public health literature. Even though this may make it easier for a researcher to analyze data and to communicate results with colleagues and patients, information from categorization cannot be fully used and, even worse, might jeopardize the results.

For example, estrogen receptor (ER) has been reported as a useful biomarker for predicting outcomes of treating breast cancer patients. The ER status is frequently described as either positive or negative. Based on a meta-analysis performed by the Early Breast Cancer Trialists' Collaborative Group (EBETCG), breast cancer patients who were ER-positive and received Tamoxifen therapy for 5 years had less chance of mortality (ratio of annual death rates= 0.66 [SE=0.04]) (EBETCG, 2005). However, the patients who were ER-negative and received Tamoxifen therapy for 5 years did not benefit from the therapy (ratio of annual death rates=1.04 [SE=0.08]).

Even though the ER status has been commonly used, the cutoff point used for dichotomizing the positive/negative status is not a universal one (Kuo, 2000; Althuis et al, 2004). As a consequence, the summarized effects of raloxifene on the ER-positive and

ER-negative cancer from one large scale multi-center clinical trial are not appropriate (Kuo, 2000).

Covariates in all of the discussed examples in Section 1.2 were categorized. Therefore, they all share the problems of losing information and possibly jeopardizing the results. More details and discussions about the impacts on the results will be discussed in Section 1.3.5.

1.3.2 Choice of Cutoff Point(s) and Inconsistency

For the purpose of categorizing subjects into groups for analysis and clinical interpretation, cutoff points need to be decided by the researcher when planning or conducting the study. As described in Section 1.1, different methods have been used to find the cutoff points. However, there is no consensus on finding "the" best cutoff points for categorizing a quantitative variable when each study is conducted individually.

One of the methods of choosing a cutoff point to dichotomize a covariate is the "minimum p-value" method (Altman et al, 1994). The term of "optimum p-value" is also used. The rationale for using this approach is to choose a cutoff point which can discriminate the collected data best to show a statistically significant difference. Therefore, the minimum p-value is a common criterion for choosing a cutoff point. This approach is data-driven. As a consequence, different studies use different cutoff points for the same covariate. Furthermore, the process of finding the cutoff point involves the "multiple comparison" problem. Therefore, the reported p-value should be adjusted to prevent Type I error inflation.

Quantile is another commonly means for selecting cutoff points. For example, median has been a good candidate to dichotomize a covariate. Use of quantiles can prevent the problems from multiple comparisons. However, quantiles are still data-driven. Therefore, they differ from one study to another. Consequently, when performing a meta-analysis, the inconsistent cutoff points between studies introduce problems for summarizing the effects.

From the example in Section 1.2.1, the minimum p-value method was used by three studies on finding cutoff points (included one used the classification and regression trees [CART] method); three studies report that medians were used as the cutoff points. However, two studies do not report how the cutoff point was chosen and two studies do not report their cutoff point for dichotomizing the covariate.

If a single method for identifying the cutoff points is used for all of the studies, the distribution of the cutoff point could be used to estimate the covariate distribution. For example, if the mean of each study is used as the cutoff point and the covariate is normally distributed with mean μ and standard deviation σ in each study, the distribution of the cutoff points follows the Central Limit Theorem and has a mean of μ and standard deviation of $\sigma / \sqrt{n}$, where n is the sample size of each study. Therefore, the distribution of the cutoff point could be used to estimate the underlying covariate distribution. However, when multiple methods are used for finding the cutoff points, the resulting cutoff points from studies is a mixture of distributions and might not be able to provide any useful information.

1.3.3 Mixture Distribution of the Reference Group

When an individual subject in a study can be classified as one of two types of status based on the exposure, either exposed or not exposed, the "no-exposure" group is frequently used as the reference group to assess the dose-response association. However, in some cases, the subjects in the no-exposure group and the low-level exposure group are combined as a group for analysis. For example, Chang and colleagues (2004) grouped the subjects into four groups with equal numbers in each group (example in Section 1.2.3). Therefore, the non-drinker and the low-level drinker groups were mixed together as the reference group. Without finding out the number of non-drinkers, this study cannot be used for meta-analysis using current methodology, if the other studies used the non-drinker as the reference group.

1.3.4 Statistics Reported in the Literature

In printed publications of study results, the space devoted to biostatistics is very limited in medical and public health journals as well as other professional journals. Therefore, the information reported by the researcher is based on its importance related to the outcomes. Unfortunately, detailed statistics describing the covariate distribution are usually not available from the publication.

The following statistics are available from the four examples in Section 1.2 in most of the included studies for meta-analysis and are usually available from the published study:

1. The cutoff points for grouping the quantitative covariate
2. Numbers of subjects in each group
3. Measured effect

Therefore, the development of statistical methods should focus on using only this limited information.

1.3.5 Inappropriate Performance of Meta-Analysis

When inconsistent cutoff points exist between studies and no appropriate statistical methods are available, there are three possible impacts from the inconsistency:

1) Incorrect effect estimated from ignoring the inconsistency

   The effect in one study cannot be compared with the effect in another study if the dichotomized cutoff points are not the same between studies. Therefore, when performing a meta-analysis by using studies with different cutoff points, the estimated effect is incorrect. The examples described in Section 1.2.1 present this problem.

2) Fewer studies may be included in the meta-analysis, as a result of acknowledging the inconsistency

   Without existing statistical methods on estimating the covariate distribution when only one cutoff point is used to categorize a covariate, the studies with a single cutoff point cannot be included for meta-analysis. The examples in Section 1.2.2 encountered this problem and need to exclude some relevant studies.

3) Failure to perform meta-analysis because of acknowledging the inconsistency

   When the inconsistency exists but no appropriate method can be used to overcome the concern, systematic reviews are usually presented instead of performing meta-analysis. As a consequence, the results are harder to communicate and to be used for practicing evidence-based medicine.

1.4 Methods for Modeling Categorized Data

When a continuous variable is categorized into several groups, four types of approaches are available for modeling this continuous variable:

1) Summary Statistics

   When the mean and standard deviation are reported, both summary statistics can be used to estimate the assumed continuous distribution.

2) Expectation and Maximization (EM) Algorithm

   Categorized data can be modeled by using the EM algorithm. When raw data in one of the groups are available, a continuous variable can be estimated (Dempster et al, 1977). When the data are reported in double-truncated form, the EM algorithm can also be used to model the lognormal distribution (McLaren et al, 1986).

3) Goodness-of-Fit

   The parameters can be estimated by fitting a continuous distribution to the observed numbers in each group by minimizing chi square (Hartemink et al, 2006).

4) Probability Plotting

   Based on the cumulative probability and cutoff points, the parameters can be estimated by the linearized association. When the grouped data are from a normal distribution, this method estimates the parameters very well (Chêne and Thompson, 1996). However, their method works only for normally distributed data.

No method for handling the dichotomized quantitative covariate with inconsistent cutoff points between studies is currently available. The possible reason could be that the status of the categorization (e.g. positive / negative) is used for meta-analysis without

considering the inconsistency. Also when only one study is of interest, estimation of the underlying distribution is not possible because of the limited information.

Current methods only take care of modeling a continuous distribution in one study. By the criteria of meta-analysis, all of the included studies should have similar characteristics. Therefore, estimation of parameters for each study should be able to improve if the distribution can be estimated by using information from all of the similar studies.

If the raw data are available, methods for estimating the characteristics of a distribution in a mixture distribution have been developed. However, when only grouped data are available, the current methods for raw data cannot be used.

The following table summarizes what has been done for handling some types of the studies. This table also describes the role of the proposed approaches: filling the gap and improving the current approach.

Table 1.4.1 Methods on Modeling the Categorized Distribution

Type of Studies	**Existing Methods**	**Our Approach**
Dichotomized covariate with inconsistent cutoff points from different studies	N/A	Works
Categorized covariate including dichotomized cutoff points from different studies	N/A	Works
Categorized normally distributed covariate excluded dichotomized cutoff points in one study or from different studies	Chêne and Thompson (CT) method estimates individual study	Improves the Chêne and Thompson method by using all studies together
Covariate with distribution other than normal in one study or from different studies	CT suggests transformation	Estimate parameters using quantiles via weighted linear regression approach (gamma distribution) or weighted goodness-of-fit
Combined no-exposure and low-level exposure subjects into one group and also categorizing a quantitative covariate in one study	N/A	Estimate the proportions of the no-exposure and low-level exposure subjects as well as the underlying categorized covariate

In addition, even though the influence of categorization on the efficiency of estimating the parameters of a continuous variable has been studied, the impact on estimating the association between the outcome variable and the predictor variable (either this categorized covariate serves as the predictor or a confounding variable) is still not well understood.

1.5 Summary of the Dissertation

The overall goal of this dissertation research is to develop new approaches by which the distribution of a continuous covariate can be estimated from studies included in a meta-analysis by using the limited information reported in the published manuscript. The estimated distribution can be used to answer some new questions of interest and improve the summarized effects from relevant studies in a meta-analysis. This research also evaluates the impact from categorization on the parameter estimation, either on the parameters of a continuous variable or on the association between the outcome variable and the predictor variable.

In Chapter 2, we will review literature to discuss the existing methodology on the relevant topics and their limitations.

In Chapter 3, a novel linear model approach for estimating the underlying normal distribution from different studies with inconsistent cutoff points for dichotomizing a covariate will be introduced and evaluated. The same rationale is further extended to estimate parameters from different studies with multiple inconsistent cutoff points.

We also discuss the use of linear model approach on estimating the categorized gamma distribution. Due to the characteristics of a gamma distribution, we proposed a numerical iteration algorithm to estimate parameters based on a property of the incomplete gamma distribution. We will associate the linear model approach to a goodness-of-fit approach and discuss its application to parameter estimation from grouped data.

Because this approach works for a categorized covariate, we also propose a goodness-of-fit approach to estimate the proportion exceeding zero as well as the underlying distribution of a categorized covariate from a mixture distribution.

At the end of Chapter 3, we will use the developed methods to estimate the underlying distribution of examples described in Section 1.2. We will also perform a new meta-analysis by including the studies which were excluded from the meta-analysis described in Section 1.2.2.

In Chapter 4, the robustness and efficiency of linear model approach will be evaluated by using simulation studies. The impact on estimating the parameters of a normal distribution from the number of studies used, number of subjects in each study, characteristics of the underlying distribution, and variation of the cutoff point, and the number of cutoff points will be evaluated. The use of the median and mean as the cutoff point will also be discussed.

In Chapter 5, we will discuss the use of a multinomial maximum likelihood approach to estimating the underlying distribution of a categorized continuous covariate. We will demonstrate both the analytic and numerical approaches. Simulation studies will also be used to compare the relative efficiency.

In Chapter 6, we evaluate the efficiency of categorizing dose in a dose-response relationship. The relative efficiency of the categorization, from dichotomization to many cutoff points, will be demonstrated by using the results from simulation studies. We will also discuss the impact from the dose-response association and from the characteristics of the underlying distribution.

In Chapter 7, we will discuss the impact from categorizing a continuous covariate on the estimation of treatment effect. We will discuss the general analytical equation for the asymptotic relative efficiency. We will also use the equation to replicate a published study by providing details of the completed calculation.

Chapter 2

Literature Review

2.1 Methods for Estimating the Covariate Distribution from Categorized Variable

When the value of a continuous variable is available, the exposure parameter can be estimated by using the method of maximum likelihood, the method of moments, and probability plotting. We will discuss each method in turn.

2.1.1 Method of Maximum Likelihood

When a continuous variable is categorized, it is natural to treat the categorized variable as the outcome of a multinomial distribution. Therefore, the parameters of this multinomial distribution can be estimated by using the method of maximum likelihood.

2.1.2 Method of Moments

The method of moments provides an alternative approach to estimate the parameters. However, as described in Section 1.3.4, the summary statistics are not always available from the published studies. Without those statistics, we cannot use this approach.

2.1.3 Probability Plotting Approach

One of the parameter estimation methods is the method of probability plotting. The rationale is that the association between the measurements and the corresponding cumulative probability function is able to be expressed as a simple linear regression by

using the parameters of the underlying distribution as the coefficients after appropriate transformation.

Chernoff and Lieberman (1954, 1956) applied this approach and discussed using normal probability paper for finding the optimum estimates of mean and standard deviation of a normal distribution.

This characteristic has been used for the specific types of probability paper from which the parameters of a distribution can be estimated. Probability papers are available for use to estimate the parameters for the normal distribution, lognormal distribution, exponential distribution and Weibull distribution (e.g. Weibull.com, 2008).

Even though probability plotting has been used as a standard method of parameter estimation, most of its application was seen on using the raw data, not grouped data.

2.1.3.1 Chêne and Thompson Approach

The method of probability plotting has been used to estimate the parameters of a covariate within a meta-analysis study by Chêne and Thompson (1996). Instead of using the original measurements, they used the cutoffs of all of the categories and the corresponding cumulative proportions of subjects to fit a linear regression line under the assumption that the covariate follows a normal distribution. Following is a more detailed description of their approach.

Let a continuous variable X be categorized into k groups by the cutoffs x_j, j=1, 2, … , k-1. N_j is the number of subject in each group, j=1, 2, … , k. The total number of subjects is $N = \sum_{j=1}^{k} N_j$. Therefore, the cumulative proportion is $P_j = \sum_{i=1}^{j} N_i / N$. That is, P_j is the proportion of the subjects that had the measurements less than the cutoff x_j and $P_k = 1$.

Based on the underlying normal distribution, we have $z_j = \Phi^{-1}(p_j)$ as the normal derivate which corresponds to P_j, j=1, 2, … , k, and Φ is the cumulative standard normal distribution function. As a result, if the continuous variable X is normal, a plot of cutoffs x_j against normal derivates z_j should be linear.

When we regress x_j on z_j, the estimated intercept m and slope s are the estimates of the mean μ and the standard deviation σ of the normal distribution, respectively. The regression analysis should be weighted inversely proportional to the variance of the quantiles.

However, there are two limitations from their method. First, their approach depends on the assumption of normality. That is, their method could work only for the normal distribution or for another distribution which can be transformed to be normally distributed. Secondly, when a study has only one cutoff point, their weighted regression approach cannot be used.

Even though the method of probability plotting used by Chêne and Thompson (1996) is related to meta-analysis, their application is to estimate the effect in each study to derive the dose-response association, but not to estimate the covariate distribution based on all of the studies.

2.1.4 Comparisons on Estimation Methods

Because the current statistical methods for meta-analysis are developed for summarizing effects, no methods are available for estimating the covariate distribution from all of the studies included in a meta-analysis. That is, the probability plotting

method has not been used for estimating the covariate distribution based on all studies in a meta-analysis.

The comparison of the previous mentioned approaches for estimating the parameters of a distribution are summarized in the following table:

Characteristics	Estimating Method		
	Maximum Likelihood	**Moments**	**Probability Plotting**
Assumption for Covariate	Need	Need	Need
Information from data	Observed value	Observed value	Observed grouped value and cumulative proportion (quantile)
Sample Size for better result	Large	Large	Large However, limited number of value (with quantile) can work
Used for meta-analysis	Not yet	Not yet	Not yet
Suitability for meta-analysis	No	No	Yes

Instead of using the estimation approach, one may attempt to contact the researcher of each study in a meta-analysis to get the individual patient data to estimate the parameters, when the quantitative covariate is reported as grouped covariate. That is, once the raw data are available, the appropriate statistical methods can be used to estimate the effect from covariate on the outcome variable (Stewart et al, 1993, Stewart et al, 1995). But obtaining the original datasets from all of the studies is generally not feasible.

2.2 Methods for Modeling a Mixture Distribution containing Excess Zeroes

From the example in Section 1.2.3, the categorized covariate might contain excess zero measurements, that is, it may contain more zeros than one would expect from a log-normal or gamma distribution. When the existence of excess zeros is not under consideration while assessing the association between the covariate and the outcome variable, the estimated association is in question. Therefore, in order to correctly account for the excess zeros, a statistical method needs to be applied to estimate the proportion of excess zeros and also estimate the parameters of the distribution of continuous covariate.

The zero inflated Poisson distribution approach will be briefly discussed in Section 2.2.1. Because there is also a possibility that the covariate is from the Tweedie families in which a high proportion of zero is observed, we will discuss the Tweedie families in Section 2.2.2.

2.2.1 Zero Inflated Poisson Distribution Approach

The zero inflated Poisson distribution can be considered as a special case of mixture distribution. It consists of a one-point distribution (zero) and a Poisson distribution. A parameter P can be used to model the probability of zero and also model the probability of Poisson distribution. To estimate the parameter P and the parameter of Poisson distribution λ, the maximum likelihood approach is used.

2.2.2 Tweedie Families

The Tweedie family of densities belongs to the class of exponential dispersion models. It is a two-parameter distribution in which a linear exponential family and a dispersion parameter. (Dunn and Smith, 2001, 2005).

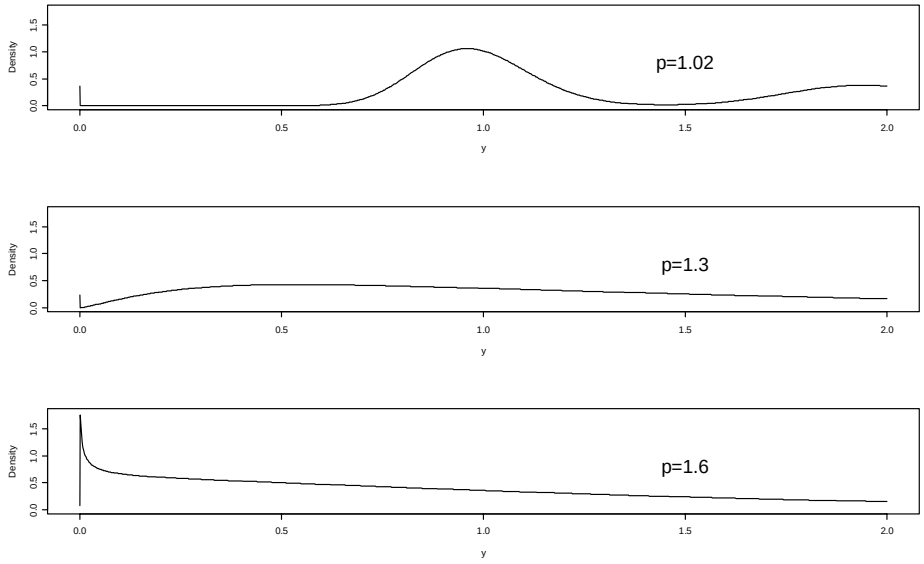
When a random variable Y follows an exponential dispersion model, the density function can be written as:

$$P_Y(y;\mu,\phi) = a_p(y,\phi)\exp\{[y\theta - \kappa(\theta)]/\phi\}$$

$$= b_p(y,\phi)\exp\{-d(y,\mu)/(2\phi)\}$$

where the mean is $\mu = \mathrm{E}[Y] = \kappa'(\theta)$, ϕ>0 is the dispersion parameter, θ is the canonical parameter, $d(y,\mu)$ is the unit deviance, and $\kappa(\theta)$ is the cumulant function. The power variance function $\mathrm{V}[\mu] = \phi\mu^p$ characterizes the Tweedie family densities, where $p \in (-\infty, 0] \cup [1, \infty)$ is the index which determines the distribution.

When the parameter 1<p<2, the density function contains a mass at zero. Therefore, this family could be used to model the covariate which containing a mass of zero. Three examples of Tweedie families are shown in the following figures by using the `tweedie` library in R.



2.3 Effect of Categorizing a Continuous Covariate on the Efficiency

From a bivariate normal population, when a continuous variable is dichotomized at the mean, it reduces the correlation coefficient from r to $0.789r$ (Cohen, 1983). When both continuous variables are dichotomized, the correlation coefficient becomes $0.637r$. Therefore, the statistical power has been reduced.

Categorizing a continuous covariate increases the variance of the treatment effect when assessing the association between a treatment variable and the outcome variable, while controlling for the continuous covariate. In a study by Morgan and Elashoff (Morgan and Elashoff, 1986), the authors evaluated the influence from categorization on modeling the survival time. They derived an analytical solution on calculating the asymptotic relative efficiency (ARE) from categorization. Under the assumption that there is no treatment effect, categorizing a covariate increases the variance of treatment effect estimate. Therefore, the ARE of treatment effect estimate reduced. The reduction of the ARE depends on the parameter of gamma distribution. It also depends on the

number of cutoff points. However, their equation was developed under the assumption that there is no association between the treatment effect and the survival time and the covariate is serving as the confounding variable. This equation is derived for survival time which is exponentially distributed. Therefore, the influence on the logistic regression model and the effect on the association between covariate and the outcome variable are unavailable. See Section 7.1 for more details.

Categorizing a continuous covariate could result in a biased estimate. Chen and colleagues investigated the biases from dichotomizing the age variable on the odds ratio (Chen et al, 2007). They used simulated data to demonstrate that when age is dichotomized, the estimated odds ratio is biased. The bias happens when age is used either as a risk factor or as a confounder in the model of assessing the association via the logistic regression models. Their study pointed out that the biased odds ratio is a result of dichotomization. However, they did not evaluate the impacts from different distributions.

Chapter 3

Parameter Estimation for Categorized Exposure Variables in Meta-Analysis of Disease/Exposure Epidemiological Studies

When a continuous covariate is dichotomized, the estimated odds ratio depends on the cutoff point. Therefore, if two different studies use different cutoff points to assess the association, the results are not comparable. As a consequence, when performing a meta-analysis, the synthesized association based on the categorized status does not represent the real association.

As discussed in Chapter 1, the original distribution of the continuous covariate cannot usually be determined from the published data. Only the number of subjects in each group and the cutoff points are available. Therefore, in order to perform a meta-analysis to summarize the association between covariate and the outcome by using a common cutoff point from studies with different cutoff points, estimating the underlying covariate distribution is the first step. However, there is no existing method can be used to estimate the underlying distribution if a covariate is dichotomized. Therefore, a study with a dichotomized covariate is usually excluded from a meta-analysis.

In order to estimate the underlying distribution for unifying the exposure status, we propose two novel approaches to convert inconsistent cutoff points which are used in a meta-analysis into useful information.

The first approach is based on the linearization of parameters, that is, the method of probability plotting, or the linear model approach. This approach is based on the assumption that the parameters of the underlying distribution can be linearized.

The second approach is based on the method of goodness-of-fit. This method will work even if not all of the parameters can be linearized.

We will use the normal and gamma distributions as a basis for summarizing studies for a meta-analysis.

3.1 Normal Model

In this section, we use a normal distribution as the underlying distribution of a categorized covariate in each study. We start from using the case when each study has only one cutoff point to categorize the covariate. Then the same approach will be extended to allow for both single and multiple cutoff points for categorizing covariates.

3.1.1 Single Cutoff Point in a Study

Let X be the covariate which is distributed as a normal distribution with mean μ and standard deviation σ. The density function of X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

where $-\infty<x<\infty$, $-\infty<\mu<\infty$, $\sigma>0$.

The cumulative density function of X is

$$F(x) = \int_{-\infty}^{x} f(t)dt = \Phi(\frac{x-\mu}{\sigma})$$

where Φ is the cumulative standard normal distribution function.

The cutoff point X_i of the i^{th} study has a corresponding cumulative probability $P_i=F(X_i)$. Let $z_i = \frac{X_i - \mu}{\sigma}$, then $z_i = \Phi^{-1}(P_i)$.

Based on the association, we have $z_i = \frac{X_i - \mu}{\sigma} = \Phi^{-1}(P_i)$. The association between the cutoff point and the cumulative probability can be then expressed as:

$$X_i = \mu + \sigma \times \Phi^{-1}(P_i) \qquad (3.1.1)$$

For notation, we can use C_i to replace X_i to indicate cutoff point, so that (3.1.1) becomes

$$C_i = \mu + \sigma \times \Phi^{-1}(P_i) \qquad (3.1.2)$$

Therefore, when we regress the cutoff points C_i on $\Phi^{-1}(P_i)$, the estimates of the intercept and the slope from the linear model (Equation 3.1.2) are the estimates of the mean and standard deviation of the underlying normal distribution, respectively.

The numbers of subjects in each dichotomized group are usually reported. Let N_{i1} be the number of the subject which have the value less or equal to cutoff point C_i and N_i be the total number of the subject in the i^{th} study. Therefore, let $P_i = N_{i1} / N_i$ and the Equation 3.1.2 can be expressed as:

$$C_i = \mu + \sigma \times \Phi^{-1}(N_{i1} / N_i) + \varepsilon_i \qquad (3.1.3)$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.

3.1.2 Single or Multiple Cutoff Points in a Study

The model described in the previous section can be generalized to estimate the parameters of the underlying normal distribution when the covariate in each study is categorized by any number of cutoff points, if the underlying distribution is a normal distribution with mean μ and standard deviation σ.

Let X be the covariate which is distributed as a normal distribution with mean μ and standard deviation σ. The j^{th} cutoff point C_{ij} of the i^{th} study has a corresponding cumulative probability P_{ij}. Let $z_{ij} = \frac{C_{ij} - \mu}{\sigma}$. Then $z_{ij} = \Phi^{-1}(P_{ij})$ where Φ is the cumulative standard normal distribution function.

Based on the association, we have $z_{ij} = \frac{C_{ij} - \mu}{\sigma} = \Phi^{-1}(P_{ij})$. The association between the cutoff point and the cumulative probability can be then expressed as:

$$C_{ij} = \mu + \sigma \times \Phi^{-1}(P_{ij}) \qquad (3.1.4)$$

Therefore, when we regress the cutoff points C_{ij} on $\Phi^{-1}(P_{ij})$ (the normal deviate of the cumulative probability P_{ij}), the estimates of the intercept and the slope from the linear model (Equation 3.1.4) are the estimates of the mean and standard deviation of the underlying normal distribution, respectively.

The numbers of subjects in each categorized group are usually reported. Let N_{ij} be the number of the subjects which have the value less or equal to cutoff point C_{ij} and N_i be the total number of the subject in the i^{th} study. Thus, $P_{ij} = N_{ij} / N_i$ and the Equation 3.1.4 can be expressed as:

$$C_{ij} = \mu + \sigma \times \Phi^{-1}(N_{ij} / N_i) + \varepsilon_{ij} \qquad (3.1.5)$$

where ε_{ij} follows a normal distribution with mean 0 and standard deviation $\sigma_{\varepsilon_i}^{\ 2}$

We can further generalize the linear model to handle the situation where the covariate in each study follows a normal distribution with its own mean and standard deviation. That is,

$$C_{ij} = \mu_i + \sigma_i \times \Phi^{-1}(N_{ij} / N_i) + \varepsilon_{ij} \qquad (3.1.6)$$

where ε_{ij} follows a normal distribution with mean 0 and standard deviation $\sigma_{\varepsilon_i}^2$, $\mu_i = \mu + m_j$, m_j follows a normal distribution with mean 0 and standard deviation σ_m^2, and $\sigma_i = \sigma + s_i$ where s_i follows a normal distribution with mean 0 and standard deviation σ_s^2. Therefore, the mean and standard deviation for the i^{th} study is μ_i and σ_i, respectively.

3.2 Gamma Model

The gamma distribution is frequently used to model data because its shape and scale parameters can be used flexibly. However, the linear model approach we used for normal distribuiton cannot be applied to the gamma distribution. An alternative appraoch will be discussed next.

3.2.1 Single Cutoff Point in a Study

Let *X* be a random variable with a gamma distribuiton with mean $\alpha\beta$ and variance $\alpha\beta^2$, where the shape parameter is α and the scale parameter is β. The density function of the gamma distribution is

$$f(x) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}$$

where x≥0, α >0, β >0, and Γ is the gamma function which has the formula

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt$$

For the single cutoff point C_i in the i^{th} study, its association with the corresponding cumulative proportion P_i is

$$C_i = F^{-1}(\alpha, \beta; P_i) \qquad (3.2.1)$$

where $F(\alpha, \beta)$ is the cumulative probability of a gamma distribution with shape parameter of α and scale parameter of β.

Standardizaiton was tried to make the parameters and the cumulative proportions independent. However, the gamma distribution cannot be standardized by using the same transformation that was used for the normal distribution.

In order to linearize the association, we use the following property of the gamma distribution.

Property

If X be a random variable with gamma distribuiton of shape parameter α and scale parameter β, $\frac{X}{\beta}$ has a gamma distribution with shape parameter α and scale parameter 1. That is, $\frac{X}{\beta}$ has an incomplete gamma distribution with shape parameter α

Therefore, this association in equation (3.2.1) can be further transformed by using the property of incomplete gamma distribution. That is,

$$\frac{C_i}{\beta} = F^{-1}(\alpha, 1; P_i)$$

$$\Rightarrow \qquad C_i = \beta \times F^{-1}(\alpha, 1; P_i) \qquad (3.2.2)$$

Based on the incomplete gamma distribtion, this associaiton becomes a linear relationship when the shape parameter α is known.

The equation (3.2.2) can be further extended to become a liner model of parameters. That is, the intercept is 0 and the slope corresponds to the scale parameter β, given the shape parameter α. Therefore, the linear regression model can be expressed as

$$C_i = 0 + \beta \times F^{-1}(\alpha, 1; P_i) \qquad (3.2.3)$$

Let N_i be the number of subject in the i^{th} study and N_{i1} be the number of subject in the $_i^{th}$ group which X value is less than or equal to the cutoff point C_i. That is, $P_i=N_{i1}/N_i$. Therefore, equation (3.2.3) becomes

$$C_i = 0 + \beta \times F^{-1}(\alpha, 1; N_{i1} / N_i) + \varepsilon_i \qquad (3.2.4)$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.

However, the shape parameter α cannot be estimated directly from this linear regression model. The shape parameter α needs to be appropriately assigned.

Based on the characteristic that this regression line (3.2.3) or (3.2.4) goes through origin, the criteria used for estimating the shape parameter α is that the shape estimate $\hat{\alpha}$ can result in an intercept which is closest to the origin among all of the possible α estimates. That is,

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{k} | \beta \times F^{-1}(\alpha, 1; P_i) - C_i | \qquad (3.2.5)$$

where C_i is cutoff point of the i^{th} study and P_i is the cumulative probabilities which corresponding to the cutoff point C_i.

3.2.2 Single or Multiple Cutoff Points in a Study

The model described in the previous section can be generalized to estimate the parameters of the underlying gamma distribution when the covariate in each study is categorized by any number of cutoff points, if the underlying distribution is a gamma distribution with shape parameter of α and scale parameter of β.

Let N_{ij} be the number of the subjects which have the value less or equal to cutoff point C_{ij} and N_i be the total number of the subject in the i^{th} study. Thus, $P_{ij} = N_{ij} / N_i$ and the Equation 3.2.3 can be generalized as:

$$C_{ij} = 0 + \beta \times F^{-1}(\alpha, 1; P_{ij}) \qquad (3.2.6)$$

and Equation 3.2.4 can ge generalized as:

$$C_{ij} = 0 + \beta \times F^{-1}(\alpha, 1; N_{ij} / N_i) + \varepsilon_{ij} \qquad (3.2.7)$$

where $\varepsilon_{ij} \sim N(0, \sigma_{\varepsilon i}^{\,2})$.

3.3 Algorithms for Normal Model

Based on the linear model for normal distribution discussed in Section 3.1, the mean and standard deviation of a normal distribution can be estimated by using the linear model approach.

3.3.1 Single Cutoff Point in a Study

In order to take into account the sample size in each group, a weighted linear regression analysis is performed. The weight of the i^{th} study is

$$weight_i = \frac{N_i}{p_i(1-p_i)}$$

which is the inverse of the variance. The weighting provides more weights to the probabilities which are in the tails than to those in the middle.

3.3.2 Single or Multiple Cutoff Points in a Study

When a study has more than one cutoff point, the cumulative probabilities are associated with each other. Therefore, in order to take into account the association within each study when estimating the intercept and the slope from the linear model approach, the analysis can be performed by using the function of linear regression which handles repeated/correlated measurements. In `R` language (R Development Core Team, 2009), the `lmer` function from the `lme4` Package (Bates and Maechler, 2009) can be used to estimate the parameters when takes into account the correlated measurements ($P_{i\cdot}$) within each study.

3.3.3 Covariate Estimation from Categorized Covariates with Similar Distributions

It is possible that all of the studies are sampled from the same underlying covariate distribution. However, it is also possible that the studies are sampled from covariate distributions which have different means and standard deviations.

For example, studies can be sampled from the normal distribution with the same standard deviation but different means. The studies can also be sampled from the normal distribution with different means and different standard deviations.

When all of the studies have dichotomized covariates, the estimation can only work under the assumption that all of them have the same standard deviation but different means. However, when each study has at least 2 cutoff points, the estimation can work under the assumptions that each study has its own mean and standard deviation.

To estimate the parameters assuming a distribution with common parameters across studies, the random effects linear regression method which accommodates repeated measurements can be used. In `R`, the `lmer` function from the `lme4` Package can fit such a model.

When the mean of each study and the common variance of all studies have been estimated, we can estimate the expected numbers of subjects in newly defined categories, or transform the odds ratio of the dichotomized status into the dose-response association. After that, we are able to use existing methods to combine all of the dose-response associations to re-calculate the overall dose-response association.

3.3.4 Property of the Parameter Estimates

The estimates of intercept and slope from the linear regression model are the maximum likelihood estimates. Therefore, the mean and standard deviation estimated from the linear regression model are also the maximum likelihood estimates. As a consequence, the estimated parameters have the properties of maximum likelihood estimates.

3.4 Algorithms for the Gamma Model

Based on the gamma model discussed in 3.2, a conventional linear regression model cannot be used directly because one of the parameters that needs to be specified but indeed needs to be estimated. Therefore, in order to use the linear model approach, this challenge needs to be resolved.

3.4.1 Iteration using Linear Regression Modeling

We propose a numerical iteration algorithm to estimate the shape and scale parameters simulanteously.The algorithm is the following:

1. Find a criteria of acceptable accuracy improvement to the shape estimate, such as 10^{-5}.
2. Assign initial shape estimate.
3. Use the cutoff points and cumulative proportions to perform linear regression analysis to estimate intercept.
4. When the estimated intercept is greater than 0, choose a smaller shape estimate to continue.

5. When the estimated intercept is less than 0, choose a larger shape estimate to continue.
6. Continue until convergence is achieved.

Weighted linear regression can provide improved estimates. When there is only one cutoff point in each study, the weight of the i^{th} study is based on the sample size N_i and the inverse of the product of the cumulative probability and its difference with 1, that is,

$$weight_i = \frac{N_i}{p_i(1-p_i)}$$

The weighting provides more weight to the probabilities which are in the tails of the distribution.

3.4.2 Goodness-of-Fit Approach

Based on the model described previously, the measurement of distance between observed values and expected values of cutoff points can be expressed as the sum of the squares of the distance, that is,

$$Q(\underset{\sim}{\hat{\theta}}; c) = \sum_{i=1}^{k} (c_i - \hat{c}_i)^2$$

$$= \sum_{i=1}^{k} (c_i - F^{-1}(\underset{\sim}{\hat{\theta}}; p_i))^2$$

where $\underset{\sim}{c}$ is the vector of cutoff points, $\underset{\sim}{\hat{\theta}}$ is the vector of parameter estimates,

$p_i = P(x \le c_i)$, and F is the cumulative density function with parameter $\underset{\sim}{\theta}$.

By minimizing this equation, the parameter estimates can be derived. In order to improve the estimation, the weighted distance will be use. That is,

$$Q_w(\underset{\sim}{\hat{\theta}};\underset{\sim}{c}) = \sum_{i=1}^{k} \frac{(c_i - F^{-1}(\underset{\sim}{\hat{\theta}}; p_i))^2}{p_i(1-p_i)}$$

In order to minimize the distance $Q_w(\underset{\sim}{\hat{\theta}};\underset{\sim}{c})$, conventional analytical or numerical approaches are available. For the gamma distribution, a closed form cannot be obtained. Therefore, the parameters may be approximated numerically by using optimization algorithms.

When using the linear association between the cutoff points and their corresponding cumulative probabilities, the criterion is the same as the method of least squares. That is,

$$Q(\underset{\sim}{x};\hat{\theta}) = \sum_{i=1}^{n} (x_i - \hat{x}_i)^2$$

$$= \sum_{i=1}^{n} (x_i - [\hat{\beta}_0 + \hat{\beta}_1 \times \Phi^{-1}(p_i)])^2$$

When weights are used to improve the estimation, the criteria become weighted least squares.

3.5 Covariate Estimation from a Categorized Covariate Containing Excess Zero

From the Examples in Section 1.2.3, we see that the existence of a categorized covariate containing excess zero posts challenges to the data analysis. Therefore, we propose methods for accommodating categorized covariates with excess zeros.

3.5.1 Model

Let $X \geq 0$ be the variable of interest, where X is a mixture of a positive continuous distribution and a degenerate one at zero in a single study. This can also be considered as a zero-inflated distribution. Let π be the proportion of zeros in the population, and θ be the vector of parameters of the continuous variable. The probability density function of X can be expressed as:

$$h_x(x \mid \pi, \theta) = \pi * I(x;0) + \{(1-\pi) * \{1 - I(x;0)\} * f(x \mid \theta)\}$$

where $I(x;0)$ is an indicator function. $I(x;0)=1$ if x=0, and $I(x;0)=0$ if otherwise. The likelihood function is

$$L(\pi, \theta \mid x_i) = \prod_{i=1}^{n} \pi^{I(x_i;0)} (1-\pi)^{1-I(x_i;0)} f(x_i \mid \theta)^{1-I(x_i;0)} .$$

The cumulative density function of $f(X \mid \theta)$ is $F(X \mid \theta)$. When $X>0$, the cumulative proportion p_i of x_i from $f(X \mid \theta)$ can be expressed as $p_i = P(\theta; X \leq x_i) = F(\theta; x_i)$. Therefore, the observed x_i can be expressed as the inverse of the cumulative function, that is, $x_i = F^{-1}(\theta; p_i)$.

Let n be the total number of subjects sampled from the population. Among them, m subjects have measurement of zero. After sorting the measurements of X in the ascending order, these n measurements are classified into k ordered groups based on the defined cutoff points or the specified percentage of the total sample size for each group. Let n_i be the number of subject in the i^{th} group, i=1, …, k. $\sum_{i=1}^{k} n_i = n$. All of the 0s are classified only into the first group. That is, $0 \leq m \leq n_1$.

The boundaries of the i^{th} group are $[\, c_{i-1}, c_i \,]$. When i=1, the lower bound $c_{i-1} = c_0$ =0. When i=k, the upper bound $c_k = \infty$. When only the boundaries (cutoff points) are reported, the density and cumulative density function becomes $f(c_i \mid \theta)$ and $F(c_i \mid \theta)$, respectively, i=1, …, k.

When the outcome of interested is the proportion of zero, π, in the population, the point estimate of the population proportion is $\hat{\pi} = \frac{m}{n}$. If only the subjects with 0 measurements are classified into the first group, (that is, $m = n_1$), then $\hat{\pi} = \frac{m}{n} = \frac{n_1}{n}$. However, m might not be available from the reported group data and thus needs to be estimated. The only available information about m is its range, that is, $0 \le m \le n_1$.

When the first group contains m zeros and $n_1 - m$ measurements of the continuous variable, given that m is unknown, the continuous distribution based on the known numbers of n_2 through n_k is considered as truncated data. Even though the exact number of truncated observations is unknown, the range of this truncated number in the first group is available, that is, $0 \le n_1 - m \le n_1$. Therefore, when m is known, $n_1 - m$ is also known, and vise versa.

The observed probability is calculated based on the number of subjects in each group. Let $P_{o.i}$ be the observed probability of the i^{th} group, $P_{o.i} = \frac{n_i}{n}$. When the existence of m measurements of zero is excluded from calculating the observed probability of the continuous variable, the observed probability in the i^{th} group becomes $P^*_{o.i} = \frac{n^*_i}{n - m}$, where $n^*_1 = n_1 - m$ for i=1, and $n^*_i = n_i$ if i=2, …, k.

When the parameter of this continuous variable θ can be estimated, the expected probability of the i^{th} group based on the points estimate $\hat{\theta}$ is $P_{e.i} = \int_{C_{i-1}}^{C_i} f(X \mid \hat{\theta}) dx$.

3.5.2 Parameter Estimation of a Single Distribution from a Study

The case where π=0 is a special case of the density function $h_x(x \mid \pi, \theta)$. That is, $h_x(x \mid 0, \theta) = f(x \mid \theta)$. Therefore, this problem becomes one of estimating the parameter of a single distribution. We can use the approaches described previously to estimate the parameters based on the assumption of the underlying distribution.

3.5.3 Parameter Estimation of a Mixture Distribution from a Study

Let the mixture distribution contain the value of zero and a positive continuous distribution. By proposing an appropriate assumption for the continuous variable, both the proportion π of zeros and the parameter of the continuous variable can be estimated simultaneously by using the principle of goodness-of-fit.

3.5.3.1 Naïve Goodness-of-Fit Score

To quantify the deviation, a naïve goodness-of-fit score is proposed as the following:

$$G(C, n \mid \hat{\pi}, \hat{\theta}) = \sum_{i=1}^{k} (P_{o.i} - P_{e.i})^2$$

$$= \sum_{i=1}^{k} \left[\frac{n^*_i}{n - n\hat{\pi}} - \int_{C_{i-1}}^{C_i} f(x \mid \hat{\theta}) dx \right]^2$$

Based on the principle of goodness-of-fit, $G(x,n\,|\,\pi,\theta)$ should equal 0 if the data are from the mixture distribution of the parameters of π and θ.

Therefore, when estimating both parameters, $G(x,n\,|\,\pi,\theta)$ should reach the minimum if both estimates are the best estimates. The possible values for $\hat{\pi}=\frac{\hat{m}}{n}$ is the range of m, that is, $0\le m\le n_1$. Therefore, the best estimates are:

$$
\begin{aligned}
(\hat{\pi},\hat{\theta}) &= \arg\min_{0\le m_j\le n_1} G(x,n\,|\,\hat{\pi}_j,\hat{\theta}_j) \\
&= \arg\min_{0\le m_j\le n_1} G(x,n\,|\,\hat{m}_j,\hat{\theta}_j) \\
&= \arg\min_{0\le m_j\le n_1} \sum_{i=1}^{k}(P_{o.ij}-P_{e.ij})^2 \\
&= \arg\min_{0\le m_j\le n_1} \sum_{i=1}^{k}\left(\frac{n^*_i}{n-n\hat{\pi}_j}-\int_{C_{i-1}}^{C_i} f(x\,|\,\hat{\theta}_j)dx\right)^2 \\
&= \arg\min_{0\le m_j\le n_1} \sum_{i=1}^{k}\left(\frac{n^*_i}{n-m_j}-\int_{C_{i-1}}^{C_i} f(x\,|\,\hat{\theta}_j)dx\right)^2
\end{aligned}
$$

Where $m_j=j$, j=0, …, n_1. $\hat{\theta}_j$ is the best estimate when using $m_j=j$ and the values of n_i, j=2, …, k.

The above equation can also be simply expressed by:

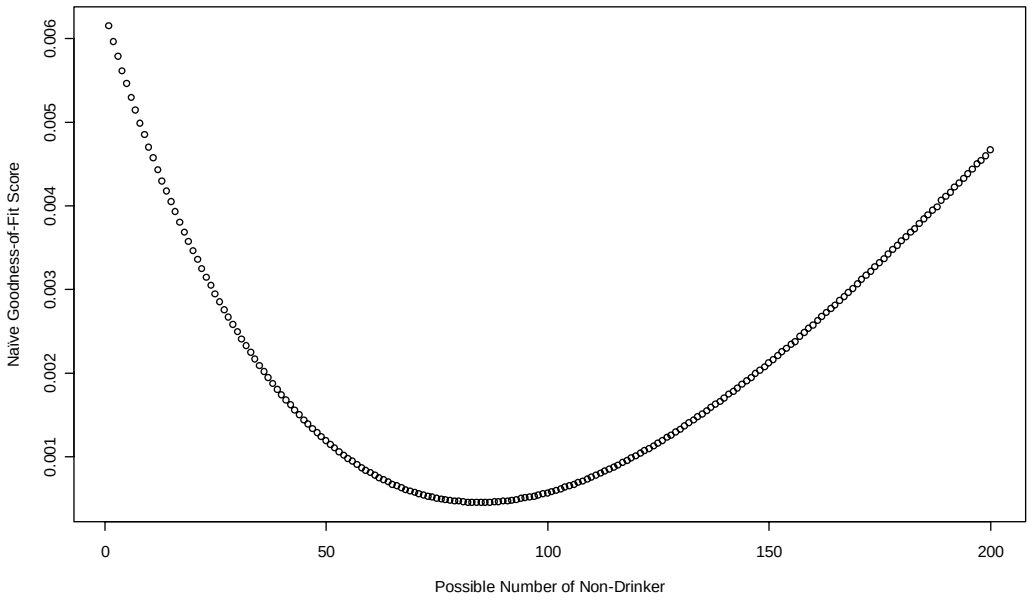
$$
(\hat{m},\hat{\theta})=\arg\min_{(m,\theta)}\sum_{i=1}^{k}\left(\left(\frac{n^*_i}{n-m}-\int_{C_{i-1}}^{C_i} f(x\,|\,\theta)dx\right)\right)^2
$$

The minimum value could be obtained by using a numerical approach.

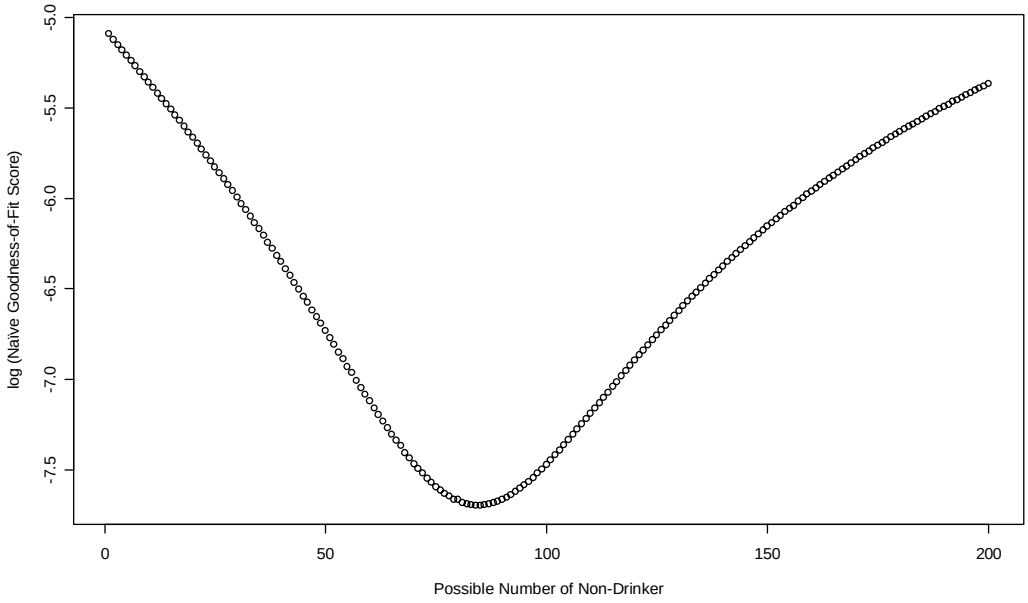
We used the numerical approach for this dissertation research. One example of calculating the goodness-of-fit scores from the possible values of m are shown in Figure 3.5.1. Panel a) shows the possible numbers of non-drinker and their corresponding goodness-of-fit scores. For a better separation of those values, the log transformations of goodness-of-fit scores which presented in Panel a) are shown in Panel b). From these plots, we can see that the lowest goodness-of-fit score corresponds to the number of 85. That is, we used n_1=200 and obtained m=85 from this example.

Figure 3.5.1 Values of a Naïve Goodness-of-Fit Score

a) Original Scale of the score



b) Log transformation of the scores in Panel a)



To take into account the impact from probabilities for improving the estimation, this score can be weighted by the probability of each interval. The weight could be $weight_i = \frac{1}{p_i(1-p_i)}$ which is proportional to the inverse of the variance of a multinomial distribution.

The Pearson Chi-square goodness-of-fit is also a candidate for the goodness-of-fit score. However, the chi-square goodness-of-fit score depends on the sample size. Therefore, when all of the possible sample size will be used to make comparisons, the impact from sample size might jeopardize the estimation.

3.5.4 Global Goodness-of-Fit Score on Testing Distribution Assumption

With very limited information from the grouped data, all of the estimates are based on the distribution assumption. However, the existing goodness-of-fit test can only conclude whether or not the data fit the hypothesized distribution.

We propose using the naïve goodness-of-fit score as a test to test the possible distributions. The distribution with the lowest global goodness-of-score has a higher probability to be the appropriate assumption.

3.6 Generalization for Meta-Analysis

Based on the proposed methods (Sections 3.1 through 3.5), we are able to estimate the parameters of the underlying distribution of the categorized covariate.

For instance, if a covariate follows a normal distribution and multiple cutoff points are used in studies included in a meta-analysis, we can use the mixed-effect weighted linear regression model to estimate the individual mean and standard deviation from each study.

When the underlying covariate distribution has been estimated and performing a meta-analysis based on categorized status is necessary, we can use the estimates and the chosen cutoff point(s) to perform a new meta-analysis. For example, if the purpose of a meta-analysis is to compare high-value group with low-value group, the first step is using the estimates and cutoff point to calculate the expected probability in each group. After that, we use the total number of subjects in each study to find the expected number in each group. Then we can further use the estimated numbers to calculate the association within each study based on this chosen and consistent cutoff point. Then a new meta-analysis can be performed by using those comparable associations.

3.7 Application of Proposed Approaches to Examples

In this section, we present the results from using our methods to analyze examples in the sequence of being described in Section 1.2.

3.7.1 Estimation of the Underlying Distribution under the Gamma Distribution Assumption from Studies with Single Cutoff Point

From the published studies, cathepsin-D level in breast cancer patient is not normally distributed. Foekens and colleagues (1999) studied cathepsin-D level measured by immunoradiometic assay (IMRA) from 2,810 breast cancer patients. They found that the mean (standard deviation) level was 58 (48) pmol mg^{-1} protein. The median level was 47 pmol mg^{-1} and the range was 0-902 pmol mg^{-1}. The authors used a log transformation to make the values closer to a normal distribution.

