SCORE TEST FOR TESTING ETIOLOGIC HETEROGENEITY IN TWO-STAGE POLYTOMOUS LOGISTIC REGRESSION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SAYGIN KARAGÜLLE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

SEPTEMBER 2013

Approval of the thesis:

SCORE TEST FOR TESTING ETIOLOGIC HETEROGENEITY IN TWO-STAGE POLYTOMOUS LOGISTIC REGRESSION

submitted by **SAYGIN KARAGÜLLE** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. İnci Batmaz _____
Head of Department, **Statistics**

Assist. Prof. Dr. Zeynep Kalaylıoğlu _____
Supervisor, **Statistics Dept., METU**

Examining Committee Members:

Prof. Dr. İnci Batmaz
Statistics Dept., METU _____

Assist. Prof. Dr. Zeynep Kalaylıoğlu
Statistics Dept., METU _____

Assoc. Prof. Dr. Özlem İlk
Statistics Dept., METU _____

Assist. Prof. Dr. Ceylan Talu Yozgatılgil
Statistics Dept., METU _____

Assoc. Prof. Dr. Erdem Karabulut
Biostatistics Dept., Hacettepe University _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: SAYGIN KARAGÜLLE

Signature :

ABSTRACT

SCORE TEST FOR TESTING ETIOLOGIC HETEROGENEITY IN TWO-STAGE POLYTOMOUS LOGISTIC REGRESSION

Karagülle, Saygın

M.S., Department of Statistics

Supervisor : Assist. Prof. Dr. Zeynep Kalaylıoğlu

September 2013, 34 pages

Two-stage polytomous logistic regression was proposed by Chatterjee [11] as an effective tool to analyze epidemiological data when disease subtype information is available. In this modeling approach, a classic logistic regression is employed in the first level of the model. In the second level, the first-stage regression parameters are modeled as a function of some contrast parameters in a somehow similar spirit of an ANOVA model. This modeling also enables a practical way of estimating the heterogeneity in the probabilities of occurrence of different subtypes given a certain covariate set. However, the only way of testing for significance of the heterogeneity is the Wald test, so an alternative test has yet to be developed. In this context, the aim is to develop a score test and examine both the asymptotic and finite sample properties of the test. The simulation results showed that a minimum average expected subtype frequency, depending on the number of disease subtypes and total sample size, must be attained for the asymptotic distribution of the score test to hold. For the cases in which it is implausible to make asymptotic distribution assumption, through an extensive Monte Carlo simulation study, use of permutation test-based critical values were suggested.

Keywords: Categorical Response, Poltomous Logistic Regression, Score Test, Two-Stage Regression

ÖZ

İKİ BASAMAKLI ÇOKLU LOJİSTİK REGRESYONDA ETİYOLOJİK HETEROJENLİĞİ TEST ETMEK İÇİN SKOR TEST

Karagülle, Saygın
Yüksek Lisans, İstatistik Bölümü
Tez Yöneticisi : Yrd. Doç. Dr. Zeynep Kalaylıoğlu

Eylül 2013 , 34 sayfa

Hastalık alttür bilgisinin mecvut olduğu epidemiyolojik verilerin analizinde etkili bir araç olan ve Chatterjee [11] tarafından geliştirilen iki basamaklı lojistik regresyon kullanılmaktadır. Bu modelleme yaklaşımının ilk aşamasında, klasik lojistik regresyon modeli kurulmaktadır. İkinci aşamada ise, birinci basamaktaki regresyon katsayıları tek yönlü ANOVA'yı andıran bir tarzda bir takım kontrast parametreleriyle modellenmektedir. Bu model aynı zamanda, ilglenilen hastalığa ait farklı alttürlerin olabilirliklerinin heterojenliğini tahmin eden pratik bir yöntem de sağlamaktadır. Bu heterojenliğin test edilmesi için mevcut yöntem olan Wald teste alternatif başka bir testin geliştirilmesine ihtiyaç vardır. Bu bağlamda amaç, bir skor testi geliştirmek ve bu testin hem asimtotik hem de sonlu örneklem özelliklerini iki basamaklı lojistik regresyon çerçevesinde ortaya koymaktır. Yapılan simülasyon çalışmalarının neticesinde, skor testin asimtotik dağılımına yaklaşması için hastalık alttür sayısına ve toplam örneklem büyüklüğüne bağlı olarak değişen bir beklenen asgari alttür sıklık değerinin sağlanması gerekmektedir. Asimtotik yaklaşımın kullanılamayacağı durumlarda, permütasyon tabanlı bir test ile elde edilecek kritik değerlerin kullanılması tavsiye edilmektedir.

Anahtar Kelimeler: Çoklu Lojistik Regresyon, İki Basamaklı Lojistik Regresyon, Kategorili Yanıt Değişkeni, Skor Testi

To my beloved family...
for their everlasting love and support.

ACKNOWLEDGMENTS

First of all, I would like to express my deepest appreciation to my thesis supervisor Assist. Prof. Zeynep Kalaylıoğlu for the continuous support of my master's study and research, for her patience, motivation, enthusiasm, and immense knowledge. Without her invaluable guidance and persistent help this thesis would not have been possible. I could not have imagined having a better advisor and mentor for my master's thesis study. It has been a great honor for me to be a student of her and work with her.

I would like to present my grateful thanks to my examining committee members: Prof. İnci Batmaz, Assoc. Prof. Özlem İlk, Assist. Prof. Ceylan Talu Yozgatlıgil and Assoc. Prof. Erdem Karabulut for their valuable time to review this thesis and their constructive comments and suggestions.

I would also like to thank the Scientific and Technological Research Council of Turkey (TUBİTAK) for their financial support and the Department of Computer Engineering, Middle East Technical University, for providing the HPC resources.

I owe my special thanks to all of my instructors for their encouragement, consistent guidance and support. Also, special thanks to my friends in the department for their everlasting support and companionship.

Finally, I would like to express my gratitude to my family for unconditional support, understanding and love that they always gave me. Their support encouraged me throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF ABBREVIATIONS

MLE	Maximum Likelihood Estimate
PCL	Pseudo-Conditional-Likelihood

CHAPTER 1

INTRODUCTION

The main focus of this thesis study is multi-level categorical response variables constructed by cross-classification of their characteristics and two-stage regression models used to understand the relationship between independent variables and aforementioned type of response variable. It is possible to come across with such categorical response variables in a variety of different fields. To illustrate, firms applying to banks for loan can be categorized (subtyped) based on their *loan purpose* (debt payment/project financing/shipping) and *desired loan type* (bridge/facility/lease/term) and this results in a nominal response variable of 12 categories describing the feature of desired loan. Similarly, a response variable defining the breast cancer can be constructed based on cancer characteristics such as *tumor size* (small/medium/large) and *nodal status* (yes/no) and this results in a 6 level nominal outcome variable with subtypes: *(small,yes), (small,no), (medium,yes), (medium,no), (large,yes),* and *(large,no).* This way, instead of working with a dichotomous response variable which allows only two options (absence or presence of the breast cancer), one can gain insight into the association between the breast cancer risk factors and the risk of the breast cancer at the disease characteristic level through the cross-classification of the characteristics. The main question of interest in such studies: e.g. what are the effects of certain firm characteristics (such as profitability, credit ranking) on the loan feature? What are the effects of certain breast cancer risk factors (such as number of full term births, family history) on the type of the breast cancer?

The most common modeling strategies for relating this type of response data to the covariates are constructing either (i) separate logistic regressions for each characteristic or (ii) only one large polytomous logistic regression model in which the response variable consists of cross-classification of characteristic levels. For the breast cancer study, for example, the first approach suggests one polytomous logistic regression for tumor size and one binary logistic regression for nodal status. However, these two regression models are treated independent of each other; therefore, it carries the risk of ignoring the relationships that are naturally present among the characteristics. That is, it is unreasonable to assume independence between the characteristics of tumor. Contrary to the first approach, the second approach takes inherent relation between the charac-

teristics into account by suggesting one large polytomous logistic regression in which the response variable has the following levels: *(small,yes)*, *(small,no)*, *(medium,yes)*, *(medium,no)*, *(large,yes)*, and *(large,no)*. However, the major drawback of this approach is that dimension of the parameter space gets larger as the number of subtypes increases. One another drawback may also arise, especially in small-scale studies, due to not having enough observations for some subtypes to estimate the corresponding regression coefficients. Since both existing approaches have serious problems, a two-stage polytomous logistic regression model was developed by Chatterjee [11]. In the first stage, a classical polytomous logistic regression model is employed. In the second stage, the regression coefficients of the first stage model (i.e. βs) are modeled as a function of some constant parameters (i.e. θs) in a somehow similar spirit of an ANOVA model. In the second stage model, each θ represents the effect of a certain covariate on a certain level of a certain characteristic relative to its effect on the reference level of the same characteristic. In order to gain a better insight into the advantages and the usefulness of this two-stage model, let's consider the breast cancer study again. As explained previously, the response variable consists of six subtypes. Therefore, even a polytomous logistic regression model with one covariate results in a parameter space of size 12. However, when the two-stage model is employed, the regression parameters of the first-stage model are expressed as a function of θs so that the number of parameters to be estimated in the new parameter space, consisting of θs, reduces to 10 from 12. It is clear that one of the main advantages of two-stage modeling is that it reduces the dimension of the parameter space. Also, as the number of subtypes increases, this reduction becomes more pronounced. In addition, estimated values of the θs enable a practical way of examining the etiologic heterogeneity in the probabilities of occurrences of different subtypes given a certain covariate set. The concept of etiologic heterogeneity can be defined as the covariate effect between two disease subtypes, which have two different levels for a given characteristic, but share the same level for all the remaining characteristics. That is, for example, it becomes possible to directly calculate the probability of effect of family history on tumor being large relative to tumor being small. What is more, all the hypotheses related with the etiologic heterogeneity can be expressed by just using individual θs instead of functions of βs. For example, the hypothesis that whether the effect of family history is different for different tumor sizes (i.e. tumor being large relative to being small vs. tumor being medium relative to being small) can be expressed by just using the θs. It is obvious that the two-stage modeling outweighs the classical binary and polytomous logistic regression models since making such kind of an inference is impossible unless a two-stage model is employed. The only way of testing the aforementioned type of hypotheses is by using Wald's test, but it has some major drawbacks. Since an alternative to Wald's test does not exist in the literature, it is yet to be developed. The goal of this thesis study is to develop a score test for testing etiologic heterogeneity and examine both the asymptotic and empirical properties of this test.

The rest of this thesis is structured as follows. Chapter 2 starts with a general description of the two-stage model. The way how the first-stage regression parameters are expressed in terms of the second-stage parameters and the estimation procedure of these parameters are presented. Having provided a general motivation for the two-stage model, the type of hypothesis to be tested is illustrated and the score test is developed in this chapter. In Chapter 3, the simulation study conducted to investigate the asymptotic and finite sampling properties of the score test is explained and the results are discussed. Finally, in Chapter 4, the contributions of this thesis study are outlined and discussed.

CHAPTER 2

METHODOLOGY

2.1 Two-Stage Polytomous Logistic Regression Model

2.1.1 Definition of the Model

In statistics, regression analysis is one of the most important tools used to examine relationship between a response variable and a number of predictor variables. The choice of type of the regression model to be fitted is determined according to the scale of the response variable. If it is categorical and in nominal or ordinal scale, the response variable is regressed on predictor variables through a proper link function, such as logit, probit etc. (see Hosmer and Lemeshow [4], and Kleinbaum and Klein [5]). This type of regression models are known as logistic regression models and they are categorized based on the number of categories of the response variable can assume. For a categorical response variable assuming only two categories (i.e. binary or dichotomous), binary logistic regression models can be used. An extended form of binary logistic regression, known as polytmous (multinomial) logistic regression, developed by McFadden [3] and it handles categorical response variables with more than two categories.

Assume a multi-level categorical response variable whose levels are constructed by cross-classification of its K characteristics. If each characteristic k has M_k levels, then $M = M_1 \times M_2 \times \cdots \times M_k$ response categories (subtypes) can be defined for the response variable of interest. In such a study of size N, each subject's response category Y_i can be represented by a value from the set $\{0, 1, 2, ..., M\}$, where the value 0 is used for subjects in control group. If $\mathbf{X}_i$ is the covariate vector for the ith subject, then a polytomous logistic regression model can be written as

$$P(Y_i = m \mid \boldsymbol{X}_i) = \frac{exp(\alpha_m + \boldsymbol{X}_i^T \boldsymbol{\beta}_m)}{1 + \sum_{m=1}^{M} exp(\alpha_m + \boldsymbol{X}_i^T \boldsymbol{\beta}_m)} \quad , \quad m = 1, 2, ..., M \tag{2.1}$$

where α_m and $\boldsymbol{\beta}_m$ represent the intercept parameter and $P \times 1$ vector of regression coefficients corresponding to response category m, respectively. Here, $exp(\boldsymbol{\beta}_m)$ gives the ratio of the odds of being in subtype m versus the reference category, i.e. con-

trol group, for a 1 unit change in one of the covariates while holding the remaining covariates constant.

For a polytomous logistic regression model with characteristics mentioned as above, the number of regression parameters to be estimated are $(M_1 \times M_2 \times \cdots \times M_k \times P)+1$. It is clear that as the number of subtypes increases, the dimension of the parameter space also increases. In addition, estimation problems may easily arise due to subtypes with insufficient number of observations. Therefore, constructing one large polytomous logistic regression model in which the response variable consists of cross-classification of its characteristic's levels has major drawbacks. To overcome these problems, Chatterjee [11] has developed a new model so that the number of parameters to be estimated become fewer in size. For simplicity, let's assume that a single covariate is of interest and there exist M subtypes for the response. Thus, there should be M regression coefficients, $\beta_1, \beta_2, ..., \beta_M$, each associated with each of the M subtypes. Since each certain combination of K characteristics constitutes a certain subtype, any regression coefficient, β_m, can explicitly be represented by its associated characteristic's levels and thus, they can be re-parameterized as a linear function of θs as follows:

$$\beta_m = \{\beta_{i_1 i_2 \ldots i_k}\}_{i_1=1,i_2=1,\ldots,i_k=1}^{M_1,M_2,\ldots,M_k} = \theta^{(0)} + \sum_{k_1=1}^{K} \theta^{(1)}_{k_1(i_{k_1})} + \sum_{k_1=1}^{K} \sum_{k_2>k_1}^{K} \theta^{(2)}_{k_1 k_2 (i_{k_1} i_{k_2})} + \cdots + \theta^{(K)}_{12\ldots K(i_1 i_2 \ldots i_K)} \tag{2.2}$$

Here, $\theta^{(0)}$ is the coefficient specific to the reference response subtype. A reference response subtype always consists of cross-classification of response characteristics so that each characteristic is in its own reference level. The remaining terms of the form $\theta^{(k)}, k = 1, 2, ..., K$, represent kth-order contrasts. Except for $\theta^{(0)}$, the coefficients representing the reference level for each characteristic K must be regarded as zero.

If it is assumed that the second-order and higher contrast are equal to zero, then (2.2) becomes:

$$\beta_m = \{\beta_{i_1 i_2 \ldots i_k}\}_{i_1=1,i_2=1,\ldots,i_k=1}^{M_1,M_2,\ldots,M_k} = \theta^{(0)} + \sum_{k_1=1}^{K} \theta^{(1)}_{k_1(i_{k_1})} \tag{2.3}$$

In (2.3), each $\theta^{(1)}_{k_1(i_{k_1})}$ represents the effect of covariate on the i_{k_1}-th level of the characteristic k_1 relative to its effect on the reference level of the same characteristic. In other words, $\theta^{(1)}_{k_1(i_{k_1})}$ is the log-odds ratio of having i_{k_1}-th level of the characteristic k_1 to the reference level of the same characteristic for a one unit change in the covariate.

As for being an illustrative example for this re-parameterization procedure, let's recall the example on breast cancer given in Chapter 1. The characteristics defining the breast cancer are *tumor size* (small/medium/large) and *nodal status* (yes/no). As explained previously, 6 response categories can be constructed by cross-classifying the levels of tumor size and nodal status. Let tumor size being small and nodal status being no be the reference levels for each characteristic. If the reference level for tumor size (i.e tumor being small) is coded as 1, then tumor being medium and being large can be coded as 2 and 3, respectively. Similarly, if the reference level for nodal status being no is coded as 1, then the other level, nodal status being yes, can be coded as 2. After these coding scheme, following first-stage regression coefficients are obtained: $\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}$, and they can be re-parameterized in terms of θs as shown in Table- 2.1.

Table 2.1: Re-parameterization of the first-stage parameters

m	β	**Tumor Size**	**Nodal Status**	**Re-parameterization**
1	$\beta_1 = \beta_{00}$	small(=1)	no(=1)	$\theta^{(0)} + \theta^{(1)}_{1(1)} + \theta^{(1)}_{2(1)}$
2	$\beta_2 = \beta_{01}$	small(=1)	yes(=2)	$\theta^{(0)} + \theta^{(1)}_{1(1)} + \theta^{(1)}_{2(2)}$
3	$\beta_3 = \beta_{10}$	medium(=2)	no(=1)	$\theta^{(0)} + \theta^{(1)}_{1(2)} + \theta^{(1)}_{2(1)}$
4	$\beta_4 = \beta_{11}$	medium(=2)	yes(=2)	$\theta^{(0)} + \theta^{(1)}_{1(2)} + \theta^{(1)}_{2(2)}$
5	$\beta_5 = \beta_{20}$	large(=3)	no(=1)	$\theta^{(0)} + \theta^{(1)}_{1(3)} + \theta^{(1)}_{2(1)}$
6	$\beta_6 = \beta_{21}$	large(=3)	yes(=2)	$\theta^{(0)} + \theta^{(1)}_{1(3)} + \theta^{(1)}_{2(2)}$