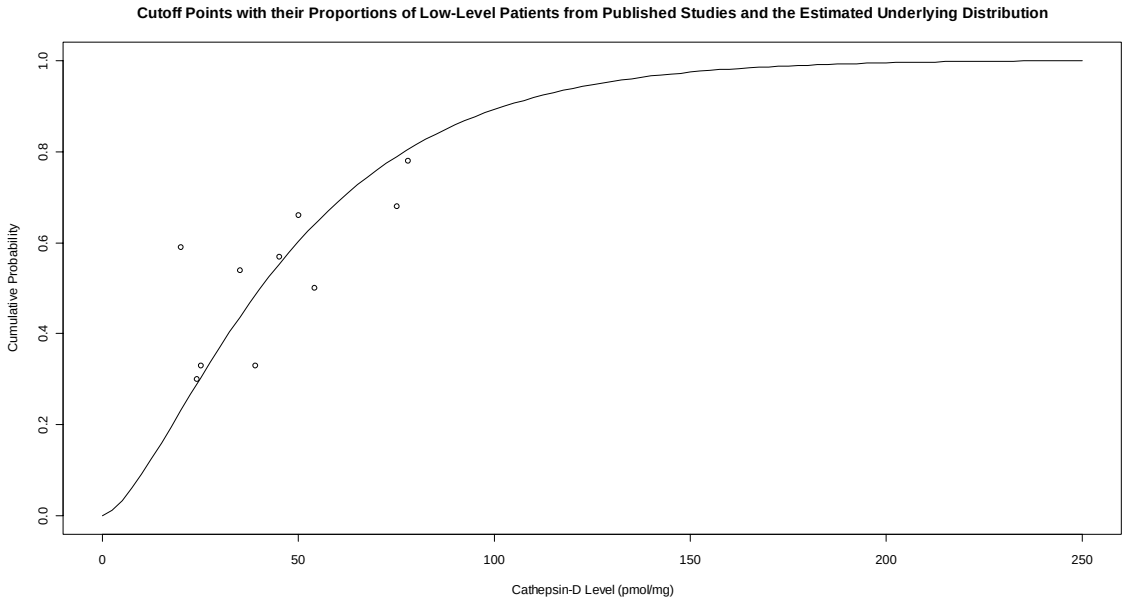
Even though the log transformation is easy to apply, a gamma distribution can be more flexible than the log-normal distribution. We assume that the underlying distribution is gamma distributed for all the studies included in Ferrandina's study (1997) described in Section 1.2.1. By using the proposed approach described in Section 3.5.1.2 and 10 cutoff points from 10 studies, the estimated underlying gamma distribution has a shape parameter of 1.612738 and scale parameter of 31.25377. Therefore, the estimated mean is 50.40414 and estimated standard deviation is 39.69029. By comparing the mean (58) and standard deviation (48) described in the study by Foekens and colleagues (1999), our estimates are reasonable.

Of note, there was a transcriptional error in Table 1 on "Positive (%)" for Kute et (1992): "28" should be "67". Two cutoff points were used by Spyratos (1989). We only

used “45” for our analysis because it was also used in the meta-analysis. The author indicated that the results were unchanged when use anther cutoff point “70”.

The association between the data and estimated distribution is shown in Figure 3.7.1. From the graph, one data point (Pujol 1993) is considered as being a special one. However, the estimated distribution fits data well.

Figure 3.7.1 Data from a Meta-Analysis and the Estimated Underlying Distribution



3.7.2 Estimating the Underlying Distribution under the Normal Distribution Assumption from Studies with Different Number of Cutoff Points

Due to excluding three studies, the result of a meta-analysis assessing the association between BMI and Barrett's esophagus performed by Kamat and colleagues (2009) is of concern. However, our concern can be resolved by using the method developed from this dissertation research.

We start by outlining our approach of performing a new meta-analysis to include all of the studies:

1. We need to estimate the underlying distribution of the studies which used different cutoff points to classify patients into weight status and did not report the mean and standard deviation.
2. We use the estimated parameters to estimate the number of subjects in each category based on the common cutoff points.
3. We can use the number of subject in each category and exposure status to perform meta-analysis.

We assume that BMI follows a normal distribution (Penman et al, 2006). Therefore, we can estimate the underlying distributions by using the cutoff points and their corresponding cumulative probabilities. We assume that each study has its own mean and standard deviation. In order to estimate the individual parameters, we use the weighted linear regression approach with the random effect modeling technique, where the intercept and slope are random. The estimation was performed by using the `lmer` function of the `lme4` package in the `R` language (R Development Core Team, 2009). The regression model was weighted by the sample size of each study and the inverse of the

product of cumulative probability and its difference with 1. For comparison, we also used the weighted linear regression approach to estimate parameters study by study. The estimated mean and standard deviation of each study are shown in Table 3.7.1.

Table 3.7.1 Estimation of mean and standard deviation for underlying normal distribution of studies included in the meta-analysis of BMI and Barrett's esophagus

First Author (Year)	BMI (kg/m^2) category	Reported BMI		Estimated BMI (random effect)		Estimated BMI (individual)	
		Case	Control	Case	Control	Case	Control
Gerson (2002)	≤25 > 25	20.05 (5.7)*	20.1 (3.6)	25.14 (1.75)	21.46 (3.29)	*Cannot estimate*	*Cannot estimate*
Bu (2006)	<22 22–24.9 25–29.9 >30	N/A	N/A	27.39 (4.26)	25.46 (4.53)	27.32 (4.33)	25.43 (4.57)
Ronkainen (2005)	<30 ≥30	N/A	N/A	26.94 (3.41)	25.72 (4.34)	*Cannot estimate*	*Cannot estimate*
Corley (2006)	<30 >30	N/A	N/A	27.47 (3.94)	26.73 (4.56)	*Cannot estimate*	*Cannot estimate*
Johansson (2007)	</23.6 23.6-26.6 >/26.6	N/A	N/A	N/A	N/A	N/A	N/A
Corley (2007)	<25.0 25.0–27.4 27.5–29.9 30.0–34.9 >35.0	29.5 (6.1)	28.9 (5.3)	29.20 (5.56)	28.65 (4.90)	29.23 (5.63)	28.66 (4.89)
Gerson (2007)	< 18.5 18.4–24.9 25–29.9 > 30	28 (5)	27.8 (5.5)	27.33 (5.33)	26.94 (5.42)	27.24 (5.38)	26.98 (5.47)
Stein (2005)	<25 25–30 >30	29.8 (5.6)	28.0 (6.0)	28.59 (4.19)	27.80 (4.92)	29.01 (3.93)	27.84 (5.08)
Veugelers (2006)	< 20 ≥20 and < 25 ≥25 and < 30 ≥30	N/A	N/A	28.38 (4.08)	27.93 (3.89)	28.64 (4.20)	27.84 (3.82)
Edelstein (2007)	<25 25–29.99 ≥30	N/A	N/A	28.79 (4.60)	27.21 (4.67)	28.99 (4.62)	27.23 (4.73)
El-Serag (2005)	<25 25–30 >30	27 (6)	24 (5)	27.49 (4.16)	24.60 (4.40)	26.78 (6.32)	23.60 (5.41)
Smith (2005)	<18.5 18.5-24.9 25-29.9 ≥30	N/A	N/A	28.60 (4.75)	27.62 (4.43)	28.96 (5.13)	27.69 (4.03)

*: Values are expressed as "mean (standard deviation)".

From the random effect model, we were able to estimate the mean and standard deviation of three studies in which only one cutoff point was used to classify the BMI. That is, we are able to add three more studies to perform a meta-analysis. If we use the weighted linear regression approach study-by-study, parameters of those three studies cannot be estimated. In addition, our approach uses data from all of the studies for estimation. Therefore, when comparing the results with those which estimated by using a single study, our approach can improve the efficiency.

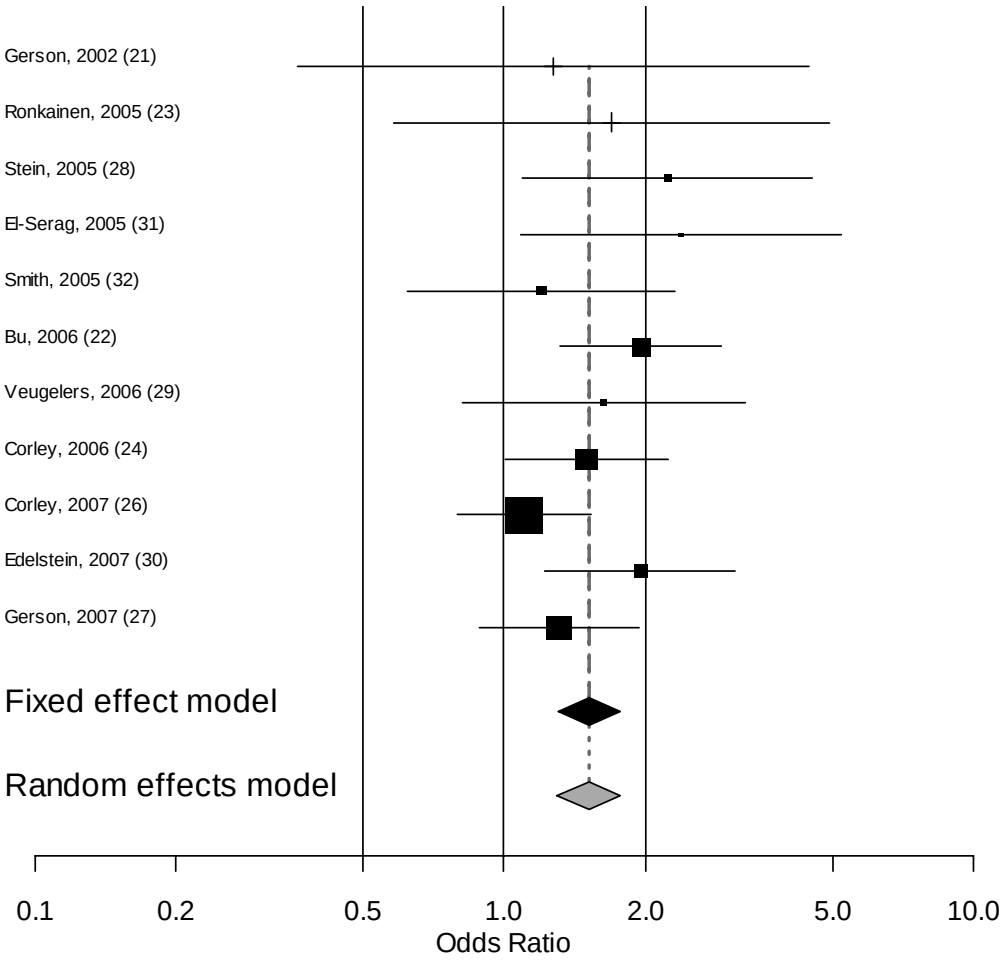
After estimating the parameters, we performed a meta-analysis based on a cutoff point of BMI=25 kg/m^2 from the estimated distribution of each individual study. In order to do so, we performed the sample size re-estimation in the two studies (Ronkainen et al 2005 and Corley et al, 2006) which had only one cutoff point at BMI=30 kg/m^2. Based on the parameters and cutoff points, we calculated the expected probabilities in each BMI category. Then we used the total number of subjects in the case and control groups to calculate the expected number. When the new numbers of subjects were derived, we used the `metabin` function of the `meta` package (Schwarzer, 2009) in the `R` language to perform a meta-analysis and calculate the summary odds ratio. The re-estimated odds ratio from the random effect model is 1.5113 with 95% confidence interval of [1.2965, 1.7617]. The output is shown in Table 3.7.2 on the following page.

We also generated a forest plot by using the `plot.meta` function of the `meta` package in the `R` language to shown the individual odds ratios and the summary odds ratio. The forest plot is shown in Figure 3.7.2.

Table 3.7.2 Output of the improved meta-analysis of the association between BMI and Barrett's esophagus.

```
                          OR            95%-CI %W(fixed) %W(random)
Gerson, 2002 (21)     1.2696  [0.3635; 4.4341]      1.53       1.50
Ronkainen, 2005 (23)  1.6905  [0.5830; 4.9019]      2.03       2.07
Stein, 2005 (28)      2.2281  [1.0966; 4.5272]      4.45       4.68
El-Serag, 2005 (31)   2.3784  [1.0812; 5.2316]      2.94       3.78
Smith, 2005 (32)      1.1978  [0.6220; 2.3066]      5.98       5.47
Bu, 2006 (22)         1.9533  [1.3120; 2.9079]     12.92      14.85
Veugelers, 2006 (29)  1.6275  [0.8157; 3.2471]      4.66       4.93
Corley, 2006 (24)     1.4993  [1.0043; 2.2382]     14.40      14.65
Corley, 2007 (26)     1.1029  [0.7935; 1.5329]     24.91      21.69
Edelstein, 2007 (30)  1.9463  [1.2235; 3.0962]      9.52      10.91
Gerson, 2007 (27)     1.3098  [0.8869; 1.9344]     16.67      15.46

Number of trials combined: 11

                          OR            95%-CI      z  p.value
Fixed effects model   1.5167  [1.3015; 1.7675] 5.3357 < 0.0001
Random effects model  1.5113  [1.2965; 1.7617] 5.2787 < 0.0001

Quantifying heterogeneity:
tau^2 = 0; H = 1 [1; 1.57]; I^2 = 0% [0%; 59.6%]

Test of heterogeneity:
    Q d.f.  p.value
 9.84   10   0.4544

Method: Mantel-Haenszel method
```

Figure 3.7.2 Forest plot of the association between increased BMI (≥25 kg/m^2) and Barrett's esophagus.



Of note, we performed this meta-analysis without considering the impacts from using the estimated underlying distributions. However, in order to take into account the re-estimation, the existing methods (such as re-sampling) could be evaluated, or new approaches should be developed.

In addition, when we used the random-effect regression approach, we assumed that the association between error term and variance estimate are negligible. However, the association between those two terms will be further investigated to improve the estimation.

In summary, our approach allowed estimation of the parameters of the studies which containing only one cutoff point, and re-estimation of the numbers of subject in the newly defined categories. As a result, we were able to add three more studies to the meta-analysis (11 studies vs. 8 studies). Inclusion of those additional studies reduced the standard error of odds ratio from originally reported 0.096 to 0.078. Therefore, our approach contributes to improving the accuracy of estimation.

3.7.3 Estimate the Proportion of zero and Underlying Distribution under Gamma Distribution Assumption from Covariate containing Excess Zeros

When a categorized variable is from a measurement which is a mixture of excess zeros and a continuous covariate, it is a challenge to use the limited but convoluted information to perform meta-analysis. However, our proposed approach provides a useful tool to overcome the constraint.

We used two studies (assessing the association between tea consumption and endometrial cancer, described in Section 1.2.3) to demonstrate the process of estimating a proportion of zeros and parameters of an underlying distribution. The expected amounts of tea consumption were also calculated for each group based on the estimated parameters. For comparison purpose, we assumed gamma and lognormal distributions.

When we used the data from Zheng's study (1996), under the gamma distribution assumption, we estimated that 125 patients (50%=125/249) did not consume tea. However, if we change the underlying distribution assumption to be lognormal, the estimated number became 139 (58%=139/249).

When we used the data from Goodman's study (1997), we estimated that about 48% (=406/844) of the patients did not drink tea, under the gamma distribution assumption. However, if we use the lognormal distribution assumption, the estimated proportion became 45% (=384/844). The results are summarized in Table 3.7.3.

Because of the different results between different underlying distribution assumptions, we will further investigate the use of goodness-of-fit scores on evaluating the assumption.

In summary, by using our approaches, we were able to estimate the proportion of excess zeros and the parameters of the underlying distribution.

Table 3.7.3 Estimated numbers of patient and expected consumption amounts in each tea consumption category

First Author (Year)	Group	Tea Consumption (g)		Gamma Assumption		Lognormal Assumption	
		Lower Limit	Upper Limit	No. of Subject	Mean Amount (g)	No. of Subject	Mean Amount (g)
Zheng (1996)	1	0	0	*125	0	*139	0
	2	0	33.86	*16	16.77	*2	25.83
	3	33.86	237	63	121.32	63	130.86
	4	237	474	29	335.55	29	331.70
	5	474	Inf	16	16.77	16	25.83
Goodman (1997)	1	0	0	*406	0	*384	0
	2	0	34	*16	21.16	*38	20.92
	3	34	237	211	135.69	211	117.83
	4	237	Inf	211	442.91	211	931.86
Jain (2000)	1	0	0	139	0	139	0
	2	0	250	215	100.45	215	119.15
	3	250	500	92	360.94	92	356.27
	4	500	Inf	106	921.73	106	1230.05

*: Estimated number of patients. The reported number is the sum of the estimated numbers in Groups 1 and 2.

3.8 Conclusions

We have proposed linear model approach that was the weighted linear regression model for estimating the mean and standard deviation of a dichotomized normal distribution by regressing the cutoff points on the normal deviates of the cumulative proportions of subjects under the cutoff points from different studies with inconsistent cutoff points. This approach works not only for the studies with dichotomized covariates, but also works for studies with different number of categories.

By using this approach, we can apply the random effect modeling techniques to improve the estimation. As shown in Section 3.7.2, we used one published meta-analysis to demonstrate that our approach not only can summarize the common association by using re-estimated association but also can add more studies to be in a meta-analysis and result in a more accurate result.

When the underlying distribution is gamma, by applying the property of the incomplete gamma distribution, we can use the linear model approach via numerical iteration. We also used this approach to estimate the underlying gamma distribution based on 10 studies in which one cutoff point was used to dichotomize the status of a biomarker.

The linear model approach was associated with the goodness-of-fit approach. We further used the goodness-of-fit approach to estimate the proportion of excess zeros in a mixture distribution.

Chapter 4

Bias and Efficiency: Simulations

When estimating covariate parameters from grouped data, the characteristics of a study might impact the estimation. Those characteristics may also impact the estimation when using all of the studies included in a meta-analysis. In order to assess how the characteristics influence the estimation, we conducted simulation studies based on the scenario that included studies have only single cutoff point or more than one cutoff point.

4.1 Covariate Estimation from Dichotomized Distributions

We evaluated the impacts on covariate estimation from the number of studies, the numbers of subjects in each included study, the characteristics of the underlying distribution, for a range of chosen cutoff points, and using mean or median.

4.1.1 Impact from Number of Studies on Parameter Estimation

We performed simulation studies to evaluate the impact from the number of studies on the parameter estimation. The efficiency of estimation was evaluated by using the relative efficiency (RE), which is the variance of estimates from the raw data divided by the variance of estimates from the weighted linear regression approach.

We allowed the number of studies to range from 2 to 30. Two underlying distributions were used: normal distributions with mean of 100 and standard deviations of 10 and 15. The number of subjects in each study was 1,000. Each condition was performed 1,000 times. The means of the estimates and standard deviation estimates were

calculated and compared with the assigned mean. The standard deviations of the mean and standard deviation estimates were calculated and used to compare with the standard deviations of the estimates from using the raw data.

The graphical results are shown in Figure 4.1.1 and Figure 4.1.2 on the following page. The numerical results are shown in the Tables A.1 through A.4 in Appendix A. From the results, we found that the mean estimate was robust with three or more studies. The mean of the mean estimates are similar between both approaches and close to the mean parameter. The relative efficiency of the mean estimates increases with the number of studies. There was a significant jump from 5 studies (0.5569) to 6 studies (0.6840). After that, the relative efficiency is about 72% with the highest value of 0.7657.

When we compare the results from using different values of the standard deviation, we find no differences in the association between RE and number of studies.

The results for the standard deviation estimates are analogous to those from the mean estimate. However, the relative efficiencies are much smaller than the mean estimates. The maximum RE is 0.2470.

From the results, we found that the number of studies impacts the relative efficiency when estimate the mean and standard deviation from dichotomized studies. However, the variation of a normal distribution does not impact the relative efficiency.

Figure 4.1.1 Relative efficiency of the mean estimates

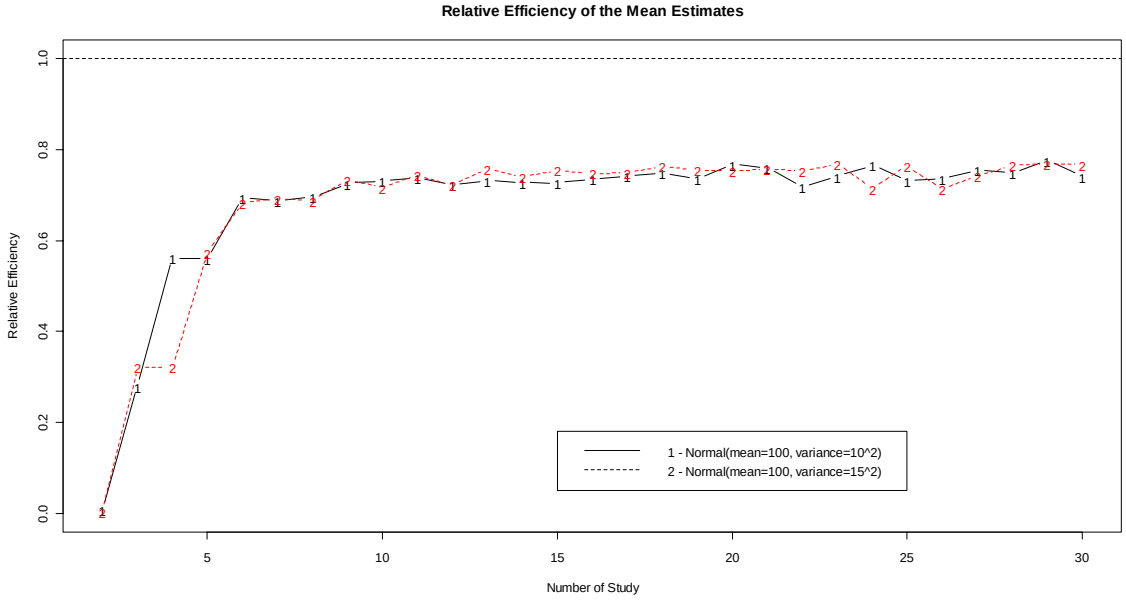
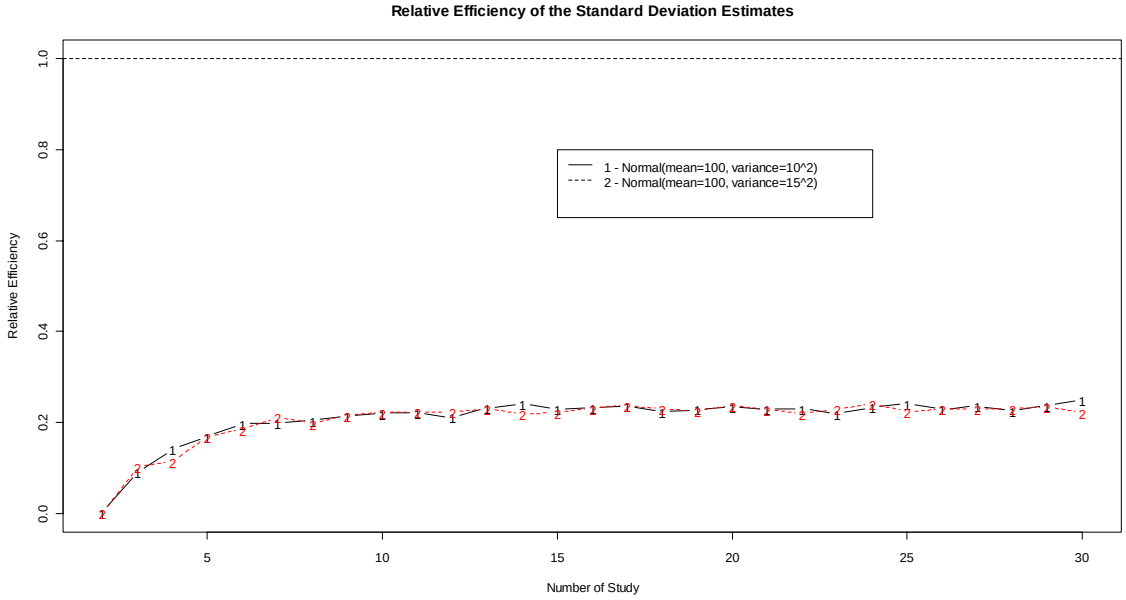


Figure 4.1.2 Relative efficiency of the standard deviation estimates



4.1.2 Impact of the Number of Subjects on the Parameter Estimation

We performed simulation studies to evaluate the impact of the sample size in each study on the parameter estimation. The efficiency of estimation was evaluated by using the relative efficiency, which is the variance of estimates from the raw data divided by the variance of estimates from the weighted linear regression approach.

We used the numbers of subjects in each study as 100, 1,000 and 10,000. The numbers of studies ranged from 2 to 30. Two underlying distributions were used: normal distributions with mean of 100 and standard deviations of 10 and 15. Simulations for each condition were performed 1,000 times. The means of mean and standard deviation estimates were calculated and compared with the assigned mean. The standard deviation of the mean and standard deviation estimates were calculated and used to compare with the standard deviations of the estimates from using the raw data. The graphical results from standard deviation of 10 are shown in Figure 4.1.3 and Figure 4.1.4 on the following pages. The numerical results are shown in Tables A.1 through A.10 in Appendix A.

The mean estimates from the weighted linear regression approach are robust. The results are similar to the estimates calculated from the raw data. The RE's increase with the number of studies. Overall, the REs are similar after the number of studies reached 6. The sample size in each study did not show significant impact to the REs.

The standard deviation estimates showed the same trend as the mean estimates. However, the REs are significantly lower than the REs of the mean estimates. The sample size in each study did not show significant impact to the REs.

When we compared the results between different standard deviation values (10 vs. 15), the results are similar.

Figure 4.1.3 Relative efficiency of the mean estimates

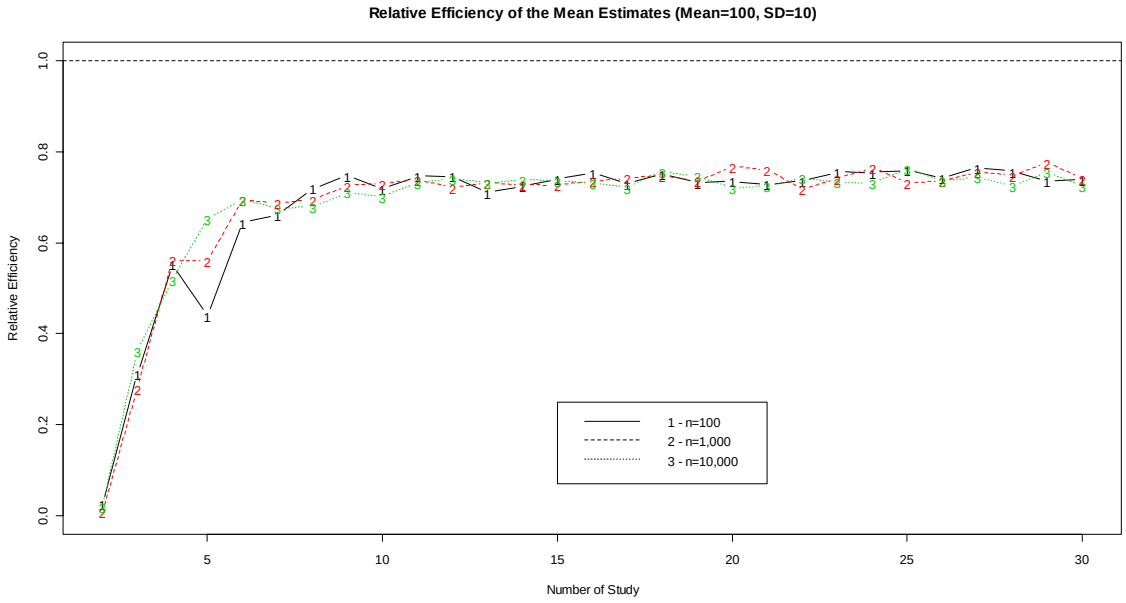
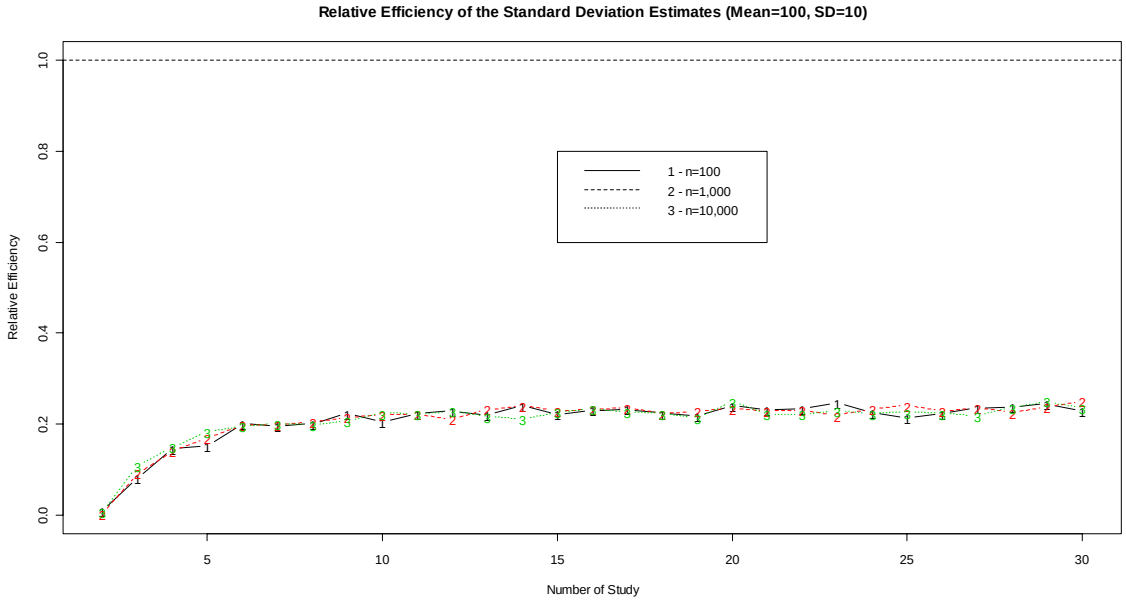


Figure 4.1.4 Relative efficiency of the standard deviation estimates



4.1.3 Impact of the Distribution on the Parameter Estimation

We performed simulation studies to evaluate the impact of the standard deviation on the parameter estimation. The efficiency of the estimation was evaluated by using the relative efficiency, which is the variance of estimates from the raw data divided by the variance of estimates from the weighted linear regression approach.

Two underlying distributions were used: normal distributions with mean of 100 and standard deviations of 10 and 15. We used 10,000 subjects in each study. The numbers of studies ranged from 2 to 30. Each simulation was performed 1,000 times. The means of the mean and standard deviation estimates were calculated and compared with the assigned mean. The standard deviation of the mean and standard deviation estimates were calculated and used to compare with the standard deviations of the estimates from using the raw data. The graphic results are shown in Figure 4.1.5 and Figure 4.1.6 on the following page. The numerical results are shown in Tables A.7 through A.10 in Appendix A.

The mean estimates from the weighted linear regression approach are robust when different values of standard deviation were used. The results are similar to the estimates calculated from the raw data. The relative efficiencies from both standard deviation values are similar. They all increase with the number of studies used for estimating parameters.

Figure 4.1.5 Relative efficiency of the mean estimates

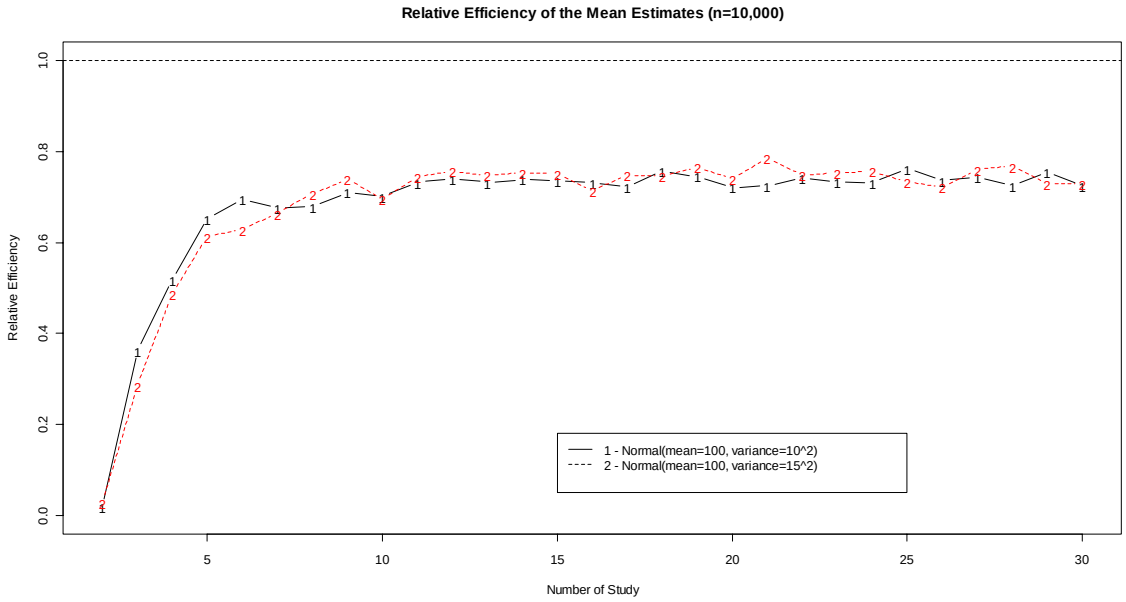
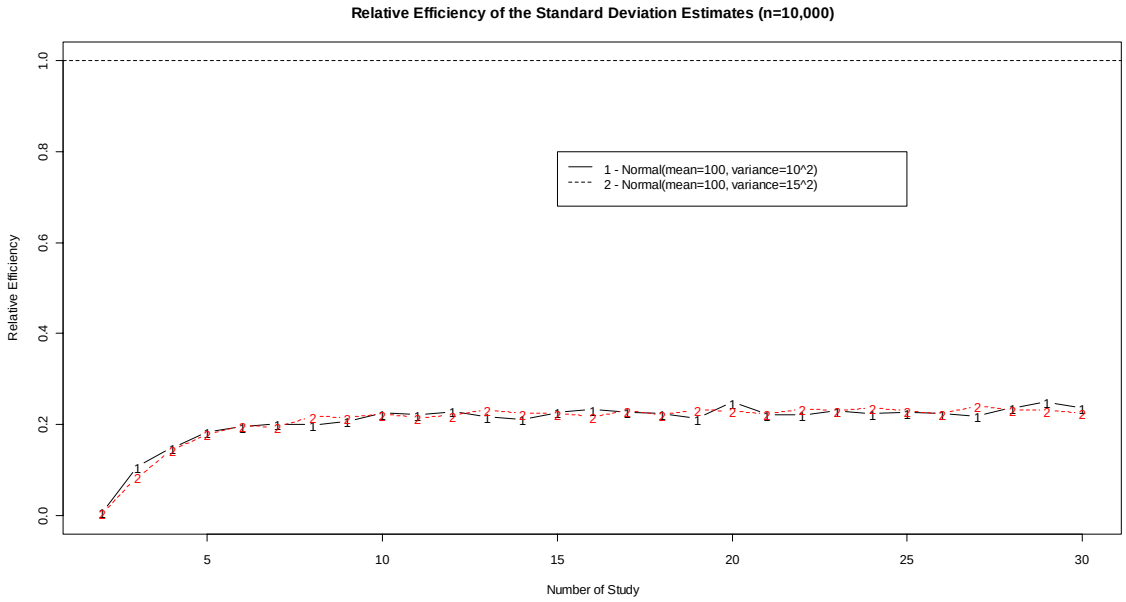


Figure 4.1.6 Relative efficiency of the standard deviation estimates



4.1.4 Impact of the Range of Cutoff Points on the Parameter Estimation

We performed simulation studies to evaluate the impact of the range of cutoff points on the parameter estimation. As before, the efficiency of estimation was evaluated by using the relative efficiency.

We used two different ranges of the cutoff point: one from 20% to 80%; and one from 35% to 65%, which is half the variation of the previous one. The numbers of subjects in each study were 1,000 or 10,000. The numbers of studies ranged from 2 to 30. We used normal distributions with mean of 100 and standard deviation of 15. Each simulation was performed 1,000 times. The means of the mean and standard deviation estimates were calculated and compared with the assigned mean. The standard deviation of the mean and standard deviation estimates were calculated and used to compare with the standard deviations of the estimates from using the raw data. The graphical results are shown in Figures 4.1.7 through Figure 4.1.10 on the next pages. The numerical results are shown in Tables A.11 through A.14 in Appendix A.

The mean estimates were similar when a different range of cutoff points were used. However, when the cutoff points had a wider range (20% to 80%), the mean estimates had a smaller RE than the cutoff points which had narrower range (35% to 65%).

However, the REs of the standard deviation estimates showed a reverse association. That is, when the cutoff points were from wider range, the REs of the standard deviation estimates are larger.

Different sample sizes did not show significant difference on the trends.

Figure 4.1.7 Relative efficiency of the mean estimates

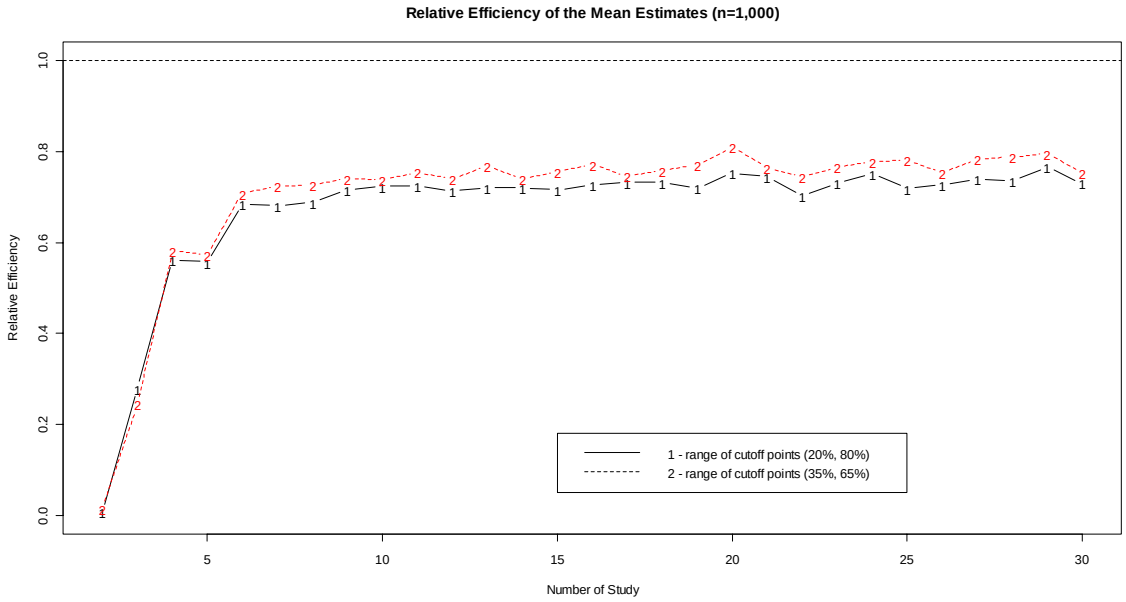


Figure 4.1.8 Relative efficiency of the standard deviation estimates

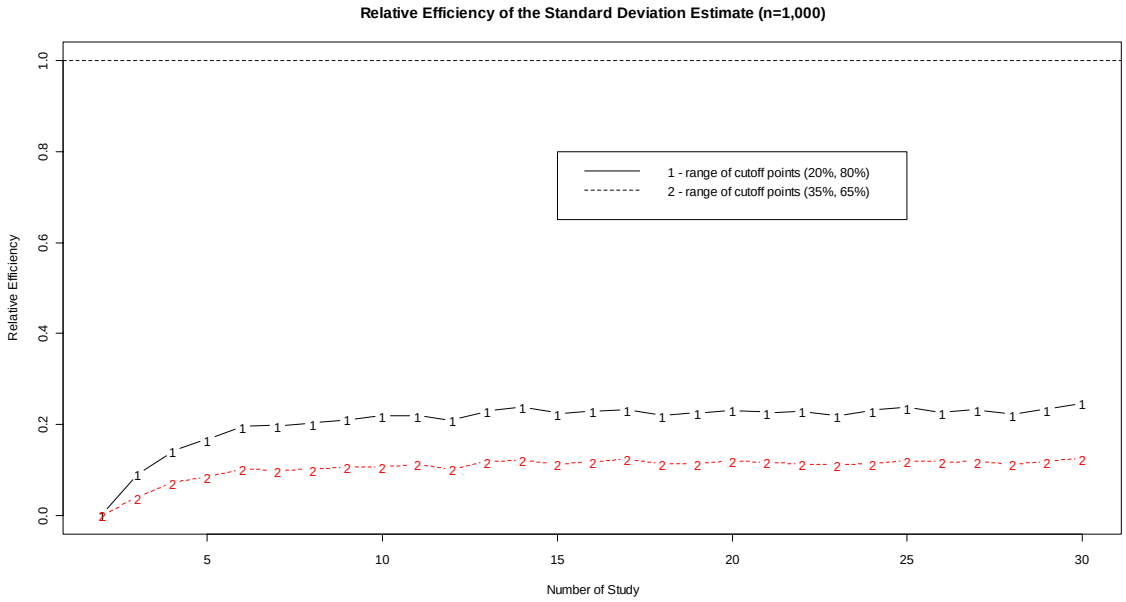


Figure 4.1.9 Relative efficiency of the mean estimates

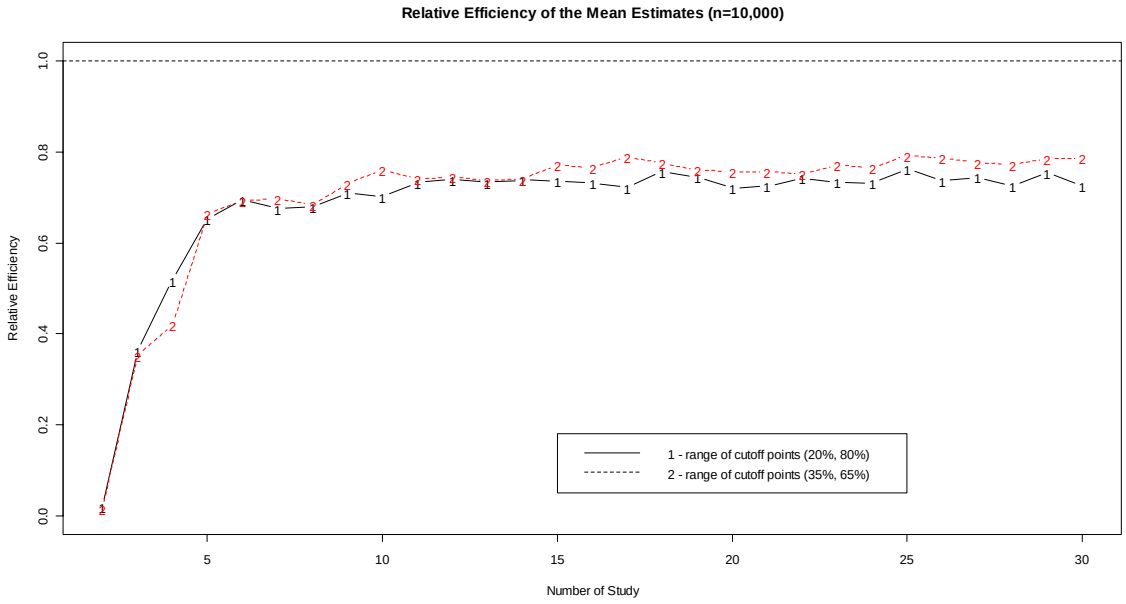
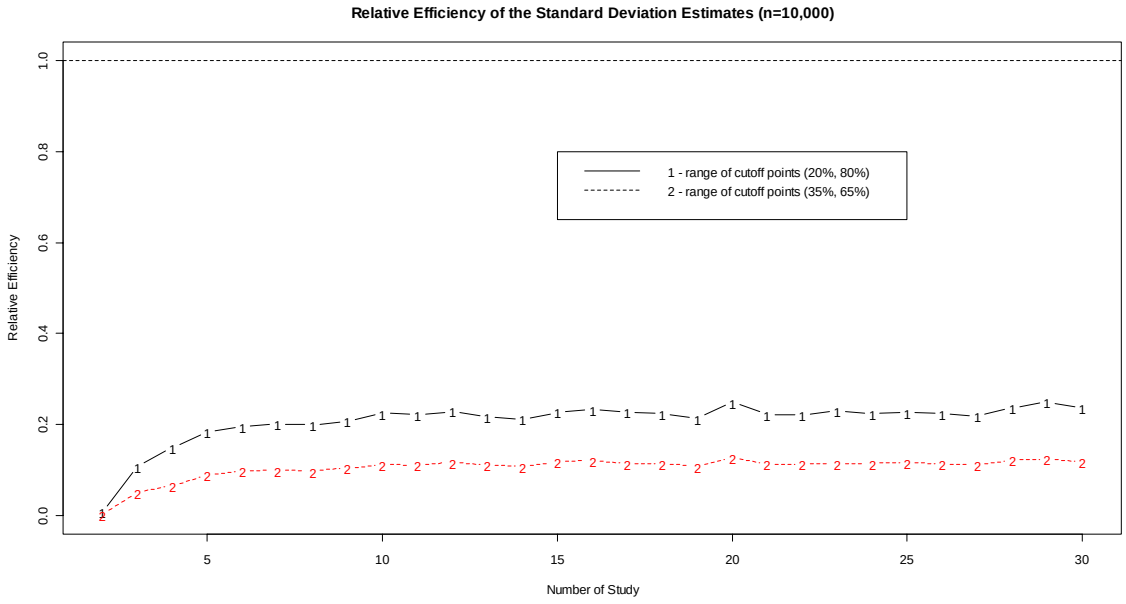


Figure 4.1.10 Relative efficiency of the standard deviation estimates



4.1.5 Using Median or Mean as Cutoff Point on the Parameter Estimation

It is possible that all of the studies included in a meta-analysis use the median or mean as the cutoff point to dichotomize the continuous covariate. Therefore, we performed simulation studies to evaluate the robustness and efficiency of estimation by using the sampling distribution of the median and the mean.

We chose a normal distribution with mean of 100 and standard deviation of 10 as the continuous covariate and generated 10,000 data points for each study. The median and mean were identified as the cutoff point for dichotomizing the covariate. We used the mean of all of the cutoff points in each simulation as the mean estimate. For a given number of studies included in a simulation, we summarize the sampling distribution of the mean estimate. In the simulations, the number of studies ranged from 2 to 30. We performed 1,000 simulations based on each condition. The RE was calculated by using the variance of mean estimate based on the mean-cutoff point divided by the variance of mean estimate based on the median-cutoff point. The graphical results are shown in Figure 4.1.11 and Figure 4.1.12 on the next page. The numerical results are shown in Table

When either the median or mean was used as the cutoff point, the mean of the sampling distribution is close to the assigned parameter 100. They are similar across all ranges of the number of studies included in a simulation. The variances of the mean estimates decrease with the number of studies included. However, the REs did not change with the numbers of studies.

The estimated standard deviation from the mean-cutoff point is close to the assigned standard deviation parameter of 10 when the number of study is larger than 6. However,

the estimated standard deviation from the median-cutoff point is larger than 10, and larger than 12 when the number of studies larger than 6. The REs did not change with the numbers of studies.

Figure 4.1.11 Relative efficiency of the mean estimates

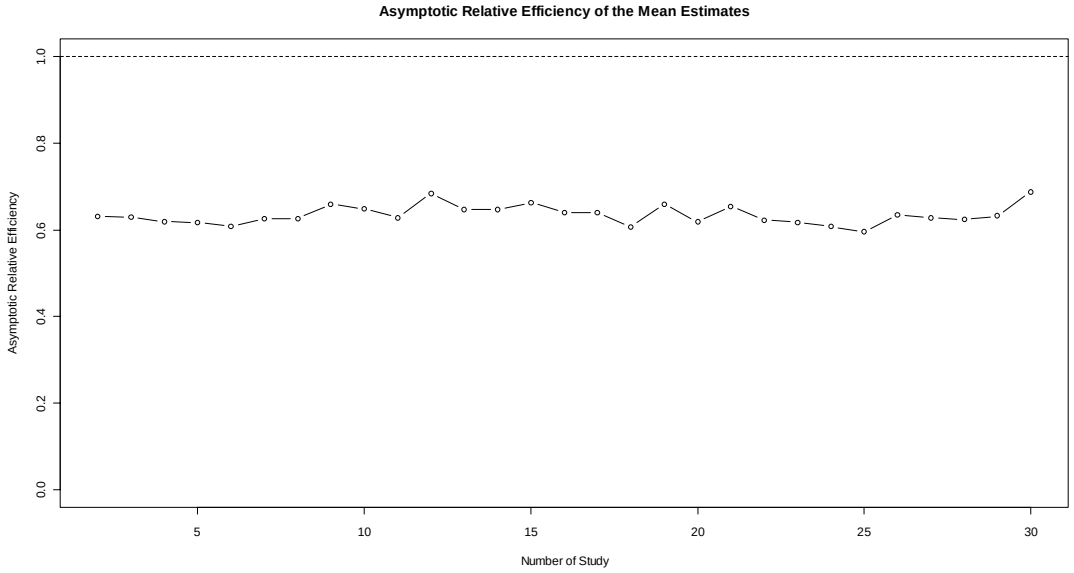
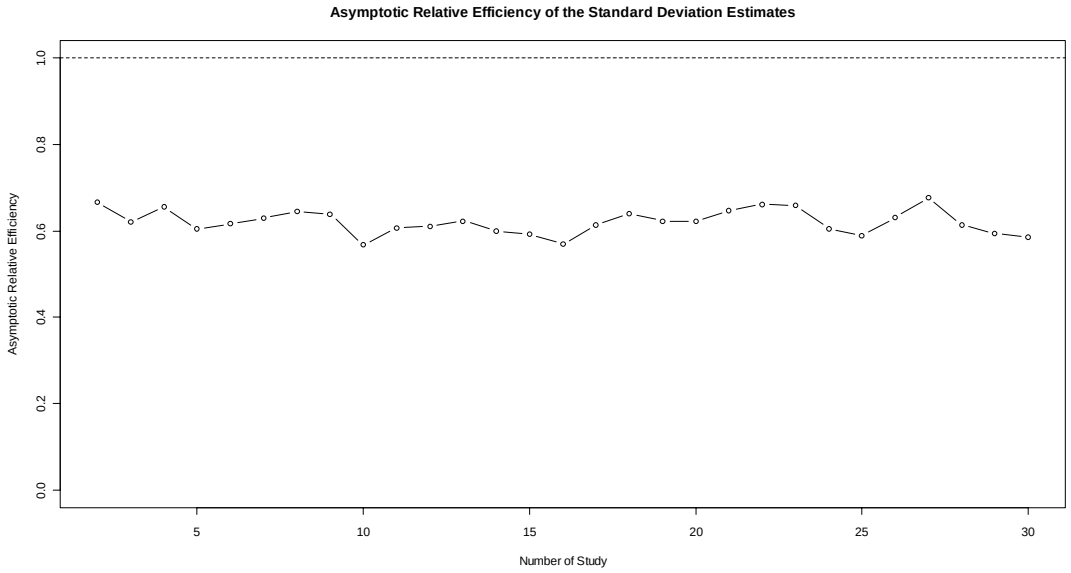


Figure 4.1.12 Relative efficiency of the standard deviation estimates



4.2 Covariate Estimation from Categorized Distributions

The impacts from different numbers of cutoff points in a study, and the numbers of cutoff points and the number of studies use for analysis are discussed.

4.2.1 Impacts from Number of Cutoff Points in a Study on Estimation

It is known that the estimates can be improved when the number of group increases. However, little is known about the improvement by using the Chêne and Thompson approach (Chêne and Thompson, 1996).

We performed simulation studies to evaluate the improvement. Two normal distributions with mean of 100 and standard deviations of 10 and 15 were chosen. The sample size of a study was 1,000. The numbers of group were ranged from 3 to 30. For each group, equal numbers of subjects were used. That is, the cutoff points were tercile, quartile, quintile, etc from the generated data. Each condition was repeated 1,000 times. The relative efficiency was calculated by using the variance of the estimate from raw data divided by the variance of estimate from Chêne and Thompson approach. The graphic results are shown in Figures 4.2.1 through 4.2.4. The numeric results are shown in Table A.17 through A.24 in Appendix A.

From the results, we found that the REs of mean estimates are similar between two normal distributions which have the same mean but different standard deviations. There was no significant difference between moderate and large sample sizes. There were no significant differences between the numbers of groups.

Similar associations between REs of standard deviation estimates and the sample size, the number of group and the distribution were also found.

Figure 4.2.1 Relative efficiency of the mean estimates

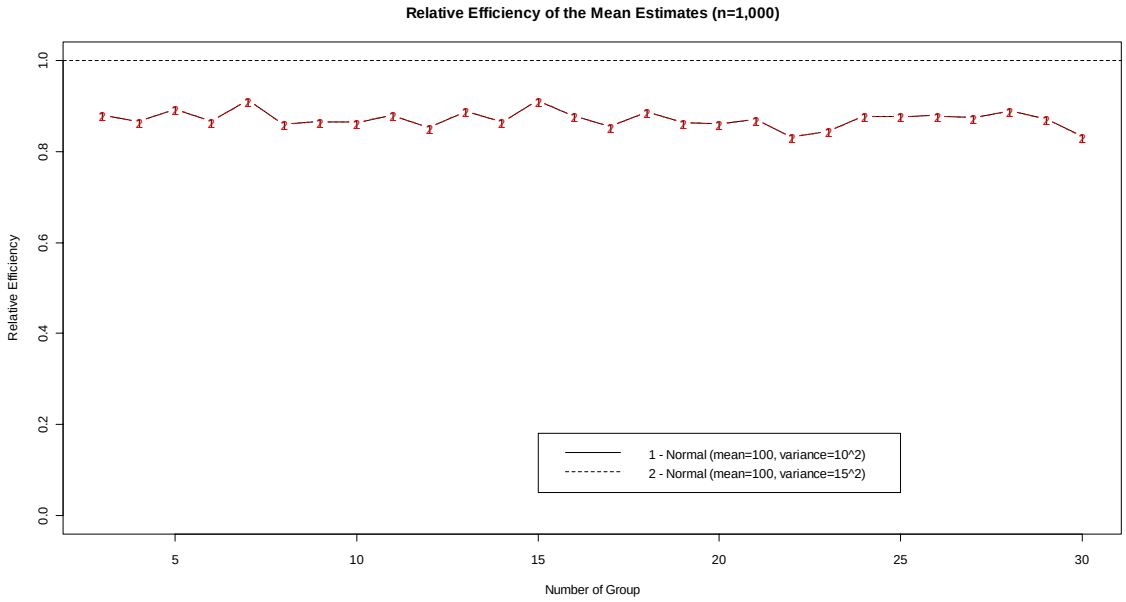


Figure 4.2.2 Relative efficiency of the standard deviation estimates

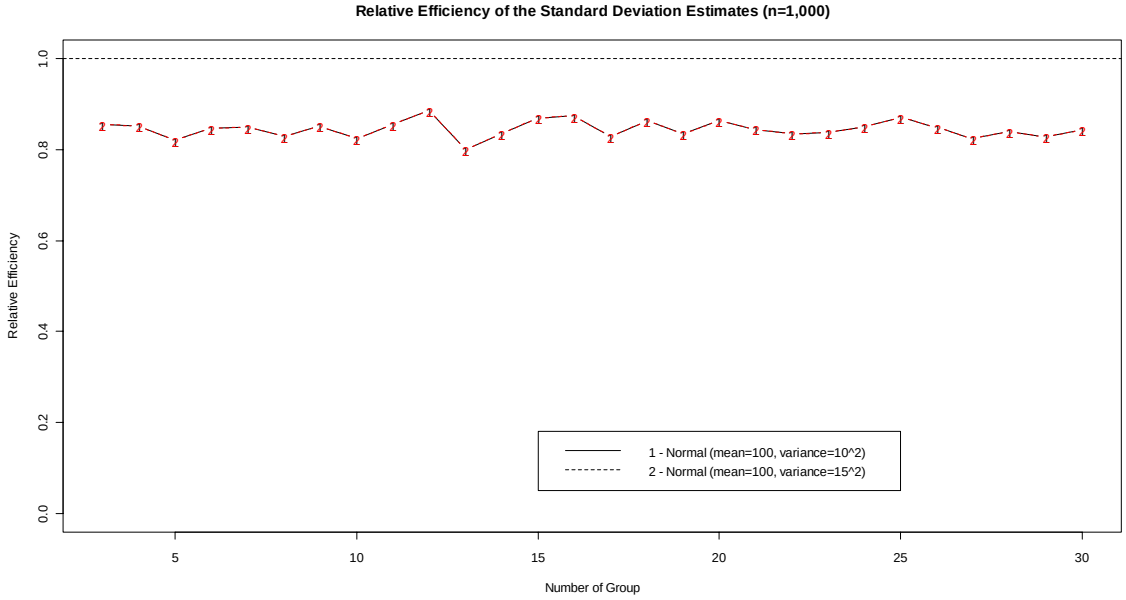


Figure 4.2.3 Relative efficiency of the mean estimates

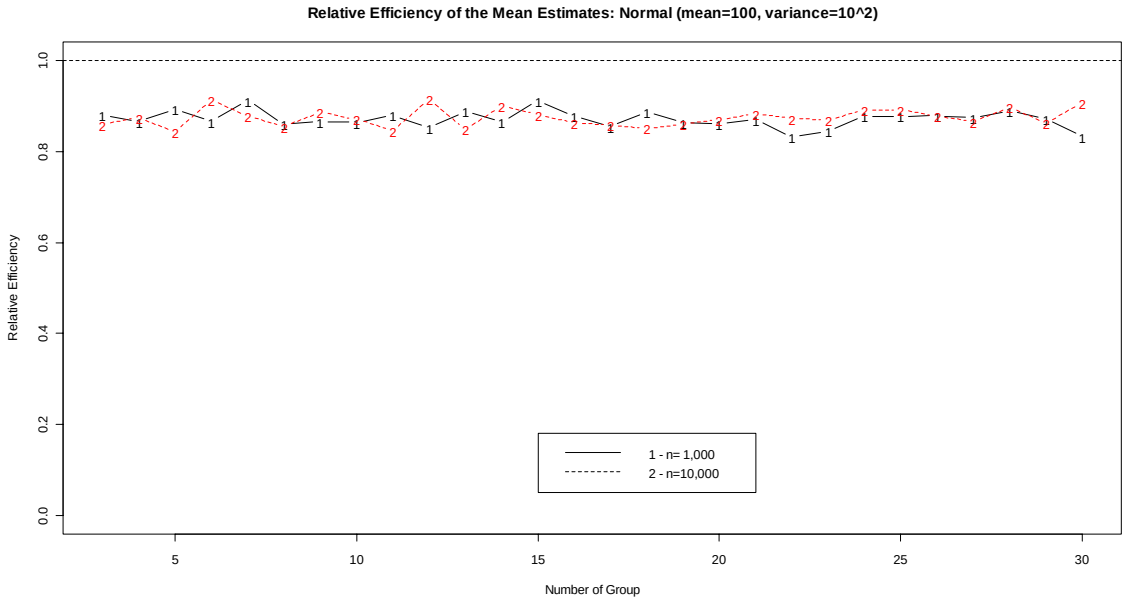
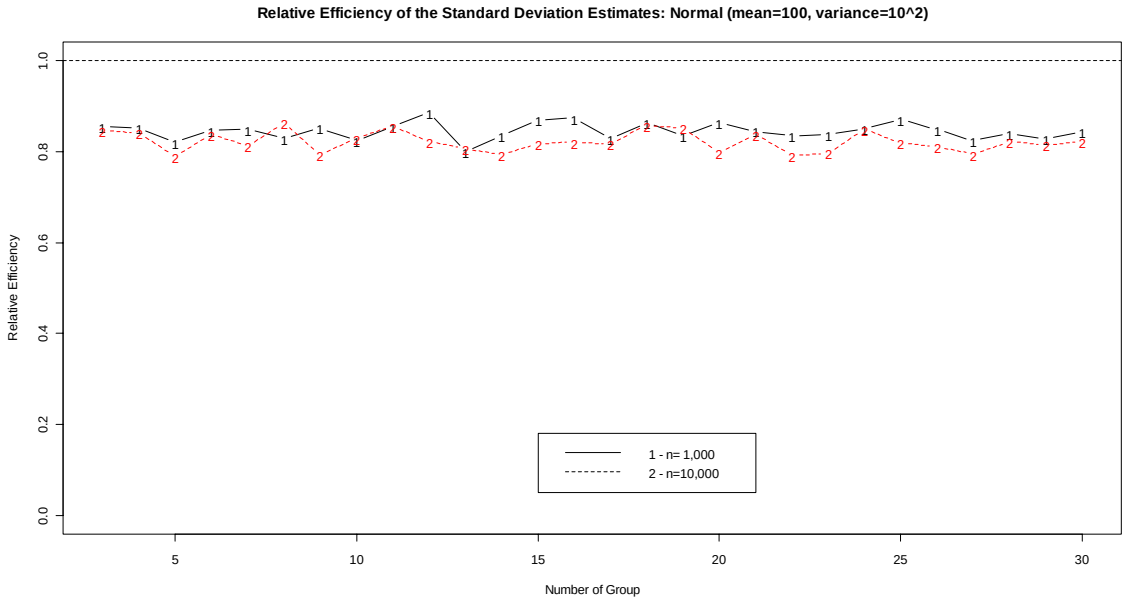


Figure 4.2.4 Relative efficiency of the standard deviation estimates



4.2.2 Impact of Number of Cutoff Points and Number of Studies

When estimating the distribution of a covariate from many studies, using all of the raw data should generate the best estimate. However, when only the grouped data are available, using the weighted linear regression approach can still yield useful information. However, there will be some loss of robustness and efficiency when using grouped data resulting from assuming common parameters.

We performed simulation studies to evaluate the improvement. A normal distribution with mean of 100 and standard deviation of 10 was used. The sample size of a study was 1,000 or 10,000. The numbers of groups ranged from 3 to 10. For each group, equal numbers of subjects were used. That is, the cutoff points were tercile, quartile, quintile, etc from the generated data. The number of studies ranged from 2 to 30. Each condition was repeated 1,000 times.

The `lmer` function from `lme4` Package of `R` language was used to perform the weighted linear regression approach. Due to the nature of correlated measurements in each study, the mixed effect model was used to assign the random effects for the intercept and the slope. The relative efficiency was calculated by using the variance of the estimate from the raw data divided by the variance of the fixed-effect estimates from the weighted linear regression approach. The graphical results are shown Figures 4.2.5 through 4.2.8 on the following pages. The numerical results are shown in Tables A.25 through A.32 in Appendix A.

Given the same number of studies, the REs of the mean estimates increase with the numbers of groups (from 3 to 7 groups) in each study. However, there is no significant

difference between 7 groups and 10 groups. Given the same number of group in each study, the REs did not change significantly with the increase in the number of studies.

When we compare the REs of the mean estimates for the sample size of 1,000 vs. 10,000 in each study, we did not find any significant difference.

The REs of the standard deviation estimates increase with the number of groups (from 3 to 10 groups) in each study. However, as with the mean estimate, the REs did not change significantly with the increase of numbers of studies.

We also compare the REs of the standard deviation estimates between different sample sizes, and the results are similar.

Figure 4.2.5 Relative efficiency of mean estimates

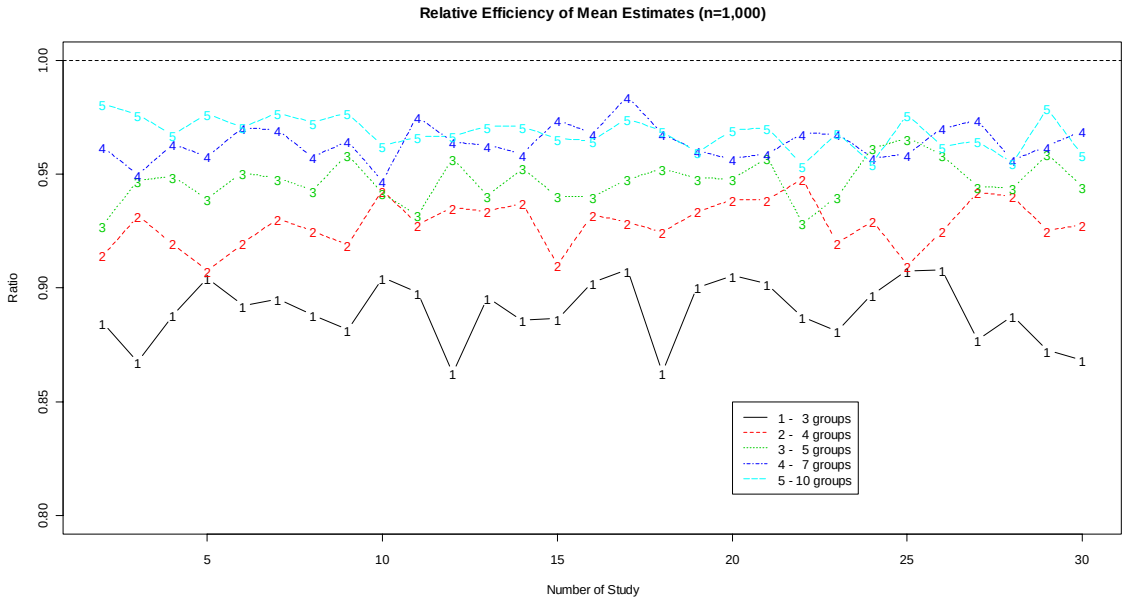


Figure 4.2.6 Relative efficiency of mean estimates

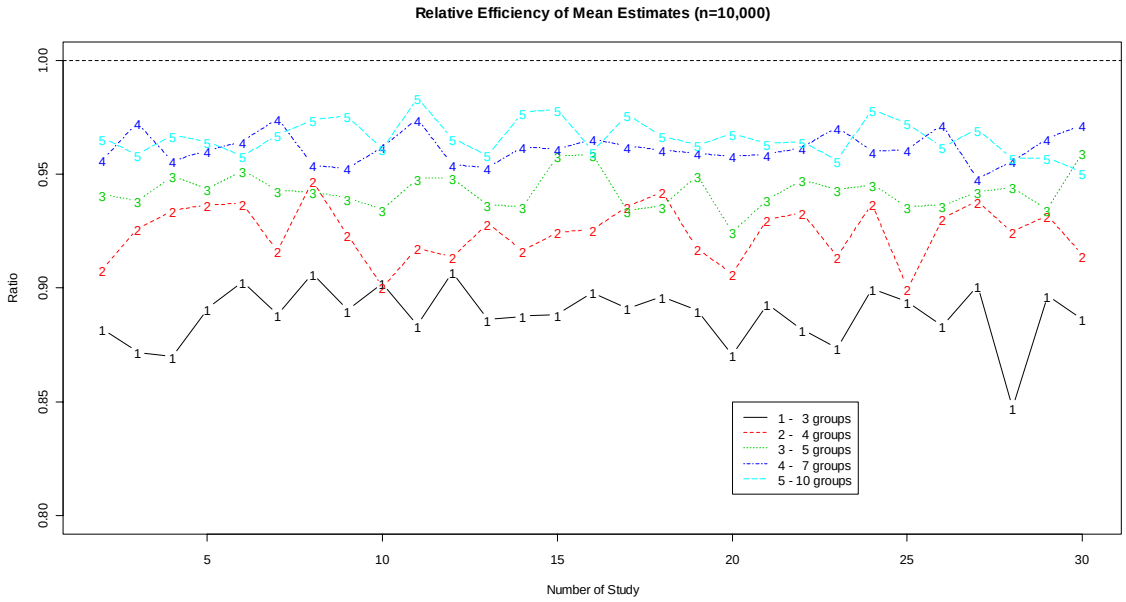


Figure 4.2.7 Relative efficiency of standard deviation estimates

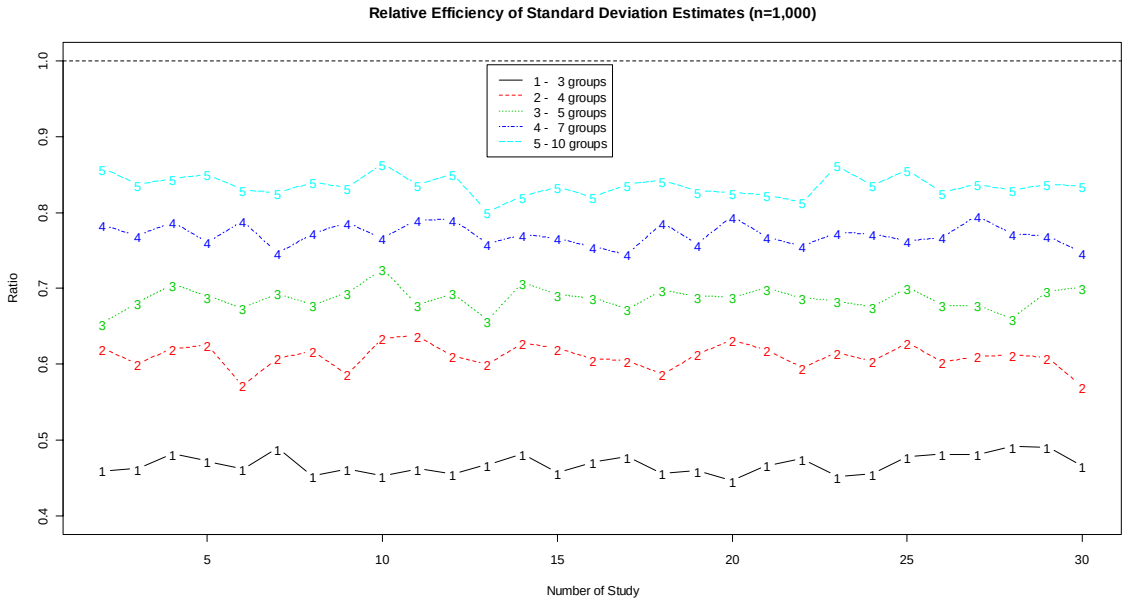
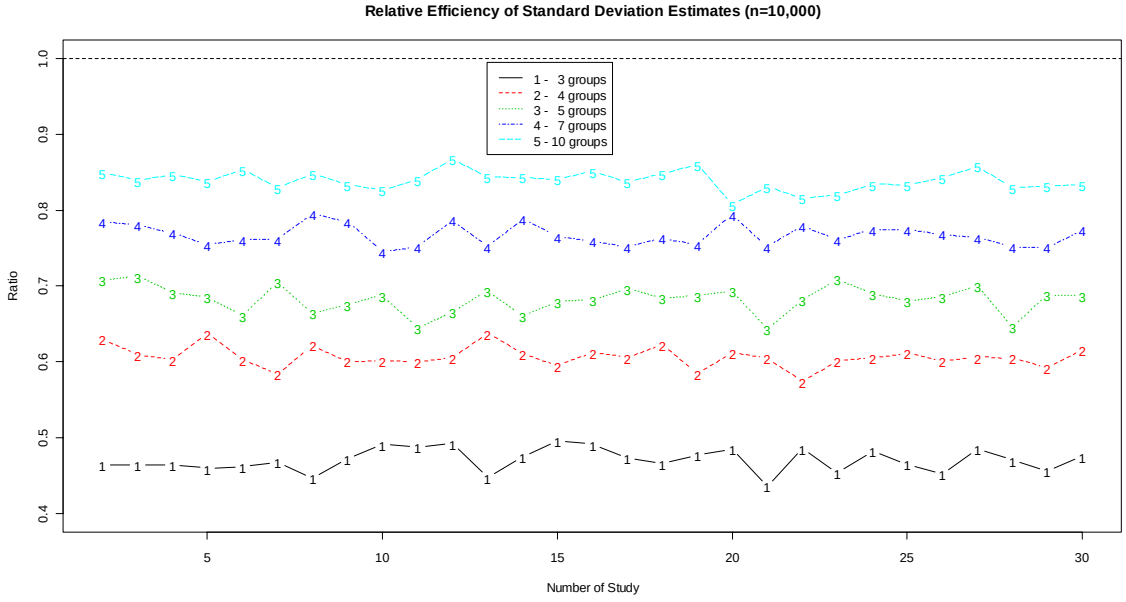


Figure 4.2.8 Relative efficiency of standard deviation estimates



4.3 Conclusions

In summary, our simulation studies demonstrate that sample size within a study does not impact the relative efficiency. The number of cutoff points and the number of studies included for parameter estimation impact the efficiency but not the robustness. When estimating parameters from studies containing a single cutoff point, the gain of efficiency increases rapidly if the number of studies less than 6. But the efficiency does not show improvement when the number of studies is more than 10. In addition, using mean as the cutoff point has a better efficiency than using median as the cutoff point.

Chapter 5

Computing Asymptotic Relative Efficiency Using the Multinomial Distribution

When a continuous variable is categorized into groups, it is natural to consider these groups as multinomial distribution. Therefore, the parameters of the underlying distribution can be estimated by using the multinomial maximum likelihood approach.

5.1 Model

Let $X_1, X_2, \ldots X_n$ be independently and identically distributed (iid) with density function $f_\xi(X_i)$ and cumulative distribution function $F_\xi(X_i)$, i=1, …, n. ξ is the scale parameter of this distribution, and n is the number of observation.

Let $X^*_{ij} = I(C_{j-1} \le X_i < C_j)$, j=1, …, m, C_j and C_{j-1} is the upper and lower bound of the j^{th} interval, respectively; C_0 is the minimum value of the distribution; m is the number of group. That is,

$$X^*_{ij} = \begin{cases} 1 & \text{if } C_{j-1} \le X_i < C_j \\ 0 & \text{otherwise} \end{cases}$$

Therefore,

$$P_\xi(X^*_{ij} = 1) = \int_{C_{j-1}}^{C_j} f_\xi(X)\, dX = F_\xi(C_j) - F_\xi(C_{j-1}) \qquad (5.1.1)$$

When only X^*_{ij}'s are available, we can only use X^*_{ij}'s and (C_j, C_{j-1}) to estimate ξ.

Given data X^*_{ij} from m intervals, the likelihood function of ξ:

$$L_c(\xi \mid X^*_{ij}) = \prod_{j=1}^{m} P_\xi(X^*_{ij} = 1)^{n_j} \qquad (5.1.2)$$

where $n = \sum_{j=1}^{m} n_j$, n_j is the number of observation in the j^{th} group.

Let $l_c(\xi) = l_c(\xi \mid X^*_{ij}) = \log L_c(\xi \mid X^*_{ij})$ and $P_j = P_\xi(X^*_{ij} = 1)$

Therefore,

$$l_c(\xi) = \log L_c(\xi \mid X^*_{ij})$$

$$= log\left[\prod_{j=1}^{m} P_\xi(X^*_{ij} = 1)^{n_j}\right] = \sum_{j=1}^{m} n_j log P_\xi(X^*_{ij} = 1)$$

$$= \sum_{j=1}^{m} n_j log P_j$$

Let $l_c'(\xi) = \frac{\partial}{\partial \xi} l_c(\xi)$

$$l_c''(\xi) = \frac{\partial^2}{\partial \xi^2} l_c(\xi)$$

$$P_j' = \frac{\partial}{\partial \xi} P_j$$

$$P_j'' = \frac{\partial^2}{\partial \xi^2} P_j$$

Then

$$l_c'(\xi) = \sum_{j=1}^{m} n_j \frac{P_j'}{P_j} \qquad (5.1.3)$$

$$l_c''(\xi) = \sum_{j=1}^{m} n_j \left(\frac{P_j'}{P_j}\right)'$$

$$= \sum_{j=1}^{m} n_j \frac{P_j P_j'' - (P_j')^2}{P_j^2} \qquad (5.1.4)$$

To find the maximum likelihood estimate $\hat{\xi}_c$, we can solve the equation (5.1.3) equals to 0, that is, the solution to $l_c''(\xi)=0$ is the maximum likelihood estimate.

Based on the second derivative (5.1.4), the expected Fisher information

$$I(\xi_c) = -E[l_c''(\xi)] = Var^{-1}(\hat{\xi}_c)$$

Therefore, the variance of the maximum likelihood estimate from the grouped data can be calculated by using the expected Fisher information, that is,

$$Var(\hat{\xi}_c) = I^{-1}(\xi_c) \tag{5.1.5}$$

When the data in original values are available, the likelihood function is

$$L(\xi \mid X_i) = \prod_{i=1}^{n} f_{\xi}(X_i)$$

By using the procedures described previously, we can find the maximum likelihood estimate $\hat{\xi}$ by solving the first derivative of log-likelihood function $l(\xi) = logL(\xi \mid X_i)$. We can also calculate the variance of the maximum likelihood estimate from the original data by using the second derivative of the log-likelihood function, that is,

$$Var(\hat{\xi}) = I^{-1}(\xi) = -E[l''(\xi)]^{-1}$$

After calculating the variances of the maximum likelihood estimate from both original and grouped data, the asymptotic relative efficiency (ARE) can be calculated as

$$\text{ARE} = \frac{Var(\hat{\xi})}{Var(\hat{\xi}_c)} = \frac{I(\xi_c)}{I(\xi)} = \frac{-E[l_c''(\xi)]}{-E[l''(\xi)]} \tag{5.1.6}$$

5.2 Maximum Likelihood Estimation of A Categorized Exponential Distribution

When X is a random variable following exponential distribution with parameter ξ, the density function can be written as

$$f(x) = P_{\xi}(X) = \xi e^{-\xi x}$$

and the cumulative distribution $F(X) = F_{\xi}(X) = 1 - e^{-\xi X}$.

Therefore, for the j^{th} interval,

$$\begin{aligned}
P_j &= F(C_j) - F(C_{j-1}) \\
&= 1 - e^{-\xi C_j} - (1 - e^{-\xi C_{j-1}}) = e^{-\xi C_{j-1}} - e^{-\xi C_j} \\
P_j' &= - C_{j-1} e^{-\xi C_{j-1}} + C_j e^{-\xi C_j} \\
P_j'' &= (- C_{j-1} e^{-\xi C_{j-1}} + C_j e^{-\xi C_j})' \\
&= C_{j-1}^2 e^{-\xi C_{j-1}} + C_j^2 e^{-\xi C_j} \\
P_j P_j'' &= (e^{-\xi C_{j-1}} - e^{-\xi C_j})(C_{j-1}^2 e^{-\xi C_{j-1}} + C_j^2 e^{-\xi C_j}) \\
&= C_{j-1}^2 e^{-2\xi C_{j-1}} - C_{j-1}^2 e^{-\xi C_j} e^{-\xi C_{j-1}} - C_j^2 e^{-\xi C_j} e^{-\xi C_{j-1}} + C_j^2 e^{-2\xi C_j} \\
(P_j')^2 &= (- C_{j-1} e^{-\xi C_{j-1}} + C_j e^{-\xi C_j})^2 \\
&= C_{j-1}^2 e^{-2\xi C_{j-1}} - 2 C_{j-1} C_j e^{-\xi C_{j-1}} e^{-\xi C_j} + C_j^2 e^{-2\xi C_j} \\
P_j P_j'' - (P_j')^2 &= C_{j-1}^2 e^{-2\xi C_{j-1}} - C_{j-1}^2 e^{-\xi C_j} e^{-\xi C_{j-1}} - C_j^2 e^{-\xi C_j} e^{-\xi C_{j-1}} + C_j^2 e^{-2\xi C_j} \\
&\qquad - C_{j-1}^2 e^{-2\xi C_{j-1}} + 2 C_{j-1} C_j e^{-\xi C_{j-1}} e^{-\xi C_j} - C_j^2 e^{-2\xi C_j} \\
&= - e^{-\xi(C_j + C_{j-1})} (C_j^2 - 2 C_{j-1} C_j + C_{j-1}^2) \\
&= - e^{-\xi(C_j + C_{j-1})} (C_j - C_{j-1})^2
\end{aligned}$$

$$\frac{P_j P_j'' - (P_j')^2}{P_j^2} = \frac{-e^{-\xi(C_j + C_{j-1})}(C_j - C_{j-1})^2}{(e^{-\xi C_{j-1}} - e^{-\xi C_j})^2}$$

Consequently, based on (5.1.4),

$$l_c''(\xi) = \sum_{j=1}^{m} n_j \frac{P_j P_j'' - (P_j')^2}{P_j^2}$$

$$= \sum_{j=1}^{m} n_j \frac{-e^{-\xi(C_j + C_{j-1})}(C_j - C_{j-1})^2}{(e^{-\xi C_{j-1}} - e^{-\xi C_j})^2}$$

To calculate the expected Fisher Information,

$$I(\xi_c) = -E[l_c''(\xi)]$$

$$= -E\left[\sum_{j=1}^{m} n_j \frac{-e^{-\xi(C_j + C_{j-1})}(C_j - C_{j-1})^2}{(e^{-\xi C_{j-1}} - e^{-\xi C_j})^2}\right]$$

$$= -\sum_{j=1}^{m} \frac{-e^{-\xi(C_j + C_{j-1})}(C_j - C_{j-1})^2}{(e^{-\xi C_{j-1}} - e^{-\xi C_j})^2} E[n_j] \qquad \text{(because } C_j\text{'s are known)}$$

$$= -\sum_{j=1}^{m} \frac{-e^{-\xi(C_j + C_{j-1})}(C_j - C_{j-1})^2}{(e^{-\xi C_{j-1}} - e^{-\xi C_j})^2} E[nP_j]$$

$$= -\sum_{j=1}^{m} \frac{-e^{-\xi(C_j + C_{j-1})}(C_j - C_{j-1})^2}{(e^{-\xi C_{j-1}} - e^{-\xi C_j})^2} E[n\{F(C_j) - F(C_{j-1})\}]$$

$$= -\sum_{j=1}^{m} \frac{-e^{-\xi(C_j + C_{j-1})}(C_j - C_{j-1})^2}{(e^{-\xi C_{j-1}} - e^{-\xi C_j})^2} E[n\{e^{-\xi C_{j-1}} - e^{-\xi C_j}\}]$$

$$= -n\sum_{j=1}^{m} \frac{-e^{-\xi(C_j + C_{j-1})}(C_j - C_{j-1})^2}{(e^{-\xi C_{j-1}} - e^{-\xi C_j})^2} E[e^{-\xi C_{j-1}} - e^{-\xi C_j}]$$

$$= -n\sum_{j=1}^{m} \frac{-e^{-\xi(C_j + C_{j-1})}(C_j - C_{j-1})^2}{(e^{-\xi C_{j-1}} - e^{-\xi C_j})^2} \int_{C_{j-1}}^{C_j} (e^{-\xi C_{j-1}} - e^{-\xi C_j}) \frac{\xi e^{-\xi x}}{e^{-\xi C_{j-1}} - e^{-\xi C_j}} dx$$

$$= -n \sum_{j=1}^{m} \frac{-e^{-\xi(C_j+C_{j-1})}(C_j - C_{j-1})^2}{(e^{-\xi C_{j-1}} - e^{-\xi C_j})^2} \int_{C_{j-1}}^{C_j} \xi e^{-\xi x} dx$$

$$= -n\,\xi \sum_{j=1}^{m} \frac{-e^{-\xi(C_j+C_{j-1})}(C_j - C_{j-1})^2}{(e^{-\xi C_{j-1}} - e^{-\xi C_j})^2} \int_{C_{j-1}}^{C_j} e^{-\xi x} dx$$

$$= -n\,\xi \sum_{j=1}^{m} \frac{-e^{-\xi(C_j+C_{j-1})}(C_j - C_{j-1})^2}{(e^{-\xi C_{j-1}} - e^{-\xi C_j})^2} \left[-\frac{1}{\xi} e^{-\xi x} \Big|_{C_{j-1}}^{C_j} \right]$$

$$= -n\,\xi \sum_{j=1}^{m} \frac{-e^{-\xi(C_j+C_{j-1})}(C_j - C_{j-1})^2}{(e^{-\xi C_{j-1}} - e^{-\xi C_j})^2} \left[-\frac{1}{\xi} (e^{-\xi C_j} - e^{-\xi C_{j-1}}) \right]$$

("-" and ξ can be cancelled out)

$$= n \sum_{j=1}^{m} \frac{-e^{-\xi(C_j+C_{j-1})}(C_j - C_{j-1})^2}{(e^{-\xi C_{j-1}} - e^{-\xi C_j})^2} (e^{-\xi C_j} - e^{-\xi C_{j-1}})$$

[$-(e^{-\xi C_j} - e^{-\xi C_{j-1}})$ can be cancelled out]

$$= n \sum_{j=1}^{m} \frac{e^{-\xi(C_j+C_{j-1})}(C_j - C_{j-1})^2}{e^{-\xi C_{j-1}} - e^{-\xi C_j}}$$

In summary, the expected Fisher Information from n observations with m groups:

$$I(\xi_c) = n \sum_{j=1}^{m} \frac{e^{-\xi(C_j+C_{j-1})}(C_j - C_{j-1})^2}{e^{-\xi C_{j-1}} - e^{-\xi C_j}}$$

When $m = 2$

When data from an exponential distribution are categorized into two groups, we have m=2, $C_2 = \infty$, C_0=0.

The expected Fisher information from this type of categorization,

$$I(\xi_c) = n \sum_{j=1}^{2} \frac{e^{-\xi(C_j + C_{j-1})}(C_j - C_{j-1})^2}{e^{-\xi C_{j-1}} - e^{-\xi C_j}}$$

$$= n \left[\frac{e^{-\xi(C_1 + C_0)}(C_1 - C_0)^2}{e^{-\xi C_0} - e^{-\xi C_1}} + \frac{e^{-\xi(C_2 + C_1)}(C_2 - C_1)^2}{e^{-\xi C_1} - e^{-\xi C_2}} \right]$$

$$\left(\text{because } \frac{e^{-\xi(C_2 + C_1)}(C_2 - C_1)^2}{e^{-\xi C_1} - e^{-\xi C_2}} \to 0,\ e^{-\xi C_0} = 1 \right)$$

$$= n \frac{e^{-\xi C_1} C_1^{\,2}}{1 - e^{-\xi C_1}}$$

Because $P_1 = 1 - e^{-\xi C_1}$

$$P_2 = 1 - P_1 = e^{-\xi C_1}$$

$$C_1^{\,2} = \left[-\frac{1}{\xi} log(1 - P_1) \right]^2$$

Therefore,

$$I(\xi_c) = n \frac{e^{-\xi C_1} C_1^{\,2}}{1 - e^{-\xi C_1}}$$

$$= n \frac{P_2}{P_1} \left[-\frac{1}{\xi} \log(1 - P_1) \right]^2$$

$$= n \frac{P_2}{P_1} \frac{1}{\xi^2} \left[\log(1 - P_1) \right]^2$$

$$= \frac{n}{\xi^2} \frac{P_2}{P_1} [log(1 - P_1)]^2$$

By using the original observed data without categorization, the expected Fisher information

$$I(\xi) = \frac{n}{\xi^2}.$$

Therefore, the asymptotic relative efficiency (ARE) from categorizing observed data in continuous scale into two groups:

$$\text{ARE} = \frac{Var(\hat{\xi})}{Var(\hat{\xi}_c)} = \frac{I(\xi_c)}{I(\xi)} = \frac{\frac{n}{\xi^2}\frac{P_2}{P_1}[log(1-P_1)]^2}{\frac{n}{\xi^2}}$$

$$= \frac{P_2}{P_1}[log(1-P_1)]^2$$

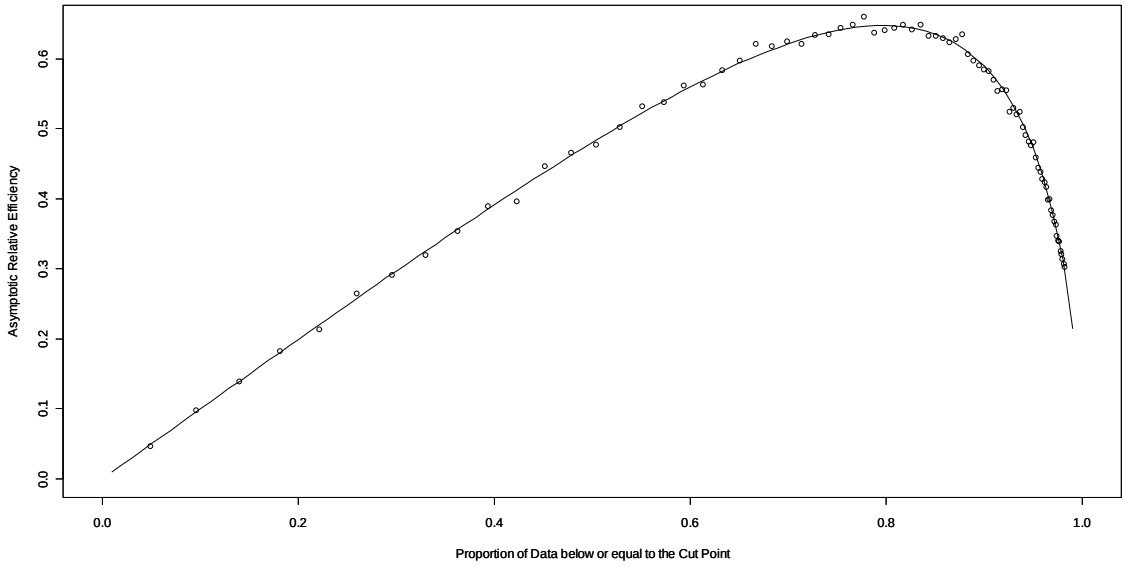
$$= \frac{1-P_1}{P_1}[log(1-P_1)]^2 \qquad (5.2.1)$$

The followings are the comparisons between using the original observations vs. the dichotomized observations from an exponential distribution:

Maximum Likelihood Estimate	**Complete Data**	**Categorized Data**
Parameter Estimate	$\hat{\xi} = \frac{1}{\bar{x}}$	$\hat{\xi}_c = \frac{1}{c_1} log \frac{1}{1-P_1}$
Variance of Parameter	$\text{var}(\hat{\xi}) = \frac{\xi^2}{n}$	$\text{var}(\hat{\xi}_c) = \frac{\xi^2}{n}\frac{P_1}{1-P_1}[log(1-P_1)]^{-2}$

A simulation study was conducted by using data from an exponential distribution with ξ=1. A total of 10,000 simulations were performed and 10,000 data points were used in each simulation. By comparing the estimated relative efficiency calculated from the data and the asymptotic relative efficiency calculated from the equation of $\frac{1-P_1}{P_1}[log(1-P_1)]^2$ (solid line), they are agree with each other in Figure 5.2.1.

Figure 5.2.1 Comparison between asymptotic relative efficiency based on equation (5.2.1) and estimated relative efficiency via simulation using exponential distribution with ξ=1.



When $m = 3$

When data from an exponential distribution are categorized into three groups, we have m=3, $C_3 = \infty$, and C_0=0.

The expected Fisher information from this type of categorization,

$$I(\xi_c) = n \sum_{j=1}^{3} \frac{e^{-\xi(C_j + C_{j-1})}(C_j - C_{j-1})^2}{e^{-\xi C_{j-1}} - e^{-\xi C_j}}$$

$$= n \left[\frac{e^{-\xi(C_1 + C_0)}(C_1 - C_0)^2}{e^{-\xi C_0} - e^{-\xi C_1}} + \frac{e^{-\xi(C_2 + C_1)}(C_2 - C_1)^2}{e^{-\xi C_1} - e^{-\xi C_2}} + \frac{e^{-\xi(C_3 + C_2)}(C_3 - C_2)^2}{e^{-\xi C_2} - e^{-\xi C_3}} \right]$$

$$\left(\frac{e^{-\xi(C_3 + C_2)}(C_3 - C_2)^2}{e^{-\xi C_2} - e^{-\xi C_3}} \to 0,\ e^{-\xi C_0} = 1 \right)$$

$$= n \left[\frac{e^{-\xi C_1} C_1^{\,2}}{1 - e^{-\xi C_1}} + \frac{e^{-\xi(C_2 + C_1)}(C_2 - C_1)^2}{e^{-\xi C_1} - e^{-\xi C_2}} \right]$$

Because $P_1 = 1 - e^{-\xi C_1}$

$e^{-\xi C_1} = 1 - P_1$

$C_1^{\,2} = [-\frac{1}{\xi} log(1 - P_1)]^2$

$e^{-\xi(C_2 + C_1)} = e^{-\xi C_2} e^{-\xi C_1} = [1 - F(C_2)][1 - F(C_1)]$

$= (1 - P_1 - P_2)(1 - P_1)$

$(C_2 - C_1)^2 = [-\frac{1}{\xi} log(1 - P_1 - P_2) + \frac{1}{\xi} log(1 - P_1)]^2$

$= \frac{1}{\xi^2} [log \frac{1 - P_1 - P_2}{1 - P_1}]^2$

$e^{-\xi C_1} - e^{-\xi C_2} = P_2$

Therefore,

$$I(\xi_c) = n\left[\frac{e^{-\xi C_1}C_1^{\,2}}{1-e^{-\xi C_1}} + \frac{e^{-\xi(C_2+C_1)}(C_2-C_1)^2}{e^{-\xi C_1}-e^{-\xi C_2}}\right]$$

$$= \frac{n}{\xi^2}\left\{\frac{1-P_1}{P_1}[\log(1-P_1)]^2 + \frac{(1-P_1-P_2)(1-P_1)}{P_2}\left[\log\frac{1-P_1-P_2}{1-P_1}\right]^2\right\}$$

By comparing with the variance from continuous data,

$$\text{ARE} = \frac{1-P_1}{P_1}[\log(1-P_1)]^2 + \frac{(1-P_1-P_2)(1-P_1)}{P_2}\left[\log\frac{1-P_1-P_2}{1-P_1}\right]^2$$

When $m = 4$

When data from an exponential distribution are categorized into four groups, we have m=4, $C_4=\infty$, and C_0=0.

The expected Fisher information from this type of categorization,

$$I(\xi_c) = n\sum_{j=1}^{4}\frac{e^{-\xi(C_j+C_{j-1})}(C_j-C_{j-1})^2}{e^{-\xi C_{j-1}}-e^{-\xi C_j}}$$

$$= n\left[\frac{e^{-\xi(C_1+C_0)}(C_1-C_0)^2}{e^{-\xi C_0}-e^{-\xi C_1}} + \frac{e^{-\xi(C_2+C_1)}(C_2-C_1)^2}{e^{-\xi C_1}-e^{-\xi C_2}} + \frac{e^{-\xi(C_3+C_2)}(C_3-C_2)^2}{e^{-\xi C_2}-e^{-\xi C_3}}\right.$$

$$\left. + \frac{e^{-\xi(C_4+C_3)}(C_4-C_3)^2}{e^{-\xi C_3}-e^{-\xi C_4}}\right]$$

$$= \frac{n}{\xi^2}\left\{\frac{(1-P_1)(1-0)}{P_1}[\log(1-P_1)]^2 + \frac{(1-P_1-P_2)(1-P_1)}{P_2}\left[\log\frac{1-P_1-P_2}{1-P_1}\right]^2\right.$$

$$\left. + \frac{(1-P_1-P_2-P_3)(1-P_1-P_2)}{P_3}\left[\log\frac{1-P_1-P_2-P_3}{1-P_1-P_2}\right]^2\right\}$$

By comparing with the variance from continuous data,

$$\text{ARE}= \frac{(1-P_1)(1-0)}{P_1}[\log(1-P_1)]^2 + \frac{(1-P_1-P_2)(1-P_1)}{P_2}\left[\log\frac{1-P_1-P_2}{1-P_1}\right]^2$$

$$+\frac{(1-P_1-P_2-P_3)(1-P_1-P_2)}{P_3}\left[\log\frac{1-P_1-P_2-P_3}{1-P_1-P_2}\right]^2$$

When $m = k$

When data from an exponential distribution are categorized into m groups, we have m=k, $C_k = \infty$, and C_0=0.

After deducting from the previous conditions, the expected Fisher information from categorization,

$$I(\xi_c) = n\sum_{j=1}^{k}\frac{e^{-\xi(C_j+C_{j-1})}(C_j-C_{j-1})^2}{e^{-\xi C_{j-1}}-e^{-\xi C_j}}$$

$$= \frac{n}{\xi^2}\sum_{h=1}^{k-1}\frac{(1-\sum_{g=1}^{h}P_g)(1-\sum_{g=0}^{h-1}P_g)}{P_h}\left[\log\frac{1-\sum_{g=1}^{h}P_g}{1-\sum_{g=0}^{h-1}P_g}\right]^2$$

where P_0=0.

By comparing with the variance from continuous data,

$$\text{ARE} = \sum_{h=1}^{k-1}\frac{(1-\sum_{g=1}^{h}P_g)(1-\sum_{g=0}^{h-1}P_g)}{P_h}\left[\log\frac{1-\sum_{g=1}^{h}P_g}{1-\sum_{g=0}^{h-1}P_g}\right]^2 \qquad (5.2.2)$$

5.3 Numerical Approach for Getting the MLE

When an analytical solution is available, it is easy to assess the asymptotic relative efficiency. However, not all of the likelihood functions have an analytical solution. Therefore, the evaluation of estimated relative efficiency needs to use a numerical approach.

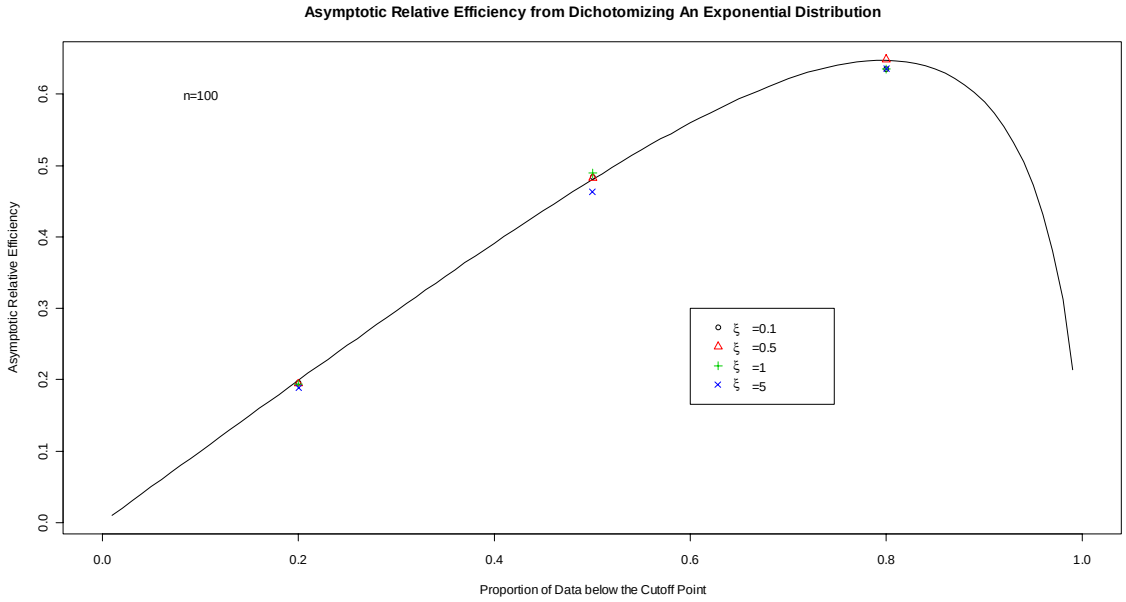
In the previous section, we used exponential distribution to derive the analytical form for the maximum likelihood estimate. We also derived the equation for calculating the asymptotic relative efficiency based on the number of category. We also use the simulation studies to demonstrate that the estimated relative efficiency is consistent with the results based on our equation.

5.4 Simulation Approach for Getting the Relative Efficiency of Exponential Distributions

In order to evaluate the impacts from parameter and sample size in a study, simulation studies were performed to assess the estimated relative efficiency and the ARE.

When n=100 was used for 10,000 simulation, the results are shown in Figure 5.4.1. From the results, we found that the parameter of an exponential distribution does not impact the estimated relative efficiency. The simulation results are consistent with the asymptotic relative efficiency calculated by using equation 5.2.2.