In this re-parameterization, $\theta^{(0)}$ is the coefficient specific to the reference disease subtype, i.e. tumor being small and nodal status being yes. The coefficients $\theta^{(1)}_{1(1)}, \theta^{(1)}_{1(2)}, \theta^{(1)}_{1(3)}$ are specific to tumor size being small, being medium, and being large respectively. Since each second-stage coefficient θ shows the effect of a certain covariate on a specific level of a characteristic relative to its effect on the reference level of the same characteristic, the θ coefficients associated with the reference level of each characteristic needs to be set at zero, except for $\theta^{(0)}$. Here for the characteristic tumor size, $\theta^{(1)}_{1(2)}$ represents the effect of a certain covariate, e.g. smoking history, on tumor being medium relative to its effect on tumor being small. Similarly, $\theta^{(1)}_{1(3)}$ represents the effect of a certain covariate, e.g. smoking history, on tumor being large relative to its effect on tumor being small. However, $\theta^{(1)}_{1(1)}$ needs to be set at zero. In the same manner, nodal status being no and being yes are represented by the coefficients $\theta^{(1)}_{2(1)}$ and $\theta^{(1)}_{2(2)}$, respectively. $\theta^{(1)}_{2(2)}$ represents the effect of a certain covariate, e.g. smoking history, on nodal status being yes relative to its effect on nodal status being no. However, $\theta^{(1)}_{2(1)}$ need to be set at zero as it is for $\theta^{(1)}_{1(1)}$. Thus, ignoring the second-stage parameters which are set at zero, this re-parameterization can also be represented by means of transformation matrix $\boldsymbol{Z}$ such that $\boldsymbol{\beta} = \boldsymbol{Z\theta}$ as shown in (2.4).

$$\boldsymbol{\beta} = \boldsymbol{Z} \times \boldsymbol{\theta} \Rightarrow \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} \theta^{(0)} \\ \theta^{(1)}_{1(2)} \\ \theta^{(1)}_{1(3)} \\ \theta^{(1)}_{2(2)} \end{bmatrix} \tag{2.4}$$

2.1.2 Estimation of the Parameters

For a polytomous logistic regression model as shown in (2.1), there exist two types of parameters to be estimated. They are namely the intercept, α_m, and regression, β_m, parameters. The regression parameters are crucial for the usual odds ratio interpretation when comparing a certain response subtype with the reference subtype, and thus they are of value in terms of scientific point of view. However, since the intercept parameters provide information only about the baseline likelihood of occurrence of different subtypes, they are not of scientific interest, and thus they can be regarded as nuisance parameters. As mentioned previously, the dimension of parameter space gets large in the existence of large number of response subtypes. In such a case, maximum likelihood estimation method remains incapable of handling these huge number of unknown parameters, a great part of which is constituted by the intercept parameters. At this stage, a semi-parametric approach which focuses only on the regression parameters and ignores the nuisance intercept parameters would be needed. Chatterjee [11] proposed an approach for this purpose and developed a method known as "pseudo-conditional-likelihood" (PCL). What makes PCL appealing is that it is only a function of regression parameters and does not contain any intercept parameters.

To gain a better insight into the method of PCL, consider a study in which subjects are divided into two sets C_1 and C_0 such that C_1 consists of cases and C_0 consists of controls. For each case in the set C_1, a matched set S_i can be defined such that it involves the i-th case and all the controls. Thus, for each set S_i, given that only one subject is chosen with a certain subtype from the set of cases and all the remaining subjects are chosen from the set of controls, the conditional probability of the observed configuration of the subjects of S_i can be defined as

$$L_i^c = Pr[Y_i = y_i, Y_j = 0; j \epsilon S_i, j \neq i \mid \bigcup_{k \epsilon S_i} \{Y_k = Y_i, Y_l = 0; l \epsilon S_i, l \neq k\}] \tag{2.5}$$

The expression L_i can be derived from the polytomous logistic regression model formula given in (2.1) and it does not contain any intercept parameters. Thus, an expression for the PCL of the data, which is free of intercept parameter α_{y_i}, can be obtained as given in (2.6)

$$L_{PCL} = \prod_{i \epsilon C_1} L_i^c = \prod_{i \epsilon C_1} \frac{exp(\boldsymbol{X}_i^T \boldsymbol{\beta}_{y_i})}{exp(\boldsymbol{X}_i^T \boldsymbol{\beta}_{y_i}) + \sum_{j \epsilon C_0} exp(\boldsymbol{X}_i^T \boldsymbol{\beta}_{y_i})} \ , \tag{2.6}$$

Since the regression parameters of the first-stage model are re-parameterized by the second-stage parameters θ, the PCL score equations can be defined as

$$\frac{\partial L_{PCL}}{\partial \boldsymbol{\beta}} \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}} = 0 \ . \tag{2.7}$$

The fact that $\boldsymbol{\beta} = \boldsymbol{Z\theta}$ provides a way to express the score equations as

$$\boldsymbol{Z}^T T_\beta = 0 \ , \tag{2.8}$$

where $T_\beta = (T_{\beta_1}^T, ..., T_{\beta_m}^T)^T$ and
$T_{\beta_m} = \sum_{i \epsilon C_1} I(Y_i = m) \times \left\{ \boldsymbol{X}_i - \frac{\boldsymbol{X}_i exp(\boldsymbol{X}_i^T \boldsymbol{\beta}_m) + \sum_{j \epsilon C_0} \boldsymbol{X}_j exp(\boldsymbol{X}_j^T \boldsymbol{\beta_m})}{exp(\boldsymbol{X}_i^T \boldsymbol{\beta}_m) + \sum_{j \epsilon C_0} exp(\boldsymbol{X}_j^T \boldsymbol{\beta}_m)} \right\}$

Simultaneous solution of the score equations given in (2.8) yields the vector of maximum likelihood estimates of second stage parameters, $\hat{\theta}^{(k)}$. These PCL estimators are asymptotically normal (see Chatterjee [11]).

2.2 Hypothesis of Interest

As illustrated in Chapter 1, it is possible to come across with multi-level response variables in epidemiological studies. In such studies, if it is possible to categorize the disease of interest through its defining characteristics, then data collection can be carried out by classifying the subjects under study according to the disease subtype to which they belong. Such data can be studied to examine the relationship between a disease and a group of risk factor through disease characteristic level. It also provides the opportunity of examining etiologic heterogeneity among disease subtypes (see Garcia-Closas et al. [9], Sherman et al. [8], and Erdem [10]). Dealing with the concept of etiologic heterogeneity in a epidemiological study is simply looking for an answer for the question: Is any disease subtype more likely to be associated with the effect of a certain covariate? (Chatterjee [11]). It is clear that determining the etiologically heterogeneous disease subtype is of scientific value when a number of disease characteristics are taken into consideration in a statistical analysis. However, such an information cannot be obtained with the use of standard polytomous logistic regression. Therefore, when the the problem of handling a parameter space of high dimension is taken into account as well, the need for a novel approach became increasingly apparent. Thus, the two-stage regression model, where the second-stage parameters provide a measure of degree of etiologic heterogeneity, was proposed by Chatterjee [11].

For a two-stage polytomous logistic regression model, the linear representation of the first-stage regression parameters in terms of the second-stage parameters was given in (2.2). Assuming that all the second-order and higher contrasts are zero, then the linear model given in (2.3) can be obtained. For this linear representation, consider two first-stage regression parameters of the form $\beta_{i_1,..,i_k,..,i_K}$ and $\beta_{i_1,..,i_k^{'},..,i_K}$. It is clear that only for the characteristic-k, the levels are i_k and $i_k^{'}$ but the same level of the remaining characteristics is shared for both. When the difference is taken between these two first-stage regression parameters, the expression given in (2.9) is obtained.

$$\beta_{i_1,..,i_k,..,i_K} - \beta_{i_1,..,i_k^{'},..,i_K} = \theta^{(1)}_{k(i_k)} - \theta^{(1)}_{k(i_k^{'})} \tag{2.9}$$

The difference between the terms $\theta^{(1)}_{k(i_k)} - \theta^{(1)}_{k(i_k^{'})}$ provide a measure of degree of etiologic heterogeneity associated with the levels of the kth characteristic (Chatterjee [11]). It is also important to underline that a second-stage model as given in Equation (2.3) is valid under the assumption that etiologic heterogeneity associated with one characteristic is independent of the other characteristics.

For data with disease subtype information, the hypothesis of interest is that there exists etiologic heterogeneity between the levels of the kth characteristic. Therefore, the null hypothesis deals with the absence of etiologic heterogeneity. However, the form of the alternative hypothesis is shaped according to the number of levels of the

characteristic of interest. To illustrate, let's consider characteristics differing in the number of levels.

For a characteristic with 2 levels, let $\theta^{(1)}_{k(1)}$ and $\theta^{(1)}_{k(2)}$ be its corresponding second-stage coefficients, respectively. If $\theta^{(1)}_{k(1)}$ is the coefficient associated with the reference level, then the null and alternative hypotheses can be written as given in (2.10).

$$H_0 : \theta^{(1)}_{k(2)} = 0 \quad \text{vs.} \quad H_1 : \theta^{(1)}_{k(2)} \neq 0 \ . \tag{2.10}$$

Note that the the coefficient $\theta^{(1)}_{k(1)}$ is associated with the reference level, so it needs to be set to 0.