Figure 5.4.1 Analytical and simulation results from n=100 in each study

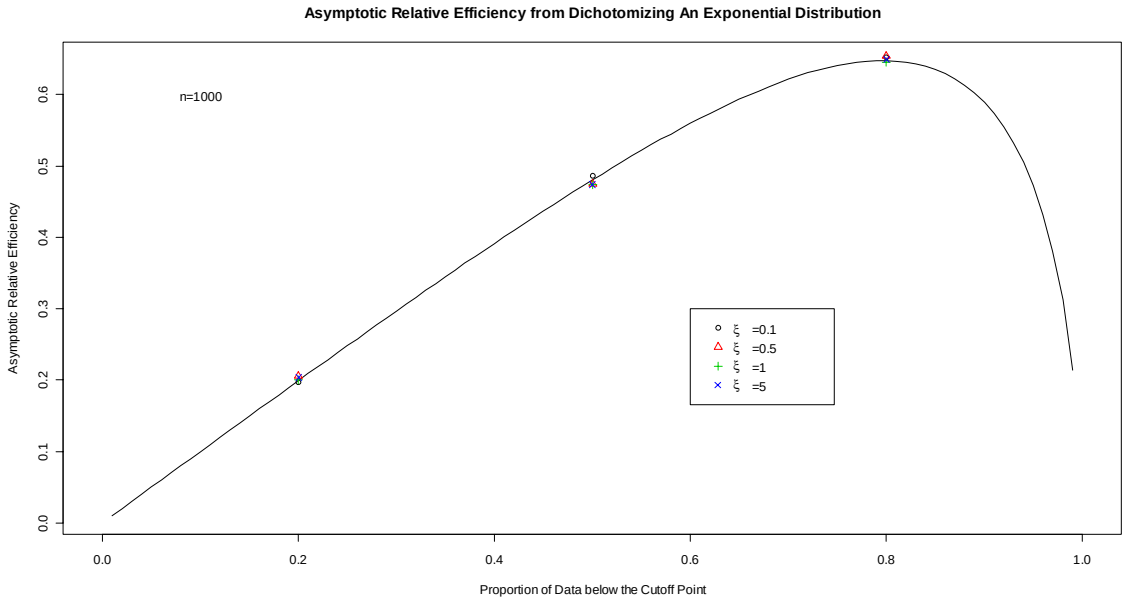


When n=1,000 was used for 10,000 simulation, the results are shown in Figure 5.4.2. From the results, we found that there is no difference on the relative efficiency between exponential distributions with different parameters. The analytical results are consistent with the simulation results.

We further compare the results based on different number of observation in each study. There is no difference between the number of observation on the relative efficiency.

In summary, based on the equation 5.2.2 and simulation studies, we found that only the cutoff point impacts the relative efficiency in the exponential distribution.

Figure 5.4.2 Analytical and simulation results from n=1,000 in each study



5.5 Conclusions

We use the maximum likelihood approach to estimate the underlying continuous covariate when it is categorized and expressed as the form of a multinomial form. The analytic approach was demonstrated by using the exponential distribution. We derive a general form to calculate the asymptotic relative efficiency based on the number of cutoff points. We also performed simulation studies to assess the potential impact to the relative efficiency. From our studies on the exponential distributions, we found that the asymptotic relative efficiency depends on only the choice of cutoff points.

Chapter 6

Efficiency of Categorizing Dose in a Dose-Response Relationship

When the parameters of a continuous variable are estimated from a categorized form, the major impact is loss of asymptotic efficiency. When the categorized variable is used for assessing the dose-response relationship, how the categorization is done impacts the efficiency of the coefficient estimation is of interest.

6.1 Model

Let Y_i be an independent and identically distributed (i.i.d.) random variable with a density function $f_\xi(y)$. Let the expected value of Y_i, μ, be a linear function of an i.i.d. continuous random variable X_i. That is, we can use the concept of generalized linear model to define

$$G(\mu) = G(E[Y]) = \boldsymbol{\beta X}$$

where G is a link function (McCullagh and Nelder, 1989), $\boldsymbol{\beta}$ is the vector of regression coefficient, and $\boldsymbol{X}$ is the vector of explanatory variables. For simplicity, we use $G(\mu) = G(E[Y]) = \beta_0 + \beta_1 X$ for this chapter. Therefore, $\beta = (\beta_0, \beta_1)$

Let $f(Y \mid X, \beta)$ be the density function of Y given X and β. When X_i is categorized into the j^{th} interval $[C_{j-1}, C_j]$, we define X^*_{ij} as:

$$X^*_{ij} = \begin{cases} 1 & \text{if } C_{j-1} \le X_i < C_j \\ 0 & \text{otherwise} \end{cases}$$

where j=1, … m, and m is the number of groups. Therefore, we have the conditional density function

$$f(y_i \mid X^*_{ij},\ \beta) = f(y_i \mid C_{j-1} \le X_i < C_j,\ \beta)$$

$$= \frac{f(y_i, C_{j-1} \le X_i < C_j \mid \beta)}{f_X(C_{j-1} \le X_i < C_j)}$$

$$= \frac{\int_{C_{j-1}}^{C_j} f(y, x \mid \beta)dx}{\int_{C_{j-1}}^{C_j} f_X(x)dx}$$

$$= \frac{\int_{C_{j-1}}^{C_j} f(y \mid x,\ \beta) f(x)dx}{\int_{C_{j-1}}^{C_j} f_X(x)dx}$$

$$= \frac{1}{P_j} \int_{C_{j-1}}^{C_j} f(y \mid x,\ \beta) f(x)dx$$

where $P_j = \int_{C_{j-1}}^{C_j} f_X(x)dx$

When we want to estimate the parameter of interest from y and x, we use the density function $f(y \mid X,\ \beta)$. Let log-likelihood function

$$l(\beta \mid y,\ x) = \sum_{i=1}^{n} l_i(\beta \mid y_i,\ x_i) = \sum_{i=1}^{n} \log f(y_i \mid x_i,\ \beta)$$

where $l_i(\beta \mid y_i,\ x_i) = \log f(y_i \mid x_i,\ \beta)$.

Based on the regression model described previously, $\beta = (\beta_0,\ \beta_1)$. Therefore, we can calculate the score function,

$$\frac{\partial l_i(\beta \mid y_i,\ x_i)}{\partial \beta} = \frac{\partial \log f(y_i \mid x_i,\ \beta)}{\partial \beta}$$

$$= \frac{\frac{\partial}{\partial \beta} f(y_i \mid x_i,\ \beta)}{f(y_i \mid x_i,\ \beta)}$$

From the result, we can also calculate the expected Fisher information which is the expected value of the product of the first derivative of the log-likelihood function, or the negative expected value of the second derivative. When we use the negative expected value of the second derivative, the second derivative is:

$$\frac{\partial^2 l_i(\beta \mid y_i,\ x_i)}{\partial \beta \partial \beta^T} = \frac{\partial}{\partial \beta^T}\left[\frac{1}{f(y_i \mid x_i,\ \beta)} \frac{\partial}{\partial \beta} f(y_i \mid x_i,\ \beta)\right]'$$

$$= \frac{f(y_i \mid x_i,\ \beta)\frac{\partial^2}{\partial \beta \partial \beta^T} f(y_i \mid x_i,\ \beta) - \frac{\partial}{\partial \beta} f(y_i \mid x_i,\ \beta)\frac{\partial}{\partial \beta^T} f(y_i \mid x_i,\ \beta)}{\left[f(y_i \mid x_i,\ \beta)\right]^2}$$

$$I_i(\beta) = E\left[-\frac{\partial^2 l_i(\beta \mid y_i,\ x_i)}{\partial \beta \partial \beta^T}\right]$$

$$= \int_{-\infty}^{\infty} \frac{\frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta}\frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta^T} - f(y_i \mid x_i,\ \beta)\frac{\partial^2 f(y_i \mid x_i,\ \beta)}{\partial \beta \partial \beta^T}}{\left[f(y_i \mid x_i,\ \beta)\right]^2} f(y_i \mid x_i,\ \beta) dy_i$$

$$= \int_{-\infty}^{\infty} \frac{\frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta}\frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta^T} - f(y_i \mid x_i,\ \beta)\frac{\partial^2 f(y_i \mid x_i,\ \beta)}{\partial \beta \partial \beta^T}}{f(y_i \mid x_i,\ \beta)} dy_i$$

$$= \begin{bmatrix} I_{i11} & I_{i12} \\ I_{i21} & I_{i22} \end{bmatrix}$$

Based on our model, we have:

$$\frac{\partial}{\partial \beta} f(y_i \mid x_i,\ \beta) = \begin{bmatrix} \frac{\partial}{\partial \beta_0} f(y_i \mid x_i,\ \beta) \\ \frac{\partial}{\partial \beta_1} f(y_i \mid x_i,\ \beta) \end{bmatrix}$$

$$\frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta}\frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta^T}$$

$$= \begin{bmatrix} \frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta_0}\frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta_0} & \frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta_0}\frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta_1} \\ \frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta_1}\frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta_0} & \frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta_1}\frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta_1} \end{bmatrix}$$

$$\frac{\partial^2 f(y_i \mid x_i,\ \beta)}{\partial \beta \partial \beta^T} = \begin{bmatrix} \frac{\partial^2 f(y_i \mid x_i,\ \beta)}{\partial \beta_0 \partial \beta_0} & \frac{\partial^2 f(y_i \mid x_i,\ \beta)}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 f(y_i \mid x_i,\ \beta)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 f(y_i \mid x_i,\ \beta)}{\partial \beta_1 \partial \beta_1} \end{bmatrix}$$

Therefore, the (a, b)th component of the information matrix is:

$$I_{iab} = \int_{-\infty}^{\infty} \frac{\frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta_{a-1}}\frac{\partial f(y_i \mid x_i,\ \beta)}{\partial \beta_{b-1}} - f(y_i \mid x_i,\ \beta)\frac{\partial^2 f(y_i \mid x_i,\ \beta)}{\partial \beta_{a-1}\partial \beta_{b-1}}}{f(y_i \mid x_i,\ \beta)} dy_i$$

When we want to estimate the parameter of interest from y and categorized x, we use the density function $f(y \mid X^*,\ \beta)$. Let log-likelihood function

$$l^*(\beta \mid y,\ x) = \sum_{i=1}^{n} l_i^*(\beta \mid y_i,\ x^*_i) = \sum_{i=1}^{n} \log f(y_i \mid x^*_i,\ \beta)$$

where $l_i^*(\beta \mid y_i,\ x_i^*) = \log f(y_i \mid x_i^*,\ \beta)$.

Therefore, we can calculate the score function,

$$\frac{\partial l_i^*(\beta \mid y_i,\ x_i^*)}{\partial \beta} = \frac{\partial \log f(y_i \mid x_i^*,\ \beta)}{\partial \beta}$$

$$= \frac{\frac{\partial}{\partial \beta} f(y_i \mid x_i^*,\ \beta)}{f(y_i \mid x_i^*,\ \beta)}$$

From the result, we can calculate the expected Fisher information which is the expected value of the product of the first derivative of the log-likelihood function, or the negative expected value of the second derivative. When we use the negative expected value of the second derivative, the second derivative is:

$$\frac{\partial^2 l_i^*(\beta \mid y_i,\ x_i^*)}{\partial\beta\partial\beta^T} = \frac{\partial}{\partial\beta^T}\left[\frac{1}{f(y_i \mid x_i^*,\ \beta)}\frac{\partial}{\partial\beta} f(y_i \mid x_i^*,\ \beta)\right]'$$

$$= \frac{f(y_i \mid x_i^*,\ \beta)\frac{\partial^2}{\partial\beta\partial\beta^T} f(y_i \mid x_i^*,\ \beta) - \frac{\partial}{\partial\beta} f(y_i \mid x_i^*,\ \beta)\frac{\partial}{\partial\beta^T} f(y_i \mid x_i^*,\ \beta)}{\left[f(y_i \mid x_i^*,\ \beta)\right]^2}$$

$$I^*_i(\beta) = E\left[-\frac{\partial^2 l_i^*(\beta \mid y_i,\ x_i^*)}{\partial\beta\partial\beta^T}\right]$$

$$= \int_{-\infty}^{\infty} \frac{\frac{\partial f(y_i \mid x_i^*,\ \beta)}{\partial\beta}\frac{\partial f(y_i \mid x_i^*,\ \beta)}{\partial\beta^T} - f(y_i \mid x_i^*,\ \beta)\frac{\partial^2 f(y_i \mid x_i^*,\ \beta)}{\partial\beta\partial\beta^T}}{\left[\int_{C_{j-1}}^{C_j} f(y_i \mid x_i) f(x_i) dx_i\right]^2} \frac{\int_{C_{j-1}}^{C_j} f(y_i \mid x_i) f(x_i) dx_i}{P_j} dy_i$$

$$= \int_{-\infty}^{\infty} \frac{\frac{\partial f(y_i \mid x_i^*,\ \beta)}{\partial\beta}\frac{\partial f(y_i \mid x_i^*,\ \beta)}{\partial\beta^T} - f(y_i \mid x_i^*,\ \beta)\frac{\partial^2 f(y_i \mid x_i^*,\ \beta)}{\partial\beta\partial\beta^T}}{P_j \int_{C_{j-1}}^{C_j} f(y_i \mid x_i) f(x_i) dx_i} dy_i$$

$$= \begin{bmatrix} I^*_{i11} & I^*_{i12} \\ I^*_{i21} & I^*_{i22} \end{bmatrix}$$

Based on our model, we have:

$$\frac{\partial}{\partial\beta} f(y_i \mid x_i^*,\ \beta) = \begin{bmatrix} \frac{\partial}{\partial\beta_0} f(y_i \mid x_i^*,\ \beta) \\ \frac{\partial}{\partial\beta_1} f(y_i \mid x_i^*,\ \beta) \end{bmatrix}$$

$$\frac{\partial f(y_i \mid x_i^*, \beta)}{\partial \beta} \frac{\partial f(y_i \mid x_i^*, \beta)}{\partial \beta^T}$$

$$= \begin{bmatrix} \frac{\partial f(y_i \mid x_i^*, \beta)}{\partial \beta_0} \frac{\partial f(y_i \mid x_i^*, \beta)}{\partial \beta_0} & \frac{\partial f(y_i \mid x_i^*, \beta)}{\partial \beta_0} \frac{\partial f(y_i \mid x_i^*, \beta)}{\partial \beta_1} \\ \frac{\partial f(y_i \mid x_i^*, \beta)}{\partial \beta_1} \frac{\partial f(y_i \mid x_i^*, \beta)}{\partial \beta_0} & \frac{\partial f(y_i \mid x_i^*, \beta)}{\partial \beta_1} \frac{\partial f(y_i \mid x_i^*, \beta)}{\partial \beta_1} \end{bmatrix}$$

$$\frac{\partial^2 f(y_i \mid x_i^*, \beta)}{\partial \beta \partial \beta^T} = \begin{bmatrix} \frac{\partial^2 f(y_i \mid x_i^*, \beta)}{\partial \beta_0 \partial \beta_0} & \frac{\partial^2 f(y_i \mid x_i^*, \beta)}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 f(y_i \mid x_i^*, \beta)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 f(y_i \mid x_i^*, \beta)}{\partial \beta_1 \partial \beta_1} \end{bmatrix}$$

Therefore, the (a, b)[th] component of the information matrix is:

$$I^*_{iab} = \int_{-\infty}^{\infty} \frac{\frac{\partial f(y_i \mid x_i^*, \beta)}{\partial \beta_{a-1}} \frac{\partial f(y_i \mid x_i^*, \beta)}{\partial \beta_{b-1}} - f(y_i \mid x_i^*, \beta) \frac{\partial^2 f(y_i \mid x_i^*, \beta)}{\partial \beta_{a-1} \partial \beta_{b-1}}}{P_j \int_{C_{j-1}}^{C_j} f(y_i \mid x_i) f(x_i) dx_i} dy_i$$

When we use the data from n observations, the average expected information matrix based on the original values is

$$I(\beta) = \frac{1}{n} \begin{bmatrix} \sum_{i=1}^{n} I_{i11} & \sum_{i=1}^{n} I_{i12} \\ \sum_{i=1}^{n} I_{i21} & \sum_{i=1}^{n} I_{i22} \end{bmatrix}$$

Therefore, the variance of maximum likelihood estimator of β is

$$Var(\hat{\beta}) = \frac{I^{-1}(\beta)}{n}$$

$$= \frac{1}{\sum_{i=1}^{n} I_{i11} \sum_{i=1}^{n} I_{i22} - \sum_{i=1}^{n} I_{i12} \sum_{i=1}^{n} I_{i21}} \begin{bmatrix} \sum_{i=1}^{n} I_{i22} & -\sum_{i=1}^{n} I_{i12} \\ -\sum_{i=1}^{n} I_{i21} & \sum_{i=1}^{n} I_{i11} \end{bmatrix}$$

The equations available at www.wolframalpha.com (Weisstein, 2010) were used to calculate the inverse of a square matrix.

For the coefficient of interest β_1, the variance of maximum likelihood estimator is

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} I_{i11}}{\sum_{i=1}^{n} I_{i11} \sum_{i=1}^{n} I_{i22} - \sum_{i=1}^{n} I_{i12} \sum_{i=1}^{n} I_{i21}}$$

When we use the data from n observations, the average expected information matrix based on the categorized values is

$$I^*(\beta) = \frac{1}{n} \begin{bmatrix} \sum_{i=1}^{n} I^*_{i11} & \sum_{i=1}^{n} I^*_{i12} \\ \sum_{i=1}^{n} I^*_{i21} & \sum_{i=1}^{n} I^*_{i22} \end{bmatrix}$$

Therefore, the variance of maximum likelihood estimator of β^* is

$$Var(\hat{\beta}^*) = \frac{I^{*-1}(\beta)}{n}$$

$$= \frac{1}{\sum_{i=1}^{n} I^*_{i11} \sum_{i=1}^{n} I^*_{i22} - \sum_{i=1}^{n} I^*_{i12} \sum_{i=1}^{n} I^*_{i21}} \begin{bmatrix} \sum_{i=1}^{n} I^*_{i22} & -\sum_{i=1}^{n} I^*_{i12} \\ -\sum_{i=1}^{n} I^*_{i21} & \sum_{i=1}^{n} I^*_{i11} \end{bmatrix}$$

For the coefficient of interest β_1, the variance of maximum likelihood estimator is

$$Var(\hat{\beta}^*_1) = \frac{\sum_{i=1}^{n} I^*_{i11}}{\sum_{i=1}^{n} I^*_{i11} \sum_{i=1}^{n} I^*_{i22} - \sum_{i=1}^{n} I^*_{i12} \sum_{i=1}^{n} I^*_{i21}}$$

Therefore, we can calculate the asymptotic relative efficiency by using the variance of β_1 estimated from original observations divided by the variance estimated from the categorical observations,

$$\text{ARE} = \frac{Var(\hat{\beta}_1)}{Var(\hat{\beta}^*_1)}$$

6.2 Impact from Dichotomization on the Relative Efficiency of Coefficient

In order to assess the impacts from dichotomization on the relative efficiency of coefficient, we performed simulation studies under different conditions. The simulation studies were performed by using the R language. The relative efficiency was calculated by using the variance of the coefficient which derived from continuous covariate divided by the variance of the coefficient which derived from the categorized covariate.

6.2.1 When the Null Hypothesis is True: Coefficient =0

We assessed the influence by assigning the outcome variable *Y* as binary with two possible outcomes 0 and 1. Therefore, the dose-response association becomes:

$$\text{logit}[P(Y=1)] = \log\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = \beta_0 + \beta_1 \times X$$

When the null hypothesis is true, β_1=0.

For categorization, a value from the data was chosen as the cutoff point C_1. The dose-response association becomes:

$$\text{logit}[P(Y=1)] = \log\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = \beta_{D0} + \beta_{D1} \times X_c$$

The truncated means of each group was used as the value for estimating the coefficient β_{D1}. That is,

$$X_c = \begin{cases} \int_{-\infty}^{C_1} f(x)dx & \text{if } X_i \le C_1 \\ \int_{C_1}^{\infty} f(x)dx & \text{if } C_1 < X_i \end{cases}$$

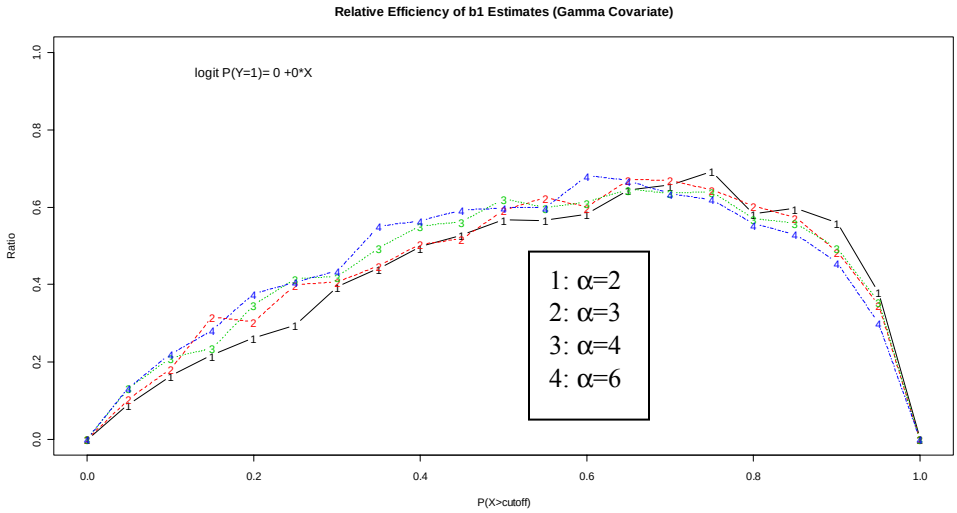
Each simulation was performed by using 20,000 data points. A total of 1,000 simulations were performed for each distribution. The `glm` function of R language was used.

6.2.1.1 Gamma Covariate

We assume that the covariate follows a gamma distribution. Without loss of generality, we assume that the scale parameter β_1 equals to 1. Four different shape parameters (2, 3, 4, 6) were used to assess the impact from the shape parameters. The graphic results are shown in Figure 6.2.1. The numerical results are shown in Table B.1 in Appendix B.

The estimate of β_1 and the estimate of β_{D1} are similar and closed to 0. From the graph, it shows that the RE changes with the choice of cutoff point. The cutoff point which has the highest RE varies with the shape parameter. When α=2, the highest RE is 69.2% when the 75th percentile was used as the cutoff point. The highest RE is 67.1% when the 65th percentile was used as the cutoff point when α=3, When α=4, the highest RE is 64.6% when the 65th percentile was used as the cutoff point. The highest RE is 68.2% when α=6 and the 60th percentile was used as the cutoff point.

Figure 6.2.1 Relative Efficiency of β_1 estimates when X is gamma distribution

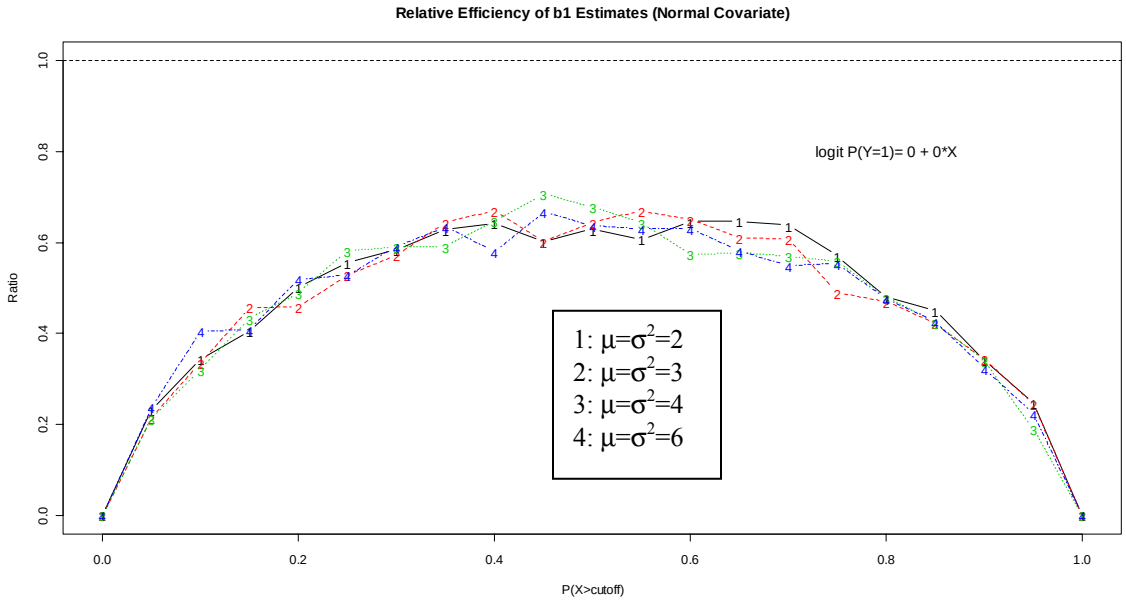


6.2.1.2 Normal Covariate

We assume that the covariate follows a normal distribution. In order to compare the results with the gamma covariate in the previous section, we used four different mean values (2, 3, 4, 6) to assess the impacts. The variance is the same as the mean value. The graphic results are shown in Figure 6.2.2. The numerical results are shown in Table B.2 in the Appendix B.

The estimate of β_1 and the estimate of β_{D1} are similar and closed to 0. From the graph, it shows that the RE changes with the choice of cutoff point. However, different from the gamma distribution, the REs are similar among parameters. The highest REs associate with the cutoff points of 45th percentile in all 4 different distributions.

Figure 6.2.2 Relative Efficiency of β_1 estimates when X is normal distribution



6.2.2 When the Null Hypothesis is False: Coefficient≠0

We were also interested in knowing the impacts from dichotomizing a covariate on a dose-response association in which the null hypothesis is false. We performed simulation studies on assessing the effects.

We assume that the underlying dose-response association is:

$$\text{logit } P(Y=1) = -4 + 1 \times X$$

6.2.2.1 Gamma Covariate

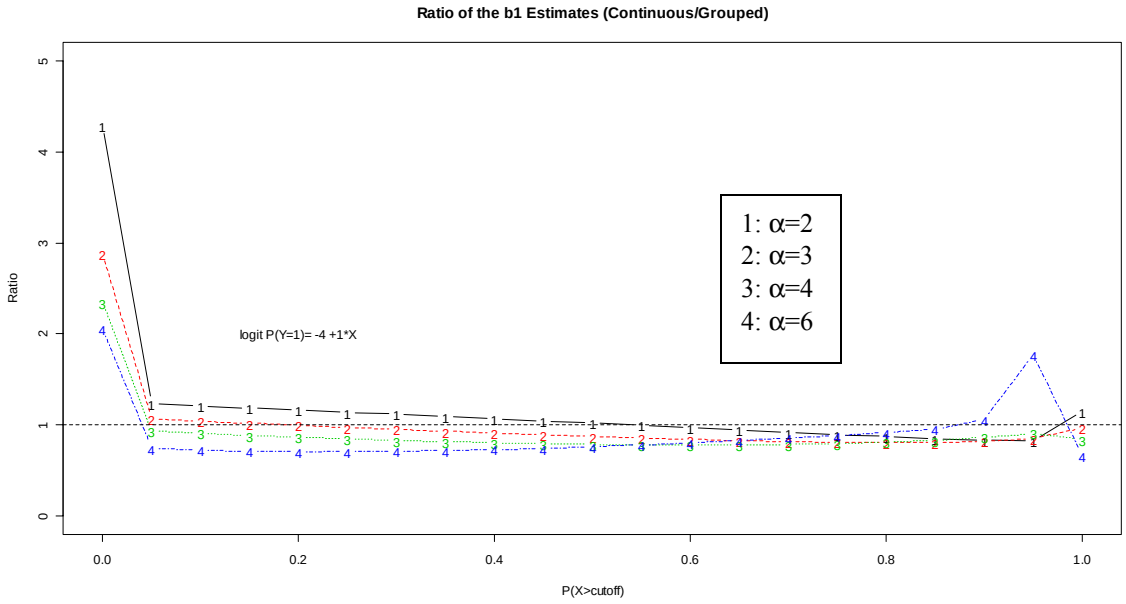
Based on the gamma covariate and this assigned association, the probability of Y=1 is about 30%.

When compare the β_1 estimate with the defined coefficient, the estimate is close to the assigned value. However, when compared the β_1 estimate with the β_{D1} estimate, the ratios changed with the cutoff point as well as the shape parameter of the gamma distribution. The results are shown in Figure 6.2.3 and the numerical results are shown in Table B.3 in Appendix B.

From the figure on the next page, we found that β_{D1} tends to be overestimated, that is, the ratio of β_1 / β_{D1} less than 1. However, when α=2, β_{D1} was underestimated when we used the cutoff points which are smaller than the median.

The overall impression from the results here is that dichotomizing a continuous covariate biases the coefficient estimation. The impact depends on the choice of cutoff points and the distribution of the covariate.

Figure 6.2.3 Ratio of the b1 estimates (continuous/grouped)



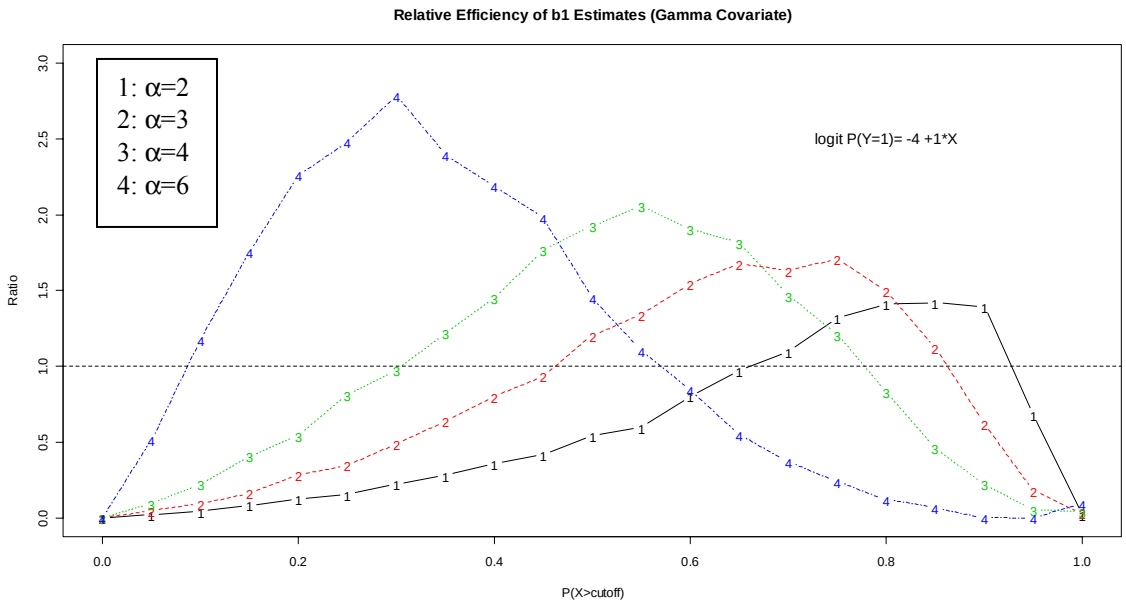
The influences from dichotomization on the RE were also evaluated. The graphic results are shown in Figure 6.2.4, and the numeric results are shown in Table B.4 in Appendix B.

From the graph, it shows that the RE changes with the choice of cutoff point and the parameter of the gamma distribution.

When the cutoff points are away from the median under the shape parameters of 3, 4, and 6, the RE tends to be smaller than 1. That is, categorization increases the variance of β_{D1}. However, when the cutoff points are closed to the median, categorization reduces the variance of β_{D1}.

The observation does not hold for gamma distribution with shape parameter equals 2. The REs larger than 1 when the cutoff points were chosen between the 70th and 90the percentiles.

Figure 6.2.4 Relative efficiency of b1 estimate when X is gamma distribution



Because the bias was found from the β_{D1} estimation, we further assess the mean square error (MSE) of both parameter estimates. The results are shown in Figures 6.2.5 through 6.2.8 by each shape parameter of the gamma distribution. For a better comparison, the MSE values are shown after logarithm transformation. The numerical results in original value are also shown in Table B.5 and Table B.6 in Appendix B.

Overall, all of the MSEs of β_{D1} are larger than the MSEs of β_1, given the same gamma distribution and the same cutoff point. The difference between the MSEs of β_1 and β_{D1} changes with the shape parameter of gamma distribution.

Figure 6.2.5 MSE of Coefficient Estimate when X distributed Gamma (2, 1)

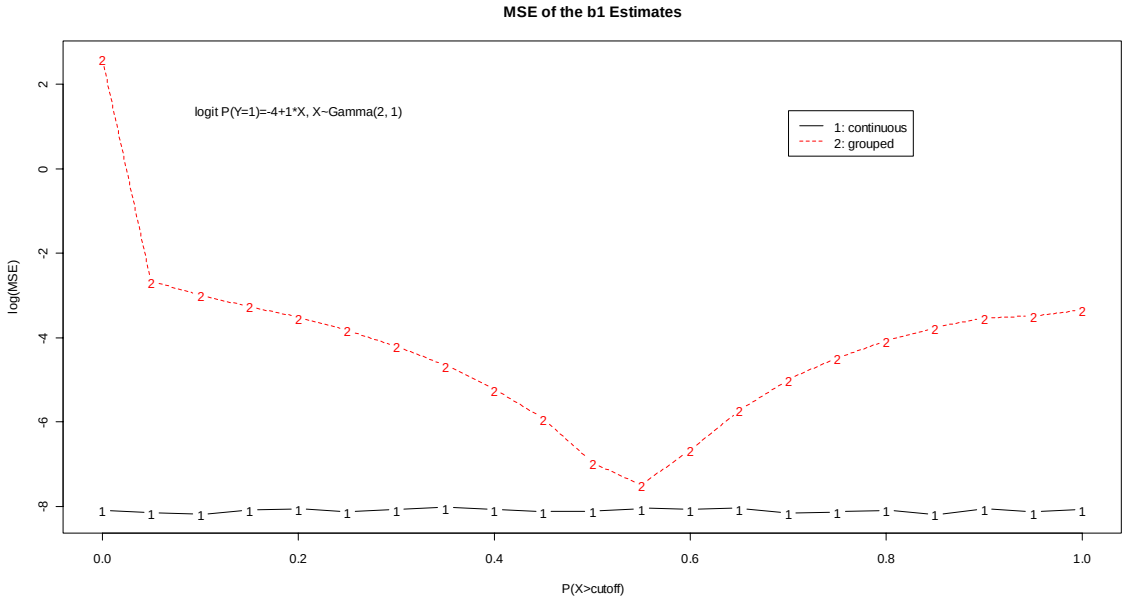


Figure 6.2.6 MSE of Coefficient Estimate when X distributed Gamma (3, 1)

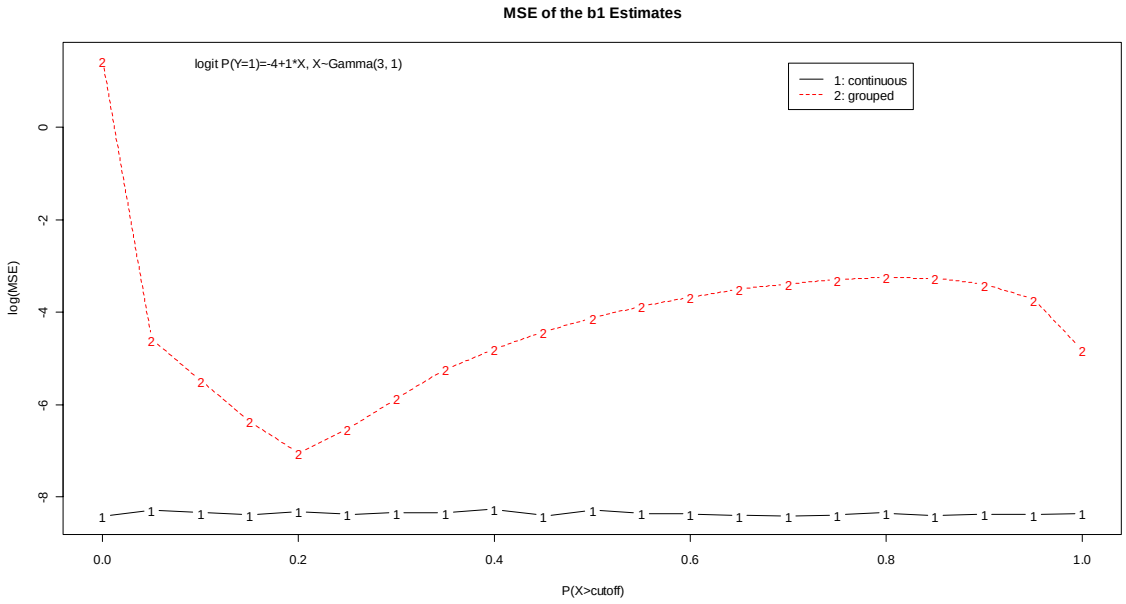


Figure 6.2.7 MSE of Coefficient Estimate when X distributed Gamma (4, 1)

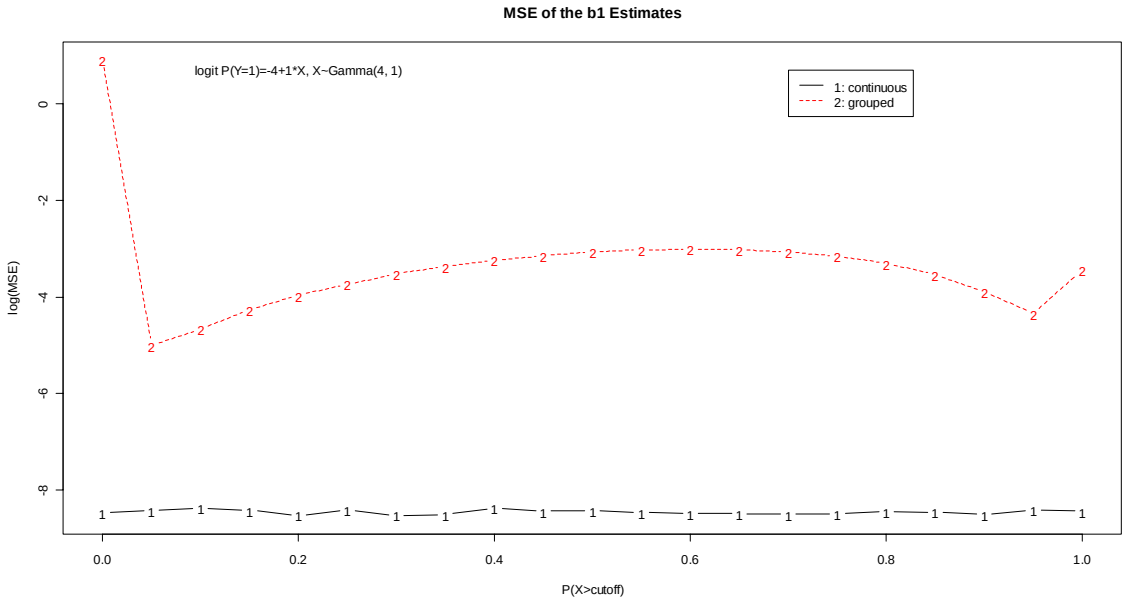
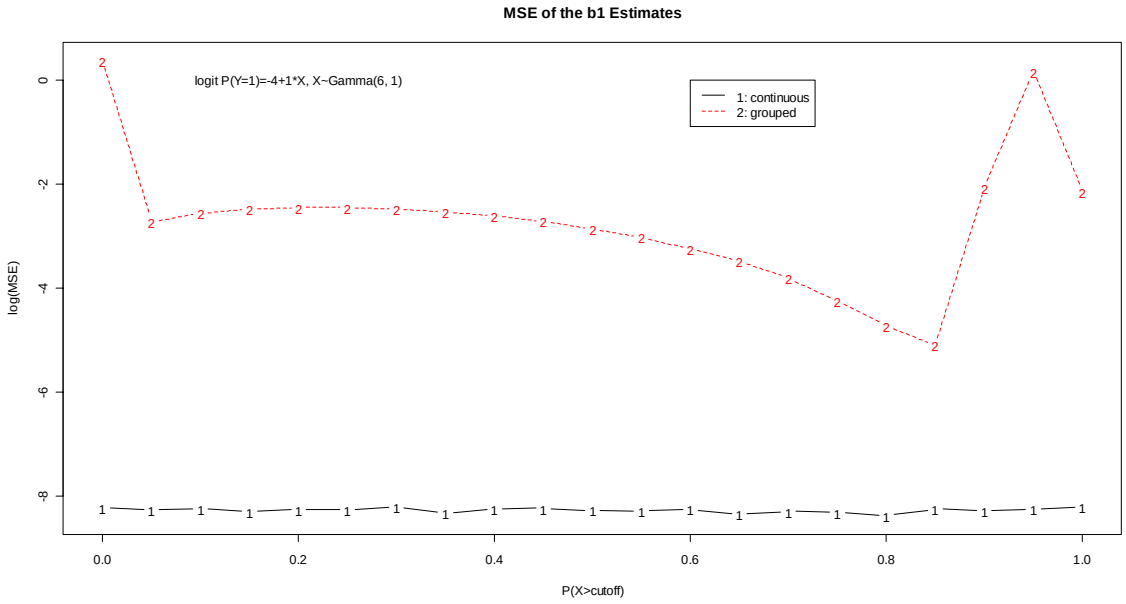


Figure 6.2.8 MSE of Coefficient Estimate when X distributed Gamma (6, 1)



6.2.2.2 Normal Covariate

We performed simulation studies on assessing the impact from dichotomized covariate with different normal distribution. We assume the same underlying dose-response association as:

$$\text{logit } P(Y=1) = -4 + 1 \times X$$

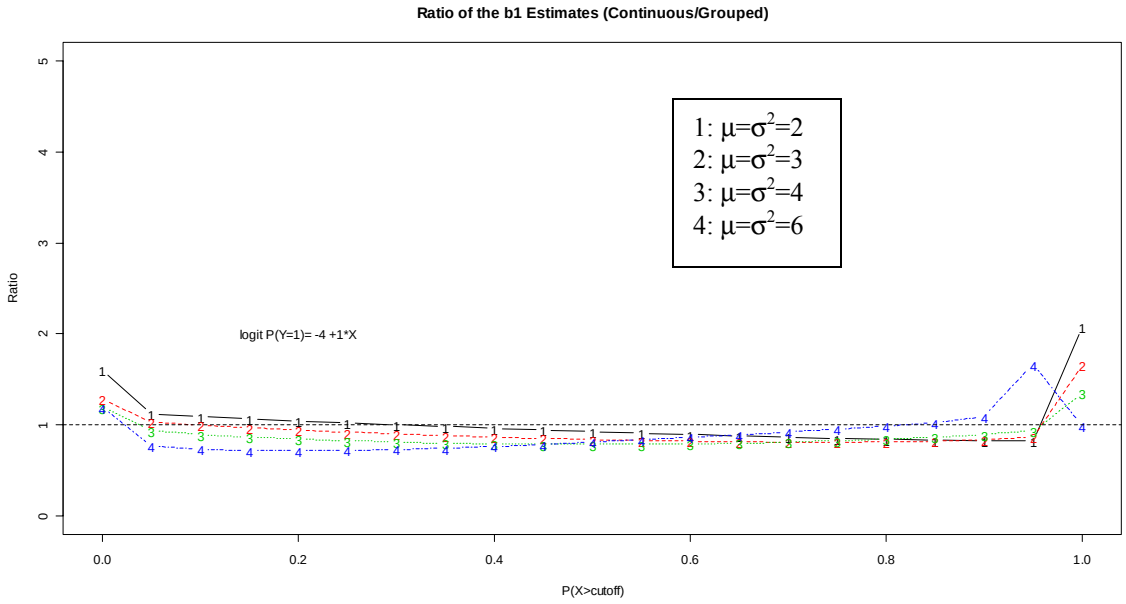
X is assumed to follow a normal distribution. Based on this assigned association, the probability of Y=1 is about 20% when the covariate is distributed as normal (μ=6, σ^2=6), and about 75% when the covariate is distributed as normal (6, 6).

When compare the β_1 estimate with the defined coefficient, the estimate is close to the assigned value. However, when compare the β_1 estimate with the β_{D1} estimate, the ratios changed with the cutoff point as well as the shape parameter of the normal distribution. The results are shown in the Figure 6.2.9. The numerical results are in Table B.7 in Appendix B.

From the figure, we found that β_{D1} tends to be overestimated, that is, the ratio of β_1 / β_{D1} less than 1. They show similar pattern as gamma covariate but have smaller ratios.

The overall impression from the results here is that dichotomizing a continuous covariate biases the coefficient estimation. The impact depends on the choice of cutoff points and the distribution of the covariate.

Figure 6.2.9 Ratio of the b1 estimates (continuous/grouped)



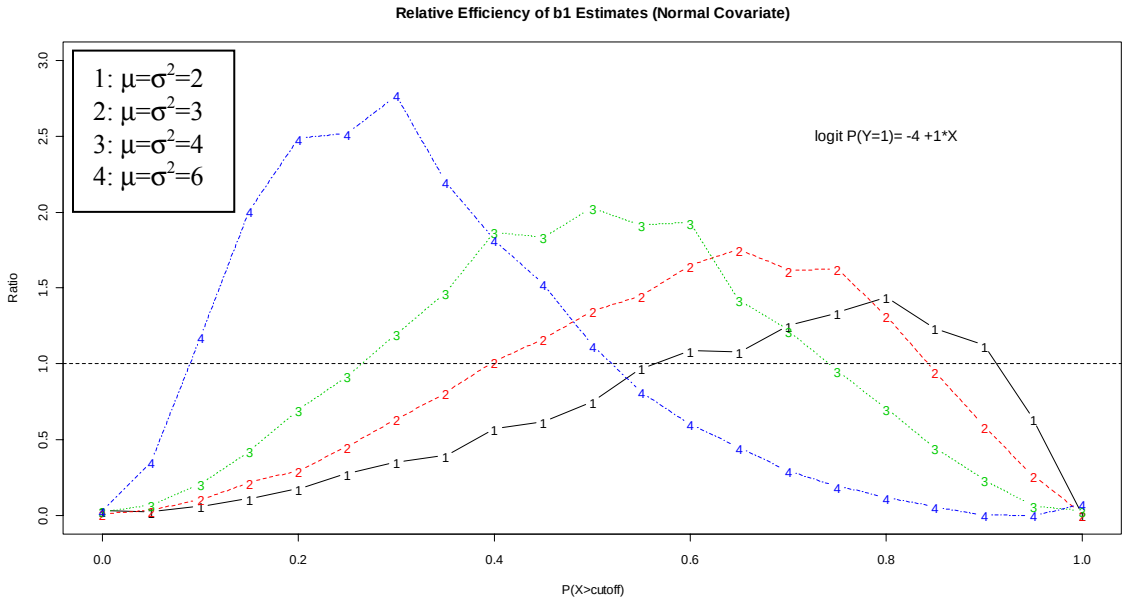
The influence from dichotomization on the RE were also evaluated. The results are shown in Figure 6.2.10. The numerical results are shown in the Table B.8 in Appendix B.

From the graph, we see that the RE changes with the choice of cutoff point and the parameter of the normal distribution.

When the cutoff points are away from the median under the means of 3, 4, and 6, the RE tends to be smaller than 1. That is, categorization increases the variance of β_{D1}. However, when the cutoff points are closed to the median, categorization reduces the variance of β_{D1}.

The observation does not hold for normal distribution with mean equals 2. The REs larger than 1 when the cutoff points were chosen between the 60th and 90the percentiles.

Figure 6.2.10 Relative Efficiency of the b1 estimates (Normal Covariate)



Because the bias was found from the β_{D1} estimation, we further assess the mean square error of both parameter estimates. The graphical results are shown in Figures 6.2.11 through 6.2.14 for each mean and standard deviation the normal distribution. For better comparison, the MSE values are shown after logarithm transformation. The numerical results are also shown in Table B.9 and B.10 in Appendix B..

Overall, all of the mean square errors (MSEs) of β_{D1} are larger than the MSEs of β_1, given the same normal distribution and the same cutoff point. The difference between the MSEs of β_{D1} and β_1 changes with the parameters of normal distribution.