Similarly, for a characteristic with 3 levels, if $\theta^{(1)}_{k(1)}$, $\theta^{(1)}_{k(2)}$ and $\theta^{(1)}_{k(3)}$ are the corresponding second-stage coefficients, then the null and alternative hypotheses can be expressed as given in (2.11).

$$H_0 : \theta^{(1)}_{k(2)} = \theta^{(1)}_{k(3)} \quad \text{vs.} \quad H_1 : \theta^{(1)}_{k(2)} \neq \theta^{(1)}_{k(3)} \ . \tag{2.11}$$

Thus, for a characteristic with m_k levels, assuming that $\theta^{(1)}_{k(1)}$, $\theta^{(1)}_{k(2)}$, ... , $\theta^{(1)}_{k(k)}$ are the corresponding second-stage parameters, then the associated null and alternative hypotheses can be defined as given in (2.12).

$$\begin{aligned} &H_0 : \theta^{(1)}_{k(2)} = \theta^{(1)}_{k(3)} = \cdots = \theta^{(1)}_{k(m_k)} \quad \text{vs.} \\ &H_1 : \theta^{(1)}_{k(i)} \neq \theta^{(1)}_{k(j)} \text{ for at least one } i, j \epsilon \{1, 2, ..., m_k\}, i \neq j \ , \end{aligned} \tag{2.12}$$

To provide an answer for the question: What do these hypotheses mean?, let's revisit the example on breast cancer. The first characteristic, tumor size, has 3 levels. Therefore, a test for etiologic heterogeneity between the levels of tumor size should be constructed as given (2.11). This type of hypothesis means that for a certain covariate, say smoking status, there is a difference between the effect of smoking status on tumor being medium relative to tumor being small and that on tumor being large relative to tumor being small. In the same manner, since the other characteristic, nodal status, consists of 2 levels, then its associated hypothesis of etiologic heterogeneity means that there is an effect of smoking status on nodal status being yes relative to nodal status being no. The form of the hypothesis should be as given in (2.10).

2.3 Score Test for Testing Etiologic Heterogeneity

A general structure of a hypothesis associated with etiologic heterogeneity for levels of a certain characteristic-k with m_k levels was presented in Section 2.2. It is important to notice that when testing for etiologic heterogeneity, the focus is only on the certain components of the parameter vector $\boldsymbol{\theta}$. That is to say, except for the second-stage parameters corresponding to characteristic of interest, those corresponding to other characteristics under consideration are left unspecified in the null hypothesis. Therefore, every null hypothesis of absence of etiologic heterogeneity is always composite.

In literature, Wald test is known as the most usual and common way of testing such hypotheses. It is a likelihood-based test and its asymptotic distribution is chi-square when the necessary regularity conditions are satisfied. In order to introduce the use of the test, let's start with assuming a characteristic with 2 levels. In this case, its corresponding hypothesis test of etiologic heterogeneity is assumed to be of the form as given in (2.10). Since the interest is only on the second-stage parameter $\theta^{(1)}_{k(2)}$, the parameter vector can be partitioned as $\boldsymbol{\theta}^T = (\theta^{(1)}_{k(2)}, \boldsymbol{\eta}^T)$, where the nuisance parameters are represented by $\boldsymbol{\eta}^T$. Assuming that $\hat{\boldsymbol{\theta}}$ denote the PCL estimates of the second-stage parameters, the information matrix $\hat{\boldsymbol{I}}_T(\hat{\boldsymbol{\theta}})$ can be partitioned as

$$\hat{\boldsymbol{I}}_T(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \hat{\boldsymbol{I}}_{T,11} \hat{\boldsymbol{I}}_{T,12} \\ \hat{\boldsymbol{I}}_{T,21} \hat{\boldsymbol{I}}_{T,22} \end{pmatrix} \quad ; \tag{2.13}$$

then the Wald test statistic T_w becomes

$$T_w = (\hat{\theta}^{(1)}_{k(2)})^2 (\hat{\boldsymbol{I}}_{T,11} - \hat{\boldsymbol{I}}_{T,12} \hat{\boldsymbol{I}}^{-1}_{T,22} \hat{\boldsymbol{I}}_{T,21}) \ , \tag{2.14}$$

and $T_w \xrightarrow{d} \chi^2_{(1)}$ as $n \to \infty$.

However, the partitioned-vector approach may not come in handy in most cases. For example, consider a characteristic with 3 levels and its corresponding null hypothesis for the test of etiologic heterogeneity $H_0 : \theta^{(1)}_{k(2)} = \theta^{(1)}_{k(3)}$. In such a case, the partitioned-vector approach necessitates a re-parameterization such that $H_0 : \theta^{(1)}_{k(2)} - \theta^{(1)}_{k(3)} = 0$ but the Wald statistic is not invariant to any kind of re-parameterization (Boos and Stefanski [2]). This is another way of saying that different ways of expressing the same question of interest produce different Wald test statistics. It is clear that when testing for the etiologic heterogeneity, it is more likely to encounter characteristics with more than two levels. Therefore, the inadequacy of the Wald test statistic is beyond doubt.

Instead of using partitioned-vector approach, another way of expressing such null hypotheses may be as $H_0 : \boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$, for some function $\boldsymbol{h}$. Here, $\boldsymbol{h}(.)$ is a vector function with matrix of first partial derivatives $\boldsymbol{H}(\boldsymbol{\theta}) = \partial \boldsymbol{h}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$. To illus-

trate this approach, let's consider again a characteristic with 3 levels and its associated hypothesis $H_0 : \theta^{(1)}_{k(2)} = \theta^{(1)}_{k(3)}$. In this case, the parameter vector becomes $\boldsymbol{\theta} = (\theta^{(1)}_{k(2)}, \theta^{(1)}_{k(3)}, \boldsymbol{\eta})^T$, where the second-stage parameters corresponding to remaining characteristics are represented by $\boldsymbol{\eta}$. One other way of re-expressing the hypothesis may be $H_0 : \theta^{(1)}_{k(2)} - \theta^{(1)}_{k(3)} = 0$ and this yields $h(\boldsymbol{\theta}) = \theta^{(1)}_{k(2)} - \theta^{(1)}_{k(3)}$ with $\boldsymbol{H(\theta)} = (1, -1, \mathbf{0})$, where the dimension of the zero vector is equal to that of $\boldsymbol{\eta}$. Based on the specification of the null hypothesis as $\boldsymbol{h(\theta) = 0}$, the following form of the Wald test statistic is obtained (Boos and Stefanski [2]).

$$T_w = \boldsymbol{h}(\hat{\boldsymbol{\theta}})^T [\boldsymbol{H}(\hat{\boldsymbol{\theta}}) \hat{\boldsymbol{I}}_T^{-1}(\hat{\boldsymbol{\theta}}) \boldsymbol{H}(\hat{\boldsymbol{\theta}})^T]^{-1} \boldsymbol{h}(\hat{\boldsymbol{\theta}}) \tag{2.15}$$

Even if the null hypothesis is specified as in the form of $\boldsymbol{h(\theta) = 0}$, the use of Wald statistic is still inconvenient. This is due to the lack of invariance of the Wald test to the choice of $\boldsymbol{h(.)}$. It is apparent that the null hypothesis may also be specified as $H_0 : \theta^{(1)}_{k(3)} - \theta^{(1)}_{k(2)} = 0$. Since this way of specification yields a different $\boldsymbol{H(\hat{\theta})}$ of the form $(-1, 1, \mathbf{0})$, thus obtaining a different Wald test statistic is inevitable.

Although the Wald statistic is easy to apply and it is very common in most of the statistical computing packages, its use in testing for etiologic heterogeneity is open to doubt. Since an alternative method for testing etiologic heterogeneity does not exist in the literature, as the major contribution, a score test is developed in this thesis study.

When considered within the scope of testing for etiologic heterogeneity, in addition to the Wald test, two other ways of testing approaches may come into the picture, namely score test and likelihood ratio test. These three statistics are likelihood based and they are asymptotically equivalent (Engle [12]). However, being asymptotically equivalent does not mean that these three test statistics can be used interchangeably (Boos and Stefanski [2]). Each test performs the best in certain situations. As discussed and illustrated previously, being not invariant to re-parameterization makes the use of Wald statistic inconvenient. On the other hand, both the score and likelihood ratio statistics are invariant to re-parameterization. However, since the likelihood ratio statistic involves MLE of parameters under both the null and alternative hypotheses, the computational demand is much more intensive for it compared to the score statistic. What is more than this, the use of likelihood ratio statistic is much more appropriate when a sequence of nested models are of interest provided that the log-likelihood corresponding to each different model has been derived. In addition to all these, finite-sample studies of these three statistics, in which chi-squared critical values are used, reveals that violation of the assumed nominal rate of Type-I error occurs at most with the Wald statistic. The most liberal one among the three is the Wald statistic and it is followed by the likelihood ratio statistic, whereas the score statistic is somewhat conservative. The similar studies also divulge that the rate of convergence to the asymptotic chi-squared distribution is much better for the score statistic when compared with the others. As a result, in the light of these findings, the score test

statistic is believed to be the best choice to test for the etiologic heterogeneity.

The score test statistic was introduced by Rao in 1948 (see Bera and Bilias [1]) and is used to test a simple null hypothesis that whether a parameter vector $\boldsymbol{\eta}$ of dimension r is equal to a pre-specified vector $\boldsymbol{\eta_0}$ or not. For such a null hypothesis, the score test statistic T_s is

$$T_s = \boldsymbol{S(\eta_0)}^T \{\boldsymbol{I}_T(\boldsymbol{\eta_0})\}^{-1} \boldsymbol{S(\eta_0)} \ , \tag{2.16}$$

where, $\boldsymbol{S(.)}$ is the score function, the first derivative of the log-likelihood function with respect to the parameters of interest.

Under the null hypothesis, it can be shown that $E[\boldsymbol{S(\eta_0)}] = \mathbf{0}$ and $Var[\boldsymbol{S(\eta_0)}] = \boldsymbol{I}_T(\boldsymbol{\eta_0})$. Thus, it follows from The Central Limit Theorem that the distribution of $\boldsymbol{S(\eta_0)}$ is asymptotically normal with parameters $\mathbf{0}$ and $\boldsymbol{I}_T(\boldsymbol{\eta_0})$, respectively. This implies that$\{\boldsymbol{I}_T(\boldsymbol{\eta_0})\}^{-1}\boldsymbol{S(\eta_0)}$ converges in distribution to normal distribution with parameters $\mathbf{0}$ and $\mathbb{I}_r$, respectively, where $\mathbb{I}$ denotes the identity matrix. Thus, it can be inferred that the score statistic T_s is asymptotically chi-squared distributed with r degrees of freedom.

In order to develop a score test for testing the presence of etiologic heterogeneity for a certain characteristic-k with m_k levels, first recall the associated null and alternative hypotheses as given below

$$\begin{aligned} &H_0 : \theta^{(1)}_{k(2)} = \theta^{(1)}_{k(3)} = \cdots = \theta^{(1)}_{k(m_k)} \quad \text{vs.} \\ &H_1 : \theta^{(1)}_{k(i)} \neq \theta^{(1)}_{k(j)} \ \text{ for at least one } i, j \epsilon \{1, 2, ..., m_k\}, i \neq j \ . \end{aligned} \tag{2.17}$$

As discussed previously, since the focus in only on a certain characteristic, not the entire parameter vector is involved in the null hypothesis. Therefore, the partitioned-vector approach can be used. The parameter vector $\boldsymbol{\theta}$ is partitioned as $\boldsymbol{\theta} = (\boldsymbol{\zeta}, \boldsymbol{\eta}) = (\theta^{(1)}_{k(2)}, \theta^{(1)}_{k(3)}, \cdots, \theta^{(1)}_{k(m_k)}, \boldsymbol{\eta})$, where η represents the nuisance second-stage parameters for this test. Based on this partitioning, it is then necessary to calculate the PCL estimates of the second-stage parameters under the null hypothesis. This can be carried out by assigning a single unknown parameter instead of all the parameters included in $\boldsymbol{\zeta}$ and then maximizing the pseudo-conditional likelihood with respect to $\boldsymbol{\eta}$ as explained in Section-2.1.2. If $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\zeta}}, \tilde{\boldsymbol{\eta}})$ denotes these PCL estimates, then the score test statistic T_s for testing etiologic heterogeneity is

$$T_s = \boldsymbol{S}(\tilde{\boldsymbol{\theta}})^T \{\boldsymbol{I}_T(\tilde{\boldsymbol{\theta}})\}^{-1} \boldsymbol{S}(\tilde{\boldsymbol{\theta}}) \ , \tag{2.18}$$

and $T_s \xrightarrow{d} \chi^2_{(m_k-2)}$ as $n \to \infty$.

In (2.18), the score function $\boldsymbol{S}(\tilde{\boldsymbol{\theta}})$ is a partitioned score function, and thus it has two components as shown in (2.19)

$$\boldsymbol{S}(\tilde{\boldsymbol{\theta}}) = \begin{pmatrix} \boldsymbol{S}_1(\tilde{\boldsymbol{\theta}}) \\ \boldsymbol{S}_2(\tilde{\boldsymbol{\theta}}) \end{pmatrix} \ . = \begin{pmatrix} \left\{ \frac{\partial}{\partial \tilde{\boldsymbol{\zeta}}} logL_{PCL}(\tilde{\boldsymbol{\theta}}) \right\}^T \\ \left\{ \frac{\partial}{\partial \tilde{\boldsymbol{\eta}}} logL_{PCL}(\tilde{\boldsymbol{\theta}}) \right\}^T \end{pmatrix} \tag{2.19}$$

It is clear that $\tilde{\boldsymbol{\theta}}$ satisfies $\boldsymbol{S}_2(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$. Therefore, the score statistic given in (2.18) can be rewritten as

$$T_s = \left(\boldsymbol{S}_1(\tilde{\boldsymbol{\theta}})^T \mathbf{0} \right) \begin{pmatrix} \tilde{\boldsymbol{I}}_{T,11} \tilde{\boldsymbol{I}}_{T,12} \\ \tilde{\boldsymbol{I}}_{T,21} \tilde{\boldsymbol{I}}_{T,22} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{S}_1(\tilde{\boldsymbol{\theta}}) \\ \mathbf{0} \end{pmatrix} \ , \tag{2.20}$$

which is also equivalent to

$$T_s = \boldsymbol{S}_1(\tilde{\boldsymbol{\theta}})^T (\tilde{\boldsymbol{I}}_{T,11} - \tilde{\boldsymbol{I}}_{T,12} \tilde{\boldsymbol{I}}_{T,22}^{-1} \tilde{\boldsymbol{I}}_{T,21})^{-1} \boldsymbol{S}_1(\tilde{\boldsymbol{\theta}}) \ . \tag{2.21}$$

Apart from the partitioned vector approach, even when the null hypothesis is specified as in the form of $\boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$, the score test statistic can still be used. To illustrate this, let's re-express the null hypothesis given in (2.17) as follows:

$$H_0 : \begin{cases} \theta^{(1)}_{k(2)} - \theta^{(1)}_{k(3)} = 0 \\ \theta^{(1)}_{k(2)} - \theta^{(1)}_{k(4)} = 0 \\ \vdots \\ \theta^{(1)}_{k(2)} - \theta^{(1)}_{k(m_k)} = 0 \end{cases} \ . \tag{2.22}$$

Based on (2.22), the null hypothesis can be specified as in the form of $\boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$ as follows:

$$H_0 : \begin{pmatrix} h_1(\boldsymbol{\theta}) \\ h_2(\boldsymbol{\theta}) \\ \vdots \\ h_{m_{k-1}}(\boldsymbol{\theta}) \end{pmatrix} = \mathbf{0} \ , \tag{2.23}$$

with $\boldsymbol{H}(\boldsymbol{\theta})$ as given in (2.24).

$$\boldsymbol{H}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial\theta^{(1)}_{k(2)}} h_1(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial\theta^{(1)}_{k(m_k)}} h_1(\boldsymbol{\theta}) & \frac{\partial}{\partial\boldsymbol{\eta}} h_1(\boldsymbol{\theta}) \\ \frac{\partial}{\partial\theta^{(1)}_{k(2)}} h_2(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial\theta^{(1)}_{k(m_k)}} h_2(\boldsymbol{\theta}) & \frac{\partial}{\partial\boldsymbol{\eta}} h_2(\boldsymbol{\theta}) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial}{\partial\theta^{(1)}_{k(2)}} h_{m_{k-1}}(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial\theta^{(1)}_{k(m_k)}} h_{m_{k-1}}(\boldsymbol{\theta}) & \frac{\partial}{\partial\boldsymbol{\eta}} h_{m_{k-1}}(\boldsymbol{\theta}) \end{bmatrix} . \tag{2.24}$$

Since $\tilde{\boldsymbol{\theta}}$ maximizes L_{PCL} subject to the constraint $\boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$, then $S(\tilde{\boldsymbol{\theta}}) - \boldsymbol{H}(\tilde{\boldsymbol{\theta}})^T \tilde{\boldsymbol{\lambda}} = 0$ and $\boldsymbol{h}(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$ are satisfied by $\tilde{\boldsymbol{\theta}}$, where $\boldsymbol{\lambda}$ is the data-dependent vector of Lagrange multipliers. As a result, replacing $S(\tilde{\boldsymbol{\theta}})$ by $\boldsymbol{H}(\tilde{\boldsymbol{\theta}})^T \tilde{\boldsymbol{\lambda}}$ in (2.18) yields the following:

$$T_s = \tilde{\boldsymbol{\lambda}}^T \boldsymbol{H}(\tilde{\boldsymbol{\theta}}) \{\boldsymbol{I}_T(\tilde{\boldsymbol{\theta}})\}^{-1} \boldsymbol{H}(\tilde{\boldsymbol{\theta}})^T \tilde{\boldsymbol{\lambda}} \ . \tag{2.25}$$

The above form of the score test is known as the Lagrange multiplier test in econometrics (Engle [12]).

As underlined before, different $\boldsymbol{H}$ matrices can be obtained depending on the choice of $\boldsymbol{h}$. However, unlike the Wald statistic, the score statistic is not affected by the choice of $\boldsymbol{h}$, and hence the resulting statistics are always the same.

CHAPTER 3

SIMULATION STUDY

In this thesis, following Monte Carlo simulation studies were conducted:

i. to compare the finite sampling characteristics of the Wald and score test.

ii. to compare the power of the two tests.