Figure 6.2.11 MSE of Coefficient Estimate when X distributed normal (2, 2)

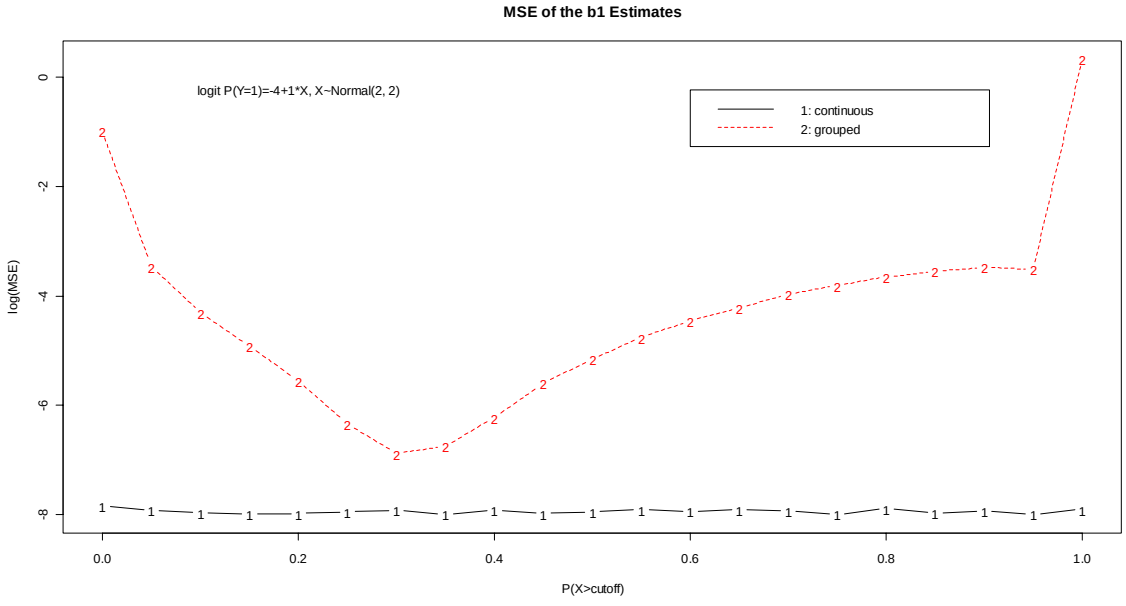


Figure 6.2.12 MSE of Coefficient Estimate when X distributed normal (3, 3)

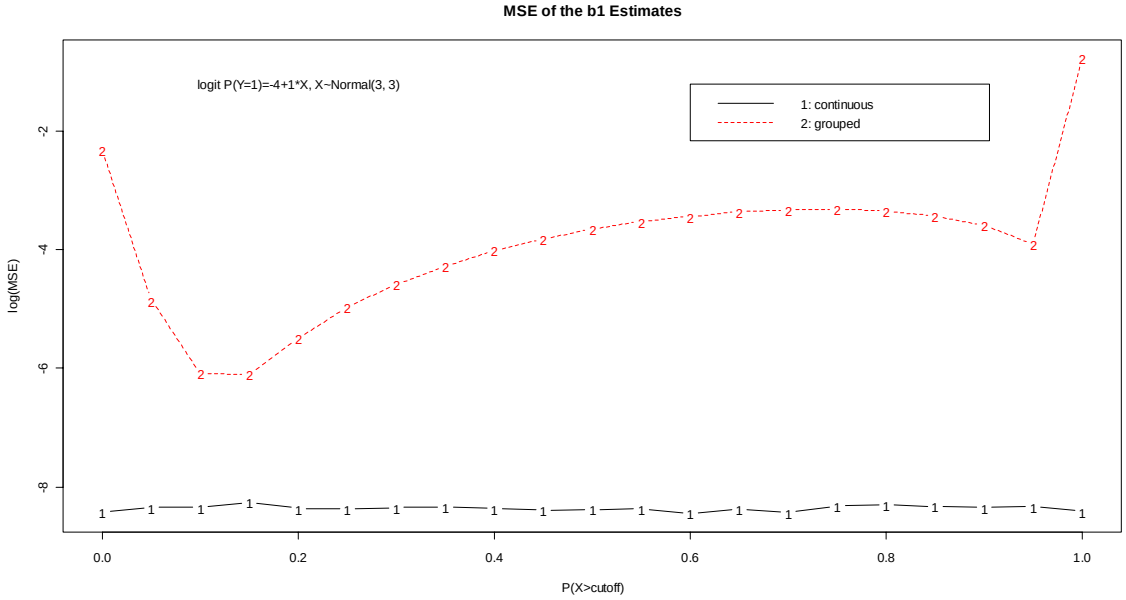


Figure 6.2.13 MSE of Coefficient Estimate when X distributed normal (4, 4)

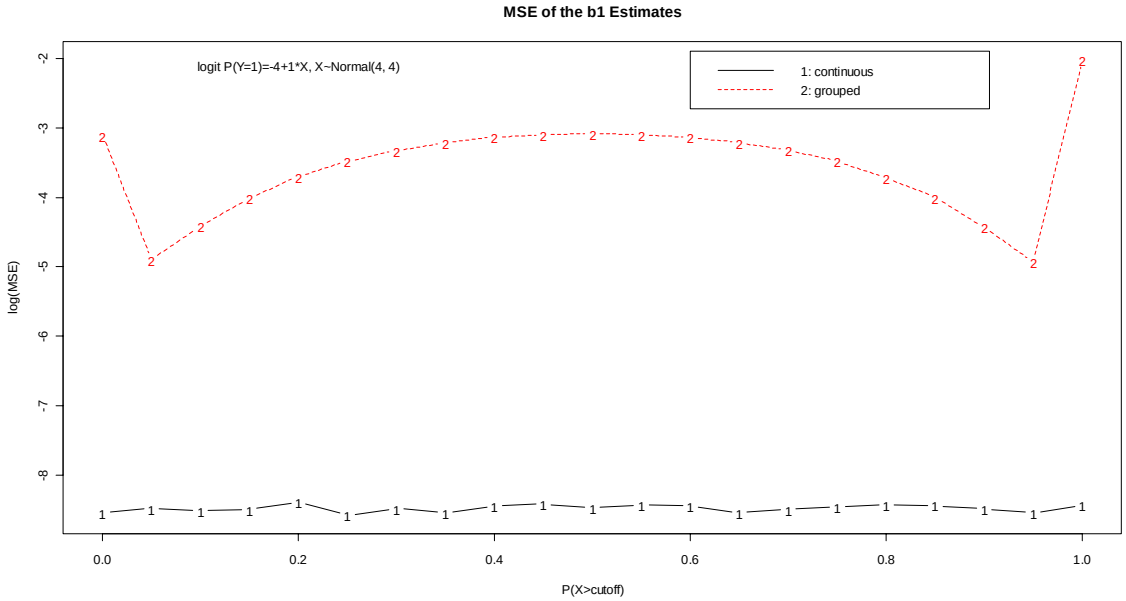
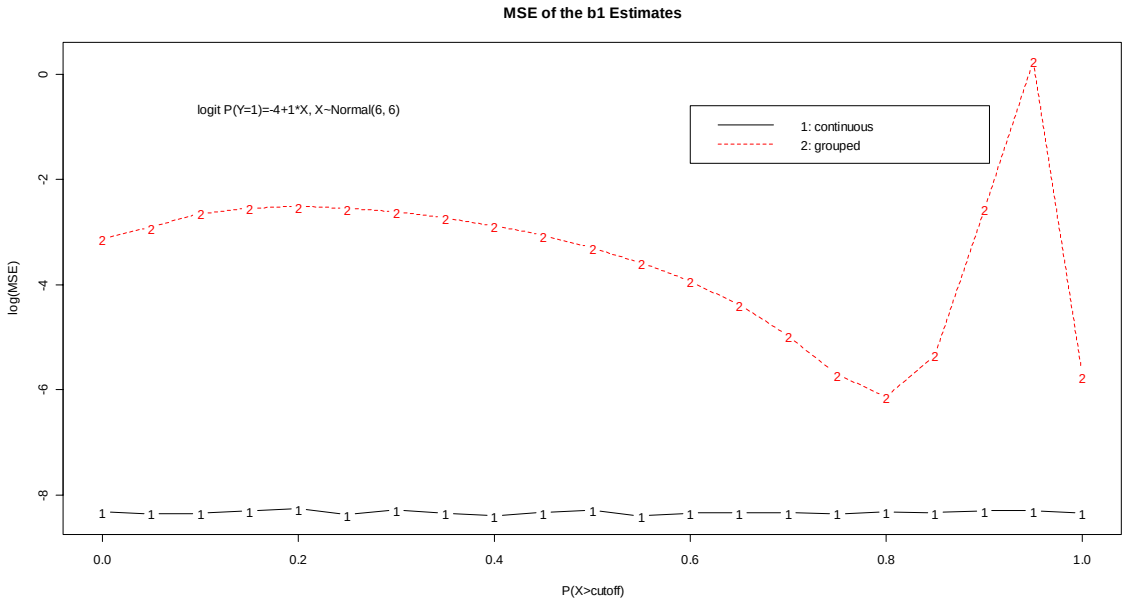


Figure 6.2.14 MSE of Coefficient Estimate when X distributed normal (6, 6)



6.3 Impact from Different Number of Cutoff Points

From the studies, we found that the efficiency or the MSE is affected by the hypothesis, the covariate distribution and the choice of cutoff points.

We further investigated the impact from using different number of cutoff points on the efficiency of estimation.

Simulation studies were conducted by using different quantiles (median, tertile, quartile and quintile) to evaluate the relative efficiency.

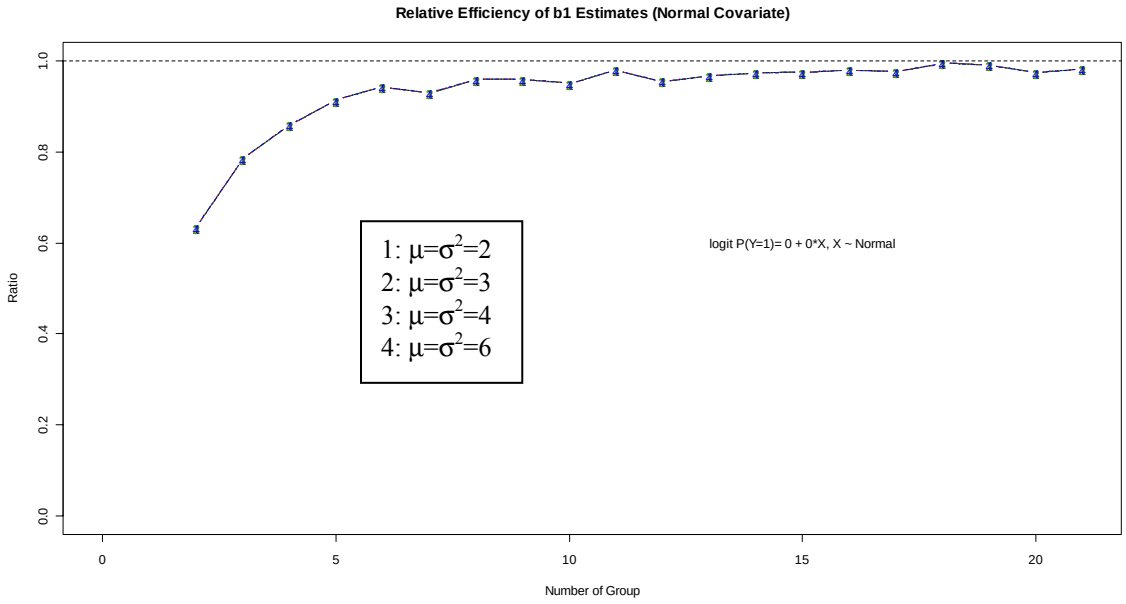
6.3.1 Normal Covariate and β_1=0

We performed simulation studies to evaluate the association between REs and the number of cutoff points. The studies were under the null hypothesis that there is no association between the covariate and the outcome variable.

The covariate was from a normal distribution with 4 different means: (2, 3, 4, 6). The variance in each distribution is the same as the mean value. The graphic results are shown in Figure 6.3.1 on the following page. The numeric results are shown in Table B.11 in Appendix B.

From the results, we found that the REs increase with the number of groups. The REs increased significantly before reaching 6 groups.

Figure 6.3.1 Relative efficiency on b1 estimate from number of group when X is normal distribution and β_1=0



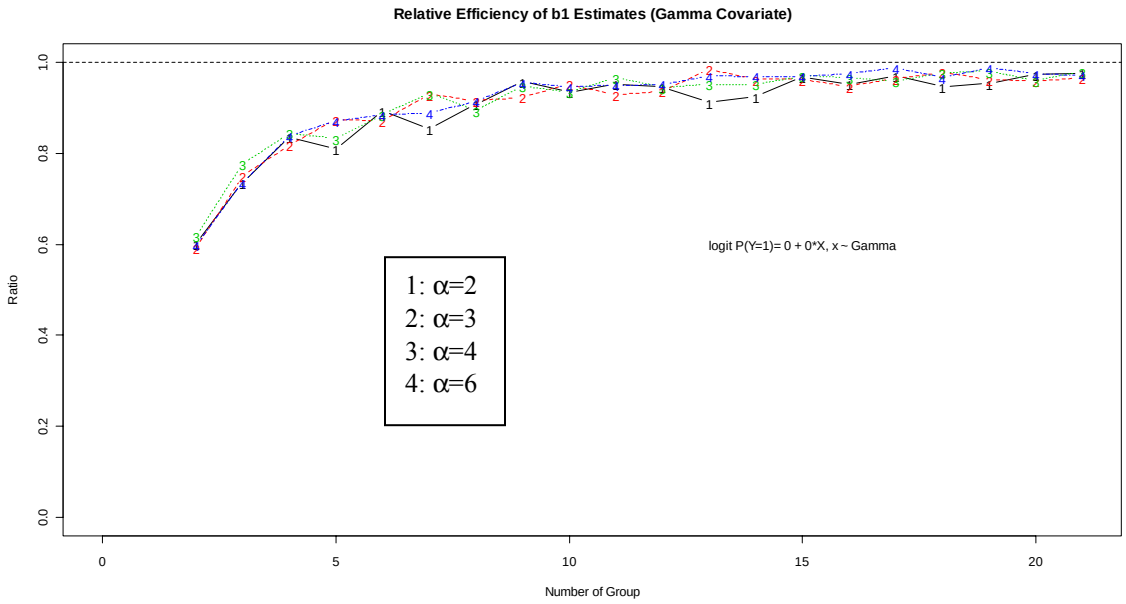
6.3.2 Gamma Covariate and β_1=0

We also performed simulation studies to evaluate the association between REs and the number of cutoff points by using the gamma distribution as the covariate. The studies were under the null hypothesis that there is no association between the covariate and the outcome variable.

The covariate was from a gamma distribution with four different shape parameters: (2, 3, 4, 6). We used 1 as the scale parameter for all of the studies. The graphic results are shown in Figure 6.3.2. The numeric results are shown in the Table B.12 in the Appendix B.

From the results, we found that the REs increase with the number of groups. The REs were larger than 90% when the number of group reached 9 in all distributions.

Figure 6.3.2 Relative efficiency on b1 estimate from number of group when X is gamma distribution and β_1=0

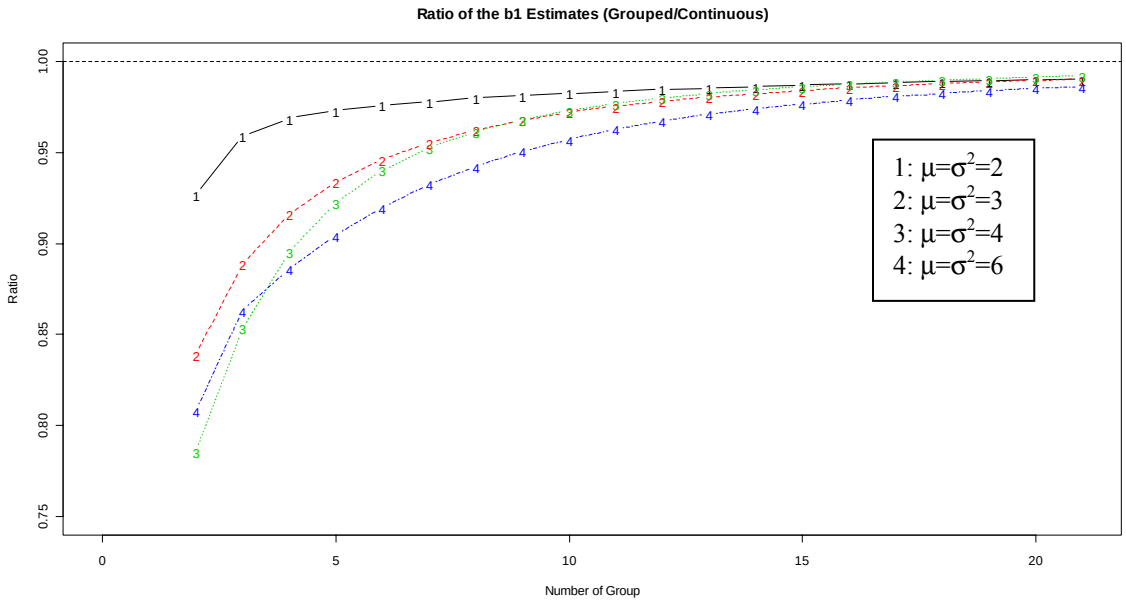


6.3.3 Normal Covariate and β≠0

We performed simulation studies to evaluate the association between REs and the number of cutoff points. The studies were under the assumption that there is an association between the covariate and the outcome variable. The slope is 1 with the intercept of -4. The covariate was from a normal distribution with 4 different means: (2, 3, 4, 6). The variance in each distribution is the same as the mean value.

We compared the coefficients estimated by either grouped or continuous covariates. The graphical results are shown in Figure 6.3.3. The numerical results are shown in Table B.13 in Appendix B. From the results, we found that the coefficients were closed to the defined association when used the continuous covariate. However, it is underestimated when the categorized covariates were used. The magnitudes of underestimation decreased with the increase of number of group.

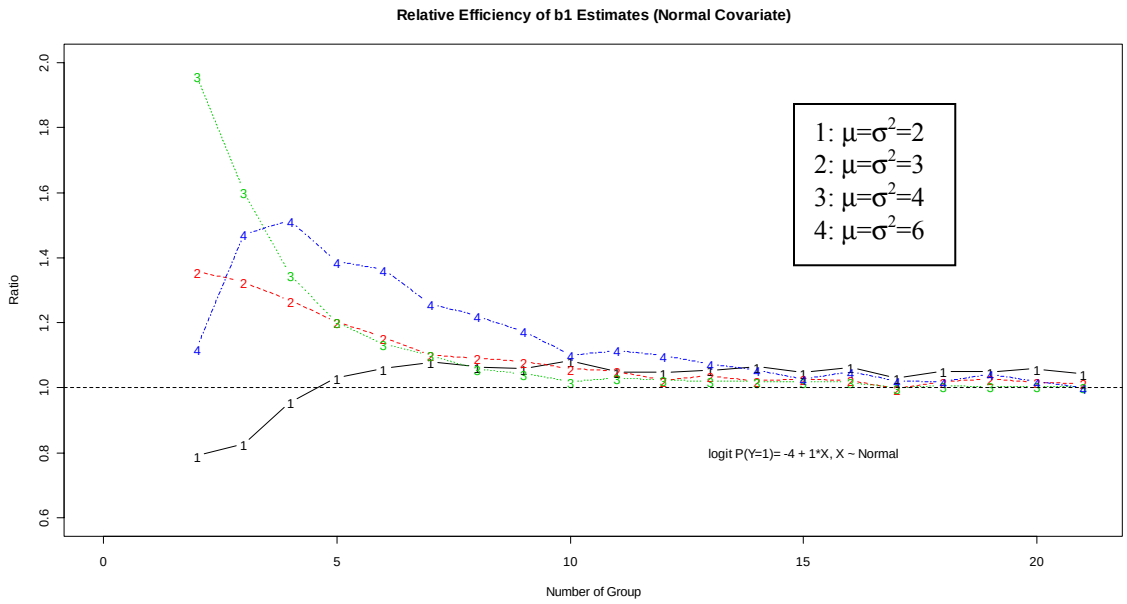
Figure 6.3.3 Ratio of the b1 Estimates (Grouped/Continuous) when X is Normal



For better understand the impact on estimation bias, we calculated the relative bias (100%x[estimate-parameter]/parameter). The relative bias depends on the parameter. The results are shown in Tables B.14 and B.15 in Appendix B.

The graphical results for relative efficiency are shown in Figure 6.3.4. The numeric results are shown in Table B.16 in Appendix B. From the results, we found that the association between REs and the number of groups change with the parameters. The same trend is that the REs approach 1 when the number of group is large.

Figure 6.3.4 Relative Efficiency of b1 Estimate when X is normal distribution and β_1=1



The MSE of each condition was also evaluated. The graphical results are shown in Figure 6.3.5 through Figure 6.3.8. The MSE values are shown in Table B.17 and B.18 in Appendix B.

Overall, when used the continuous covariate to assess the dose-response association, the MSE is smaller than using the grouped covariate.

Figure 6.3.5 MSE of Coefficient Estimate when X distributed normal (2, 2) and β_1=1

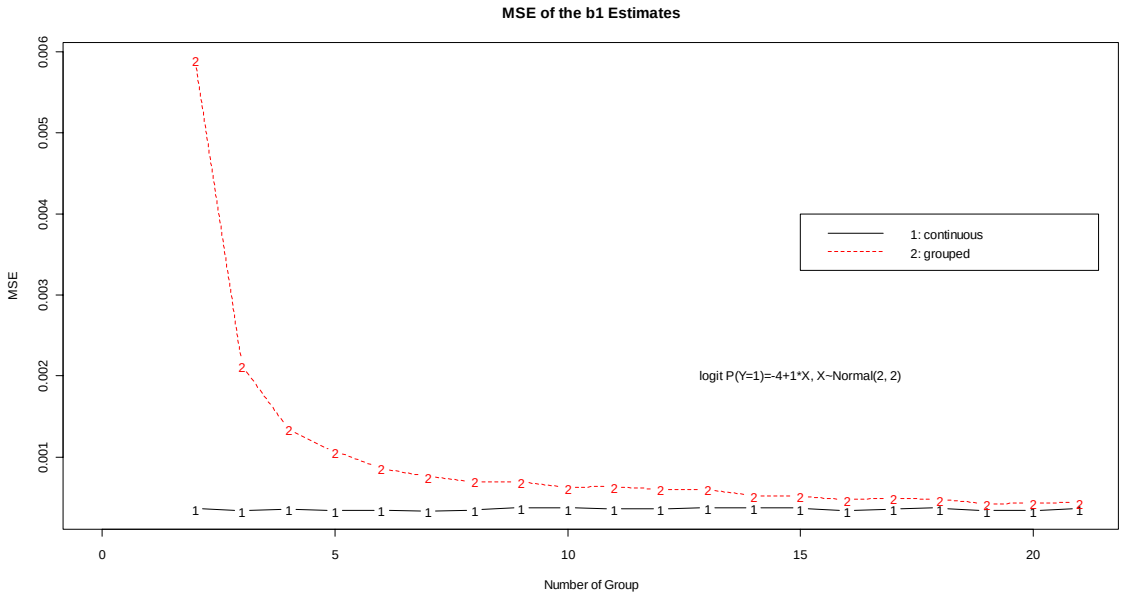


Figure 6.3.6 MSE of Coefficient Estimate when X distributed normal (3, 3) and β_1=1

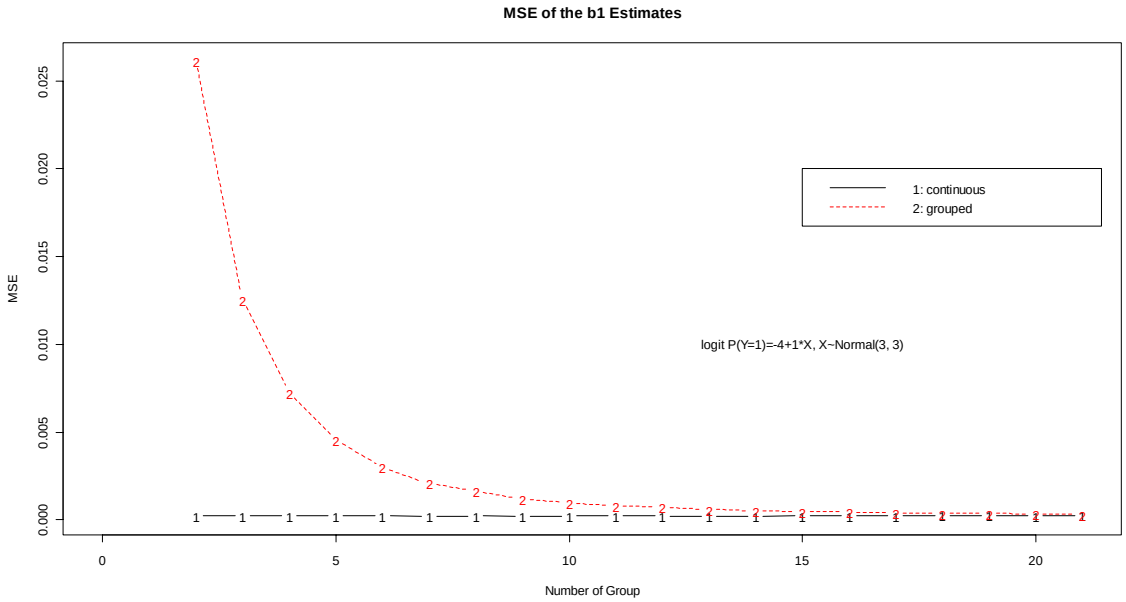


Figure 6.3.7 MSE of Coefficient Estimate when X distributed normal (4, 4) and β_1=1

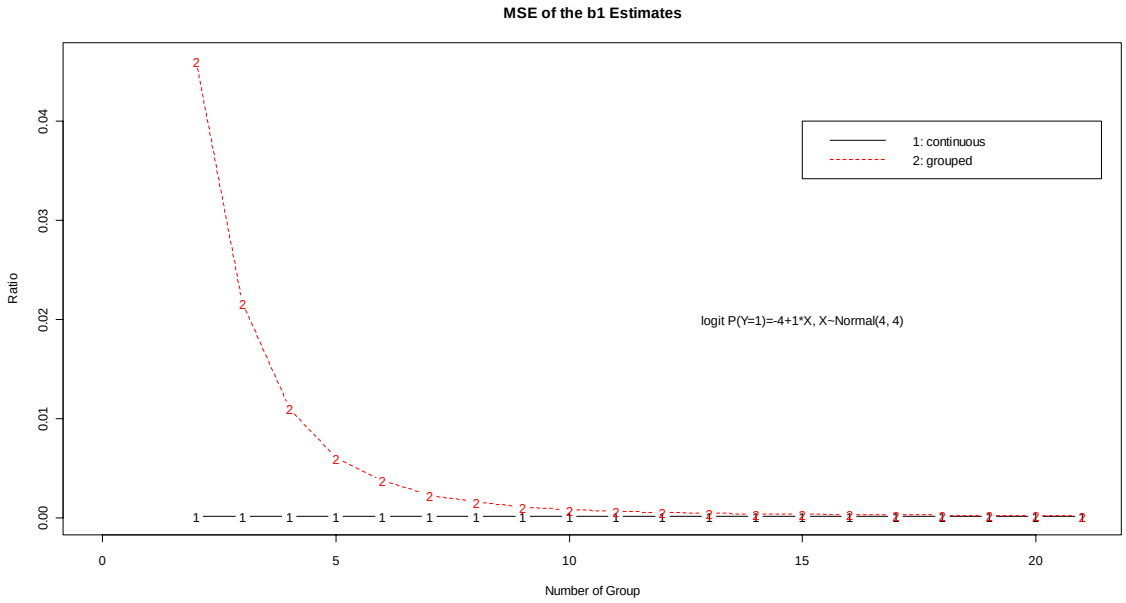
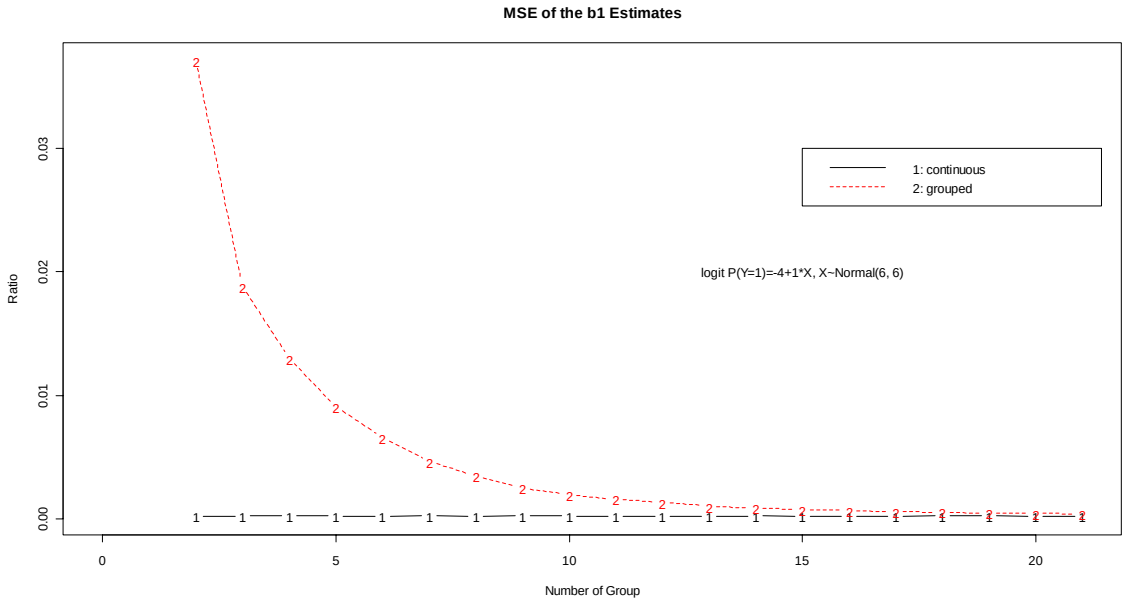


Figure 6.3.8 MSE of Coefficient Estimate when X distributed normal (6, 6) and β_1=1



6.3.4 Gamma Covariate and β≠0

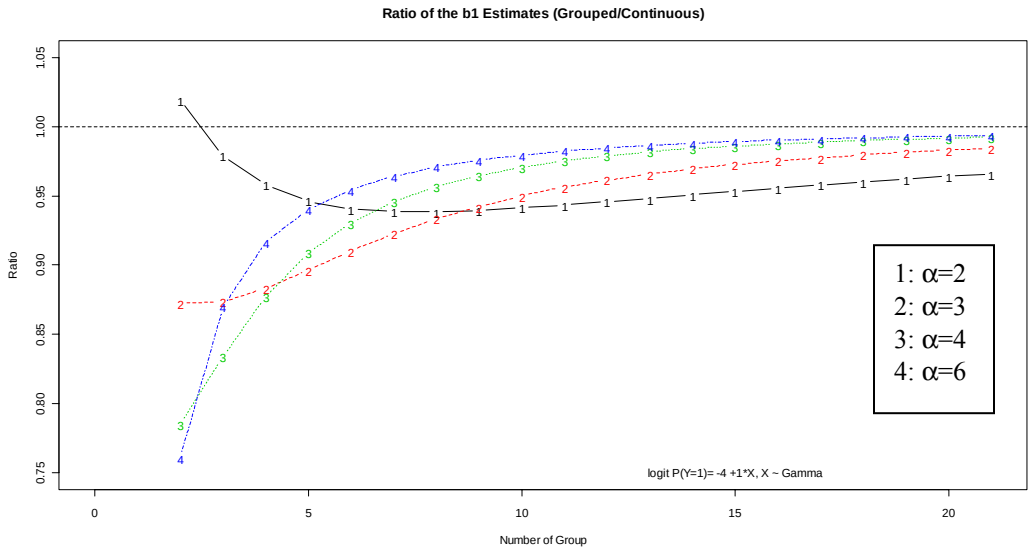
We also performed simulation studies to evaluate the association between REs and the number of cutoff points by using the gamma distribution for the covariate. The studies were under the assumption that there is an association between the covariate and the outcome variable. The slope is 1 with intercept of -4.

The covariate was from a gamma distribution with 4 different shape parameters: (2, 3, 4, 6). We used 1 as the scale parameter for all of the studies.

We evaluate the association between coefficients estimated by using the continuous and categorized covariate and the number of groups. The graphical results are shown in Figure 6.3.9. The numerical results are shown in Table B.19 in Appendix B.

Other than when α=2 at 2 groups, we found that the coefficient of dose-response association were underestimated. The magnitude of under-estimation decreases with the number of group. However, the magnitude of under-estimation is still larger than the other shape parameters when α=2 with the increased number of group.

Figure 6.3.9 Ratio of the b1 Estimates (Grouped/Continuous) when X is Gamma

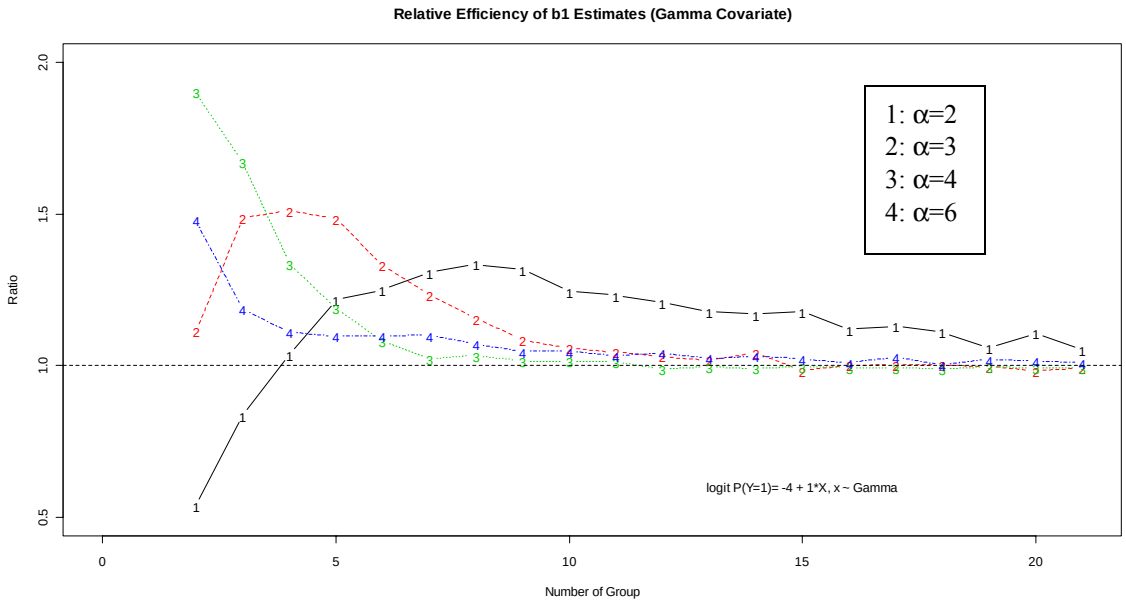


For better understand the impact on estimation bias, we calculated the relative bias (100%x[estimate-parameter]/parameter). The relative bias depends on the parameter. Almost all of the estimates from categorized covariates are under-estimated. The results are shown in Tables B.20 and B.21 in Appendix B.

The relative efficiency of coefficient estimates was also evaluated. The graphic results are shown in Figure 6.3.10. The numeric results are shown in Table B.22 in Appendix B.

From the results, we found that the association between REs and the number of groups change with the parameters. The REs were approaching 1 when the number of group increases. However, when α=2, the REs are still be away from 1, even it decreases.

Figure 6.3.10 Relative Efficiency of b1 Estimate when X is gamma distribution and β_1=1



The MSE of the estimates were also studied. The graphic results are shown in Figure 6.3.11 through Figure 6.3.14. The numeric results are shown in Table B.23 and Table B.24.

Under each gamma distribution, the MSEs of the coefficient from the continuous covariate were smaller than the MSEs from the grouped covariate. The magnitude of different on MSE changes with the shape parameter.

Figure 6.3.11 MSE of Coefficient Estimate when X distributed gamma (2, 1) and β_1=1

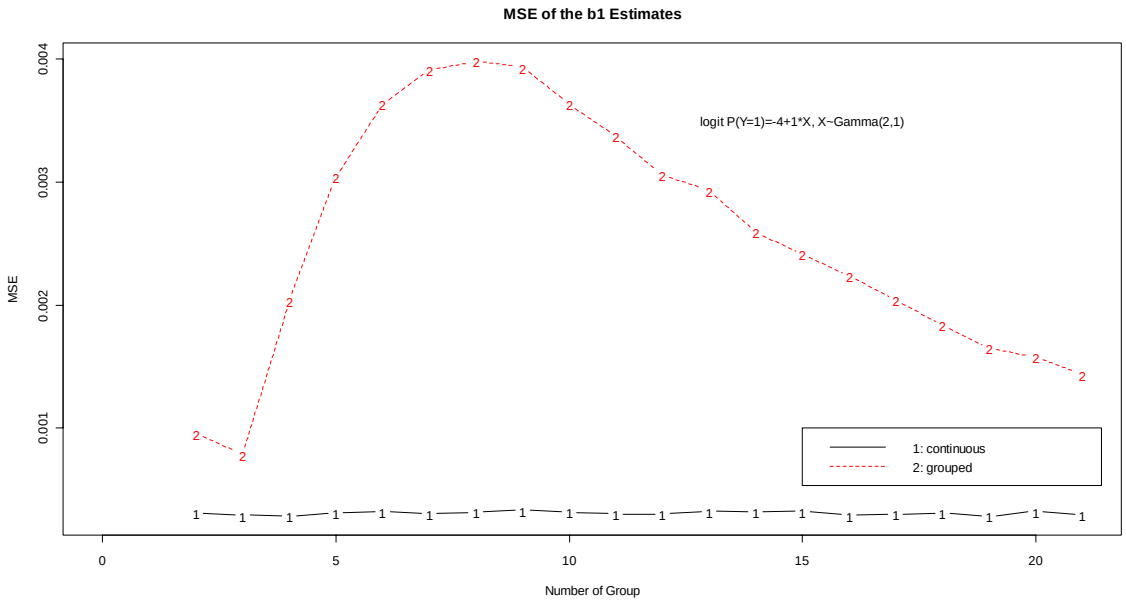


Figure 6.3.12 MSE of Coefficient Estimate when X distributed gamma (3, 1) and β_1=1

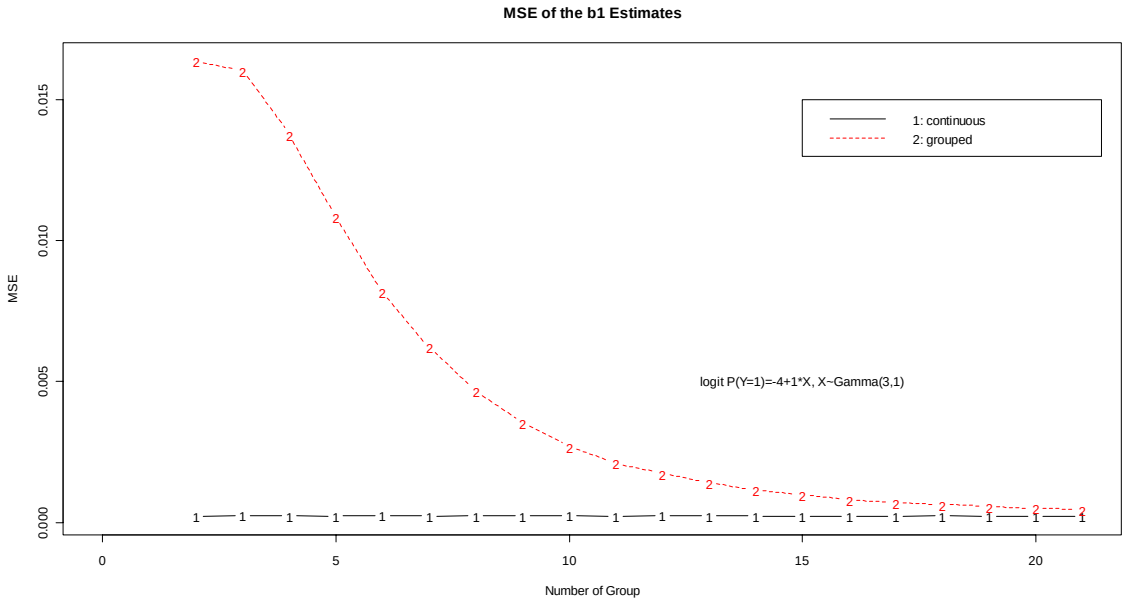


Figure 6.3.13 MSE of Coefficient Estimate when X distributed gamma (4, 1) and β_1=1

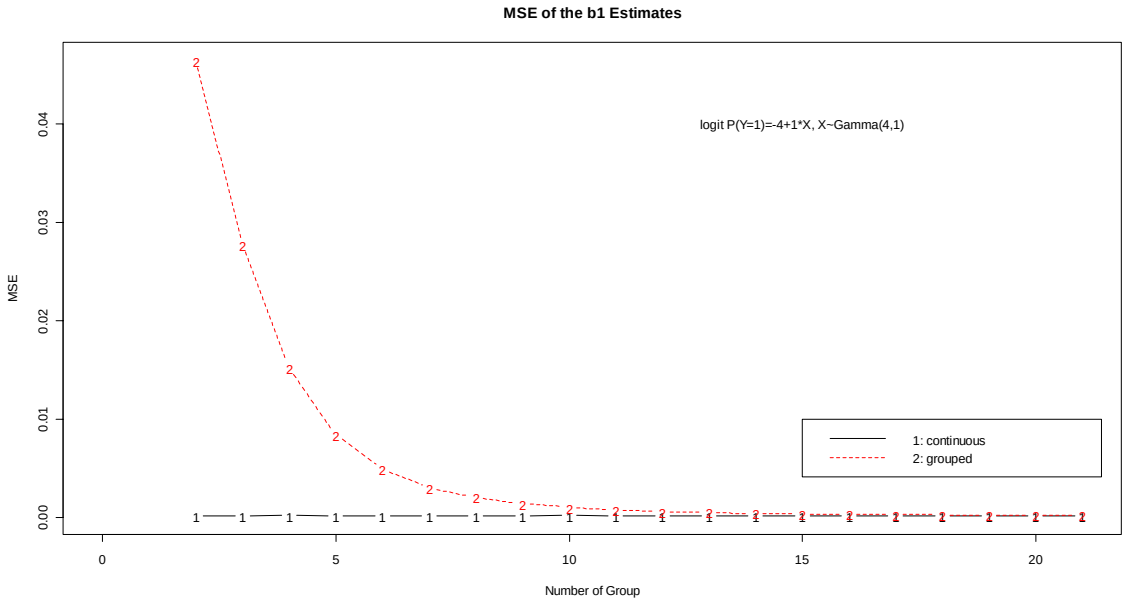
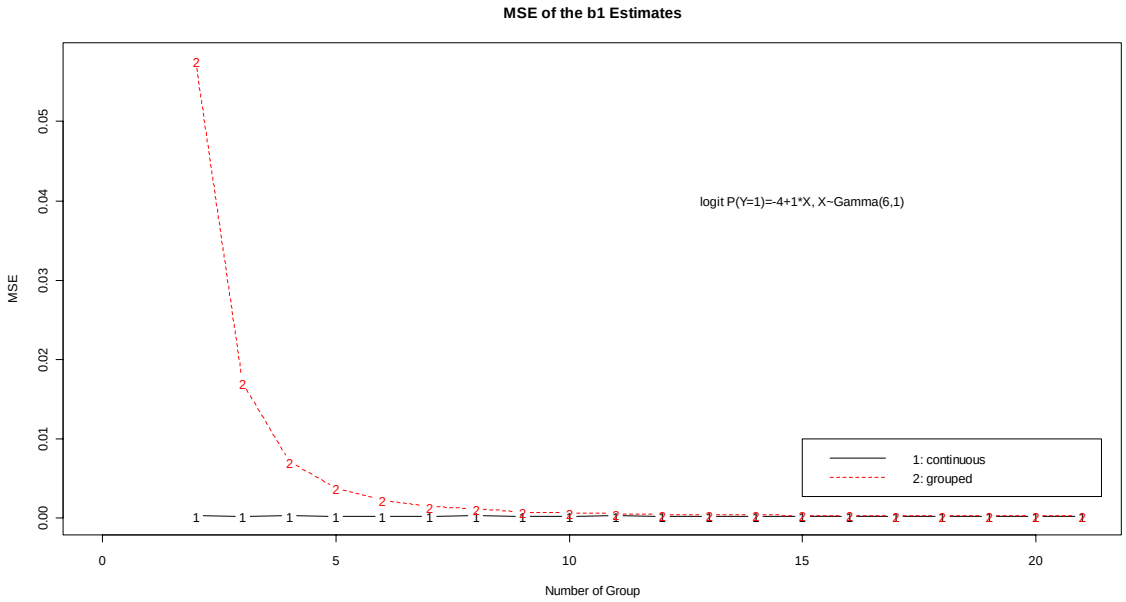


Figure 6.3.14 MSE of Coefficient Estimate when X distributed gamma (6, 1) and β_1=1



6.4 Conclusions

Our simulation studies demonstrated that when a continuous covariate is dichotomized, the relative efficiency reduced under the null hypothesis. The magnitude of reduction depends on the covariate distribution and the choice of cutoff points.

When there is an association between the covariate and the outcome variable, the coefficient estimate might be biased due to the dichotomization. The mean square error (MSE) of the estimate from dichotomized covariate is larger than the MSE of the estimate from continuous covariate.

When the number of group increases, the relative efficiency increases.

Chapter 7

Effect of Categorizing a Continuous Covariate on the Comparison of Survival Time and Dose Response

When performing a multivariate analysis, it is very often that a continuous covariate is categorized to be treated as a confounding variable. For example, when assessing the treatment effect as compared to the control group, age is a potential confounding factor which we control for. Instead of using the real age value, researchers often categorized age as age group, such as: "Young vs. Old", or "Young, Mid-age, and Old".

As described in Section 2.3, Morgan and Elashoff (1986) assessed the impacts from categorizing a gamma-distributed covariate. The choice of cutoff point, the parameter of gamma distribution and the number of cutoff point impact the asymptotic relative efficiency. However, their study was under the assumption that the null hypothesis is true, that is, there is neither effect from the main effect nor the effect from the confounding variable.

In this Chapter we will evaluate the impact on the estimation of main effect from categorizing a continuous covariate. We will assess the effect of categorization under the assumption that null hypothesis is true.

7.1 Model

Let Y_i be an independent and identically distributed (i.i.d.) random variable with a density function $f_{\xi}(y)$. Let the expected value of Y_i, μ, be a linear function of an i.i.d. continuous random variable X_i. That is, we can use the concept of generalized linear model to define

$$G(\mu) = G(E[Y]) = \boldsymbol{\beta X}$$

where G is a link function (McCullagh and Nelder, 1989), $\boldsymbol{\beta}$ is the vector of regression coefficient, and $\boldsymbol{X}$ is the vector of explanatory variables. For simplicity, we use $G(\mu) = G(E[Y]) = \beta_0 + \beta_1 Z + \beta_2 X$ for this chapter. Z is a dichotomized variable.

When β_1 is the coefficient of interest under our regression model, based on the method of maximum likelihood, the following steps of deriving the general form for calculating asymptotic relative efficiency will be similar to what we described in Chapter 6. However, it will be different because more coefficients are included in the regression model.

Let $f(Y \mid Z,\ X,\ \beta)$ be a density function of Y given Z, X and β. When X_{1i} is categorized into the j^{th} interval $[C_{j-1}, C_j]$, we define X^*_{ij} as:

$$X^*_{ij} = \begin{cases} 1 & \text{if } C_{j-1} \le X_i < C_j \\ 0 & \text{otherwise} \end{cases}$$

where j=1, … m, and m is the number of groups. Therefore, we have the conditional density function

$$f(y_i \mid Z,\ X^*_{ij},\ \beta) \quad = f(y_i \mid Z,\ C_{j-1} \le X_i < C_j,\ \beta)$$

$$= \frac{f(y_i, C_{j-1} \le X_i < C_j \mid Z,\ \beta)}{f_X(C_{j-1} \le X_i < C_j)}$$

$$= \frac{\int_{C_{j-1}}^{C_j} f(y, x \mid Z,\ \beta)dx}{\int_{C_{j-1}}^{C_j} f_X(x)dx}$$

$$= \frac{\int_{C_{j-1}}^{C_j} f(y \mid Z,\ x,\ \beta) f(x)dx}{\int_{C_{j-1}}^{C_j} f_X(x)dx}$$

$$= \frac{1}{P_j} \int_{C_{j-1}}^{C_j} f(y \mid Z,\ x,\ \beta) f(x)dx$$

where $P_j = \int_{C_{j-1}}^{C_j} f_X(x)dx$

When we want to estimate the parameter of interest from y and x, we use the density function $f(y \mid Z,\ X,\ \beta)$. Let log-likelihood function

$$l(\beta \mid z,\ y,\ x) = \sum_{i=1}^{n} l_i(\beta \mid z_i,\ y_i,\ x_i) = \sum_{i=1}^{n} \log f(y_i \mid z_i,\ x_i,\ \beta)$$

where $l_i(\beta \mid z_i,\ y_i,\ x_i) = \log f(y_i \mid z_i,\ x_i,\ \beta)$.

Based on the regression model described previously, $\beta = (\beta_0,\ \beta_1,\ \beta_2)$. Therefore, we can calculate the score function,

$$\frac{\partial l_i(\beta \mid z_i,\ y_i,\ x_i)}{\partial \beta} = \frac{\partial \log f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta}$$

$$= \frac{\frac{\partial}{\partial \beta} f(y_i \mid z_i,\ x_i,\ \beta)}{f(y_i \mid z_i,\ x_i,\ \beta)}$$

From the result, we can also calculate the expected Fisher information which is the expected value of the product of the first derivative of the log-likelihood function, or the

negative expected value of the second derivative. When we use the negative expected value of the second derivative, the second derivative is:

$$\frac{\partial^2 l_i(\beta \mid z_i,\, y_i,\, x_i)}{\partial\beta\partial\beta^T} = \frac{\partial}{\partial\beta^T}\left[\frac{1}{f(y_i \mid z_i,\, x_i,\, \beta)}\frac{\partial}{\partial\beta} f(y_i \mid z_i,\, x_i,\, \beta)\right]'$$

$$= \frac{f(y_i \mid z_i,\, x_i,\, \beta)\frac{\partial^2}{\partial\beta\partial\beta^T} f(y_i \mid z_i,\, x_i,\, \beta) - \frac{\partial}{\partial\beta} f(y_i \mid z_i,\, x_i,\, \beta)\frac{\partial}{\partial\beta^T} f(y_i \mid z_i,\, x_i,\, \beta)}{\left[f(y_i \mid z_i,\, x_i,\, \beta)\right]^2}$$

$$I_i(\beta) = E\left[-\frac{\partial^2 l_i(\beta \mid z_i,\, y_i,\, x_i)}{\partial\beta\partial\beta^T}\right]$$

$$= \int_{-\infty}^{\infty} \frac{\frac{\partial f(y_i \mid z_i,\, x_i,\, \beta)}{\partial\beta}\frac{\partial f(y_i \mid z_i,\, x_i,\, \beta)}{\partial\beta^T} - f(y_i \mid z_i,\, x_i,\, \beta)\frac{\partial^2 f(y_i \mid z_i,\, x_i,\, \beta)}{\partial\beta\partial\beta^T}}{\left[f(y_i \mid z_i,\, x_i,\, \beta)\right]^2} f(y_i \mid z_i,\, x_i,\, \beta)dy_i$$

$$= \int_{-\infty}^{\infty} \frac{\frac{\partial f(y_i \mid z_i,\, x_i,\, \beta)}{\partial\beta}\frac{\partial f(y_i \mid z_i,\, x_i,\, \beta)}{\partial\beta^T} - f(y_i \mid z_i,\, x_i,\, \beta)\frac{\partial^2 f(y_i \mid z_i,\, x_i,\, \beta)}{\partial\beta\partial\beta^T}}{f(y_i \mid z_i,\, x_i,\, \beta)} dy_i$$

$$= \begin{bmatrix} I_{i11} & I_{i12} & I_{i13} \\ I_{i21} & I_{i22} & I_{i23} \\ I_{i31} & I_{i32} & I_{i33} \end{bmatrix}$$

Based on our model, we have:

$$\frac{\partial}{\partial\beta} f(y_i \mid z_i,\, x_i,\, \beta) = \begin{bmatrix} \frac{\partial}{\partial\beta_0} f(y_i \mid z_i,\, x_i,\, \beta) \\ \frac{\partial}{\partial\beta_1} f(y_i \mid z_i,\, x_i,\, \beta) \\ \frac{\partial}{\partial\beta_2} f(y_i \mid z_i,\, x_i,\, \beta) \end{bmatrix}$$

$$\frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta}\frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta^T}$$

$$= \begin{bmatrix} \frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_0}\frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_0} & \frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_0}\frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_1} & \frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_0}\frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_2} \\ \frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_1}\frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_0} & \frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_1}\frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_1} & \frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_1}\frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_2} \\ \frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_2}\frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_0} & \frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_2}\frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_1} & \frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_2}\frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_2} \end{bmatrix}$$

$$\frac{\partial^2 f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta \partial \beta^T} = \begin{bmatrix} \frac{\partial^2 f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_0 \partial \beta_0} & \frac{\partial^2 f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_0 \partial \beta_2} \\ \frac{\partial^2 f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_1 \partial \beta_1} & \frac{\partial^2 f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_1 \partial \beta_2} \\ \frac{\partial^2 f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_2 \partial \beta_0} & \frac{\partial^2 f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_2 \partial \beta_2} \end{bmatrix}$$

Therefore, the (a, b)th component of the information matrix is:

$$I_{iab} = \int_{-\infty}^{\infty} \frac{\frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_{a-1}}\frac{\partial f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_{b-1}} - f(y_i \mid z_i,\ x_i,\ \beta)\frac{\partial^2 f(y_i \mid z_i,\ x_i,\ \beta)}{\partial \beta_{a-1}\partial \beta_{b-1}}}{f(y_i \mid z_i,\ x_i,\ \beta)} dy_i$$

When we want to estimate the parameter of interest from y and categorized x, we use the density function $f(y \mid Z,\ X^*,\ \beta)$. Let log-likelihood function

$$l^*(\beta \mid z,\ y,\ x) = \sum_{i=1}^{n} l_i^*(\beta \mid z_i,\ y_i,\ x_i^*) = \sum_{i=1}^{n} \log f(y_i \mid z_i,\ x_i^*,\ \beta)$$

where $l_i^*(\beta \mid z_i,\ y_i,\ x_i^*) = \log f(y_i \mid z_i,\ x_i^*,\ \beta)$.