For each purpose listed below, different case-control data sets were simulated under different scenarios such that they differ by the number of characteristic levels and sample size. Also, under each scenario, the disease prevalence was altered by setting different values for the second-stage parameters.

The performances of the two tests were compared under three different scenarios. For the first two scenarios, eight disease subtypes ($2 \times 2 \times 2$) were assumed and samples of size 500 and 1000 were generated. These two scenarios were considered to test for $H_0 : \theta^{(1)}_{2(2)} = 0$, where $\theta^{(1)}_{2(2)}$ is the degree of etiologic heterogeneity with respect to the second characteristic. For the third scenario, 16 disease subtypes ($4 \times 2 \times 2$) were considered and samples of size 1000 were generated. This last scenario was considered to test for the etiologic heterogeneity with respect to the first characteristic, where the null hypothesis of interest is $H_0 : \theta^{(1)}_{1(2)} = \theta^{(1)}_{1(3)} = \theta^{(1)}_{1(4)}$. Only one covariate from standard normal distribution was considered for each scenario. All the simulation experiments were run in MATLAB R2012A and the original codes of Chatterjee [11] were modified according to each purpose to be investigated.

3.1 Comparison of Finite Sampling Characteristics of the Wald and Score Tests

3.1.1 Data Generation Process

For a detailed illustration of the data generation process, let's consider the first two scenarios in which there are $M = 2 \times 2 \times 2 = 8$ disease subtypes. The data generation

process follows the same strategy for the third scenario as well, where the number of disease subtypes is $M = 4 \times 2 \times 2 = 16$ in that case.

Initially, the second-order and higher contrasts were set to zero and the coefficients corresponding to reference disease subtype (θ^0) and first-order contrasts ($\theta^{(1)}_{1(2)}, \theta^{(1)}_{2(2)}, \theta^{(1)}_{3(2)}$) were determined. Note that since the null hypothesis of interest was defined as $H_0 : \theta^{(1)}_{2(2)} = 0$, in this case, $\theta^{(1)}_{2(2)}$ was set to zero to generate the data under the null hypothesis. Having set the values for the second-stage parameters, using the additive model given in (2.3), the first-stage regression parameters $(\beta_1, ..., \beta_8)$ were obtained. To obtain the intercept parameters $(\alpha_1, ..., \alpha_8)$, the additive model given in (2.3) was extended to consider the second-order interaction terms as given below

$$\theta^{(0)} + \sum_{k_1=1}^{K} \theta^{(1)}_{k_1(i_{k_1})} + \sum_{k_1=1}^{K} \sum_{k_2>k_1}^{K} \theta^{(2)}_{k_1 k_2 (i_{k_1} i_{k_2})} \tag{3.1}$$

and the predetermined second-stage parameters were replaced by their corresponding values.

Using the first-stage model parameters and a covariate vector of size $N \times 1$ from standard normal distribution, the probability of being in the disease category-m , p_{mi}, and that of being disease-free, p_{0i}, for the i-th of N subjects were obtained as shown below

$$p_{mi} = P(Y_i = m \mid \boldsymbol{X}_i) = \frac{exp(\alpha_m + \boldsymbol{X}_i^T \boldsymbol{\beta}_m)}{1 + \sum_{m=1}^{M} exp(\alpha_m + \boldsymbol{X}_i^T \boldsymbol{\beta}_m)} \ , \tag{3.2}$$

and

$$p_{0i} = P(Y_i = 0 \mid \boldsymbol{X}_i) = \frac{1}{1 + \sum_{m=1}^{M} exp(\alpha_m + \boldsymbol{X}_i^T \boldsymbol{\beta}_m)} \ , \tag{3.3}$$

where, $m = 1, 2, ..., 8$ and $i = 1, 2, ..., N$.

Then, for each of the N subjects, a disease subtype status, including being disease-free, was randomly generated assuming a multinomial distribution with probabilities p_{mi} and p_{0i}. Thus, the generated data set of size N consists of a response variable Y_i showing the disease status of a particular subject, where $Y_i = 0, 1, 2, ..., 8$ and her/his covariate value X_i, where $X_i \sim N(0, 1)$ and $i = 1, 2, ..., N$. At the final stage, a random sample of size n was selected from the previously generated sample of size N in such a way that the diseased (case) and disease-free (control) subjects are equal in number. That is, $n = n_{case} + n_{control}$, where $n_{case} = n_{control}$. Since samples of size $n = 500$ and $n = 1000$ were studied under the first two scenarios, n_{case} was set as 250 and 500, respectively. It is crucial to highlight that the sampling scheme is based on a two-stage approach. That is to say, a random sample of size N is generated in the first stage, and then equal number of cases and controls are sampled in the second stage. With this way of sampling, N subjects are assumed to constitute a population in which a certain

percentage of them are diseased, and then a random sample of size n is sampled from this population. In addition, especially for diseases with low frequency of occurrence, the two-stage sampling gives a chance to cover sufficient number of diseased subjects in a sample of size n. For each scenario in this simulation study, N was set as 7000.

3.1.2 Calculation of Empirical Type-I Error Rate and Minimum Average Expected Subtype Frequency

Under each scenario, different second-stage parameters were set and 20000 Monte Carlo replications were conducted for each set of second-stage parameters. In each Monte Carlo replication, $N = 7000$ random samples were generated and equal number of cases and controls among them are randomly selected for a case-control sample as it is illustrated in the previous section. For each case-control sample, both Wald and score test statistics were computed and compared against the chi-squared critical value at nominal $\alpha = 0.05$. Then, empirical Type-I error rate for each statistic was obtained by calculating the proportion of values exceeding the chi-squared critical value.

In addition to empirical Type-I error rate, average expected frequency of each subtype was also calculated to examine and determine the minimum average expected subtype frequency required for the asymptotic distribution of the score test to hold. 20000 data sets were generated and for each generated data set, the probability of having disease subtype-m, where $m = 1, 2, ..., M$, and that of being disease-free were estimated from (3.2) and (3.3), respectively for each subject. Let $\widetilde{p}_{0i}^{(j)}, \widetilde{p}_{1i}^{(j)}, \widetilde{p}_{2i}^{(j)}, ..., \widetilde{p}_{mi}^{(j)}$ denote these estimates for the i-th subject in the j-th simulated data, where $i = 1, 2, ..., 7000$ and $j = 1, 2, ..., 20000$. Then, for each simulated data, the probabilities associated with each subtype were averaged to obtain an average probability of being in a certain subtype. For the j-th simulated data, for example, the average probability of being in subtype-k was calculated as $\overline{p}_{k\cdot}^{(j)} = \sum_{i=1}^{7000} \widetilde{p}_{ki}^{(j)}$, where $k = 0, 1, ..., m$. At the end, these average probabilities were also averaged for each subtype as $\overline{\overline{p}}_{k\cdot} = \sum_{j=1}^{20000} \overline{p}_{k\cdot}^{(j)}$ and the resulting values were multiplied by the total sample size n as $n \times \overline{\overline{p}}_{k\cdot}$ to obtain the average expected subtype frequencies.

For both the empirical Type-I error and minimum average expected subtype frequency calculations, different number of replications were also studied but 20000 replications were observed enough to provide stability in the resulting values.

3.1.3 Simulation Results

In this section, Monte Carlo simulation results for comparing empirical Type-I error rates of Wald and score tests when testing for etiologic heterogeneity with respect to minimum average expected subtype frequency are presented.

Tables 3.1 and 3.2 summarize the simulation results for 8 disease subtypes ($2 \times 2 \times 2$) for a random sample of size 500 and 1000, respectively. For both cases, the null hypothesis of interest is $H_0 : \theta^{(1)}_{2(2)} = 0$ and the test statistics were compared against the $\chi^2_{(1)}$ based critical value at nominal $\alpha = 0.05$. An overall conclusion that can be drawn from Table 3.1 is that neither of the tests are satisfactory enough in maintaining the nominal significance level when chi-square based critical value is used.

Table 3.1: Empirical Type I error rates (χ^2 based approach). Disease subtypes=$2 \times 2 \times 2$; $n = 500$

Minimum Average Expected Subtype Frequency	**Wald Test**	**Score Test**
25	0.0523	0.0566
24	0.0512	0.0554
22	0.0552	0.0582
15	0.0510	0.0566
13	0.0517	0.0570
12	0.0527	0.0609
11	0.0539	0.0612
10	0.0535	0.0601

Table 3.2 shows the case when the sample is increased from 500 to 1000. If the minimum size is at least 49 or more, the empirical Type-I error rate of the Wald test gets closer to the nominal level. Being not as satisfactory as for the Wald test, somehow similar conclusion seems to be valid for the score test. Nevertheless, when the minimum size gets smaller than 30, the empirical significance level is more likely to move away from the nominal level for both of the tests.

Table 3.2: Empirical Type I error rates (χ^2 based approach). Disease subtypes=$2 \times 2 \times 2$; $n = 1000$

Minimum Average Expected Subtype Frequency	**Wald Test**	**Score Test**
50	0.0510	0.0597
49	0.0560	0.0638
45	0.0610	0.0592
30	0.0570	0.0625
27	0.0660	0.0611
24	0.0630	0.0691
23	0.0670	0.0652
21	0.0690	0.0673

Table 3.3 summarizes the simulation results for 16 disease subtypes ($4 \times 2 \times 2$) for a random sample of size 1000. In this case, the null hypothesis of interest becomes $H_0 : \theta^{(1)}_{1(2)} = \theta^{(1)}_{1(3)} = \theta^{(1)}_{1(4)}$ and the test statistics were compared against the $\chi^2_{(2)}$ based critical value at nominal $\alpha = 0.05$.

Table 3.3: Empirical Type I error rates (χ^2 based approach). Disease subtypes=$4 \times 2 \times 2$; $n = 1000$

Minimum Average Expected Subtype Frequency	Wald Test	Score Test
24	0.0678	0.0683
23	0.0649	0.0655
22	0.0690	0.0690
14	0.0708	0.0711
13	0.0727	0.0730
12	0.0791	0.0795
11	0.0817	0.0820

According to Table 3.3, even when the minimum subtype size is 24, the empirical Type-I error rate is significantly higher than the nominal level for both the Wald and score tests. It is also obvious to conclude that the smaller the minimum size, the higher the empirical Type-I error rates.

Tables 3.1, 3.2 and 3.3 reveal that both the Wald and score tests approximate to chi-square distribution if and only if a minimum average expected subtype frequency is attained. In other words, convergence to the asymptotic distribution is not directly associated with the sample size. Even if the total sample size is large enough, neither test follows a chi-square distribution unless the minimum average expected subtype frequency is satisfied. It is obvious that this required minimum size depends on both the number of disease subtypes under consideration and total sample size. What is also obvious from these results is that as the number of disease subtypes increases, the sample size required to attain the minimum expected frequency also increases. As a result, all these findings make it clear that the use of the asymptotic approach for testing etiologic heterogeneity may not be always plausible and may not provide reasonable results in some cases. Therefore, a valid method should be developed to obtain the distribution of the score test statistic under null hypothesis when the asymptotic approach is questionable.

3.1.4 Small-Sample Distribution of the Score Test

As discussed in the previous section, asymptotic approach is open to doubt in the presence of a disease subtype size below the required. If this is the case, the distribution of the score test statistic under null the hypothesis of interest can be derived using a nonparametric approach. In this context, permutation test, dates back to Fisher[6], was preferred as an easily applicable non-parametric test to obtain the distribution of the score test under null hypothesis. The underlying idea behind the method is that the characteristics of a test statistic can be gained by randomly shuffling the data and calculating the test statistic for each case (Ernst [7]).

Table 3.4 illustrates how the permutation test can be applied to the first scenario consisting of eight disease subtypes ($2 \times 2 \times 2$). The first column Y is the response variable and includes all the possible disease subtypes except for being disease-free. For each disease subtype, the next three columns involve the corresponding levels for each characteristic. Since there are three characteristics each with two levels, three different indicator variables, namely Y_1, Y_2, and Y_3 can be defined for each, respectively, as shown in the last three columns. The construction of these indicator variables is based on the principal that if a characteristic is not in its reference level, then the corresponding indicator variable of that characteristic is set as 1. Recall that in this case, the null hypothesis of interest is the absence of etiologic heterogeneity associated with the second characteristic. Since the aim is to obtain the distribution of the test statistic under this null hypothesis, the data must be permuted by taking this condition into account. That is to say, the null hypothesis states that there is no etiologic heterogeneity with respect to the levels of the second characteristic. Therefore, this means that any unobserved value of the indicator variable corresponding to the second characteristic (Y_2) may also be observed with another covariate value as well as its existing one. That is the reason why Y_2 must be permuted. In addition to that, the reason why Y_2 is permuted instead of the covariate vector is that Y_2 is random and the covariate is fixed, so it is plausible to permute the random one. It is clear that for each permutation of the second indicator variable, the three indicator variables together define a new response variable Y. Thus, this newly generated response variable, along with the existing covariate vector, can be treated as another random sample of equal size. It is also crucial to note that the permutation must only be applied to diseased subjects. The reason for this is that the data must be permuted under null hypothesis and this does not concern diseased-free subjects.

Table 3.4: Illustration of the permutation test for a disease defined by three characteristics each with two levels

Y	Characteristic-1	Characteristic-2	Characteristic-3	Y_1	Y_2	Y_3
1	1	1	1	0	0	0
2	1	1	2	0	0	1
3	1	2	1	0	1	0
4	1	2	2	0	1	1
5	2	1	1	1	0	0
6	2	1	2	1	0	1
7	2	2	1	1	1	0
8	2	2	2	1	1	1

To obtain a Monte Carlo estimate of Type-I error rate when no distributional assumption is made, the following approach was used: Initially, a data set was generated based on some predefined second-stage parameters. For this generated data set, Wald (T_w) and score (T_s) statistics were calculated first. Then, the original data set was permuted as explained in the previous paragraph and permutation method-based test statistics

$T_w^{(p)}$ and $T_s^{(p)}$ were calculated from this permuted data. This procedure of permuting the original data and re-calculating the test statistics were repeated 10000 times in total and the resulting 10000 values for each test statistic were used to constitute an empirical distribution for T_w and T_s. These empirical distributions were used to obtain the permutation test-based critical values $cv_{(w)}^{(p)}$ and $cv_{(s)}^{(p)}$ for each test statistic by calculating the 95-th quantile of each empirical distribution as $0.95 = P(T_{(\cdot)}^{(p)} < cv_{(\cdot)}^{(p)})$. At the end, the test statistics obtained from the original sample T_w and T_s were compared against the permutation test-based critical values $cv_{(w)}^{(p)}$ and $cv_{(s)}^{(p)}$, respectively, to check whether a Type-I error has been made. The whole process up to here was also replicated 1000 times from the beginning with the same predefined second-stage parameters and at each replication it was checked again whether a Type-I error was committed. Finally, a rate for the number of false rejections of the null hypothesis was obtained out of 1000 replications.

When obtaining the permutation test-based critical values, different number of replications were also studied but 10000 replications for permutation together with 1000 replications for empirical Type-I error rate were observed enough to provide stability in the resulting values.

Tables 3.5 and 3.6 summarize the simulation results for 8 disease subtypes ($2\times2\times2$) for a random sample of size 500 and 1000, respectively. For both cases, the null hypothesis of interest is $H_0 : \theta_{2(2)}^{(1)} = 0$ and the test statistics were compared against 95th quantile of their empirical distributions. When compared with Table 3.1, Table 3.5 shows that the performance of the Wald test is more or less the same under both parametric and nonparametric approaches. However, the benefits of the nonparametric approach can be observed for the score test when the minimum size is below 13.

Table 3.5: Empirical Type I error rates (permutation test-based approach). Disease subtypes=$2 \times 2 \times 2$; $n = 500$

Minimum Average Expected Subtype Frequency	**Wald Test**	**Score Test**
25	0.0460	0.0460
24	0.0540	0.0540
22	0.0520	0.0520
15	0.0510	0.0500
13	0.0510	0.0520
12	0.0600	0.0610
11	0.0560	0.0560
10	0.0540	0.0540

Table 3.6 shows the case when the sample is increased from 500 to 1000. It is the counterpart of Table 3.2. It is obvious that the nominal significance level is not significantly violated even when the minimum size is below 30.

Table 3.6: Empirical Type I error rates (permutation test-based approach). Disease subtypes=$2 \times 2 \times 2$; $n = 1000$

Minimum Average Expected Subtype Frequency	**Wald Test**	**Score Test**
50	0.0400	0.0400
49	0.0560	0.0560
45	0.0670	0.0570
30	0.0643	0.0667
27	0.0654	0.0673
24	0.0570	0.0570
23	0.0540	0.0540
21	0.0530	0.0530

Table 3.7 summarizes the simulation results for 16 disease subtypes ($4 \times 2 \times 2$) for a random sample of size 1000. In this case, the null hypothesis of interest becomes $H_0 : \theta^{(1)}_{1(2)} = \theta^{(1)}_{1(3)} = \theta^{(1)}_{1(4)}$ and the test statistics were compared against 95-th quantile of their empirical distributions.

Table 3.7: Empirical Type I error rates (permutation test-based approach). Disease subtypes=$4 \times 2 \times 2$; $n = 1000$

Minimum Average Expected Subtype Frequency	**Wald Test**	**Score Test**
24	0.0570	0.0570
23	0.0410	0.0410
22	0.0490	0.0490
14	0.0520	0.0520
13	0.0470	0.0470
12	0.0380	0.0380
11	0.0460	0.0460

Recall Table 3.3, where the empirical Type-I error rates are significantly higher than the nominal significance level for both the Wald and score tests. However, it is obvious from Table 3.7 that there seems to be no risk of violating the nominal significance level provided that permutation test-based critical value is used.

3.2 Comparison of the Power of the Wald and Score Tests

3.2.1 Calculation of Power

Some of the scenarios studied for the empirical Type-I error comparison were also considered for the power study and the same data generation process was followed, except for the fact that the data sets were generated under a variety of alternative hypotheses. In general, the power study was conducted separately under two approaches: chi-square-based and permutation test-based. In this context, (i) 8 disease subtypes ($2 \times 2 \times 2$) and (ii) 16 disease subtypes ($4 \times 2 \times 2$) were studied. However, the second-stage parameters were set in such a way that the minimum and maximum values for the minimum average expected subtype frequency has been considered. Under chi-square-based approach, 1000 Monte Carlo replications were conducted and each time the test statistics were compared against the chi-square-based critical value. Under permutation test-based approach, on the other hand, the permutation test was applied for each generated data set at each of 1000 Monte Carlo replications and the test statistics were compared against 95-th quantile of their empirical distributions.