Therefore, we can calculate the score function,

$$\frac{\partial l_i^*(\beta \mid z_i,\ y_i,\ x_i^*)}{\partial \beta} = \frac{\partial \log f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta}$$

$$= \frac{\frac{\partial}{\partial \beta} f(y_i \mid z_i,\ x_i^*,\ \beta)}{f(y_i \mid z_i,\ x_i^*,\ \beta)}$$

From the result, we can calculate the expected Fisher information which is the expected value of the product of the first derivative of the log-likelihood function, or the negative expected value of the second derivative. When we use the negative expected value of the second derivative, the second derivative is:

$$\frac{\partial^2 l_i^*(\beta \mid z_i,\ y_i,\ x_i^*)}{\partial\beta\partial\beta^T} = \frac{\partial}{\partial\beta^T}\left[\frac{1}{f(y_i \mid z_i,\ x_i^*,\ \beta)}\frac{\partial}{\partial\beta} f(y_i \mid z_i,\ x_i^*,\ \beta)\right]'$$

$$= \frac{f(y_i \mid z_i,\ x_i^*,\ \beta)\dfrac{\partial^2}{\partial\beta\partial\beta^T} f(y_i \mid z_i,\ x_i^*,\ \beta) - \dfrac{\partial}{\partial\beta} f(y_i \mid z_i,\ x_i^*,\ \beta)\dfrac{\partial}{\partial\beta^T} f(y_i \mid z_i,\ x_i^*,\ \beta)}{\left[f(y_i \mid z_i,\ x_i^*,\ \beta)\right]^2}$$

$$I^*_i(\beta) = E\left[-\frac{\partial^2 l_i^*(\beta \mid z_i,\ y_i,\ x_i^*)}{\partial\beta\partial\beta^T}\right]$$

$$= \int_{-\infty}^{\infty} \frac{\dfrac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial\beta}\dfrac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial\beta^T} - f(y_i \mid z_i,\ x_i^*,\ \beta)\dfrac{\partial^2 f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial\beta\partial\beta^T}}{\left[f(y_i \mid z_i,\ x_i^*,\ \beta)\right]^2} f(y_i \mid z_i,\ x_i^*,\ \beta)\,dy_i$$

$$= \int_{-\infty}^{\infty} \frac{\dfrac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial\beta}\dfrac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial\beta^T} - f(y_i \mid z_i,\ x_i^*,\ \beta)\dfrac{\partial^2 f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial\beta\partial\beta^T}}{\left[\int_{C_{j-1}}^{C_j} f(y_i \mid z_i,\ x_i,\ \beta) f(x_i)dx_i\right]^2} \frac{\int_{C_{j-1}}^{C_j} f(y_i \mid z_i,\ x_i,\ \beta) f(x_i)dx_i}{P_j}\,dy_i$$

$$= \int_{-\infty}^{\infty} \frac{\dfrac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial\beta}\dfrac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial\beta^T} - f(y_i \mid z_i,\ x_i^*,\ \beta)\dfrac{\partial^2 f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial\beta\partial\beta^T}}{P_j\int_{C_{j-1}}^{C_j} f(y_i \mid z_i,\ x_i,\ \beta) f(x_i)dx_i}\,dy_i$$

$$= \begin{bmatrix} I^*_{i11} & I^*_{i12} & I^*_{i13} \\ I^*_{i21} & I^*_{i22} & I^*_{i23} \\ I^*_{i31} & I^*_{i32} & I^*_{i33} \end{bmatrix}$$

Based on our model, we have:

$$\frac{\partial}{\partial \beta} f(y_i \mid z_i,\ x_i^*,\ \beta) = \begin{bmatrix} \frac{\partial}{\partial \beta_0} f(y_i \mid z_i,\ x_i^*,\ \beta) \\ \frac{\partial}{\partial \beta_1} f(y_i \mid z_i,\ x_i^*,\ \beta) \\ \frac{\partial}{\partial \beta_2} f(y_i \mid z_i,\ x_i^*,\ \beta) \end{bmatrix}$$

$$\frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta} \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta^T}$$

$$= \begin{bmatrix} \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_0} \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_0} & \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_0} \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_1} & \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_0} \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_2} \\ \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_1} \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_0} & \frac{df(y_i \mid z_i,\ x_i^*,\ \beta)}{d\beta_1} \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_1} & \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_1} \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_2} \\ \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_2} \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_0} & \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_2} \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_1} & \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_2} \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_2} \end{bmatrix}$$

$$\frac{\partial^2 f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta \partial \beta^T} = \begin{bmatrix} \frac{\partial^2 f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_0 \partial \beta_0} & \frac{\partial^2 f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_0 \partial \beta_2} \\ \frac{\partial^2 f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_1 \partial \beta_1} & \frac{\partial^2 f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_1 \partial \beta_2} \\ \frac{\partial^2 f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_2 \partial \beta_0} & \frac{\partial^2 f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_2 \partial \beta_2} \end{bmatrix}$$

Therefore, the (a, b)[th] component of the information matrix is:

$$I^*_{iab} = \int_{-\infty}^{\infty} \frac{\frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_{a-1}} \frac{\partial f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_{b-1}} - f(y_i \mid z_i,\ x_i^*,\ \beta) \frac{\partial^2 f(y_i \mid z_i,\ x_i^*,\ \beta)}{\partial \beta_{a-1} \partial \beta_{b-1}}}{P_j \int_{C_{j-1}}^{C_j} f(y_i \mid z_i,\ x_i,\ \beta) f(x_i) dx_i} dy_i$$

When we use the data from n observations, the average expected information matrix based on the original values is

$$I(\beta) = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} I_{i11} & \sum_{i=1}^{n} I_{i12} & \sum_{i=1}^{n} I_{i13} \\ \sum_{i=1}^{n} I_{i21} & \sum_{i=1}^{n} I_{i22} & \sum_{i=1}^{n} I_{i213} \\ \sum_{i=1}^{n} I_{i31} & \sum_{i=1}^{n} I_{i32} & \sum_{i=1}^{n} I_{i33} \end{bmatrix}$$

Therefore, the variance of maximum likelihood estimator of β is

$$Var(\hat{\beta}) = \frac{I^{-1}(\beta)}{n}$$

$$= \frac{1}{DET}\begin{bmatrix} \sum_{i=1}^{n} I_{i33}\sum_{i=1}^{n} I_{i22} - \sum_{i=1}^{n} I_{i32}\sum_{i=1}^{n} I_{i23} & -(\sum_{i=1}^{n} I_{i33}\sum_{i=1}^{n} I_{i12} - \sum_{i=1}^{n} I_{i32}\sum_{i=1}^{n} I_{i13}) & \sum_{i=1}^{n} I_{i23}\sum_{i=1}^{n} I_{i12} - \sum_{i=1}^{n} I_{i22}\sum_{i=1}^{n} I_{i13} \\ -(\sum_{i=1}^{n} I_{i33}\sum_{i=1}^{n} I_{i21} - \sum_{i=1}^{n} I_{i31}\sum_{i=1}^{n} I_{i23}) & \sum_{i=1}^{n} I_{i33}\sum_{i=1}^{n} I_{i11} - \sum_{i=1}^{n} I_{i31}\sum_{i=1}^{n} I_{i13} & -(\sum_{i=1}^{n} I_{i23}\sum_{i=1}^{n} I_{i11} - \sum_{i=1}^{n} I_{i21}\sum_{i=1}^{n} I_{i13}) \\ \sum_{i=1}^{n} I_{i32}\sum_{i=1}^{n} I_{i21} - \sum_{i=1}^{n} I_{i31}\sum_{i=1}^{n} I_{i22} & -(\sum_{i=1}^{n} I_{i32}\sum_{i=1}^{n} I_{i11} - \sum_{i=1}^{n} I_{i31}\sum_{i=1}^{n} I_{i12}) & \sum_{i=1}^{n} I_{i22}\sum_{i=1}^{n} I_{i11} - \sum_{i=1}^{n} I_{i21}\sum_{i=1}^{n} I_{i12} \end{bmatrix}$$

where $DET = \sum_{i=1}^{n} I_{i11}(\sum_{i=1}^{n} I_{i32}\sum_{i=1}^{n} I_{i22} - \sum_{i=1}^{n} I_{i32}\sum_{i=1}^{n} I_{i23}) - \sum_{i=1}^{n} I_{i21}(\sum_{i=1}^{n} I_{i33}\sum_{i=1}^{n} I_{i12} - \sum_{i=1}^{n} I_{i32}\sum_{i=1}^{n} I_{i13})$

$$+\sum_{i=1}^{n} I_{i31}(\sum_{i=1}^{n} I_{i23}\sum_{i=1}^{n} I_{i12} - \sum_{i=1}^{n} I_{i22}\sum_{i=1}^{n} I_{i13})$$

The equations available at www.wolframalpha.com (Weisstein, 2010) were used to calculate the inverse of a square matrix.

For the coefficient of interest β_1, the variance of maximum likelihood estimator is

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} I_{i33}\sum_{i=1}^{n} I_{i11} - \sum_{i=1}^{n} I_{i31}\sum_{i=1}^{n} I_{i13}}{DET}$$

When we use the data from n observations, the average expected information matrix based on the categorized values is

$$I^*(\beta) = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} I^*_{i11} & \sum_{i=1}^{n} I^*_{i12} & \sum_{i=1}^{n} I^*_{i13} \\ \sum_{i=1}^{n} I^*_{i21} & \sum_{i=1}^{n} I^*_{i22} & \sum_{i=1}^{n} I^*_{i213} \\ \sum_{i=1}^{n} I^*_{i31} & \sum_{i=1}^{n} I^*_{i32} & \sum_{i=1}^{n} I^*_{i33} \end{bmatrix}$$

Therefore, the variance of maximum likelihood estimator of β^* is

$$Var(\hat{\beta}^*) = \frac{I^{*-1}(\beta)}{n}$$

$$= \frac{1}{DET^*}\begin{bmatrix} \sum_{i=1}^{n} I^*_{i33}\sum_{i=1}^{n} I^*_{i22} - \sum_{i=1}^{n} I^*_{i32}\sum_{i=1}^{n} I^*_{i23} & -(\sum_{i=1}^{n} I^*_{i33}\sum_{i=1}^{n} I^*_{i12} - \sum_{i=1}^{n} I^*_{i32}\sum_{i=1}^{n} I^*_{i13}) & \sum_{i=1}^{n} I^*_{i23}\sum_{i=1}^{n} I^*_{i12} - \sum_{i=1}^{n} I^*_{i22}\sum_{i=1}^{n} I^*_{i13} \\ -(\sum_{i=1}^{n} I^*_{i33}\sum_{i=1}^{n} I^*_{i21} - \sum_{i=1}^{n} I^*_{i31}\sum_{i=1}^{n} I^*_{i23}) & \sum_{i=1}^{n} I^*_{i33}\sum_{i=1}^{n} I^*_{i11} - \sum_{i=1}^{n} I^*_{i31}\sum_{i=1}^{n} I^*_{i13} & -(\sum_{i=1}^{n} I^*_{i23}\sum_{i=1}^{n} I^*_{i11} - \sum_{i=1}^{n} I^*_{i21}\sum_{i=1}^{n} I^*_{i13}) \\ \sum_{i=1}^{n} I^*_{i32}\sum_{i=1}^{n} I^*_{i21} - \sum_{i=1}^{n} I^*_{i31}\sum_{i=1}^{n} I^*_{i22} & -(\sum_{i=1}^{n} I^*_{i32}\sum_{i=1}^{n} I^*_{i11} - \sum_{i=1}^{n} I^*_{i31}\sum_{i=1}^{n} I^*_{i12}) & \sum_{i=1}^{n} I^*_{i22}\sum_{i=1}^{n} I^*_{i11} - \sum_{i=1}^{n} I^*_{i21}\sum_{i=1}^{n} I^*_{i12} \end{bmatrix}$$

where $DET^* = \sum_{i=1}^{n} I^*_{i11}(\sum_{i=1}^{n} I^*_{i32}\sum_{i=1}^{n} I^*_{i22} - \sum_{i=1}^{n} I^*_{i32}\sum_{i=1}^{n} I^*_{i23})$

$$-\sum_{i=1}^{n} I^*_{i21}(\sum_{i=1}^{n} I^*_{i33}\sum_{i=1}^{n} I^*_{i12} - \sum_{i=1}^{n} I^*_{i32}\sum_{i=1}^{n} I^*_{i13})$$

$$+\sum_{i=1}^{n} I^*_{i31}(\sum_{i=1}^{n} I^*_{i23}\sum_{i=1}^{n} I^*_{i12} - \sum_{i=1}^{n} I^*_{i22}\sum_{i=1}^{n} I^*_{i13})$$

For the coefficient of interest β_1, the variance of maximum likelihood estimator is

$$Var(\hat{\beta}^*_1) = \frac{\sum_{i=1}^{n} I^*_{i33}\sum_{i=1}^{n} I^*_{i11} - \sum_{i=1}^{n} I^*_{i31}\sum_{i=1}^{n} I^*_{i13}}{DET^*}$$

Therefore, we can calculate the asymptotic relative efficiency by using the variance of β_1 estimated from original observations divided by the variance estimated from the categorical observations,

$$\text{ARE} = \frac{Var(\hat{\beta}_1)}{Var(\hat{\beta}^*_1)}$$

The expected Fisher information can also be calculated by using the product of the first derivative of log-likelihood function.

7.1.1 Model for Survival Time

Morgan and Elashoff (1986) studied the impacts from categorizing a continuous covariate on survival time estimates using the asymptotic relative efficiency. However, their paper did not provide the full details of deriving the equation but referred to a technical report which was not available via request. Therefore, the general equation on ARE derived in the previous section was used to justify their equation and to provide more details of the steps. The original notations were modified to be consistent with the notion in this dissertation.

Define Notations and Associations

Let hazard of exponential proportional hazard model be $h = e^{\mu+\alpha Z+\beta X^*}$, where

$$z = \begin{cases} 1 & \text{if treatment group} \\ 0 & \text{if control group} \end{cases}$$

Let $\lambda = e^{\alpha}$, $X = e^{\mu+\beta X^*}$

Therefore, $h = \lambda X$. This simplifies from 3 parameters to 1 parameter.

Given the hazard function, the conditional distribution of the survival time Y given X follows an exponential distribution with the parameter of $h = \lambda X$. That is, the density function of Y given X is expressed as:

$$f(Y \mid X) = \lambda X e^{-\lambda XY}$$

When X is categorized, the density function of Y within the j^{th} interval:

$$f(Y_j \mid X_j^*) = f(Y_j \mid C_{j-1} \le X_j < C_j)$$

$$= \frac{1}{P_j} \int_{C_{j-1}}^{C_j} f(y \mid x) f(x) dx \qquad \text{where } P_j = \int_{C_{j-1}}^{C_j} f_X(x) dx$$

From $f(Y \mid X) = \lambda X e^{-\lambda XY}$,

$$\log f(Y \mid X) = \log \lambda + \log X - \lambda XY$$

Therefore, $$\frac{d}{d\lambda} \log f(Y \mid X) = \frac{1}{\lambda} - XY$$

$$\frac{d^2}{d\lambda^2} \log f(Y \mid X) = \frac{-1}{\lambda^2}$$

By using the result above to the first derivative,

$$\frac{d}{d\lambda} \log f(Y \mid X) = \frac{1}{f(Y \mid X)} \frac{d}{d\lambda} f(Y \mid X) = \frac{1}{\lambda} - XY$$

After rearranging the later parts of equation, we got

$$\frac{d}{d\lambda} f(Y \mid X) = (\frac{1}{\lambda} - XY) f(Y \mid X)$$

Calculate Fisher Information

When we estimate λ from Y with continuous covariate X, the likelihood function

$$L_1(\lambda \mid y) = f(y) = \int_0^\infty f(y \mid x) f(x) dx$$

We assume that there is no censoring for the survival time.

Let $l_1(\lambda) = \log L_1(\lambda \mid y) = \log f(y)$

$$= \log \int_0^\infty f(y \mid x) f(x) dx$$

$$l_1'(\lambda) = \frac{1}{f(y)} f'(y)$$

$$= \frac{1}{f(y)} \frac{d}{d\lambda} \int_0^\infty f(y \mid x) f(x) dx$$

(differentiating under an integral sign)

$$= \frac{1}{f(y)} \int_0^\infty \left[\frac{d}{d\lambda} f(y \mid x) \right] f(x) dx$$

[because $\frac{d}{d\lambda} f(Y \mid X) = (\frac{1}{\lambda} - XY) f(Y \mid X)$]

$$= \frac{1}{f(y)} \int_0^\infty \left[(\frac{1}{\lambda} - XY) f(Y \mid X) \right] f(x) dx$$

[replace $f(y)$ with $\int_0^\infty f(y \mid x) f(x) dx$]

$$= \frac{\int_0^\infty \left[(\frac{1}{\lambda} - XY) f(Y \mid X) \right] f(x) dx}{\int_0^\infty f(y \mid x) f(x) dx}$$

We use the expected value of the product of first derivative of log-likelihood function to calculate the Fisher information:

$$E[l_1'(\lambda)^2] = \int_0^\infty \left[\frac{\int_0^\infty \left[(\frac{1}{\lambda} - XY) f(Y \mid X) \right] f(x) dx}{\int_0^\infty f(y \mid x) f(x) dx} \right]^2 f(y) dy$$

[replace $f(y)$ with $\int_0^\infty f(y \mid x) f(x) dx$]

$$= \int_0^\infty \left[\frac{\int_0^\infty \left[(\frac{1}{\lambda} - XY) f(Y \mid X) \right] f(x) dx}{\int_0^\infty f(y \mid x) f(x) dx} \right]^2 \left[\int_0^\infty f(y \mid x) f(x) dx \right] dy$$

[cancel out $\int_0^\infty f(y \mid x) f(x) dx$]

$$= \int_0^\infty \frac{\left[\int_0^\infty \left[(\frac{1}{\lambda} - XY) f(Y \mid X) \right] f(x) dx \right]^2}{\int_0^\infty f(y \mid x) f(x) dx} dy$$

When we estimate λ from Y with categorized covariate X, the likelihood function

$$L_1^*(\lambda \mid y^*) = f(y^*) = \frac{\int_{C_{j-1}}^{C_j} f(y \mid x) f(x) dx}{P_j}$$

Let $l_1^*(\lambda) = \log L_1^*(\lambda \mid y^*) = \log f(y^*)$

$$l_1^{*\prime}(\lambda) = \frac{1}{f(y^*)} f'(y^*)$$

$$= \frac{1}{f(y^*)} \frac{d}{d\lambda} \left[\frac{1}{P_j} \int_{C_{j-1}}^{C_j} f(y \mid x) f(x) dx \right]$$

(differentiating under an integral sign)

$$= \frac{1}{f(y^*)}\left\{\frac{1}{P_j}\int_{C_{j-1}}^{C_j}\left[\frac{d}{d\lambda} f(y \mid x)\right] f(x)dx\right\}$$

[because $\frac{d}{d\lambda} f(y \mid x) = \left(\frac{1}{\lambda} - xy\right) f(y \mid x)$]

$$= \frac{1}{f(y^*)}\frac{1}{P_j}\int_{C_{j-1}}^{C_j}\left[\left(\frac{1}{\lambda} - xy\right) f(y \mid x)\right] f(x)dx$$

[replace $f(y^*)$ with $\frac{\int_{C_{j-1}}^{C_j} f(y \mid x) f(x)dx}{P_j}$]

$$= \frac{\frac{1}{P_j}\int_{C_{j-1}}^{C_j}\left(\frac{1}{\lambda} - xy\right) f(y \mid x) f(x)dx}{\frac{1}{P_j}\int_{C_{j-1}}^{C_j} f(y \mid x) f(x)dx}$$

$$= \frac{\int_{C_{j-1}}^{C_j}\left(\frac{1}{\lambda} - xy\right) f(y \mid x) f(x)dx}{\int_{C_{j-1}}^{C_j} f(y \mid x) f(x)dx}$$ (after canceling out $\frac{1}{P_j}$)

We use the expected value of the product of first derivative of log-likelihood function to calculate the Fisher information:

$$E[l_1^{*}{}'(\lambda)^2] = \int_0^{\infty}\left[\frac{\int_{C_{j-1}}^{C_j}\left(\frac{1}{\lambda} - xy\right) f(y \mid x) f(x)dx}{\int_{C_{j-1}}^{C_j} f(y \mid x) f(x)dx}\right]^2 f(y^*)dy$$

[replace $f(y^*)$ with $\frac{\int_{C_{j-1}}^{C_j} f(y \mid x) f(x)dx}{P_j}$]

$$= \int_0^{\infty} \left[\frac{\int_{C_{j-1}}^{C_j} \left(\frac{1}{\lambda} - xy \right) f(y \mid x) f(x) dx}{\int_{C_{j-1}}^{C_j} f(y \mid x) f(x) dx} \right]^2 \frac{\int_{C_{j-1}}^{C_j} f(y \mid x) f(x) dx}{P_j} dy$$

[cancel out $\int_{C_{j-1}}^{C_j} f(y \mid x) f(x) dx$]

$$= \int_0^{\infty} \frac{\left[\int_{C_{j-1}}^{C_j} \left(\frac{1}{\lambda} - xy \right) f(y \mid x) f(x) dx \right]^2}{P_j \int_{C_{j-1}}^{C_j} f(y \mid x) f(x) dx} dy$$

Assess Asymptotic Relative Efficiency

Let $E_j^{-1} = E[l_1^*{}'(\lambda)^2]^{-1}$ be the variance of maximum likelihood estimate derived from Y when X is in the j^{th} interval (that is, $C_{j-1} \le X_j < C_j$). Let n_j be the number of observation in the j^{th} group, and $n = \sum_{j=1}^{g} n_j$.

From each observation, the asymptotic relative efficiency:

$$\text{ARE} = \frac{E_{e_i}^{-1}}{E_j^{-1}} = \frac{E_j}{E_{e_i}}$$

where $E_{e_i}^{-1} = E[l_1'(\lambda)^2]^{-1}$ is the variance of the MLE derived from the i^{th} observation when X in its original scale, i=1, …, n.

Therefore, when MLE is derived from all of the n observations,

$$\text{ARE} = E_c = \frac{\sum_{j=1}^{g} n_i E_i}{\sum_{i=1}^{n} E_{e_i}}$$

When λ=1

If under the null hypothesis that there is no treatment effect, that is, α=0, and consequence, $\lambda = e^{\alpha Z}$ =1, the expected Fisher information

$$E_{e_i} = Variance_e^{-1} = E[l_1'(\lambda)^2] = -E[l_1''(\lambda)] = -E[-\lambda^2] = 1$$

Therefore,

$$\text{ARE} = \frac{\sum_{j=1}^{g} n_j E_j}{\sum_{i=1}^{n} E_{e_i}} = \frac{\sum_{j=1}^{g} n_j E_j}{n \times 1} = \sum_{j=1}^{g} \frac{n_j}{n} E_j = \sum_{j=1}^{g} P_j E_j$$

Where $$E_j = \int_0^{\infty} \frac{\left[\int_{C_{j-1}}^{C_j} \left(\frac{1}{\lambda} - xy\right) f(y \mid x) f(x) dx\right]^2}{P_j \int_{C_{j-1}}^{C_j} f(y \mid x) f(x) dx} dy$$

(replace λ with 1)

$$= \int_0^{\infty} \frac{\left[\int_{C_{j-1}}^{C_j} (1 - xy) f(y \mid x) f(x) dx\right]^2}{P_j \int_{C_{j-1}}^{C_j} f(y \mid x) f(x) dx} dy$$

The results here confirm what Morgan and Elashoff (1986) derived.

7.2 Impact from Dichotomization on the Asymptotic Relative Efficiency of the Treatment Effect

In order to evaluate the influence from dichotomization on the asymptotic relative efficiency of the treatment effect, we performed simulation studies under different conditions. The simulation studies were formed by using the R language. We calculate the asymptotic relative efficiency by using the variance of the coefficient which derived from controlling for the categorized covariate divided by the variance of the coefficient which derived from controlling for the continuous covariate.

7.2.1 When the Null Hypothesis is True: Coefficient=0

In order to compare the results from the study by Morgan and Elashoff (Morgan and Elashoff, 1986), we assign the outcome variable *Y* as survival time which follows an exponential distribution with the association of:

$$\log Y = \beta_0 + \beta_1 \times Z + \beta_2 X$$

When the null hypothesis is true, we have $\beta_1 = 0$ and $\beta_2 = 0$

When *X* is categorized, a data value was chosen as the cutoff point. The association becomes:

$$\log Y = \beta_{D0} + \beta_{D1} \times Z + \beta_{D1} X_D$$

The truncated means of each group was used as the value for the estimate the coefficient of treatment effect, β_{D1}.

Simulation studies were conducted to assess the impacts from dichotomizing a continuous confounding variable on the ARE of treatment effect. Each simulation is performed by using 20,000 data points. A total of 1,000 simulations were performed for

each distribution. The regression analysis was performed by using the `survreg` function of `survival` package (Therneau et al, 2009) in `R` language

7.2.1.1 Gamma Covariate

We assume that the continuous covariate follows a gamma distribution. Four different shape parameters (α=2, 3, 4, 6) were used. Without losing the generality and for comparing the results with Morgan's study (Morgan and Elashoff, 1986), we use the scale parameter equals to 1.

The graphic results are shown in Figure 7.2.1 and the numeric results are shown in Table 7.2.1. From the results, we found that the relative asymptotic efficiency changes with the choice of cutoff points. It also changes with the parameter of the gamma distribution.

When compare the results with the Morgan's study, our data are comparable with theirs.

Figure 7.2.1 Relative Efficiency of b1 Estimate

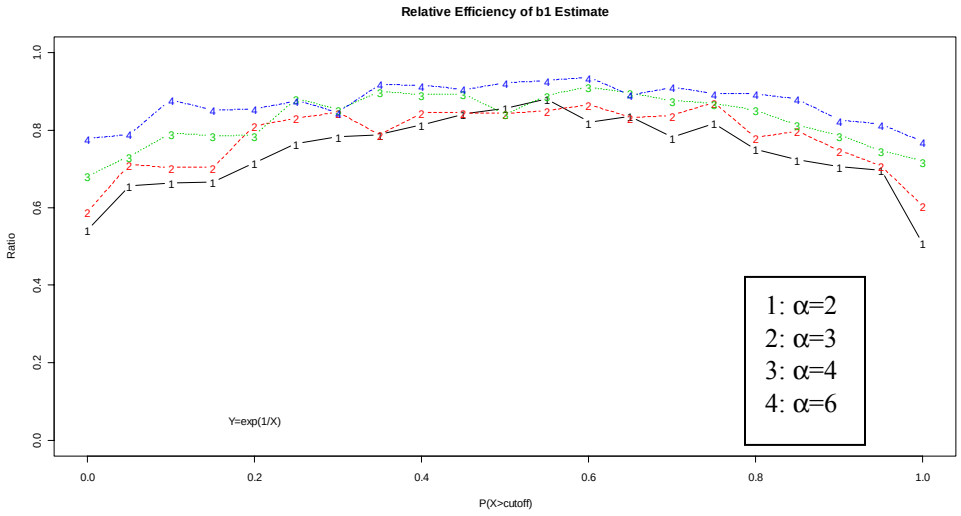


Table 7.2.1 Relative Efficiency of b1 Estimate

Percentile	α=2	α=3	α=4	α=6
0	0.54270	0.58927	0.68249	0.77813
5	0.65542	0.71122	0.73188	0.78955
10	0.66439	0.70402	0.79124	0.87839
15	0.66635	0.70392	0.78614	0.85086
20	0.71589	0.81083	0.78688	0.85473
25	0.76551	0.83148	0.88116	0.87529
30	0.78248	0.84627	0.85298	0.84644
35	0.78779	0.78857	0.89962	0.91787
40	0.81314	0.84504	0.89161	0.91463
45	0.84054	0.84446	0.89355	0.90509
50	0.85582	0.84271	0.84095	0.92123
55	0.88031	0.84978	0.88905	0.92784
60	0.82062	0.86552	0.91112	0.93534
65	0.83566	0.83241	0.89524	0.88994
70	0.78186	0.83756	0.87589	0.91054
75	0.81742	0.87149	0.86940	0.89446
80	0.75068	0.78042	0.85150	0.89345
85	0.72249	0.79706	0.81263	0.87994
90	0.70549	0.74683	0.78704	0.82557
95	0.69613	0.70665	0.74676	0.81500
100	0.50935	0.60454	0.72043	0.77078

7.3 Conclusions

Our studies demonstrated that categorizing a continuous confounding variable impacts the estimation of the treatment effect. When the null hypothesis is true, that is, no treatment effect exists, the dichotomization of a continuous covariate reduces the relative efficiency. The magnitude of influence depends on the distribution of the confounding variable and the location of the cutoff point.

When the distribution of outcome and explanatory variables are available, they can be plugged in the ARE equation for calculating ARE. If the ARE cannot be solved analytically, the estimated relative efficiency can be calculated numerically or via the simulation studies.

Chapter 8

Conclusion and Future Work

8.1 Conclusions

This dissertation research proposed a linear model approach via weighted linear regression for estimating the parameters of a dichotomized covariate from studies with inconsistent cutoff points. Because there was no method for estimating the parameters from a dichotomized covariate, this type of study is usually excluded from the meta-analysis, or sometimes inappropriately included.

This proposed approach can be extended to accommodate categorized covariates from studies with different numbers of cutoff points. By using the techniques for handling correlated data and the mixed effect model, the estimation can be improved. As a consequence, the meta-analysis from using the re-estimated dose-response association can be improved.

We also propose the goodness-of-fit approach to estimate parameters from categorized variables included in a meta-analysis. This approach can also be used to estimate the proportion of excess zeros when a mixture distribution consists of a combination of true zeros and a continuous variable.

This dissertation also investigated the impact from categorization on the estimation. We found that the impact depends on the covariate distribution and on the location and number of cutoff points. When categorizing a continuous variable, either it serves as the covariate of interest or the confounding variable, a biased association to the outcome variable could be estimated. The magnitude of bias depends on the location of the cutoff

point as well as the distribution of the covariate. Therefore, researchers should avoid categorizing a continuous variable to assess the association with outcome variable if possible.

8.2 Future Work

When the covariate is normally distributed, it is easy to use the linear model approach with the mixed effect model. However, the covariate might not be normally distributed, or it can be transformed to become normally distributed. The gamma distribution has more flexibility to model the covariate. Therefore, the use of the mixed effects model for estimating the gamma distribution will be investigated in the future.

The mixture distribution with excess zeros is common in epidemiology studies. Even though it is natural to assume that the mixture distribution comes from two distributions, one of which puts point mass at zero, it is still possible that the excess zeros is the characteristics of a distribution, such as the Tweedie family. Therefore, the potential distributions containing excess zeros will be investigated in the future.

When we assessed the impact from categorizing a covariate on estimating survival time, we only considered the case with no censoring. In order to understand the impact from censoring, we will extend the current approach to take censoring into account.

APPENDIX A

Table A.1 Mean Estimate when X ~ Normal (100, 10^2), n=1,000 in each study

Number of Study	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
2	98.9718	30.5536	99.9972	0.2285	0.0075
3	99.9980	0.6635	100.0087	0.1850	0.2788
4	99.9975	0.2822	99.9934	0.1586	0.5619
5	100.0063	0.2445	99.9996	0.1367	0.5592
6	100.0081	0.1868	100.0072	0.1294	0.6929
7	99.9900	0.1737	99.9924	0.1194	0.6876
8	100.0015	0.1604	100.0018	0.1116	0.6954
9	99.9964	0.1476	99.9982	0.1071	0.7256
10	99.9965	0.1376	99.9973	0.1006	0.7308
11	99.9976	0.1250	100.0001	0.0921	0.7371
12	99.9985	0.1298	100.0019	0.0937	0.7218
13	99.9987	0.1229	100.0004	0.0899	0.7313
14	99.9994	0.1140	99.9974	0.0829	0.7269
15	100.0004	0.1153	99.9995	0.0838	0.7264
16	100.0034	0.1054	100.0018	0.0775	0.7354
17	99.9992	0.1037	99.9997	0.0769	0.7420
18	100.0020	0.1011	100.0011	0.0756	0.7480
19	99.9966	0.0990	99.9954	0.0728	0.7357
20	100.0003	0.0922	99.9972	0.0708	0.7676
21	99.9978	0.0895	99.9978	0.0679	0.7592
22	99.9988	0.0892	100.0025	0.0641	0.7183
23	100.0020	0.0877	99.9995	0.0650	0.7410
24	100.0014	0.0868	100.0011	0.0664	0.7650
25	99.9943	0.0835	99.9963	0.0611	0.7320
26	100.0011	0.0860	100.0002	0.0632	0.7352
27	99.9983	0.0806	100.0004	0.0608	0.7552
28	100.0000	0.0807	100.0000	0.0604	0.7483
29	100.0028	0.0770	100.0019	0.0597	0.7763
30	99.9990	0.0770	100.0003	0.0571	0.7410

Table A.2 Mean Estimate when X ~ Normal (100, 15^2), n=1,000 in each study

Number of Study	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
2	100.8516	69.8133	100.0089	0.3325	0.0048
3	100.0144	0.8572	100.0018	0.2757	0.3216
4	100.0215	0.7394	100.0000	0.2381	0.3220
5	100.0148	0.3703	100.0062	0.2126	0.5742
6	100.0269	0.2842	100.0091	0.1942	0.6833
7	100.0028	0.2627	99.9923	0.1817	0.6916
8	100.0078	0.2513	99.9998	0.1730	0.6884
9	99.9938	0.2217	100.0010	0.1626	0.7333
10	99.9987	0.2119	99.9964	0.1518	0.7161
11	99.9993	0.1993	100.0016	0.1481	0.7434
12	100.0066	0.1880	100.0028	0.1357	0.7217
13	99.9984	0.1805	100.0050	0.1369	0.7584
14	100.0031	0.1720	100.0021	0.1273	0.7402
15	100.0001	0.1584	100.0007	0.1194	0.7541
16	100.0021	0.1565	100.0039	0.1169	0.7469
17	100.0025	0.1583	100.0001	0.1185	0.7486
18	99.9964	0.1427	99.9997	0.1089	0.7627
19	100.0035	0.1417	100.0042	0.1069	0.7545
20	100.0003	0.1410	99.9986	0.1061	0.7526
21	99.9991	0.1332	100.0016	0.1007	0.7560
22	99.9960	0.1356	100.0010	0.1020	0.7523
23	99.9880	0.1310	99.9904	0.1006	0.7683
24	99.9940	0.1313	99.9953	0.0939	0.7148
25	99.9991	0.1253	100.0001	0.0957	0.7639
26	99.9987	0.1249	100.0014	0.0891	0.7132
27	100.0006	0.1241	100.0001	0.0922	0.7429
28	99.9998	0.1185	100.0009	0.0906	0.7645
29	100.0006	0.1192	99.9992	0.0917	0.7688
30	99.9962	0.1145	99.9979	0.0879	0.7674

Table A.3 Standard Deviation Estimate when X ~ Normal (100, 10^2), n=1,000 in each study

Number of Study	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
2	8.1255	78.6075	9.9970	0.1518	0.0019
3	10.0144	1.4126	9.9994	0.1293	0.0915
4	9.9642	0.7831	9.9995	0.1108	0.1415
5	9.9972	0.6052	9.9986	0.1021	0.1687
6	9.9901	0.4735	10.0032	0.0936	0.1977
7	9.9819	0.4164	9.9986	0.0831	0.1996
8	9.9672	0.3849	9.9966	0.0788	0.2046
9	9.9863	0.3506	10.0026	0.0754	0.2149
10	9.9818	0.3202	10.0003	0.0707	0.2208
11	9.9809	0.3063	9.9946	0.0679	0.2216
12	9.9842	0.2922	10.0037	0.0616	0.2107
13	9.9870	0.2719	9.9993	0.0631	0.2321
14	9.9914	0.2516	9.9991	0.0608	0.2417
15	9.9793	0.2512	9.9993	0.0575	0.2288
16	9.9688	0.2460	9.9992	0.0574	0.2333
17	9.9804	0.2362	9.9978	0.0557	0.2357
18	9.9841	0.2379	9.9997	0.0531	0.2234
19	9.9825	0.2256	9.9960	0.0515	0.2284
20	9.9890	0.2132	9.9994	0.0501	0.2349
21	9.9898	0.2070	10.0011	0.0474	0.2289
22	9.9945	0.2004	10.0001	0.0463	0.2312
23	9.9846	0.2076	10.0005	0.0455	0.2192
24	9.9768	0.1950	10.0002	0.0456	0.2337
25	9.9910	0.1876	10.0014	0.0454	0.2420
26	9.9884	0.1944	10.0017	0.0444	0.2283
27	9.9794	0.1785	9.9993	0.0422	0.2366
28	9.9836	0.1826	9.9989	0.0413	0.2261
29	10.0003	0.1730	9.9996	0.0411	0.2378
30	9.9776	0.1707	10.0000	0.0428	0.2509

Table A.4 Standard Deviation Estimate when X ~ Normal (100, 15^2), n=1,000 in each study

Number of Study	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
2	17.4307	142.9001	14.9867	0.2325	0.0016
3	14.8506	1.9126	15.0028	0.1946	0.1017
4	15.0086	1.4164	15.0040	0.1634	0.1154
5	14.9394	0.9158	15.0035	0.1546	0.1688
6	14.9978	0.7427	14.9996	0.1380	0.1858
7	15.0124	0.5956	15.0087	0.1259	0.2114
8	14.9851	0.5811	15.0013	0.1152	0.1982
9	14.9711	0.5213	15.0007	0.1124	0.2157
10	14.9698	0.4798	14.9965	0.1068	0.2225
11	15.0015	0.4565	14.9967	0.1018	0.2231
12	14.9997	0.4246	15.0030	0.0951	0.2240
13	14.9815	0.4116	15.0019	0.0947	0.2300
14	14.9689	0.4148	15.0001	0.0906	0.2184
15	14.9730	0.3861	15.0022	0.0859	0.2223
16	14.9875	0.3691	15.0026	0.0854	0.2315
17	14.9847	0.3525	15.0025	0.0838	0.2378
18	14.9546	0.3517	14.9985	0.0807	0.2295
19	14.9739	0.3404	14.9950	0.0768	0.2257
20	14.9716	0.3079	14.9976	0.0732	0.2379
21	14.9732	0.3182	15.0014	0.0723	0.2272
22	14.9714	0.3055	15.0014	0.0674	0.2206
23	14.9715	0.3079	14.9962	0.0703	0.2284
24	14.9953	0.2898	14.9952	0.0697	0.2406
25	14.9758	0.2899	14.9998	0.0650	0.2241
26	14.9839	0.2847	15.0017	0.0654	0.2299
27	14.9716	0.2787	15.0009	0.0639	0.2292
28	14.9617	0.2688	15.0004	0.0619	0.2302
29	14.9894	0.2661	15.0006	0.0625	0.2349
30	14.9640	0.2666	14.9979	0.0594	0.2229

Table A.5 Mean Estimate when X ~ Normal (100, 10^2), n=100 in each study

Number of Study	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
2	100.2241	28.4943	100.0027	0.7166	0.0251
3	100.0299	1.8829	100.0037	0.5859	0.3111
4	100.0111	0.8911	100.0053	0.4917	0.5517
5	100.0152	1.0398	99.9921	0.4571	0.4396
6	99.9639	0.6474	99.9791	0.4174	0.6447
7	100.0214	0.5675	100.0349	0.3754	0.6616
8	99.9818	0.5013	99.9942	0.3602	0.7187
9	100.0214	0.4519	100.0058	0.3381	0.7482
10	99.9856	0.4416	99.9837	0.3167	0.7171
11	100.0059	0.4114	99.9985	0.3070	0.7464
12	99.9969	0.3910	99.9998	0.2916	0.7458
13	100.0158	0.3811	100.0163	0.2705	0.7097
14	99.9947	0.3658	100.0033	0.2648	0.7240
15	100.0183	0.3512	100.0133	0.2601	0.7405
16	99.9962	0.3356	100.0050	0.2528	0.7533
17	100.0079	0.3251	100.0118	0.2372	0.7297
18	100.0010	0.3239	100.0021	0.2432	0.7508
19	100.0023	0.3120	99.9998	0.2282	0.7313
20	99.9957	0.2984	99.9887	0.2193	0.7348
21	99.9966	0.3039	100.0015	0.2212	0.7281
22	99.9954	0.2823	99.9960	0.2076	0.7352
23	99.9859	0.2784	99.9882	0.2102	0.7552
24	100.0018	0.2758	100.0020	0.2078	0.7535
25	100.0072	0.2616	100.0078	0.1987	0.7594
26	100.0140	0.2682	100.0065	0.1990	0.7420
27	99.9945	0.2600	100.0003	0.1990	0.7654
28	100.0004	0.2472	100.0010	0.1870	0.7565
29	99.9869	0.2499	100.0003	0.1839	0.7359
30	99.9976	0.2469	100.0019	0.1827	0.7397

Table A.6 Standard Deviation Estimate when X ~ Normal (100, 10^2), n=100 in each study

Number of Study	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
2	8.2912	49.5094	9.9937	0.4916	0.0099
3	9.7998	4.9195	10.0026	0.3979	0.0809
4	9.7497	2.4439	9.9865	0.3543	0.1450
5	9.9371	2.0127	9.9954	0.3079	0.1530
6	9.8483	1.4624	9.9882	0.2939	0.2010
7	9.9017	1.3747	9.9924	0.2684	0.1953
8	9.8957	1.2322	9.9989	0.2477	0.2010
9	9.8857	1.0832	9.9884	0.2422	0.2236
10	9.8804	1.0736	9.9929	0.2210	0.2058
11	9.9032	0.9504	9.9933	0.2120	0.2231
12	9.9271	0.9205	10.0034	0.2108	0.2290
13	9.9397	0.8876	9.9951	0.1951	0.2198
14	9.9295	0.8142	10.0063	0.1964	0.2412
15	9.8536	0.7700	9.9981	0.1703	0.2212
16	9.8400	0.7812	9.9967	0.1800	0.2303
17	9.8775	0.7415	9.9918	0.1722	0.2323
18	9.9012	0.7576	9.9970	0.1697	0.2240
19	9.8649	0.7359	10.0021	0.1603	0.2179
20	9.9171	0.6962	10.0021	0.1674	0.2404
21	9.8831	0.6633	10.0055	0.1536	0.2315
22	9.9232	0.6497	9.9995	0.1517	0.2335
23	9.8560	0.6191	9.9973	0.1525	0.2464
24	9.8577	0.6401	9.9915	0.1442	0.2253
25	9.8397	0.6304	9.9911	0.1342	0.2130
26	9.8364	0.6291	10.0028	0.1404	0.2232
27	9.8355	0.5796	10.0004	0.1366	0.2356
28	9.8722	0.6107	10.0076	0.1451	0.2376
29	9.8788	0.5375	9.9948	0.1316	0.2449
30	9.8442	0.5525	9.9969	0.1262	0.2284

Table A.7 Mean Estimate when X ~ Normal (100, 10^2), n=10,000 in each study

Number of Study	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
2	100.0214	3.5396	100.0035	0.0709	0.0200
3	100.0004	0.1602	100.0002	0.0582	0.3631
4	100.0031	0.0942	100.0020	0.0487	0.5172
5	100.0006	0.0692	100.0007	0.0452	0.6534
6	99.9973	0.0602	99.9973	0.0419	0.6950
7	100.0006	0.0568	100.0012	0.0384	0.6756
8	99.9989	0.0511	100.0011	0.0347	0.6794
9	100.0010	0.0473	99.9995	0.0336	0.7099
10	99.9993	0.0468	99.9994	0.0328	0.7010
11	99.9994	0.0411	100.0003	0.0301	0.7327
12	100.0020	0.0397	100.0003	0.0294	0.7398
13	99.9999	0.0373	100.0007	0.0273	0.7323
14	100.0004	0.0362	100.0002	0.0267	0.7383
15	100.0007	0.0350	100.0007	0.0257	0.7347
16	99.9978	0.0339	99.9987	0.0248	0.7316
17	100.0018	0.0343	100.0011	0.0247	0.7222
18	100.0009	0.0316	100.0006	0.0239	0.7565
19	99.9992	0.0309	99.9994	0.0230	0.7452
20	100.0004	0.0304	100.0001	0.0219	0.7205
21	99.9990	0.0299	99.9993	0.0216	0.7241
22	100.0000	0.0284	99.9996	0.0211	0.7421
23	99.9991	0.0281	99.9993	0.0206	0.7330
24	99.9996	0.0277	99.9993	0.0203	0.7315
25	100.0008	0.0266	100.0007	0.0203	0.7624
26	99.9999	0.0265	99.9998	0.0195	0.7357
27	99.9996	0.0264	100.0002	0.0196	0.7441
28	99.9998	0.0260	99.9999	0.0189	0.7244
29	100.0006	0.0247	100.0010	0.0187	0.7552
30	99.9996	0.0255	99.9995	0.0185	0.7244

Table A.8 Standard Deviation Estimate when X ~ Normal (100, 10^2), n=10,000 in each study

Number of Study	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
2	9.9783	6.8849	10.0021	0.0505	0.0073
3	10.0052	0.3798	10.0004	0.0410	0.1080
4	9.9963	0.2343	9.9998	0.0350	0.1495
5	9.9995	0.1747	9.9997	0.0320	0.1834
6	10.0023	0.1455	9.9984	0.0285	0.1959
7	9.9983	0.1329	10.0010	0.0267	0.2011
8	9.9969	0.1235	10.0000	0.0246	0.1989
9	9.9926	0.1131	10.0005	0.0235	0.2073
10	10.0076	0.0992	10.0008	0.0223	0.2245
11	9.9960	0.0954	9.9992	0.0212	0.2221
12	10.0044	0.0897	9.9992	0.0205	0.2285
13	9.9958	0.0878	10.0001	0.0191	0.2174
14	10.0023	0.0867	10.0003	0.0183	0.2107
15	9.9946	0.0805	9.9989	0.0183	0.2270
16	9.9987	0.0773	10.0000	0.0180	0.2333
17	9.9988	0.0754	9.9993	0.0171	0.2267
18	9.9961	0.0755	9.9987	0.0169	0.2243
19	9.9969	0.0711	9.9997	0.0152	0.2135
20	9.9978	0.0675	9.9997	0.0168	0.2488
21	10.0008	0.0674	10.0005	0.0150	0.2222
22	9.9996	0.0672	9.9998	0.0149	0.2214
23	9.9961	0.0629	10.0000	0.0144	0.2293
24	10.0011	0.0639	10.0003	0.0143	0.2241
25	9.9962	0.0610	9.9996	0.0138	0.2263
26	9.9979	0.0615	9.9999	0.0137	0.2234
27	10.0016	0.0610	9.9999	0.0134	0.2190
28	9.9964	0.0579	10.0004	0.0137	0.2364
29	10.0013	0.0541	10.0003	0.0135	0.2498
30	10.0005	0.0556	10.0004	0.0131	0.2353

Table A.9 Mean Estimate when X ~ Normal (100, 15^2), n=10,000 in each study

Number of Study	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
2	99.7464	3.5852	100.0023	0.1036	0.0289
3	99.9933	0.3013	99.9998	0.0860	0.2856
4	100.0002	0.1511	99.9976	0.0737	0.4873
5	100.0000	0.1102	100.0017	0.0675	0.6128
6	100.0009	0.0963	100.0004	0.0605	0.6282
7	99.9938	0.0830	99.9972	0.0551	0.6641
8	99.9996	0.0740	100.0004	0.0524	0.7076
9	99.9941	0.0691	99.9975	0.0511	0.7397
10	99.9970	0.0659	99.9979	0.0458	0.6947
11	100.0021	0.0596	100.0002	0.0443	0.7437
12	99.9935	0.0595	99.9969	0.0450	0.7567
13	99.9983	0.0563	99.9994	0.0421	0.7470
14	99.9983	0.0540	100.0006	0.0407	0.7532
15	99.9988	0.0529	99.9998	0.0397	0.7514
16	99.9984	0.0501	99.9989	0.0357	0.7127
17	100.0033	0.0463	100.0026	0.0347	0.7481
18	100.0007	0.0478	99.9988	0.0357	0.7463
19	99.9977	0.0453	99.9990	0.0347	0.7653
20	99.9986	0.0461	99.9998	0.0341	0.7391
21	100.0014	0.0427	100.0009	0.0336	0.7863
22	100.0003	0.0409	100.0008	0.0306	0.7482
23	99.9987	0.0418	100.0001	0.0315	0.7537
24	100.0001	0.0404	100.0006	0.0305	0.7558
25	100.0007	0.0408	100.0004	0.0299	0.7330
26	99.9990	0.0412	99.9996	0.0297	0.7212
27	100.0011	0.0388	100.0000	0.0295	0.7602
28	100.0020	0.0382	100.0012	0.0293	0.7676
29	99.9999	0.0378	99.9993	0.0276	0.7290
30	100.0003	0.0363	100.0004	0.0265	0.7282

Table A.10 Standard Deviation Estimate when X ~ Normal (100, 15^2), n=10,000 in each study

Number of Study	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
2	15.3940	12.6314	14.9999	0.0760	0.0060
3	15.0026	0.7221	15.0010	0.0613	0.0849
4	14.9893	0.3600	15.0016	0.0521	0.1446
5	15.0032	0.2630	14.9977	0.0472	0.1794
6	15.0042	0.2253	14.9989	0.0443	0.1967
7	14.9836	0.1950	15.0012	0.0380	0.1950
8	14.9942	0.1777	14.9971	0.0387	0.2176
9	14.9948	0.1637	15.0002	0.0353	0.2157
10	14.9952	0.1570	15.0003	0.0349	0.2226
11	15.0004	0.1467	15.0022	0.0317	0.2157
12	14.9978	0.1388	14.9993	0.0305	0.2199
13	14.9935	0.1287	15.0003	0.0299	0.2324
14	14.9970	0.1231	15.0001	0.0276	0.2240
15	15.0000	0.1214	15.0003	0.0272	0.2238
16	15.0029	0.1179	15.0004	0.0256	0.2174
17	14.9996	0.1127	15.0011	0.0260	0.2305
18	15.0008	0.1136	15.0009	0.0252	0.2215
19	14.9956	0.1045	15.0000	0.0243	0.2329
20	15.0042	0.1013	15.0006	0.0231	0.2283
21	14.9990	0.1002	14.9996	0.0225	0.2245
22	15.0018	0.0976	15.0002	0.0228	0.2339
23	15.0002	0.0969	15.0011	0.0223	0.2301
24	14.9982	0.0918	15.0015	0.0218	0.2376
25	15.0005	0.0936	15.0009	0.0214	0.2290
26	14.9977	0.0904	15.0002	0.0202	0.2240
27	14.9922	0.0851	14.9996	0.0205	0.2411
28	15.0016	0.0861	14.9999	0.0200	0.2321
29	14.9946	0.0856	15.0004	0.0198	0.2315
30	14.9997	0.0833	14.9994	0.0187	0.2250

Table A.11 Mean Estimate when X ~ Normal (100, 15^2), n=1,000 in each study, range of cutoff points (35%, 65%)

Number of Study	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
2	99.1300	23.3524	99.9958	0.3427	0.0147
3	100.0118	1.1344	100.0130	0.2775	0.2446
4	99.9936	0.4087	99.9901	0.2379	0.5820
5	100.0011	0.3583	99.9995	0.2051	0.5724
6	100.0152	0.2742	100.0108	0.1941	0.7081
7	99.9898	0.2472	99.9886	0.1791	0.7245
8	100.0018	0.2305	100.0028	0.1673	0.7260
9	99.9928	0.2167	99.9974	0.1606	0.7413
10	99.9985	0.2046	99.9959	0.1508	0.7373
11	99.9970	0.1832	100.0002	0.1382	0.7542
12	99.9967	0.1901	100.0028	0.1406	0.7393
13	99.9976	0.1755	100.0006	0.1349	0.7683
14	99.9998	0.1681	99.9961	0.1243	0.7396
15	99.9987	0.1664	99.9993	0.1256	0.7548
16	100.0047	0.1508	100.0027	0.1162	0.7709
17	99.9993	0.1548	99.9995	0.1154	0.7454
18	100.0005	0.1497	100.0017	0.1134	0.7576
19	99.9932	0.1419	99.9931	0.1092	0.7697
20	100.0038	0.1312	99.9959	0.1062	0.8090
21	100.0013	0.1334	99.9967	0.1019	0.7641
22	99.9995	0.1293	100.0037	0.0962	0.7438
23	100.0016	0.1274	99.9992	0.0975	0.7651
24	100.0028	0.1281	100.0017	0.0996	0.7778
25	99.9906	0.1174	99.9945	0.0916	0.7808
26	99.9968	0.1259	100.0003	0.0949	0.7532
27	99.9983	0.1166	100.0006	0.0913	0.7829
28	100.0021	0.1150	100.0000	0.0906	0.7877
29	100.0040	0.1126	100.0029	0.0896	0.7956
30	99.9988	0.1136	100.0005	0.0856	0.7536

Table A.12 Standard Deviation Estimate when X ~ Normal (100, 15^2), n=1,000 in each study, range of cutoff points (35%, 65%)

Number of Study	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
2	13.6639	132.4090	14.9955	0.2277	0.0017
3	15.0308	4.8479	14.9992	0.1939	0.0400
4	14.8997	2.2767	14.9993	0.1662	0.0730
5	14.9895	1.7913	14.9978	0.1531	0.0855
6	14.9714	1.3750	15.0047	0.1404	0.1021
7	14.9491	1.2502	14.9979	0.1247	0.0998
8	14.9469	1.1642	14.9949	0.1181	0.1015
9	14.9876	1.0556	15.0039	0.1130	0.1071
10	14.9945	0.9872	15.0004	0.1060	0.1074
11	14.9696	0.9054	14.9919	0.1018	0.1124
12	14.9506	0.9057	15.0055	0.0924	0.1020
13	14.9660	0.7941	14.9989	0.0946	0.1192
14	14.9903	0.7521	14.9987	0.0912	0.1213
15	14.9670	0.7648	14.9990	0.0862	0.1127
16	14.9556	0.7240	14.9987	0.0861	0.1189
17	14.9677	0.6794	14.9967	0.0835	0.1229
18	14.9719	0.7052	14.9995	0.0797	0.1130
19	14.9755	0.6789	14.9941	0.0773	0.1138
20	14.9705	0.6249	14.9991	0.0751	0.1202
21	14.9601	0.6046	15.0016	0.0710	0.1175
22	14.9919	0.6151	15.0002	0.0695	0.1130
23	14.9732	0.6093	15.0007	0.0683	0.1120
24	14.9443	0.6014	15.0003	0.0684	0.1137
25	15.0005	0.5644	15.0020	0.0681	0.1207
26	14.9785	0.5648	15.0025	0.0666	0.1179
27	14.9503	0.5326	14.9990	0.0634	0.1190
28	14.9902	0.5481	14.9983	0.0619	0.1129
29	15.0022	0.5185	14.9995	0.0617	0.1190
30	14.9687	0.5097	15.0001	0.0642	0.1260

Table A.13 Mean Estimate when X ~ Normal (100, 15^2), n=10,000 in each study, range of cutoff points (35%, 65%)

Number of Study	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
2	100.1221	7.3368	100.0053	0.1063	0.0145
3	100.0000	0.2479	100.0003	0.0873	0.3521
4	100.0071	0.1737	100.0030	0.0731	0.4208
5	100.0022	0.1022	100.0010	0.0679	0.6641
6	99.9957	0.0906	99.9960	0.0628	0.6930
7	100.0010	0.0827	100.0017	0.0575	0.6960
8	99.9980	0.0761	100.0016	0.0521	0.6845
9	100.0015	0.0691	99.9993	0.0504	0.7295
10	99.9987	0.0649	99.9991	0.0492	0.7592
11	99.9987	0.0610	100.0005	0.0452	0.7408
12	100.0011	0.0592	100.0005	0.0441	0.7449
13	100.0003	0.0557	100.0010	0.0410	0.7369
14	100.0021	0.0542	100.0003	0.0401	0.7397
15	99.9992	0.0500	100.0011	0.0385	0.7702
16	99.9959	0.0486	99.9980	0.0372	0.7661
17	100.0030	0.0471	100.0017	0.0371	0.7876
18	100.0010	0.0462	100.0009	0.0358	0.7753
19	99.9988	0.0455	99.9991	0.0346	0.7594
20	99.9991	0.0434	100.0002	0.0328	0.7559
21	99.9987	0.0429	99.9990	0.0324	0.7566
22	99.9999	0.0421	99.9994	0.0317	0.7517
23	99.9981	0.0400	99.9990	0.0308	0.7704
24	100.0001	0.0398	99.9990	0.0304	0.7645
25	100.0013	0.0384	100.0011	0.0305	0.7922
26	99.9989	0.0372	99.9997	0.0292	0.7858
27	100.0003	0.0379	100.0003	0.0294	0.7767
28	100.0007	0.0367	99.9999	0.0283	0.7711
29	100.0012	0.0357	100.0015	0.0280	0.7852
30	99.9992	0.0353	99.9993	0.0277	0.7866

Table A.14 Standard Deviation Estimate when X ~ Normal (100, 15^2), n=10,000 in each study, range of cutoff points (35%, 65%)

Number of Study	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
2	15.08039	27.25224	15.00313	0.075679	0.002777
3	15.05245	1.225707	15.00058	0.061537	0.050205
4	15.00094	0.7943	14.99968	0.052554	0.066164
5	14.99005	0.543391	14.99953	0.04807	0.088462
6	15.00029	0.433492	14.99766	0.042754	0.098626
7	15.01692	0.403515	15.00147	0.040088	0.099347
8	15.00442	0.37633	15.00006	0.036846	0.09791
9	14.97574	0.337763	15.00077	0.035185	0.10417
10	15.01372	0.298008	15.00124	0.0334	0.112077
11	14.98485	0.286037	14.99879	0.031765	0.111051
12	15.01185	0.26228	14.9988	0.030751	0.117244
13	14.98813	0.25909	15.00022	0.028636	0.110527
14	15.00499	0.253839	15.00051	0.02739	0.107904
15	14.99594	0.231362	14.9984	0.027421	0.118521
16	14.99729	0.223948	15.00003	0.027059	0.120829
17	14.99448	0.224679	14.99896	0.025629	0.11407
18	14.99232	0.224367	14.99803	0.025417	0.113281
19	14.99345	0.209682	14.99952	0.022757	0.10853
20	14.99555	0.198599	14.99959	0.025187	0.126826
21	14.99815	0.199238	15.00081	0.022464	0.112748
22	15.00164	0.197478	14.99977	0.02232	0.113028
23	14.99821	0.189961	14.99993	0.021639	0.113911
24	14.99924	0.187404	15.00042	0.021494	0.114691
25	14.99593	0.178632	14.9994	0.020695	0.115852
26	14.99373	0.182879	14.99985	0.020607	0.112682
27	14.9968	0.179617	14.99983	0.020048	0.111613
28	14.99128	0.168198	15.00053	0.02053	0.122058
29	15.0009	0.163529	15.00048	0.020262	0.123903
30	15.00214	0.165736	15.00053	0.019641	0.118507

Table A.15 Mean Estimates from median-cutoff point and mean-cutoff point when X ~ Normal (100, 10^2), n=10,000 in each study

Number of Study	Mean Estimate from Median-Cutoff point		Mean Estimate from Mean-Cutoff point		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
2	99.9980	0.0888	100.0008	0.0705	0.6304
3	100.0022	0.0734	100.0030	0.0583	0.6294
4	99.9990	0.0639	99.9980	0.0502	0.6188
5	100.0021	0.0599	100.0023	0.0470	0.6165
6	100.0000	0.0527	99.9980	0.0411	0.6077
7	100.0006	0.0477	99.9991	0.0377	0.6261
8	100.0002	0.0438	100.0003	0.0346	0.6251
9	100.0018	0.0414	100.0016	0.0336	0.6591
10	100.0001	0.0404	99.9994	0.0326	0.6485
11	99.9988	0.0375	99.9990	0.0297	0.6280
12	100.0011	0.0360	100.0006	0.0298	0.6838
13	99.9986	0.0344	99.9997	0.0277	0.6472
14	99.9983	0.0336	99.9976	0.0270	0.6458
15	99.9991	0.0316	99.9997	0.0257	0.6630
16	100.0005	0.0316	99.9990	0.0253	0.6395
17	99.9998	0.0303	99.9995	0.0242	0.6394
18	99.9988	0.0312	99.9990	0.0243	0.6063
19	100.0006	0.0278	100.0012	0.0226	0.6594
20	99.9996	0.0282	100.0000	0.0222	0.6178
21	99.9986	0.0278	99.9990	0.0225	0.6545
22	99.9998	0.0265	99.9995	0.0209	0.6226
23	99.9992	0.0260	99.9998	0.0204	0.6164
24	99.9995	0.0261	99.9999	0.0204	0.6080
25	100.0007	0.0264	100.0008	0.0204	0.5954
26	99.9997	0.0245	99.9996	0.0195	0.6349
27	100.0000	0.0240	100.0002	0.0190	0.6277
28	100.0003	0.0236	99.9999	0.0186	0.6230
29	99.9996	0.0236	100.0007	0.0187	0.6317
30	99.9991	0.0226	99.9991	0.0187	0.6875

Table A.16 Standard Deviation Estimate s from median-cutoff point and mean-cutoff point when X ~ Normal (100, 10^2), n=10,000 in each study

Number of Study	Standard Deviation Estimate from Median-Cutoff Point		Standard Deviation Estimate from Mean-Cutoff Point		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
2	10.1353	7.5867	8.1254	6.1909	0.6659
3	11.4406	5.8694	9.0714	4.6219	0.6201
4	11.7084	4.8933	9.2707	3.9591	0.6546
5	11.9174	4.4343	9.4320	3.4471	0.6043
6	11.8999	3.8750	9.4840	3.0415	0.6161
7	12.2227	3.5069	9.5669	2.7796	0.6282
8	12.0349	3.2674	9.6560	2.6247	0.6453
9	12.0136	3.0376	9.5153	2.4264	0.6381
10	12.1475	2.9243	9.7529	2.2023	0.5671
11	12.2114	2.7932	9.7078	2.1755	0.6066
12	12.1568	2.6645	9.7489	2.0814	0.6102
13	12.2300	2.4663	9.7569	1.9460	0.6226
14	12.2785	2.4559	9.7504	1.9018	0.5997
15	12.2898	2.4329	9.8590	1.8727	0.5925
16	12.3290	2.2825	9.8666	1.7214	0.5688
17	12.5282	2.2504	10.0135	1.7613	0.6126
18	12.4027	2.1509	9.9356	1.7213	0.6404
19	12.3656	2.1537	9.9180	1.6994	0.6226
20	12.2455	2.0101	9.7688	1.5843	0.6212
21	12.4630	1.9095	9.9222	1.5362	0.6472
22	12.3586	1.8608	9.8501	1.5133	0.6614
23	12.5015	1.8415	9.9646	1.4939	0.6581
24	12.4192	1.9138	9.8462	1.4884	0.6049
25	12.4619	1.7909	9.9119	1.3748	0.5892
26	12.4091	1.7004	9.9154	1.3498	0.6301
27	12.3594	1.7053	9.8508	1.4029	0.6768
28	12.4276	1.7362	9.9106	1.3602	0.6138
29	12.4414	1.6779	9.9552	1.2925	0.5934
30	12.4135	1.7057	9.9128	1.3048	0.5852

Table A.17 Mean Estimate when X ~ Normal (100, 10^2), n=1,000 in each study

Number of Group	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
3	100.0114	0.3474	100.0107	0.3259	0.8800
4	99.9864	0.3391	99.9849	0.3155	0.8657
5	100.0164	0.3372	100.0128	0.3186	0.8928
6	99.9809	0.3232	99.9835	0.3008	0.8660
7	100.0016	0.3294	99.9974	0.3145	0.9118
8	99.9916	0.3391	99.9867	0.3144	0.8597
9	99.9997	0.3317	99.9997	0.3086	0.8655
10	100.0177	0.3395	100.0120	0.3155	0.8636
11	99.9950	0.3482	100.0030	0.3265	0.8795
12	99.9922	0.3252	99.9972	0.3002	0.8521
13	100.0022	0.3413	100.0072	0.3216	0.8878
14	100.0122	0.3353	100.0178	0.3118	0.8649
15	100.0028	0.3437	100.0009	0.3281	0.9111
16	99.9937	0.3609	99.9964	0.3381	0.8775
17	100.0061	0.3531	100.0094	0.3265	0.8547
18	99.9993	0.3243	99.9968	0.3055	0.8874
19	99.9934	0.3382	99.9954	0.3143	0.8634
20	100.0014	0.3544	100.0046	0.3288	0.8604
21	100.0095	0.3420	100.0096	0.3189	0.8692
22	100.0082	0.3271	100.0092	0.2983	0.8315
23	99.9904	0.3515	99.9955	0.3231	0.8450
24	99.9918	0.3318	99.9901	0.3109	0.8777
25	99.9997	0.3413	99.9993	0.3196	0.8771
26	100.0124	0.3295	100.0087	0.3087	0.8781
27	100.0040	0.3534	100.0000	0.3303	0.8737
28	99.9833	0.3353	99.9903	0.3162	0.8896
29	99.9887	0.3507	99.9830	0.3274	0.8713
30	100.0062	0.3398	100.0082	0.3101	0.8324

Table A.18 Mean Estimate when X ~ Normal (100, 10^2), n=10,000 in each study

Number of Group	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
3	99.9992	0.1106	99.9988	0.1025	0.8586
4	100.0040	0.1052	100.0047	0.0984	0.8749
5	99.9997	0.1125	99.9968	0.1033	0.8426
6	100.0029	0.1090	100.0027	0.1042	0.9135
7	100.0021	0.1087	100.0023	0.1018	0.8774
8	100.0035	0.1066	100.0042	0.0985	0.8546
9	99.9983	0.1073	99.9977	0.1011	0.8863
10	99.9972	0.1093	99.9967	0.1020	0.8707
11	99.9976	0.1090	100.0008	0.1002	0.8450
12	99.9967	0.1043	99.9968	0.0997	0.9139
13	100.0018	0.1109	100.0023	0.1022	0.8489
14	100.0039	0.1060	100.0018	0.1006	0.9010
15	100.0009	0.1076	100.0018	0.1010	0.8809
16	100.0058	0.1108	100.0056	0.1029	0.8626
17	100.0006	0.1082	99.9999	0.1002	0.8576
18	99.9985	0.1076	100.0005	0.0992	0.8510
19	99.9989	0.1059	99.9981	0.0982	0.8604
20	99.9967	0.1017	99.9950	0.0948	0.8694
21	99.9962	0.1109	99.9953	0.1042	0.8825
22	99.9943	0.1069	99.9966	0.0998	0.8731
23	100.0011	0.1095	100.0025	0.1020	0.8688
24	99.9948	0.1014	99.9945	0.0958	0.8916
25	99.9967	0.1069	99.9974	0.1010	0.8917
26	99.9966	0.1044	99.9978	0.0978	0.8776
27	100.0074	0.1074	100.0059	0.0998	0.8646
28	99.9990	0.1087	100.0010	0.1030	0.8975
29	99.9937	0.1095	99.9940	0.1016	0.8619
30	100.0042	0.1030	100.0032	0.0981	0.9077

Table A.19 Mean Estimate when X ~ Normal (100, 15^2), n=1,000 in each study

Number of Group	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
3	100.0171	0.5211	100.0160	0.4888	0.8800
4	99.9796	0.5087	99.9774	0.4733	0.8657
5	100.0245	0.5057	100.0192	0.4779	0.8928
6	99.9714	0.4848	99.9753	0.4511	0.8660
7	100.0024	0.4941	99.9961	0.4718	0.9118
8	99.9874	0.5086	99.9800	0.4716	0.8597
9	99.9996	0.4975	99.9995	0.4628	0.8655
10	100.0266	0.5093	100.0180	0.4733	0.8636
11	99.9925	0.5222	100.0045	0.4898	0.8795
12	99.9883	0.4878	99.9958	0.4503	0.8521
13	100.0032	0.5120	100.0109	0.4824	0.8878
14	100.0183	0.5029	100.0267	0.4677	0.8649
15	100.0042	0.5156	100.0014	0.4921	0.9111
16	99.9906	0.5414	99.9946	0.5072	0.8775
17	100.0092	0.5297	100.0140	0.4897	0.8547
18	99.9990	0.4864	99.9952	0.4582	0.8874
19	99.9901	0.5073	99.9931	0.4714	0.8634
20	100.0022	0.5317	100.0069	0.4932	0.8604
21	100.0142	0.5130	100.0143	0.4783	0.8692
22	100.0123	0.4907	100.0138	0.4475	0.8315
23	99.9855	0.5272	99.9933	0.4846	0.8450
24	99.9877	0.4978	99.9852	0.4663	0.8777
25	99.9995	0.5119	99.9989	0.4794	0.8771
26	100.0186	0.4942	100.0130	0.4631	0.8781
27	100.0060	0.5301	100.0000	0.4955	0.8737
28	99.9750	0.5029	99.9854	0.4743	0.8896
29	99.9830	0.5261	99.9745	0.4911	0.8713
30	100.0093	0.5097	100.0122	0.4651	0.8324

Table A.20 Mean Estimate when X ~ Normal (100, 15^2), n=10,000 in each study

Number of Group	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
3	99.9987	0.1659	99.9982	0.1537	0.8586
4	100.0060	0.1578	100.0071	0.1476	0.8749
5	99.9996	0.1687	99.9953	0.1549	0.8426
6	100.0043	0.1636	100.0041	0.1563	0.9135
7	100.0031	0.1630	100.0034	0.1527	0.8774
8	100.0053	0.1599	100.0062	0.1478	0.8546
9	99.9975	0.1610	99.9965	0.1516	0.8863
10	99.9957	0.1640	99.9951	0.1530	0.8707
11	99.9964	0.1635	100.0012	0.1503	0.8450
12	99.9950	0.1564	99.9953	0.1496	0.9139
13	100.0027	0.1664	100.0035	0.1533	0.8489
14	100.0058	0.1590	100.0027	0.1509	0.9010
15	100.0014	0.1614	100.0027	0.1515	0.8809
16	100.0087	0.1662	100.0084	0.1544	0.8626
17	100.0008	0.1623	99.9998	0.1503	0.8576
18	99.9978	0.1614	100.0007	0.1489	0.8510
19	99.9984	0.1589	99.9972	0.1474	0.8604
20	99.9950	0.1525	99.9925	0.1422	0.8694
21	99.9943	0.1663	99.9930	0.1563	0.8825
22	99.9915	0.1603	99.9949	0.1498	0.8731
23	100.0016	0.1642	100.0037	0.1531	0.8688
24	99.9922	0.1521	99.9917	0.1436	0.8916
25	99.9950	0.1604	99.9961	0.1514	0.8917
26	99.9949	0.1566	99.9967	0.1467	0.8776
27	100.0110	0.1611	100.0089	0.1498	0.8646
28	99.9985	0.1630	100.0014	0.1544	0.8975
29	99.9906	0.1642	99.9909	0.1524	0.8619
30	100.0063	0.1544	100.0048	0.1471	0.9077

Table A.21 Standard Deviation Estimate when X ~ Normal (100, 10^2), n=1,000 in each study

Number of Group	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
3	9.9717	0.2357	10.0016	0.2178	0.8545
4	9.9610	0.2406	9.9871	0.2221	0.8525
5	9.9645	0.2399	9.9913	0.2172	0.8198
6	9.9754	0.2300	10.0037	0.2117	0.8470
7	9.9700	0.2508	9.9965	0.2311	0.8486
8	9.9651	0.2439	9.9920	0.2219	0.8280
9	9.9591	0.2395	9.9908	0.2210	0.8515
10	9.9738	0.2414	10.0007	0.2190	0.8230
11	9.9763	0.2466	10.0078	0.2281	0.8550
12	9.9865	0.2379	10.0118	0.2239	0.8856
13	9.9714	0.2404	9.9987	0.2151	0.8006
14	9.9764	0.2454	10.0071	0.2242	0.8346
15	9.9755	0.2405	10.0021	0.2244	0.8700
16	9.9616	0.2387	9.9889	0.2230	0.8732
17	9.9682	0.2563	9.9925	0.2333	0.8284
18	9.9590	0.2437	9.9856	0.2264	0.8631
19	9.9757	0.2368	10.0045	0.2162	0.8336
20	9.9670	0.2471	9.9961	0.2296	0.8634
21	9.9678	0.2411	9.9960	0.2214	0.8436
22	9.9702	0.2494	9.9954	0.2278	0.8342
23	9.9606	0.2461	9.9848	0.2251	0.8366
24	9.9726	0.2444	10.0056	0.2254	0.8505
25	9.9629	0.2521	9.9943	0.2353	0.8715
26	9.9834	0.2423	10.0078	0.2231	0.8479
27	9.9594	0.2318	9.9895	0.2104	0.8235
28	9.9596	0.2575	9.9853	0.2360	0.8398
29	9.9769	0.2452	10.0056	0.2230	0.8274
30	9.9748	0.2455	10.0059	0.2254	0.8431

Table A.22 Standard Deviation Estimate when X ~ Normal (100, 10^2), n=10,000 in each study

Number of Group	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
3	9.9984	0.0750	10.0005	0.0690	0.8465
4	9.9974	0.0777	9.9991	0.0712	0.8401
5	9.9981	0.0769	10.0004	0.0683	0.7889
6	10.0012	0.0744	10.0040	0.0681	0.8379
7	9.9975	0.0790	10.0002	0.0712	0.8121
8	9.9957	0.0768	9.9974	0.0714	0.8636
9	9.9963	0.0768	10.0000	0.0684	0.7917
10	10.0014	0.0800	10.0033	0.0729	0.8287
11	10.0010	0.0763	10.0030	0.0705	0.8555
12	9.9993	0.0775	9.9987	0.0702	0.8216
13	10.0015	0.0797	10.0043	0.0716	0.8065
14	9.9954	0.0782	9.9998	0.0697	0.7932
15	9.9974	0.0793	9.9987	0.0717	0.8167
16	9.9958	0.0785	9.9990	0.0711	0.8206
17	9.9990	0.0810	10.0014	0.0732	0.8158
18	9.9926	0.0775	9.9966	0.0718	0.8575
19	9.9939	0.0786	9.9959	0.0725	0.8514
20	9.9947	0.0765	9.9960	0.0683	0.7973
21	9.9931	0.0766	9.9953	0.0701	0.8377
22	9.9948	0.0787	9.9989	0.0700	0.7904
23	9.9975	0.0773	9.9996	0.0689	0.7963
24	9.9899	0.0785	9.9934	0.0723	0.8485
25	9.9974	0.0796	9.9986	0.0720	0.8189
26	9.9965	0.0760	9.9979	0.0684	0.8104
27	9.9987	0.0782	10.0000	0.0696	0.7937
28	9.9983	0.0795	10.0012	0.0721	0.8216
29	9.9992	0.0802	10.0027	0.0723	0.8141
30	9.9972	0.0793	9.9991	0.0719	0.8215

Table A.23 Standard Deviation Estimate when X ~ Normal (100, 15^2), n=1,000 in each study

Number of Group	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
3	14.9575	0.3535	15.0024	0.3268	0.8545
4	14.9416	0.3608	14.9807	0.3332	0.8525
5	14.9468	0.3598	14.9870	0.3258	0.8198
6	14.9631	0.3450	15.0056	0.3175	0.8470
7	14.9551	0.3763	14.9947	0.3466	0.8486
8	14.9476	0.3659	14.9880	0.3329	0.8280
9	14.9386	0.3593	14.9861	0.3315	0.8515
10	14.9607	0.3621	15.0010	0.3285	0.8230
11	14.9645	0.3700	15.0117	0.3421	0.8550
12	14.9797	0.3569	15.0177	0.3358	0.8856
13	14.9571	0.3606	14.9981	0.3226	0.8006
14	14.9646	0.3681	15.0106	0.3363	0.8346
15	14.9632	0.3608	15.0031	0.3365	0.8700
16	14.9424	0.3580	14.9833	0.3346	0.8732
17	14.9523	0.3845	14.9888	0.3500	0.8284
18	14.9385	0.3655	14.9784	0.3396	0.8631
19	14.9636	0.3552	15.0067	0.3243	0.8336
20	14.9505	0.3706	14.9942	0.3444	0.8634
21	14.9518	0.3616	14.9940	0.3321	0.8436
22	14.9553	0.3741	14.9930	0.3417	0.8342
23	14.9409	0.3691	14.9773	0.3376	0.8366
24	14.9589	0.3666	15.0084	0.3381	0.8505
25	14.9443	0.3782	14.9915	0.3530	0.8715
26	14.9751	0.3634	15.0117	0.3346	0.8479
27	14.9391	0.3477	14.9842	0.3156	0.8235
28	14.9393	0.3862	14.9779	0.3540	0.8398
29	14.9654	0.3678	15.0084	0.3345	0.8274
30	14.9621	0.3682	15.0088	0.3381	0.8431

Table A.24 Standard Deviation Estimate when X ~ Normal (100, 15^2), n=10,000 in each study

Number of Group	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
3	14.9976	0.1124	15.0007	0.1034	0.8465
4	14.9961	0.1166	14.9986	0.1069	0.8401
5	14.9972	0.1154	15.0006	0.1025	0.7889
6	15.0018	0.1116	15.0060	0.1022	0.8379
7	14.9963	0.1185	15.0003	0.1068	0.8121
8	14.9936	0.1152	14.9960	0.1071	0.8636
9	14.9945	0.1152	15.0000	0.1025	0.7917
10	15.0021	0.1201	15.0049	0.1093	0.8287
11	15.0016	0.1144	15.0045	0.1058	0.8555
12	14.9989	0.1162	14.9980	0.1053	0.8216
13	15.0023	0.1195	15.0064	0.1073	0.8065
14	14.9931	0.1173	14.9997	0.1045	0.7932
15	14.9962	0.1190	14.9980	0.1076	0.8167
16	14.9937	0.1177	14.9985	0.1066	0.8206
17	14.9985	0.1215	15.0022	0.1098	0.8158
18	14.9890	0.1162	14.9950	0.1076	0.8575
19	14.9909	0.1179	14.9939	0.1088	0.8514
20	14.9921	0.1147	14.9941	0.1024	0.7973
21	14.9897	0.1149	14.9929	0.1051	0.8377
22	14.9923	0.1180	14.9984	0.1049	0.7904
23	14.9962	0.1159	14.9994	0.1034	0.7963
24	14.9848	0.1177	14.9901	0.1085	0.8485
25	14.9961	0.1193	14.9980	0.1080	0.8189
26	14.9947	0.1140	14.9968	0.1027	0.8104
27	14.9980	0.1173	15.0000	0.1045	0.7937
28	14.9975	0.1193	15.0017	0.1081	0.8216
29	14.9989	0.1203	15.0040	0.1085	0.8141
30	14.9958	0.1189	14.9986	0.1078	0.8215

Table A.25 Mean Estimate when X ~ Normal (100, 10^2), n=1,000 and 3 groups in each study

Number of Study	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
2	99.9937	0.2614	99.9978	0.2312	0.8846
3	100.0008	0.2154	99.9979	0.1869	0.8676
4	99.9992	0.1796	100.0003	0.1595	0.8879
5	100.0070	0.1570	100.0039	0.1420	0.9045
6	100.0053	0.1460	100.0041	0.1303	0.8922
7	99.9931	0.1334	99.9953	0.1194	0.8951
8	100.0017	0.1226	100.0032	0.1089	0.8881
9	99.9992	0.1225	100.0019	0.1080	0.8816
10	99.9994	0.1109	100.0015	0.1003	0.9045
11	100.0000	0.1064	100.0008	0.0956	0.8980
12	99.9982	0.1098	99.9990	0.0947	0.8626
13	100.0000	0.0994	99.9992	0.0891	0.8956
14	99.9982	0.0986	99.9972	0.0873	0.8859
15	100.0002	0.0923	100.0027	0.0818	0.8864
16	100.0023	0.0877	100.0028	0.0792	0.9025
17	100.0001	0.0848	100.0021	0.0770	0.9078
18	100.0005	0.0840	99.9992	0.0724	0.8628
19	99.9944	0.0810	99.9957	0.0729	0.9003
20	99.9982	0.0810	99.9982	0.0734	0.9053
21	99.9961	0.0765	99.9967	0.0690	0.9018
22	100.0028	0.0765	100.0020	0.0679	0.8873
23	99.9999	0.0749	99.9992	0.0660	0.8811
24	99.9988	0.0709	99.9981	0.0636	0.8972
25	100.0017	0.0700	100.0024	0.0635	0.9075
26	99.9972	0.0676	99.9988	0.0614	0.9082
27	100.0023	0.0680	100.0025	0.0597	0.8772
28	99.9996	0.0654	99.9996	0.0580	0.8876
29	100.0008	0.0658	100.0018	0.0574	0.8725
30	100.0015	0.0672	100.0016	0.0584	0.8685

Table A.26 Standard Deviation Estimate when X ~ Normal (100, 10^2), n=1,000 and 3 groups in each study

Number of Group	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
2	9.9912	0.3282	9.9956	0.1510	0.4600
3	9.9799	0.2706	9.9986	0.1250	0.4620
4	9.9889	0.2390	9.9996	0.1152	0.4821
5	9.9799	0.2155	10.0038	0.1017	0.4721
6	9.9805	0.1956	9.9971	0.0903	0.4615
7	9.9858	0.1688	9.9984	0.0826	0.4892
8	9.9863	0.1679	10.0017	0.0761	0.4529
9	9.9905	0.1577	10.0044	0.0728	0.4620
10	9.9863	0.1582	9.9989	0.0716	0.4528
11	9.9904	0.1445	9.9987	0.0668	0.4624
12	9.9867	0.1427	10.0027	0.0649	0.4547
13	9.9884	0.1320	10.0025	0.0618	0.4682
14	9.9952	0.1246	10.0001	0.0600	0.4812
15	9.9876	0.1251	10.0017	0.0572	0.4570
16	9.9837	0.1185	9.9986	0.0558	0.4707
17	9.9917	0.1141	10.0008	0.0545	0.4780
18	9.9786	0.1153	9.9963	0.0525	0.4558
19	9.9843	0.1084	9.9957	0.0498	0.4597
20	9.9877	0.1088	9.9993	0.0485	0.4463
21	9.9883	0.1059	9.9962	0.0494	0.4668
22	9.9845	0.1011	9.9988	0.0481	0.4755
23	9.9880	0.1034	10.0017	0.0467	0.4516
24	9.9843	0.1005	9.9979	0.0458	0.4552
25	9.9918	0.0948	10.0018	0.0453	0.4779
26	9.9925	0.0915	10.0046	0.0441	0.4816
27	9.9814	0.0956	9.9995	0.0460	0.4814
28	9.9875	0.0876	10.0012	0.0431	0.4915
29	9.9854	0.0861	9.9998	0.0422	0.4906
30	9.9874	0.0843	9.9994	0.0393	0.4658

Table A.27 Mean Estimate when X ~ Normal (100, 10^2), n=10,000 and 3 groups in each study

Number of Study	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
2	100.0023	0.0789	100.0018	0.0696	0.8821
3	99.9983	0.0672	100.0006	0.0586	0.8718
4	99.9994	0.0577	99.9998	0.0502	0.8696
5	100.0013	0.0521	100.0017	0.0464	0.8910
6	99.9977	0.0452	99.9976	0.0408	0.9029
7	99.9995	0.0428	99.9990	0.0380	0.8882
8	99.9998	0.0381	100.0003	0.0345	0.9063
9	100.0013	0.0379	100.0016	0.0338	0.8902
10	99.9988	0.0360	99.9996	0.0325	0.9021
11	99.9996	0.0347	99.9994	0.0306	0.8835
12	100.0001	0.0327	100.0002	0.0297	0.9070
13	100.0000	0.0312	99.9999	0.0277	0.8863
14	99.9976	0.0302	99.9980	0.0268	0.8875
15	99.9988	0.0297	99.9993	0.0264	0.8884
16	99.9998	0.0281	99.9991	0.0252	0.8981
17	99.9995	0.0275	99.9992	0.0245	0.8911
18	100.0000	0.0273	99.9996	0.0245	0.8962
19	100.0007	0.0258	100.0009	0.0229	0.8903
20	99.9996	0.0253	100.0000	0.0220	0.8706
21	99.9999	0.0248	99.9992	0.0221	0.8928
22	99.9994	0.0236	99.9994	0.0208	0.8818
23	99.9996	0.0233	99.9997	0.0204	0.8737
24	100.0001	0.0219	100.0000	0.0197	0.8997
25	100.0007	0.0229	100.0008	0.0205	0.8940
26	99.9993	0.0223	99.9996	0.0197	0.8836
27	100.0005	0.0211	100.0001	0.0191	0.9010
28	100.0007	0.0214	100.0000	0.0181	0.8476
29	100.0009	0.0215	100.0004	0.0193	0.8966
30	99.9991	0.0210	99.9993	0.0186	0.8865

Table A.28 Standard Deviation Estimate when X ~ Normal (100, 10^2), n=10,000 and 3 groups in each study

Number of Group	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
2	9.9950	0.1045	9.9999	0.0485	0.4639
3	9.9968	0.0888	10.0017	0.0413	0.4648
4	9.9996	0.0738	10.0011	0.0343	0.4638
5	10.0003	0.0696	10.0003	0.0320	0.4597
6	9.9973	0.0635	9.9976	0.0293	0.4616
7	9.9991	0.0575	9.9993	0.0269	0.4675
8	10.0011	0.0550	10.0014	0.0246	0.4468
9	9.9965	0.0519	9.9999	0.0245	0.4724
10	9.9991	0.0475	10.0008	0.0233	0.4913
11	9.9969	0.0441	9.9990	0.0215	0.4880
12	9.9979	0.0414	9.9999	0.0204	0.4924
13	9.9982	0.0423	10.0005	0.0189	0.4476
14	9.9984	0.0408	10.0005	0.0194	0.4742
15	9.9993	0.0390	9.9990	0.0193	0.4957
16	9.9974	0.0366	10.0000	0.0180	0.4913
17	9.9969	0.0367	9.9991	0.0174	0.4732
18	9.9974	0.0351	9.9995	0.0164	0.4656
19	9.9987	0.0346	10.0000	0.0165	0.4775
20	9.9992	0.0330	10.0000	0.0160	0.4852
21	9.9988	0.0343	10.0003	0.0150	0.4360
22	9.9984	0.0315	9.9997	0.0153	0.4863
23	9.9985	0.0316	10.0000	0.0143	0.4535
24	10.0000	0.0312	10.0007	0.0150	0.4822
25	9.9981	0.0296	9.9993	0.0138	0.4654
26	9.9971	0.0304	9.9997	0.0137	0.4517
27	9.9989	0.0281	10.0001	0.0136	0.4850
28	9.9995	0.0293	10.0008	0.0138	0.4704
29	9.9986	0.0285	10.0005	0.0130	0.4561
30	9.9996	0.0280	10.0003	0.0133	0.4749

Table A.29 Mean Estimate when X ~ Normal (100, 10^2), n=1,000 and 10 groups in each study

Number of Study	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
2	99.9996	0.2281	99.9988	0.2237	0.9807
3	99.9997	0.1903	100.0034	0.1858	0.9763
4	100.0023	0.1624	100.0018	0.1572	0.9674
5	99.9921	0.1417	99.9933	0.1384	0.9765
6	99.9909	0.1318	99.9933	0.1279	0.9706
7	100.0028	0.1221	100.0017	0.1193	0.9771
8	99.9995	0.1135	99.9992	0.1103	0.9725
9	100.0042	0.1082	100.0031	0.1057	0.9771
10	100.0047	0.1016	100.0047	0.0978	0.9626
11	100.0032	0.0985	100.0021	0.0952	0.9664
12	100.0034	0.0928	100.0020	0.0897	0.9668
13	100.0008	0.0894	99.9991	0.0868	0.9711
14	99.9972	0.0862	99.9977	0.0837	0.9710
15	99.9977	0.0827	99.9977	0.0799	0.9656
16	100.0003	0.0833	99.9990	0.0804	0.9648
17	99.9987	0.0808	99.9988	0.0787	0.9743
18	99.9998	0.0729	100.0003	0.0706	0.9688
19	99.9993	0.0762	99.9996	0.0732	0.9596
20	99.9976	0.0707	99.9975	0.0685	0.9696
21	100.0012	0.0712	100.0015	0.0691	0.9704
22	99.9981	0.0719	99.9981	0.0686	0.9538
23	99.9994	0.0683	99.9991	0.0661	0.9683
24	99.9996	0.0660	99.9997	0.0630	0.9546
25	99.9996	0.0662	99.9997	0.0646	0.9764
26	99.9999	0.0621	100.0002	0.0598	0.9622
27	99.9983	0.0616	99.9990	0.0594	0.9645
28	99.9975	0.0592	99.9971	0.0565	0.9550
29	100.0041	0.0599	100.0046	0.0586	0.9789
30	100.0002	0.0581	99.9997	0.0556	0.9585

TableA.30 Standard Deviation Estimate when X ~ Normal (100, 10^2), n=1,000 and 10 groups in each study

Number of Group	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
2	9.9894	0.1924	10.0077	0.1651	0.8583
3	9.9910	0.1543	10.0079	0.1291	0.8372
4	9.9792	0.1362	10.0023	0.1151	0.8447
5	9.9800	0.1177	10.0001	0.1001	0.8505
6	9.9837	0.1085	10.0004	0.0900	0.8298
7	9.9844	0.1005	10.0014	0.0830	0.8256
8	9.9822	0.0958	9.9991	0.0805	0.8397
9	9.9805	0.0917	10.0010	0.0764	0.8331
10	9.9802	0.0798	9.9979	0.0690	0.8645
11	9.9817	0.0808	10.0001	0.0676	0.8361
12	9.9791	0.0764	9.9990	0.0650	0.8509
13	9.9794	0.0754	9.9996	0.0604	0.8010
14	9.9795	0.0721	9.9970	0.0592	0.8210
15	9.9814	0.0686	9.9989	0.0572	0.8343
16	9.9798	0.0692	10.0000	0.0568	0.8198
17	9.9822	0.0655	10.0005	0.0548	0.8370
18	9.9809	0.0611	10.0015	0.0515	0.8426
19	9.9809	0.0618	9.9987	0.0511	0.8278
20	9.9784	0.0600	9.9968	0.0496	0.8267
21	9.9818	0.0600	9.9999	0.0494	0.8229
22	9.9787	0.0555	9.9978	0.0452	0.8138
23	9.9816	0.0535	10.0000	0.0461	0.8620
24	9.9747	0.0563	9.9960	0.0471	0.8357
25	9.9820	0.0532	10.0010	0.0456	0.8559
26	9.9839	0.0517	10.0027	0.0427	0.8263
27	9.9820	0.0512	10.0018	0.0428	0.8376
28	9.9798	0.0501	9.9994	0.0415	0.8292
29	9.9808	0.0491	10.0002	0.0411	0.8377
30	9.9807	0.0488	10.0000	0.0407	0.8345

Table A.31 Mean Estimate when X ~ Normal (100, 10^2), n=10,000 and 10 groups in each study

Number of Study	Mean Estimate from Weighted Linear Regression Approach		Mean Estimate Calculated from Raw Data		Relative Efficiency of Mean Estimates
	Mean	SD	Mean	SD	
2	99.9991	0.0754	99.9993	0.0728	0.9658
3	100.0000	0.0600	100.0006	0.0575	0.9586
4	99.9999	0.0513	99.9999	0.0496	0.9671
5	99.9987	0.0456	99.9988	0.0439	0.9644
6	100.0006	0.0418	100.0006	0.0400	0.9578
7	99.9998	0.0370	99.9998	0.0358	0.9675
8	100.0006	0.0364	100.0006	0.0355	0.9739
9	100.0003	0.0341	100.0003	0.0333	0.9756
10	100.0003	0.0329	100.0001	0.0316	0.9613
11	100.0018	0.0303	100.0013	0.0298	0.9837
12	100.0003	0.0299	100.0003	0.0288	0.9659
13	100.0004	0.0287	100.0003	0.0275	0.9582
14	99.9991	0.0275	99.9991	0.0269	0.9771
15	99.9998	0.0261	99.9999	0.0255	0.9784
16	99.9981	0.0249	99.9986	0.0239	0.9601
17	99.9997	0.0248	99.9997	0.0242	0.9759
18	100.0005	0.0239	100.0005	0.0231	0.9668
19	99.9992	0.0243	99.9991	0.0234	0.9628
20	99.9988	0.0228	99.9990	0.0221	0.9680
21	99.9991	0.0230	99.9992	0.0222	0.9636
22	100.0006	0.0222	100.0006	0.0214	0.9645
23	99.9998	0.0215	99.9999	0.0206	0.9559
24	99.9997	0.0204	99.9999	0.0200	0.9784
25	99.9995	0.0208	99.9996	0.0202	0.9723
26	99.9998	0.0207	99.9997	0.0199	0.9620
27	99.9993	0.0198	99.9994	0.0192	0.9697
28	100.0004	0.0202	100.0005	0.0193	0.9571
29	99.9993	0.0197	99.9994	0.0189	0.9573
30	99.9988	0.0189	99.9988	0.0179	0.9507

Table A.32 Standard Deviation Estimate when X ~ Normal (100, 10^2), n=10,000 and 10 groups in each study

Number of Group	Standard Deviation Estimate from Weighted Linear Regression Approach		Standard Deviation Estimate Calculated from Raw Data		Relative Efficiency of Standard Deviation Estimates
	Mean	SD	Mean	SD	
2	9.9989	0.0611	10.0007	0.0519	0.8496
3	9.9951	0.0477	9.9980	0.0400	0.8392
4	9.9984	0.0419	10.0005	0.0355	0.8470
5	9.9967	0.0383	9.9985	0.0321	0.8379
6	9.9989	0.0348	9.9999	0.0297	0.8534
7	9.9978	0.0318	9.9994	0.0264	0.8296
8	9.9991	0.0298	10.0009	0.0253	0.8480
9	9.9975	0.0283	9.9989	0.0236	0.8336
10	9.9977	0.0270	9.9998	0.0223	0.8263
11	9.9990	0.0260	10.0006	0.0218	0.8406
12	9.9977	0.0243	9.9994	0.0211	0.8671
13	9.9990	0.0232	10.0003	0.0196	0.8439
14	9.9976	0.0220	9.9993	0.0185	0.8428
15	9.9975	0.0216	9.9992	0.0182	0.8408
16	9.9974	0.0205	9.9999	0.0175	0.8509
17	9.9982	0.0200	10.0002	0.0167	0.8368
18	9.9974	0.0199	9.9993	0.0168	0.8480
19	9.9975	0.0197	9.9992	0.0169	0.8607
20	9.9978	0.0185	10.0000	0.0149	0.8079
21	9.9981	0.0184	10.0001	0.0153	0.8307
22	9.9983	0.0182	10.0002	0.0149	0.8155
23	9.9983	0.0183	10.0003	0.0150	0.8202
24	9.9988	0.0171	10.0001	0.0143	0.8342
25	9.9979	0.0166	10.0001	0.0138	0.8333
26	9.9978	0.0170	9.9995	0.0143	0.8428
27	9.9981	0.0166	10.0000	0.0142	0.8577
28	9.9976	0.0161	10.0000	0.0134	0.8298
29	9.9982	0.0157	10.0004	0.0130	0.8313
30	9.9991	0.0153	10.0007	0.0128	0.8334

APPENDIX B

Table B.1 Relative Efficiency of β_1 estimates when X is gamma distribution

Percentile	α=2	α=3	α=4	α=6
0	4.71E-06	5.94E-06	8.76E-06	9.95E-06
5	9.08E-02	1.05E-01	1.32E-01	1.34E-01
10	1.65E-01	1.82E-01	2.10E-01	2.20E-01
15	2.18E-01	3.17E-01	2.36E-01	2.82E-01
20	2.63E-01	3.04E-01	3.47E-01	3.76E-01
25	2.95E-01	4.00E-01	4.15E-01	4.07E-01
30	3.94E-01	4.06E-01	4.20E-01	4.35E-01
35	4.42E-01	4.50E-01	4.95E-01	5.50E-01
40	4.97E-01	5.04E-01	5.52E-01	5.64E-01
45	5.27E-01	5.18E-01	5.63E-01	5.94E-01
50	5.68E-01	5.93E-01	6.22E-01	5.98E-01
55	5.66E-01	6.25E-01	6.00E-01	6.00E-01
60	5.83E-01	6.01E-01	6.13E-01	6.82E-01
65	6.45E-01	6.71E-01	6.46E-01	6.68E-01
70	6.57E-01	6.70E-01	6.37E-01	6.34E-01
75	6.92E-01	6.45E-01	6.40E-01	6.21E-01
80	5.84E-01	6.03E-01	5.71E-01	5.57E-01
85	5.97E-01	5.74E-01	5.59E-01	5.31E-01
90	5.61E-01	4.86E-01	4.95E-01	4.58E-01
95	3.81E-01	3.48E-01	3.56E-01	3.02E-01
100	2.13E-04	1.65E-04	1.39E-04	1.23E-04

Table B.2 Relative Efficiency of β_1 estimates when X is normal distribution

Percentile	$\mu=\sigma^2=2$	$\mu=\sigma^2=3$	$\mu=\sigma^2=4$	$\mu=\sigma^2=6$
0	3.48E-05	3.47E-05	3.36E-05	3.30E-05
5	2.25E-01	2.22E-01	2.17E-01	2.41E-01
10	3.20E-01	3.41E-01	3.40E-01	3.72E-01
15	4.32E-01	4.40E-01	4.63E-01	4.49E-01
20	5.23E-01	4.63E-01	4.46E-01	4.78E-01
25	5.31E-01	5.44E-01	5.39E-01	5.51E-01
30	5.88E-01	5.70E-01	5.81E-01	5.84E-01
35	6.14E-01	6.29E-01	6.13E-01	5.52E-01
40	6.31E-01	6.49E-01	6.19E-01	6.23E-01
45	6.85E-01	6.52E-01	6.59E-01	6.74E-01
50	6.28E-01	6.36E-01	6.40E-01	6.65E-01
55	6.04E-01	6.44E-01	6.67E-01	6.73E-01
60	6.08E-01	6.43E-01	6.63E-01	6.34E-01
65	5.83E-01	5.69E-01	6.05E-01	6.25E-01
70	5.39E-01	5.97E-01	5.85E-01	5.91E-01
75	5.54E-01	5.72E-01	5.23E-01	5.45E-01
80	4.96E-01	4.50E-01	4.70E-01	4.92E-01
85	4.36E-01	4.31E-01	4.20E-01	4.34E-01
90	3.46E-01	3.25E-01	3.25E-01	3.25E-01
95	2.38E-01	2.28E-01	2.36E-01	2.02E-01
100	4.79E-05	5.39E-05	5.07E-05	5.83E-05

Table B.3 Ratio of the b1 estimates (continuous/grouped)

Percentile	α=2	α=3	α=4	α=6
0	4.2866	2.8803	2.3430	2.0553
5	1.2360	1.0690	0.9337	0.7446
10	1.2094	1.0410	0.9085	0.7220
15	1.1854	1.0177	0.8831	0.7110
20	1.1635	0.9957	0.8632	0.7059
25	1.1404	0.9721	0.8458	0.7061
30	1.1161	0.9516	0.8290	0.7105
35	1.0921	0.9295	0.8147	0.7179
40	1.0671	0.9107	0.8027	0.7291
45	1.0436	0.8914	0.7927	0.7434
50	1.0188	0.8738	0.7850	0.7604
55	0.9944	0.8558	0.7803	0.7801
60	0.9691	0.8408	0.7776	0.8025
65	0.9450	0.8266	0.7792	0.8260
70	0.9191	0.8156	0.7851	0.8524
75	0.8941	0.8072	0.7944	0.8833
80	0.8697	0.8032	0.8085	0.9166
85	0.8475	0.8051	0.8305	0.9515
90	0.8297	0.8180	0.8608	1.0580
95	0.8261	0.8479	0.9047	1.7718
100	0.0186	0.0152	0.0142	0.0161

Table B.4 Relative efficiency of b1 estimate when X is gamma distribution

Percentile	α=2	α=3	α=4	α=6
0	1.22E-04	3.13E-04	3.30E-04	8.11E-04
5	2.17E-02	4.73E-02	9.69E-02	5.14E-01
10	4.62E-02	9.66E-02	2.23E-01	1.17E+00
15	8.40E-02	1.64E-01	4.08E-01	1.76E+00
20	1.29E-01	2.87E-01	5.47E-01	2.26E+00
25	1.56E-01	3.47E-01	8.15E-01	2.48E+00
30	2.24E-01	4.87E-01	9.75E-01	2.79E+00
35	2.80E-01	6.38E-01	1.22E+00	2.40E+00
40	3.62E-01	8.00E-01	1.45E+00	2.19E+00
45	4.17E-01	9.35E-01	1.77E+00	1.98E+00
50	5.42E-01	1.21E+00	1.93E+00	1.45E+00
55	5.99E-01	1.34E+00	2.06E+00	1.10E+00
60	8.03E-01	1.54E+00	1.90E+00	8.45E-01
65	9.72E-01	1.68E+00	1.81E+00	5.51E-01
70	1.10E+00	1.63E+00	1.47E+00	3.72E-01
75	1.32E+00	1.71E+00	1.21E+00	2.41E-01
80	1.41E+00	1.50E+00	8.29E-01	1.21E-01
85	1.42E+00	1.12E+00	4.61E-01	6.74E-02
90	1.39E+00	6.25E-01	2.23E-01	2.08E-03
95	6.83E-01	1.82E-01	5.50E-02	4.56E-04
100	2.38E-02	3.22E-02	4.22E-02	9.65E-02

Table B.5 MSE of b1 Estimate When X is Gamma Distributed (Part 1/2)

	α=2		α=3	
Percentile	Continuous	Grouped	Continuous	Grouped
0	0.000309	13.350150	0.000223	4.250004
5	0.000289	0.069041	0.000251	0.010174
10	0.000282	0.050011	0.000240	0.004195
15	0.000307	0.038506	0.000229	0.001721
20	0.000320	0.029780	0.000246	0.000871
25	0.000298	0.021792	0.000230	0.001461
30	0.000316	0.014965	0.000238	0.002838
35	0.000329	0.009581	0.000239	0.005364
40	0.000313	0.005434	0.000256	0.008214
45	0.000299	0.002676	0.000223	0.012017
50	0.000300	0.000947	0.000251	0.016138
55	0.000322	0.000566	0.000234	0.021042
60	0.000314	0.001282	0.000231	0.025527
65	0.000322	0.003293	0.000225	0.030188
70	0.000290	0.006822	0.000222	0.034046
75	0.000298	0.011469	0.000227	0.037139
80	0.000309	0.017059	0.000238	0.038896
85	0.000279	0.023326	0.000223	0.038237
90	0.000320	0.029231	0.000231	0.033330
95	0.000294	0.030593	0.000228	0.024216
100	0.000315	0.963224	0.000234	0.969821

Table B.6 MSE of b1 Estimate When X is Gamma Distributed (Part 2/2)

	α=4		α=6	
Percentile	Continuous	Grouped	Continuous	Grouped
0	0.000210	2.438724	0.000268	1.446342
5	0.000220	0.006674	0.000256	0.065586
10	0.000232	0.009393	0.000266	0.077219
15	0.000222	0.014078	0.000251	0.083436
20	0.000199	0.019030	0.000259	0.086306
25	0.000223	0.023846	0.000259	0.086287
30	0.000199	0.029458	0.000273	0.084168
35	0.000202	0.034429	0.000243	0.079345
40	0.000231	0.039005	0.000262	0.073502
45	0.000219	0.042982	0.000266	0.065823
50	0.000220	0.046246	0.000256	0.057252
55	0.000212	0.048399	0.000253	0.048624
60	0.000207	0.049326	0.000260	0.039173
65	0.000207	0.048570	0.000239	0.030713
70	0.000204	0.046249	0.000252	0.022244
75	0.000205	0.042568	0.000247	0.014392
80	0.000216	0.036798	0.000230	0.008846
85	0.000213	0.029066	0.000263	0.006128
90	0.000203	0.020321	0.000255	0.126015
95	0.000223	0.012986	0.000263	1.172839
100	0.000219	0.971867	0.000272	0.968068

Table B.7 Ratio of the b1 estimates (continuous/grouped)

Percentile	α=2	α=3	α=4	α=6
0	1.6008	1.2818	1.1922	1.1858
5	1.1202	1.0284	0.9322	0.7673
10	1.0884	0.9951	0.8940	0.7337
15	1.0655	0.9670	0.8672	0.7189
20	1.0434	0.9435	0.8447	0.7155
25	1.0199	0.9201	0.8258	0.7198
30	1.0003	0.9005	0.8113	0.7301
35	0.9806	0.8820	0.7999	0.7446
40	0.9614	0.8658	0.7917	0.7630
45	0.9439	0.8517	0.7876	0.7849
50	0.9269	0.8390	0.7858	0.8082
55	0.9094	0.8286	0.7872	0.8343
60	0.8940	0.8200	0.7917	0.8610
65	0.8791	0.8141	0.8001	0.8899
70	0.8643	0.8104	0.8106	0.9222
75	0.8514	0.8098	0.8255	0.9540
80	0.8397	0.8129	0.8432	0.9867
85	0.8301	0.8211	0.8663	1.0297
90	0.8249	0.8353	0.8965	1.1061
95	0.8280	0.8615	0.9340	1.6303
100	0.0182	0.0149	0.0141	0.0163

Table B.8 Relative efficiency of b1 estimates (normal covariate)

Percentile	α=2	α=3	α=4	α=6
0	2.86E-02	2.95E-02	2.92E-02	3.35E-02
5	2.04E-02	3.12E-02	6.62E-02	3.61E-01
10	6.28E-02	1.07E-01	2.36E-01	1.16E+00
15	1.05E-01	2.02E-01	4.10E-01	2.03E+00
20	1.74E-01	3.14E-01	7.00E-01	2.30E+00
25	2.28E-01	4.20E-01	9.45E-01	2.78E+00
30	3.28E-01	6.14E-01	1.25E+00	2.61E+00
35	4.15E-01	7.73E-01	1.42E+00	2.31E+00
40	5.78E-01	9.92E-01	1.72E+00	1.91E+00
45	6.37E-01	1.31E+00	1.94E+00	1.53E+00
50	7.93E-01	1.46E+00	2.04E+00	1.11E+00
55	9.35E-01	1.53E+00	1.97E+00	7.89E-01
60	1.04E+00	1.61E+00	1.74E+00	6.55E-01
65	1.16E+00	1.69E+00	1.52E+00	3.98E-01
70	1.30E+00	1.88E+00	1.31E+00	2.93E-01
75	1.43E+00	1.54E+00	9.31E-01	1.79E-01
80	1.45E+00	1.27E+00	6.18E-01	1.30E-01
85	1.29E+00	9.54E-01	4.16E-01	5.99E-02
90	1.13E+00	6.17E-01	1.98E-01	2.95E-03
95	7.05E-01	2.70E-01	6.16E-02	2.74E-04
100	3.32E+04	2.09E+05	7.70E+08	5.13E+04

Table B.9 MSE of b1 Estimate When X is Normal Distributed (Part 1/2)

	α=2		α=3	
Percentile	Continuous	Grouped	Continuous	Grouped
0	0.000365	0.372945	0.000241	0.087598
5	0.000341	0.031141	0.000227	0.008079
10	0.000356	0.013496	0.000236	0.002218
15	0.000341	0.007505	0.000238	0.002249
20	0.000348	0.003968	0.000232	0.003850
25	0.000334	0.001901	0.000219	0.006724
30	0.000347	0.001058	0.000229	0.010302
35	0.000373	0.001263	0.000220	0.014021
40	0.000369	0.002068	0.000225	0.018178
45	0.000360	0.003793	0.000246	0.022056
50	0.000360	0.005904	0.000228	0.026162
55	0.000372	0.008765	0.000213	0.029674
60	0.000377	0.011449	0.000224	0.032337
65	0.000371	0.014868	0.000237	0.034515
70	0.000341	0.018713	0.000246	0.036211
75	0.000358	0.022508	0.000229	0.036422
80	0.000371	0.025943	0.000252	0.034834
85	0.000338	0.028905	0.000252	0.032225
90	0.000340	0.031077	0.000244	0.027524
95	0.000362	0.030034	0.000251	0.019886
100	0.000360	0.963821	0.000245	0.970372

Table B.10 MSE of b1 Estimate When X is Normal Distributed (Part 2/2)

	α=4		α=6	
Percentile	Continuous	Grouped	Continuous	Grouped
0	0.000206	0.044152	0.000235	0.041477
5	0.000212	0.007820	0.000240	0.054765
10	0.000207	0.012079	0.000255	0.070841
15	0.000196	0.017941	0.000243	0.078631
20	0.000208	0.024575	0.000227	0.080969
25	0.000196	0.030307	0.000257	0.078191
30	0.000207	0.035761	0.000233	0.072690
35	0.000203	0.039803	0.000244	0.064925
40	0.000196	0.043149	0.000241	0.055987
45	0.000197	0.045039	0.000230	0.046361
50	0.000215	0.045948	0.000239	0.037129
55	0.000205	0.045467	0.000220	0.027718
60	0.000206	0.043195	0.000244	0.019611
65	0.000207	0.040017	0.000234	0.012560
70	0.000201	0.036152	0.000238	0.006922
75	0.000202	0.030464	0.000235	0.003368
80	0.000213	0.024764	0.000266	0.002191
85	0.000204	0.018256	0.000252	0.005115
90	0.000208	0.011660	0.000236	0.091188
95	0.000202	0.007532	0.000225	1.215998
100	0.000204	0.971915	0.000228	0.967728

Table B.11 Relative efficiency on b1 estimate from number of group when X is normal distribution and β_1=0

Number of Group	$\mu=\sigma^2=2$	$\mu=\sigma^2=3$	$\mu=\sigma^2=4$	$\mu=\sigma^2=6$
2	0.632973	0.632973	0.632973	0.632973
3	0.783756	0.783756	0.783756	0.783756
4	0.858662	0.858662	0.858662	0.858662
5	0.913601	0.913601	0.913601	0.913601
6	0.943288	0.943288	0.943288	0.943288
7	0.929323	0.929348	0.929326	0.929408
8	0.959068	0.959068	0.959068	0.959068
9	0.959444	0.959444	0.959444	0.959444
10	0.950355	0.950355	0.950355	0.950355
11	0.980091	0.980091	0.980091	0.980091
12	0.954304	0.954304	0.954304	0.954304
13	0.967829	0.967829	0.967829	0.967829
14	0.972806	0.972773	0.972787	0.972848
15	0.974872	0.974872	0.974872	0.974872
16	0.980262	0.980262	0.980262	0.980262
17	0.975882	0.975882	0.975882	0.975882
18	0.995099	0.995099	0.995099	0.995099
19	0.989947	0.989947	0.989947	0.989947
20	0.974273	0.974273	0.974273	0.974273
21	0.980936	0.980936	0.980936	0.980936

Table B.12 Relative efficiency on b1 estimate from number of group when X is gamma distribution and β_1=0

Number of Group	α=2	α=3	α=4	α=6
2	0.602226	0.591823	0.616974	0.597496
3	0.735532	0.749335	0.776034	0.735533
4	0.836449	0.817649	0.844314	0.838709
5	0.808836	0.874016	0.831937	0.871128
6	0.893362	0.871254	0.885381	0.884897
7	0.853849	0.929339	0.931094	0.888490
8	0.906813	0.915437	0.893216	0.914642
9	0.957387	0.922636	0.947836	0.955014
10	0.934467	0.951236	0.934374	0.946619
11	0.952020	0.928079	0.966346	0.949622
12	0.945630	0.936720	0.945290	0.950909
13	0.911925	0.984948	0.951046	0.970245
14	0.924299	0.962164	0.951295	0.968892
15	0.967639	0.962381	0.969942	0.968874
16	0.950975	0.947187	0.966331	0.974517
17	0.970348	0.964186	0.957854	0.987098
18	0.946252	0.978006	0.977071	0.966032
19	0.953593	0.962844	0.980462	0.988214
20	0.972284	0.958170	0.959311	0.974372
21	0.975960	0.965445	0.979491	0.972156

Table B.13 Ratio of coefficient estimates (grouped/continuous) when X is normal distribution with β_1=1

Number of Group	$\mu=\sigma^2=2$	$\mu=\sigma^2=3$	$\mu=\sigma^2=4$	$\mu=\sigma^2=6$
2	0.926681	0.838928	0.785486	0.808072
3	0.958785	0.888751	0.853228	0.863190
4	0.968758	0.916049	0.895266	0.886000
5	0.973095	0.933692	0.922251	0.904515
6	0.975938	0.946323	0.940306	0.919876
7	0.977700	0.955345	0.952754	0.932368
8	0.979973	0.962286	0.961448	0.942295
9	0.981228	0.967762	0.968161	0.950815
10	0.982417	0.972049	0.972946	0.957146
11	0.983682	0.975299	0.976841	0.962746
12	0.984665	0.978163	0.980042	0.967076
13	0.985336	0.980352	0.982363	0.971008
14	0.986246	0.982324	0.984461	0.974059
15	0.987134	0.984003	0.986117	0.976521
16	0.987707	0.985518	0.987571	0.978843
17	0.988459	0.986686	0.988814	0.980808
18	0.988925	0.987759	0.989772	0.982428
19	0.989321	0.988662	0.990622	0.983872
20	0.989862	0.989495	0.991429	0.985207
21	0.990261	0.990237	0.992120	0.986274

Table B.14 Relative bias (%) of b1 Estimate When X is Normal Distributed and β_1=1 (Part 1/2)

Number of Group	$\mu=\sigma^2=2$		$\mu=\sigma^2=3$	
	Continuous	Grouped	Continuous	Grouped
2	0.024210	12.088232	0.064599	3.008807
3	0.016207	8.991750	-0.005057	-0.437108
4	0.041069	6.483900	-0.004197	-3.234582
5	0.073995	4.399897	-0.001584	-5.770390
6	0.020568	2.172906	-0.048743	-8.026858
7	0.026288	0.075181	-0.033962	-9.958982
8	-0.011998	-1.775618	0.055739	-11.661895
9	0.157323	-3.677346	0.013895	-13.394193
10	-0.087803	-5.641564	0.083678	-14.749009
11	0.068305	-7.248370	0.029763	-16.051712
12	0.000747	-9.046930	0.064639	-17.090785
13	0.066112	-10.606756	0.091120	-17.880462
14	0.062082	-12.035360	0.016605	-18.605011
15	-0.075877	-13.621133	0.044006	-18.905652
16	0.099693	-14.811567	0.043004	-18.982375
17	0.009376	-16.016976	-0.048670	-18.686523
18	0.032334	-16.893825	-0.006649	-17.913719
19	-0.002707	-17.493432	0.021908	-16.505343
20	-0.006247	-17.162058	0.021830	-13.865672
21	0.027132	107.564746	0.052064	65.783786

Table B.15 Relative bias (%) of b1 Estimate When X is Normal Distributed and β_1=1 (Part 2/2)

Number of Group	$\mu=\sigma^2=4$		$\mu=\sigma^2=6$	
	Continuous	Grouped	Continuous	Grouped
2	-0.006314	-6.681240	0.061543	-23.208905
3	0.016119	-10.523160	0.031070	-26.613262
4	0.062825	-13.281300	0.060721	-28.051758
5	-0.028715	-15.614080	0.094465	-28.409253
6	0.006067	-17.475190	-0.021076	-28.003912
7	0.002653	-18.875270	0.046035	-26.969621
8	0.041526	-20.009840	-0.082580	-25.521156
9	-0.016787	-20.858810	0.069459	-23.624921
10	0.058646	-21.231450	-0.029643	-21.566399
11	0.045114	-21.407200	0.073377	-19.127173
12	0.035791	-21.223260	0.049239	-16.531941
13	0.009491	-20.818120	0.059954	-13.787998
14	0.048567	-19.996570	0.033561	-10.946520
15	-0.008001	-18.943660	0.029736	-7.796396
16	-0.028513	-17.470870	-0.074416	-4.576388
17	0.049163	-15.452730	0.046846	-1.231657
18	-0.021508	-13.340900	0.020970	2.481098
19	0.050114	-10.515920	0.053527	9.498429
20	0.109454	-6.423840	-0.027545	66.564187
21	-0.040524	35.310010	0.039047	-0.891348

Table B.16 Relative Efficiency of b1 Estimate when X is normal distribution and β_1=1

Number of Group	$\mu=\sigma^2=2$	$\mu=\sigma^2=3$	$\mu=\sigma^2=4$	$\mu=\sigma^2=6$
2	0.789408	1.357256	1.959999	1.118941
3	0.826222	1.325828	1.603100	1.471997
4	0.954804	1.267808	1.346772	1.514136
5	1.031443	1.202605	1.201126	1.387244
6	1.058861	1.155322	1.133963	1.363870
7	1.078289	1.100753	1.099271	1.258350
8	1.063556	1.090893	1.057659	1.220951
9	1.058631	1.080747	1.042884	1.173933
10	1.083413	1.058562	1.019419	1.100287
11	1.048045	1.050269	1.028884	1.114834
12	1.046369	1.022789	1.023239	1.097673
13	1.054078	1.036442	1.020799	1.073307
14	1.064095	1.020840	1.019368	1.053499
15	1.046106	1.026586	1.017101	1.027128
16	1.062564	1.023194	1.016656	1.048065
17	1.029762	0.996010	1.000024	1.022271
18	1.051212	1.019618	1.007583	1.017931
19	1.048922	1.025703	1.003193	1.042088
20	1.057994	1.017229	1.003457	1.019201
21	1.043659	1.013659	1.001829	0.998636

Table B.17 MSE of b1 Estimate When X is Normal Distributed and β_1=1 (Part 1/2)

Number of Group	$\mu=\sigma^2=2$		$\mu=\sigma^2=3$	
	Continuous	Grouped	Continuous	Grouped
2	0.000365	0.005892	0.000241	0.026120
3	0.000341	0.002123	0.000227	0.012536
4	0.000356	0.001344	0.000236	0.007263
5	0.000341	0.001063	0.000238	0.004567
6	0.000348	0.000861	0.000232	0.003006
7	0.000334	0.000761	0.000219	0.002086
8	0.000347	0.000696	0.000229	0.001640
9	0.000373	0.000693	0.000220	0.001185
10	0.000369	0.000622	0.000225	0.000979
11	0.000360	0.000635	0.000246	0.000821
12	0.000360	0.000602	0.000228	0.000711
13	0.000372	0.000597	0.000213	0.000613
14	0.000377	0.000521	0.000224	0.000508
15	0.000371	0.000513	0.000237	0.000467
16	0.000341	0.000476	0.000246	0.000463
17	0.000358	0.000497	0.000229	0.000414
18	0.000371	0.000475	0.000252	0.000367
19	0.000338	0.000420	0.000252	0.000371
20	0.000340	0.000432	0.000244	0.000349
21	0.000362	0.000436	0.000251	0.000324

Table B.18 MSE of b1 Estimate When X is Normal Distributed and β_1=1 (Part 2/2)

Number of Group	$\mu=\sigma^2=4$		$\mu=\sigma^2=6$	
	Continuous	Grouped	Continuous	Grouped
2	0.000206	0.046029	0.000235	0.037072
3	0.000212	0.021713	0.000240	0.018849
4	0.000207	0.011082	0.000255	0.013002
5	0.000196	0.006107	0.000243	0.009071
6	0.000208	0.003821	0.000227	0.006570
7	0.000196	0.002338	0.000257	0.004646
8	0.000207	0.001674	0.000233	0.003456
9	0.000203	0.001136	0.000244	0.002528
10	0.000196	0.000866	0.000241	0.001983
11	0.000197	0.000703	0.000230	0.001583
12	0.000215	0.000603	0.000239	0.001326
13	0.000205	0.000522	0.000220	0.001043
14	0.000206	0.000414	0.000244	0.000888
15	0.000207	0.000387	0.000234	0.000746
16	0.000201	0.000363	0.000238	0.000693
17	0.000202	0.000312	0.000235	0.000572
18	0.000213	0.000304	0.000266	0.000542
19	0.000204	0.000284	0.000252	0.000486
20	0.000208	0.000272	0.000236	0.000459
21	0.000202	0.000252	0.000225	0.000406

Table B.19 Ratio of the b1 Estimates (Grouped/Continuous) when X is Gamma

Number of Group	α=2	α=3	α=4	α=6
2	0.538934	1.115097	1.900088	1.479086
3	0.835964	1.486166	1.671579	1.186595
4	1.036162	1.511150	1.336038	1.110818
5	1.216012	1.484386	1.188747	1.096935
6	1.249096	1.330601	1.080591	1.097876
7	1.306316	1.234233	1.023546	1.100560
8	1.333497	1.155529	1.033785	1.071400
9	1.316352	1.085153	1.014674	1.047394
10	1.245376	1.058066	1.014441	1.047665
11	1.231750	1.044032	1.012308	1.032720
12	1.207681	1.030055	0.989094	1.041582
13	1.178206	1.016784	0.995866	1.024382
14	1.169281	1.041118	0.990392	1.029707
15	1.178980	0.984117	0.996368	1.022008
16	1.118731	1.001034	0.993881	1.010020
17	1.130139	1.002035	0.993271	1.028169
18	1.109827	1.002380	0.988895	1.001657
19	1.057578	0.997430	0.995651	1.021392
20	1.107137	0.983383	0.992077	1.014425
21	1.052215	0.991636	0.992487	1.010509

Table B.20 Relative bias (%) of b1 Estimate when X distributed gamma and β_1=1 (Part 1/2)

Number of Group	α=2		α=3	
	Continuous	Grouped	Continuous	Grouped
2	0.039646	1.945214	0.037483	-12.710024
3	0.001801	-2.097025	0.080687	-12.593880
4	0.012021	-4.196667	0.038147	-11.664959
5	0.111929	-5.281037	0.030364	-10.333097
6	0.149258	-5.814004	0.055888	-8.931995
7	0.050398	-6.070428	-0.038128	-7.758571
8	0.031630	-6.120909	-0.008654	-6.664467
9	-0.040190	-6.071076	-0.009578	-5.778615
10	0.047586	-5.818264	0.050242	-4.943386
11	0.066930	-5.595393	0.008279	-4.350076
12	0.109756	-5.302293	-0.006707	-3.862787
13	0.029759	-5.155518	-0.029714	-3.438841
14	0.105950	-4.822279	-0.014019	-3.051120
15	0.060783	-4.627905	-0.000933	-2.725684
16	-0.004610	-4.447812	0.033614	-2.428386
17	-0.010856	-4.214506	0.051002	-2.186830
18	0.066449	-3.948025	0.002956	-2.011342
19	0.053363	-3.725874	-0.016788	-1.864116
20	0.002737	-3.584759	0.050984	-1.633071
21	0.022393	-3.398225	0.065098	-1.503699

Table B.21 Relative bias (%) of b1 Estimate when X distributed gamma and β_1=1 (Part 2/2)

Number of Group	α=4		α=6	
	Continuous	Grouped	Continuous	Grouped
2	-0.038531	-21.517350	0.037214	-23.962889
3	-0.008899	-16.601059	0.030594	-12.976040
4	0.014220	-12.270009	0.072915	-8.290840
5	0.059583	-9.078351	0.049226	-5.944941
6	0.015911	-6.915743	0.080180	-4.477068
7	0.084315	-5.328592	0.050308	-3.527107
8	-0.010027	-4.314802	-0.065214	-2.964737
9	0.021431	-3.495912	0.076258	-2.329070
10	0.026536	-2.892042	0.007837	-2.038269
11	0.036124	-2.408456	0.040824	-1.711033
12	0.026610	-2.067047	0.097378	-1.438895
13	-0.007312	-1.811574	-0.017713	-1.351765
14	0.064992	-1.518809	0.049445	-1.144802
15	0.084581	-1.311502	0.000627	-1.065110
16	0.017921	-1.227623	0.081667	-0.884351
17	-0.033496	-1.142356	0.118364	-0.749516
18	0.047270	-0.955601	0.006123	-0.785418
19	0.043713	-0.865328	0.114945	-0.616270
20	-0.015831	-0.834383	0.065439	-0.606060
21	0.086488	-0.670719	0.021924	-0.590615

Table B.22 Relative Efficiency of b1 Estimate when X is gamma distribution and β_1=1

Number of Group	α=2	α=3	α=4	α=6
2	1.019048	0.872573	0.785129	0.760088
3	0.979012	0.873357	0.834064	0.869973
4	0.957918	0.883014	0.877175	0.916423
5	0.946131	0.896397	0.908675	0.940088
6	0.940456	0.910171	0.930695	0.954464
7	0.938823	0.922766	0.945917	0.964244
8	0.938494	0.933436	0.956948	0.970986
9	0.939667	0.942304	0.964834	0.975965
10	0.941369	0.950089	0.970822	0.979541
11	0.943415	0.956420	0.975563	0.982489
12	0.945939	0.961437	0.979069	0.984652
13	0.948163	0.965899	0.981956	0.986657
14	0.950770	0.969625	0.984172	0.988063
15	0.953142	0.972752	0.986051	0.989343
16	0.955566	0.975388	0.987547	0.990348
17	0.957959	0.977633	0.988908	0.991332
18	0.959882	0.979858	0.989976	0.992085
19	0.962228	0.981524	0.990914	0.992696
20	0.964126	0.983168	0.991813	0.993289
21	0.965802	0.984322	0.992435	0.993876

Table B.23 MSE of Coefficient Estimate when X distributed gamma and β_1=1

(Part 1/2)

Number of Group	α=2		α=3	
	Continuous	Grouped	Continuous	Grouped
2	0.000309	0.000951	0.000223	0.016354
3	0.000289	0.000786	0.000251	0.016029
4	0.000282	0.002033	0.000240	0.013766
5	0.000307	0.003040	0.000229	0.010831
6	0.000320	0.003635	0.000246	0.008163
7	0.000298	0.003913	0.000230	0.006206
8	0.000316	0.003983	0.000238	0.004647
9	0.000329	0.003936	0.000239	0.003560
10	0.000313	0.003636	0.000256	0.002686
11	0.000299	0.003373	0.000223	0.002106
12	0.000300	0.003059	0.000251	0.001736
13	0.000322	0.002931	0.000234	0.001412
14	0.000314	0.002593	0.000231	0.001153
15	0.000322	0.002415	0.000225	0.000971
16	0.000290	0.002238	0.000222	0.000812
17	0.000298	0.002040	0.000227	0.000705
18	0.000309	0.001836	0.000238	0.000642
19	0.000279	0.001652	0.000223	0.000571
20	0.000320	0.001574	0.000231	0.000501
21	0.000294	0.001434	0.000228	0.000456

Table B.24 MSE of Coefficient Estimate when X distributed gamma and β_1=1 (Part 2/2)

Number of Group	α=4		α=6	
	Continuous	Grouped	Continuous	Grouped
2	0.000210	0.046410	0.000268	0.057603
3	0.000220	0.027691	0.000256	0.017054
4	0.000232	0.015229	0.000266	0.007113
5	0.000222	0.008428	0.000251	0.003763
6	0.000199	0.004967	0.000259	0.002240
7	0.000223	0.003057	0.000259	0.001479
8	0.000199	0.002054	0.000273	0.001133
9	0.000202	0.001422	0.000243	0.000774
10	0.000231	0.001064	0.000262	0.000666
11	0.000219	0.000796	0.000266	0.000550
12	0.000220	0.000649	0.000256	0.000452
13	0.000212	0.000541	0.000253	0.000430
14	0.000207	0.000439	0.000260	0.000383
15	0.000207	0.000379	0.000239	0.000347
16	0.000204	0.000356	0.000252	0.000327
17	0.000205	0.000337	0.000247	0.000295
18	0.000216	0.000310	0.000230	0.000292
19	0.000213	0.000288	0.000263	0.000294
20	0.000203	0.000275	0.000255	0.000288
21	0.000223	0.000269	0.000263	0.000295

REFERENCES

1. Althuis MD, Fergenbaum JH, Garcia-Closas M, et al. Etiology of hormone receptor-defined breast cancer: a systematic review of the literature. *Cancer Epidemiol Biomarkers Prev.* 2004;13:1558-68.

2. Altman DG. Systematic reviews of studies evaluation of prognostic variables .In *Systematic Reviews in Health Care: Meta-Analysis in Context*. Egger M, Smith GD, Altman DG (eds). BMJ Publishing Group: London: 2001;228 –247.

3. Altman DG, Lausen B, Sauerbrei W, Schumacher M (1994) Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst* 86: 829–835

4. Bandera EV, Kushi LH, Gifkins DM, Moore DF, McCullough M. (2007) The association between food, nutrition, and physical activity and the risk of endometrial cancer and underlying mechanisms. In: Second Report on Food, Nutrition, Physical Activity and the Prevention of Cancer., World Cancer Research Fund International/American Institute for Cancer Research

5. Bartali B, Frongillo EA, Guralnik JM, Stipanuk MH, Allore HG, Cherubini A, Bandinelli S, Ferrucci L, Gill TM. Serum micronutrient concentrations and decline in physical function among older persons. *JAMA.* 2008;299:308-15

6. Bates D, Maechler M. (2009) lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-32. http://CRAN.R-project.org/package=lme4

7. Berlin JA, Longnecker MP, Greenland S. Meta-analyhsis of Epidemiologic Dose-Response Data. *Epidemiology* 1993;4:218-228.

8. Bu X, Ma Y, Der R, Demeester T, Bernstein L, Chandrasoma PT. Body mass index is associated with Barrett esophagus and cardiac mucosal metaplasia. *Dig Dis Sci* 2006;51:1589–94.

9. Chang ET,Smedby KE,Zhang SM,et al.Alcohol intake and risk of non-Hodgkin lymphoma in men and women. *Cancer Causes Control* 2004;15:1067 –76.

10. Chen H, Cohen P, Chen S. Biased odds ratios from dichotomization of age. *Statistics in Medicine.* 2007;26:3487–3497

11. Chêne G, Thompson SG. Methods for summarizing the risk associations of quantitative variables in epidemiologic studies in a consistent form. *Am J Epidemiol*. 1996;144:610-21.

12. Chernoff H, Lieberman GJ. Use of normal probability paper. *Journal of the American. Statistical Association.* 1954;49:778-85.

13. Chernoff H, Lieberman GJ. The Use of Generalized Probability Paper for Continuous Distributions. *The Annals of Mathematical Statistics.* 1956;27:806-818.

14. Cohen J. The cost of dichotomization. *Applied Psychological Measurement.* 1983;7:249-253.

15. Corley DA, Kubo A, Levin TR, et al. Abdominal obesity and body mass index as risk factors for Barrett's esophagus. *Gastroenterology* 2007;133:34–41.

16. Corley DA, Levin TR, Habel LA, Buffler PA. Barrett's esophagus and medications that relax the lower esophageal sphincter. *Am J Gastroenterol* 2006;101:937– 44.

17. Deeks JJ, Higgins, JPT, Altman DG, editors. Analysing data and undertaking meta-analyses In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 5.0.2 [updated September 2009]; Chapter 9. http://www.cochrane-handbook.org (accessed February 20, 2010).

18. De Laurentiis M, Arpino G, Massarelli E, et al. A Meta-Analysis on the Interaction between HER-2 Expression and Response to Endocrine Treatment in Advanced Breast Cancer. *Clinical Cancer Research* 2005;11:4741-4748.

19. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*;38:1-22, 1977

20. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986; 7: 177-188.

21. Dunn PK, Smyth GK. Tweedie Family Densities: Methods of Evaluation. In *Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling, Odense, July 2 – 6, 2001*, B. Jørgensen (eds.). International Workshop on Statistical Modelling, Odense, pp. 155-162.

22. Dunn PK, Smyth GK. Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing.* 2005;15:267-280.

23. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival. *Lancet.* 2005;365:1687-1717.

24. Edelstein ZR, Farrow DC, Bronner MP, Rosen SN, Vaughan TL. Central adiposity and risk of Barrett's esophagus. *Gastroenterology* 2007;133:403–11.

25. El-Serag HB, Kvapil P, Hacken-Bitar J, Kramer JR. Abdominal obesity and the risk of Barrett's esophagus. *Am J Gastroenterol* 2005;100:2151– 6.

26. Ferrandina G, Scambia G, Bardelli F, et al. Relationship between cathepsin-D content and disease-free survival in node-negative breast cancer patients: a meta-analysis. *British Journal of Cancer* 1997;76:661-666

27. Flegal, KM, Graubard, BI, Williamson DF, Gail MH. Cause-Specific Excess Deaths Associated With Underweight, Overweight, and Obesity. *JAMA*. 2007;2028-2037.

28. Flum DR, Koepsell T. The clinical and economic correlates of misdiagnosed appendicitis: nationwide analysis. *Arch Surg.* 2002 Jul;137(7):799-804.

29. Foekens JA, Look MP, Bolt-de Vries J, Meijer-van Gelder ME, van Putten WL, Klijn JG. Cathepsin-D in primary breast cancer: prognostic evaluation involving 2810 patients. *Br J Cancer.* 1999;79:300-7.

30. Gerson LB, Shetler K, Triadafilopoulos G. Prevalence of Barrett's esophagus in asymptomatic individuals. *Gastroenterology* 2002;123:461–7.

31. Gerson LB, Ullah N, Fass R, Green C, Shetler K, Singh G. Does body mass index differ between patients with Barrett's oesophagus and patients with chronic gastro-oesophageal reflux disease? *Aliment Pharmacol Ther* 2007;25:1079–86.

32. Goodman MT, Hankin JH, Wilkens LR, Lyu LC, McDuffie K, Liu LQ, Kolonel LN. Diet, body size, physical activity, and the risk of endometrial cancer. *Cancer Res.* 1997;57:5077-85.

33. Greenland S. Quantitative Methods in the Review of Epidemiologic Literature. *Epidemioilogic Reviews* 1987;9:1-30.

34. Hartemink N, Boshuizen HC, Nagelkerke NJD, et al. Combining Risk Estimates from Observational Studies with Different Exposure Cutpoints: A Meta-analysis on Body Mass Index and Diabetes Type 2. *Am J Epidemiol* 2006;163:1042 –1052

35. Holly EA, Lele C, Bracci PM, McGrath MS. Case-control study of non-Hodgkin's lymphoma among women and heterosexual men in the San Francisco Bay Area, California. *Am J Epidemiol.* 1999;150:375-89.

36. Jain MG, Howe GR, Rohan TE. Nutritional factors and endometrial cancer in Ontario, Canada. *Cancer Control.* 2000;7:288-96.

37. Johansson J, Hakansson HO, Mellblom L, et al. Risk factors for Barrett's oesophagus: a population-based approach. *Scand J Gastroenterol* 2007;42:148 –56.

38. Jørgensen, B. 1997. *The Theory of Dispersion Models*. Chapman and Hall, London.

39. Kamat P, Wen S, Morris J, Anandasabapathy S. Exploring the association between elevated body mass index and Barrett's esophagus: a systematic review and meta-analysis. *Ann Thorac Surg.* 2009 Feb;87(2):655-62.

40. Konecny G, Pauletti G, Pegram M, et al. Quantitative Association Between HER-2/neu and Steroid Hormone Receptors in Hormone Receptor-Positive Primary Breast Cancer. *J Natl Cancer Inst* 2003;95:142–53.

41. Kuo YH. Estrogen-Receptor Status in Breast Cancer. (Letter) *JAMA* 2000;283:338.

42. Levi F, Franceschi S, Negri E, La Vecchia C. Dietary factors and the risk of endometrial cancer. *Cancer.* 1993;71:3575-81.

43. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959; 22: 719-748.

44. McCullagh, P. and J. A. Nelder. 1989. Generalized Linear Models, Second edition. London: Chapman & Hall.

45. McLachlan GJ, Jones, PN. Fitting Mixture Models to Grouped and Truncated Data via the EM Algorithm. *Biometrics* 1988;44:571-578.

46. McLaren CE, Brittenham GM, Hasselblad V. Analysis of the volume of red blood cells: application of the expectation-Maximization algorithm to grouped data from the doubly-truncated lognormal distribution. *Biometrics* 1986;42:143-158.

47. Morgan TM, Elashoff RM. Effect of categorizing a continuous covariate on the comparison of survival Time. *JASA*. 1986;396:917-921.

48. Morton LM, Holford TR, Leaderer B, Zhang Y, Zahm SH, Boyle P, Flynn S, Tallini G, Owens PH, Zhang B, Zheng T. Alcohol use and risk of non-Hodgkin's lymphoma among Connecticut women (United States). *Cancer Causes Control.* 2003 Sep;14(7):687-94.

49. Morton LM, Zheng T, Holford TR, et al, Alcohol consumption and risk of non-Hodgkin lymphoma: a pooled analysis. *Lancet Oncol* 2005;6:469 –76

50. National Heart, Lung, and Blood Institute. Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults Web page. http://www.nhlbi.nih.gov/guidelines/obesity/ob_home.htm. Accessed February 20, 2010.

51. Penman AD, Johnson WD. The changing shape of the body mass index distribution curve in the population: implications for public health policy to reduce the prevalence of adult obesity. *Prev Chronic Dis* [serial online] 2006 Jul [Accessed on February 28, 2010]. Available from: URL: http://www.cdc.gov/pcd/issues/2006/jul/05_0232.htm.

52. Peto R. Why do we need systematic overviews of randomized trials? *Stat Med* 1987;6:233-44.

53. R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

54. Ronkainen J, Aro P, Storskrubb T, et al. Prevalence of Barrett's esophagus in the general population: an endoscopic study. *Gastroenterology* 2005;129:1825–31.

55. Schwarzer G. (2009). meta: Meta-Analysis with R. R package version 1.1-4. http://CRAN.R-project.org/package=meta

56. Smith KJ, O'Brien SM, Smithers BM, et al. Interactions among smoking, obesity, and symptoms of acid reflux in Barrett's esophagus. *Cancer Epidemiol Biomarkers Prev* 2005;14:2481– 6.

57. Stein DJ, El-Serag HB, Kuczynski J, Kramer JR, Sampliner RE. The association of body mass index with Barrett's oesophagus. *Aliment Pharmacol Ther* 2005;22:1005–10.

58. Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. Cochrane Working Group. Stat Med 1995;14:2057-79.

59. Stewart LA, Parmar MK. Meta-analysis of the literature or of individual patient data: is there a difference? Lancet 1993;341:418-22.

60. Tavani A, Gallus S, La Vecchia C, Franceschi S. Alcohol drinking and risk of non-Hodgkin's lymphoma. Eur J Clin Nutr. 2001;55:824-6.

61. Therneau T and original R port by Thomas Lumley (2009). survival: Survival analysis, including penalised likelihood.. R package version 2.35-7. http://CRAN.R-project.org/package=survival

62. Veugelers PJ, Porter GA, Guernsey DL, Casson AG. Obesity and lifestyle risk factors for gastroesophageal reflux disease, Barrett esophagus and esophageal adenocarcinoma. *Diseases of the Esophagus* 2006;19:321– 8.

63. Weibull.com. Probability Plotting Papers. http://www.weibull.com/GPaper. Accessed on January 27, 2009.

64. Weisstein EW. "Matrix Inverse." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/MatrixInverse.html

65. Willett EV, Smith AG, Dovey GJ, Morgan GJ, Parker J, Roman E. Tobacco and alcohol consumption and the risk of non-Hodgkin lymphoma. *Cancer Causes Control.* 2004;15:771-80.

66. World Health Organization. *Obesity: preventing and managing the global epidemic.* WHO Technical Report Series, No. 894. Geneva, Switzerland: World Health Organization, 1999.

67. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases* 1985; 27: 335-371.

68. Zheng W, Doyle TJ, Kushi LH, Sellers TA, Hong CP, Folsom AR. Tea consumption and cancer incidence in a prospective cohort study of postmenopausal women. *Am J Epidemiol.* 1996;144:175-82.

Curriculum Vitae

Yen-Hong Kuo

EDUCATION

1988	Bachelor of Science in Agriculture (B.S.) National Taiwan University College of Agriculture, Taipei, Taiwan
1990	Master of Science in Agriculture (M.S.) National Taiwan University College of Agriculture, Taipei, Taiwan
1996	Master of Science in Biostatistics (Sc.M.) The Johns Hopkins University School of Hygiene and Public Health, Baltimore, MD
2010	Doctor of Philosophy in Biostatistics (Ph.D.) University of Medicine and Dentistry of New Jersey School of Public Health, Piscataway, NJ.

PROFESSIONAL POSITIONS

1995-1997	Statistical Programmer/Research Assistant, Johns Hopkins University, Baltimore, MD
1996-1997	Research Statistician, Johns Hopkins University, Baltimore, MD
1997-	Biostatistician, Jersey Shore University Medical Center, Neptune, NJ

OTHER PROFESSIONAL EXPERIENCES

1999-	Adjunct Instructor, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, Piscataway, NJ
2001-2008	Member, Institutional Review Board, Jersey Shore University Medical Center, Neptune, NJ
2003-2008	Associate Chairman, Institutional Review Board, Jersey Shore University Medical Center, Neptune, NJ
2007-	Adjunct Instructor, School of Public Health, University of Medicine and Dentistry of New Jersey, Piscataway, NJ

PUBLICATIONS

1. Kuo YH. Regulation of alpha-Amylase Gene Expression by Sucrose Starvation in the Suspension-cultured Rice Cells. National Taiwan University, Master Thesis. 1990.

2. Kuo YH, Yu SM, Sheu G, Sheu YJ, Liu LF Metabolic Derepression of alpha-Amylase Gene Expression in Suspension-cultured Cells of Rice. The Journal of Biological Chemistry. 1991;266:21131-21137.

3. Yu SM, Tzou WS, Lo WS, Kuo YH, Lee HT, Wu R. Regulation of alpha-Amylase-encoding Gene Expression in Germinating Seeds and Cultured Cells of Rice. Gene. 1992;122:247-253.

4. Kuo YH. Duration Estimation from Censored Data: Application to the Preantibody Period of HIV Infection. The Johns Hopkins University, Master Thesis. 1996.

5. Chun, TW, Carruth L, Finzi D, Shen X, DiGiuseppe JA, Taylor H, Hermankova M, Chadwick,K, Margolick J, Quinn TC, Kuo YH, Brookmeyer R, Zeiger MA, Barditch-Crovo P, Siliciano RF. Quantification of Latent Tissue Reservoirs and Total Body Viral Load in HIV-1 Infection. Nature. 1997;387:183-188.

6. Kuo YH, Hamer RM Fetal Amino Acid and Enzyme Levels with Maternal Smoking. (Letter) Obstetrics & Gynecology. 1999;94:480-481.

7. Fried MD, Abel M, Pietrucha D, Kuo YH, Bal A. The Spectrum of Gastrointestinal Manifestations in Children and Adolescents with Lyme Disease. Journal of Spirochetal and Tick-borne Diseases. 1999;6:89-93.

8. Kuo YH. Estrogen-Receptor Status in Breast Cancer. (Letter) The Journal of the American Medical Association. 2000;283:338.

9. Mahoney T, Kuo YH, Topilow A, Davis JM Stage Ill Colon Cancers: Why Isn't Adjuvant Chemotherapy Offered to the Elderly? Archives of Surgery. 2000;135:182-185.

10. Thompson J, Canterino JC, Feld, SM, Stumpf PG, Kuo YH, Harrigan JT Risk Factors for Domestic Violence in Pregnant Women. Primary Care Update for OB/GYN. 2000;7:138-141.

11. Kuo YH, Kuo YL. Viral-load Kinetics and CMV Disease. (Correspondence) Lancet. 2000;356:1352-1353.

12. Kuo YH. Analysis of Hospital Length of Stay in Surgical Research. American Statistical Association 2000 Proceedings of the Section on Statistics in Epidemiology and the Health Policy Statistics Section. 2000;126-128.

13. Kuo YH, Kuo YL. Impact of Mortality Rate on the Assessment of Hospital Length of Stay in Surgical Research. 2001 Proceedings of the Annual Meeting of the American Statistical Association,[CD-ROM], Alexandria, VA: American Statistical Association.

14. Shua-Haim JR, Haim T, Shi Y, Kuo YH, Smith JM. Depression among Alzheimer's Caregivers: Identifying Risk Factors. American Journal of Alzheimer's Diseases and Other Dementias. 2001;16:353-359.

15. Kuo YH. Extrapolation of Correlation between 2 Variables in 4 General Medical Journals. The Journal of the American Medical Association. 2002;287:2815-2817.

16. Kuo YH. Arterial-wall Thickness and Impairment in ABCA1-driven Cholesterol Efflux. (Correspondence) Lancet. 2002;359:2278-2279.

17. Mathew P, Kuo YH, Vazirani B, Eng RHK, Weinstein MP. Are Three Sputum Acid-Fast Bacillus Smears Necessary for Discontinuing Tuberculosis Isolation? Journal of Clinical Microbiology. 2002;40:3482-3484.

18. Kuo YH. Use of Independent Assumption in the "Matched" Study of Obstetrics and Gynecology. 2002 Proceedings of the Annual Meeting of the American Statistical Association, Biometrics Section [CD-ROM], Alexandria, VA: American Statistical Association:1975-1977.

19. Kuo YH. Reappraisal of Neonatal Clavicular Fracture: Relationship Between Infant Size and Neonatal Mortality. (Letter) Obstetrics & Gynecology. 2003;101:202.

20. Kuo YH. Apoptosis and Renal Function in Patients with Acute Respiratory Distress Syndrome. (Letter) The Journal of the American Medical Association. 2003;290:461.

21. Kuo YH, Kuo YL. Impact of Measurement Accuracy on the Pain Assessment by Using the Visual Analog Scale: A Simulation Study. 2003 Proceedings of the Annual Meeting of the American Statistical Association, Biometrics Section [CD-ROM], Alexandria, VA: American Statistical Association: 2283-2285.

22. Kuo YH, Hill P, Warner D, Davis JM. Problems and Solutions on Illustrating Risks in the Informed Consent for Multicenter Phase III Clinical Trials. (Abstract) Clinical Trials. 2004;2:221-222.

23. Kuo YH. Impact of the Coexistence of a Composite Score and Its Components in a Multiple Logistic Regression Model on Predicting Clinical Outcomes. 2004 Proceedings of the Annual Meeting of the American Statistical Association, Biometrics Section [CD-ROM], Alexandria, VA: American Statistical Association.

24. Kuo YH, Hill P, Warner D, Davis JM. New Threats from an Old Procedure? (Abstract) Clinical Trials. 2005;2:S52

25. Kuo YH, Torres S, Davis JM. Modeling Predictors for Blood Transfusion in Patients with Small and Large Bowel Procedures. 2005 Proceedings of the Annual Meeting of the American Statistical Association, Biometrics Section [CDROM], Alexandria, VA: American Statistical Association: 263-265.

26. Torres S, Kuo YH, Morris K, Neibart R, Holtz JB, Davis JM. Intravenous iron following cardiac surgery does not increase the infection rate. Surg Infect (Larchmt). 2006; 7:361-6.

27. Gilmore, B., Kuo YH, Morris, K., Bliss-Holtz, J., Torres, S., Neibart, R., Davis, JM. The Cost of Blood Conservation Products is Cheaper than Blood. Transfusion. 2007;47:28A.

28. Ahmed N, Kuo YH. Early versus Late Tracheostomy in Patients with severe Traumatic Head Injury. Surg Infect (Larchmt). 2007;8:343-7.

29. Rahal W, Debari J, Kuo YH, Casey K, Davis JM. Is Impaired Immunity a Consequence of Surgery in Patients Infected BY The Human Immunodeficiency Virus? Surg Infect (Larchmt). 2007;8:575-80.

30. Boss, CM, Wolfe C, Kuo YH. Transforming Diabetes Self-Management or Not. Journal of Consumer Health on the Internet. 2008;12:23-36.

31. Ahmed N, Bialowas C, Kuo YH, Zawodniak L. Impact of Preinjury Anticoagulation in Patients with Traumatic Brain Injury. Sourthern Medical Journal. 2009;102:476-80.

32. Shifrin A, Xenachis C, Fay A, Matulewicz T, Kuo YH, Vernick J. 107 family members with the RET V804M proto-oncogene mutation presenting with simultaneous medullary and papillary thyroid carcinomas, rare primary hyperparathyroidism and no pheochromocytomas. Is this a new syndrome - MEN 2C? Surgery (in press)

33. Young Y, Fried LP, Kuo YH. Hip Fractures among Elderly Women: Longitudinal Comparison of Physiological Function Changes and Health Care Utilization. Journal of the American Medical Directors Association. 2010;11:100-105.