When calculating the power, different number of replications were also studied but 1000 replications were observed enough to provide stability in the resulting values.

3.2.2 Asymptotic Power Comparison

It was shown that when 8 disease subtypes ($2 \times 2 \times 2$) are of interest and the minimum average expected subtype frequency is 50, then the behavior of both the Wald and score test statistics are close to chi-square distribution. Therefore, in order to see the performance of the two tests when the asymptotic distribution assumption seems a little bit plausible, a simulation study was conducted and the results are presented in Table 3.8. As mentioned previously, the null hypothesis of interest was $H_0 : \theta^{(1)}_{2(2)} = 0$ and the test statistics were compared against the $\chi^2_{(1)}$ based critical value at nominal $\alpha = 0.05$. The first column of Table 3.8 gives the alternative hypotheses that have been considered. That is to say, each time the data were generated for a different value of $\theta^{(1)}_{2(2)}$, decreasing from 0.5 to 0.001. It can be concluded from Table 3.8 that the asymptotic power of the score test is higher in general as long as the minimum average expected subtype frequency is attained.

Table 3.8: Empirical power (χ^2 based approach). Disease subtypes=2×2×2; $n = 1000$; minimum subtype frequency=50

$\theta^{(1)}_{2(2)}$	**Score**	**Wald**
0.500	0.9986	0.9985
0.400	0.9800	0.9983
0.300	0.8813	0.8757
0.200	0.5793	0.5688
0.100	0.2064	0.2008
0.090	0.1799	0.1736
0.080	0.1534	0.1491
0.070	0.1336	0.1291
0.060	0.1145	0.1100
0.050	0.0978	0.0942
0.040	0.0867	0.0842
0.030	0.0755	0.0729
0.020	0.0675	0.0643
0.010	0.0636	0.0613
0.001	0.0604	0.0582

3.2.3 Permutation Test-based Power Comparison

Initially, a permutation test-based power comparison was conducted between Wald and score tests assuming that three disease characteristics each with two levels are of interest and the minimum average expected subtype frequency is 50.

Table 3.9 summarizes the simulation results. When compared with Table 3.8, it can be concluded that both chi-square-based and permutation test-based approach give the similar results when the required minimum average expected subtype frequency is satisfied.

Table 3.9: Empirical power (permutation test-based approach). Disease subtypes=$2 \times 2 \times 2$; $n = 1000$; minimum subtype frequency=50

$\theta^{(1)}_{2(2)}$	**Score**	**Wald**
0.500	0.9980	0.9980
0.400	0.9890	0.9890
0.300	0.8720	0.8720
0.200	0.5730	0.5730
0.100	0.2350	0.2350
0.090	0.1700	0.1700
0.080	0.1460	0.1460
0.070	0.1290	0.1290
0.060	0.1010	0.1010
0.050	0.0840	0.0840
0.040	0.0860	0.0860
0.030	0.0690	0.0690
0.020	0.0480	0.0480
0.010	0.0540	0.0540
0.001	0.0430	0.0430

As illustrated previously, chi-squared-based approach is not valid for the previously discussed scenario when the minimum average expected subtype frequency is 21, i.e. minimum. Therefore, only a permutation-based power comparison was conducted for this scenario and the simulation results are given in Table 3.10. When compared with the values given in Table 3.9, a little bit decrease in the power of the score test can be observed. An overall assessment of Tables 3.8, 3.9, and 3.10 indicate that for a sample of size 1000, as the minimum size gets smaller from 50 to 21, the decrease in the power of the score test can be minimized as much as possible by using permutation test-based approach.

Table 3.10: Empirical power (permutation test-based approach). Disease subtypes=$2\times 2\times 2$; $n = 1000$; minimum subtype frequency=21

$\theta^{(1)}_{2(2)}$	**Score**	**Wald**
0.500	0.9800	0.9820
0.400	0.9210	0.9210
0.300	0.7520	0.7550
0.200	0.4340	0.4340
0.100	0.1700	0.1700
0.090	0.1580	0.1600
0.080	0.1280	0.1280
0.070	0.1050	0.1050
0.060	0.1000	0.1000
0.050	0.0880	0.0880
0.040	0.0780	0.0790
0.030	0.0640	0.0640
0.020	0.0590	0.0590
0.010	0.0640	0.0640
0.001	0.0620	0.0630

Also recall that, as shown previously, when three disease characteristics with levels four, two and two, respectively, are under consideration, asymptotic properties cannot be applied and permutation test-based approach becomes valid. Therefore, in order to see the performance of the two tests under this scenario for which the asymptotic distribution assumption is invalid, a simulation study was conducted. As mentioned previously, the null hypothesis of interest is $H_0 : \theta^{(1)}_{1(2)} = \theta^{(1)}_{1(3)} = \theta^{(1)}_{1(4)}$ and the test statistics were compared against the corresponding permutation test-based critical value. Tables 3.11 and 3.12 summarize the simulation results for this scenario, except for the fact that the minimum size is 11 (i.e. minimum) in Table 3.11 and 26 (i.e. maximum) in Table 3.12. The first two columns of both tables jointly give the alternative hypotheses that have been considered. That is to say, on the condition that $\Delta_1 = \Delta_2$, each time the data were generated for a different value of Δ_1 and Δ_2 in, where $\Delta_1 = \theta^{(1)}_{1(2)} - \theta^{(1)}_{1(3)}$ and $\Delta_2 = \theta^{(1)}_{1(3)} - \theta^{(1)}_{1(4)}$. Tables 3.11 and 3.12 indicate that, in general, the performance of the score test gets a little bit better as the minimum subtype size changes from 11 to 26. This shows that even when the minimum subtype size is 11, the benefits of the permutation test-based approach can still be undeniable.

Table 3.11: Empirical power (permutation test-based approach). Disease subtypes=4× 2 × 2; $n = 1000$; minimum subtype frequency=11

Δ_1	Δ_2	**Score**	**Wald**
1.500	1.500	1.0000	1.0000
1.000	1.000	1.0000	1.0000
0.400	0.400	0.9990	1.0000
0.300	0.300	0.9750	0.9750
0.200	0.200	0.7630	0.7630
0.100	0.100	0.2290	0.2300
0.050	0.050	0.0710	0.0720
0.040	0.040	0.0640	0.0640
0.030	0.030	0.0550	0.0550
0.020	0.020	0.0510	0.0510
0.010	0.010	0.0500	0.0500
0.001	0.001	0.0460	0.0460

Table 3.12: Empirical power (permutation test-based approach). Disease subtypes=4× 2 × 2; $n = 1000$; minimum subtype frequency=26

Δ_1	Δ_2	**Score**	**Wald**
1.500	1.500	1.0000	1.0000
1.000	1.000	1.0000	1.0000
0.400	0.400	1.0000	1.0000
0.300	0.300	0.9840	0.9840
0.200	0.200	0.7570	0.7580
0.100	0.100	0.2600	0.2600
0.050	0.050	0.0820	0.0820
0.040	0.040	0.0650	0.0660
0.030	0.030	0.0490	0.0490
0.020	0.020	0.0430	0.0430
0.010	0.010	0.0470	0.0470
0.001	0.001	0.0530	0.0530

CHAPTER 4

DISCUSSION AND CONCLUSION

In this thesis study, as the first contribution, a score test was developed to determine whether the etiologic heterogeneity represented by the second stage parameters is significant or not. The reason why score test is considered over Wald's test are due to its general following advantages:

i. Computation of Wald test statistic requires MLEs based on unconstrained likelihood function where score test statistics requires only MLEs based on the likelihood function constrained by the conditions in the null hypothesis

ii. Matrix inversion is performed twice for the Wald test statistic and once for score test statistics.

iii. Due to (i) and (ii) optimization of the likelihood for score test has inarguable mathematical/computational advantages over Wald test.

iv. Null hypotheses of etiologic heterogeneity are always composite and require re-parameterization when the partitioned-vector approach is used. However, the Wald statistic is not invariant to any kind of re-parameterization (Boos and Stefanski [2]). In addition, even if the null hypothesis of interest is expressed as in the form of $\boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$, the choice of $\boldsymbol{h}$ also affects the value of the Wald statistic.

As the second contribution, the asymptotic and finite sample properties of the score test was investigated through an extensive Monte Carlo simulation study. The results were compared against those of the Wald test. Also, the two test statistics were compared based on statistical power under various different realistic scenarios. Results of the Monte Carlo simulations to investigate their finite sample characteristics reveal that a minimum average expected subtype frequency must be attained for the asymptotic distribution of both Wald and score tests to hold. In other words, convergence to asymptotic distribution is not directly related to total sample size. It has been observed that the minimum expected subtype frequency depends on (i) number of disease subtypes under study, and (ii) total sample size. However, permutation test was

proposed as a remedy to overcome the problem of not attaining minimum average expected subtype frequency. In this context, instead of making questionable assumptions to use chi-square-based critical values, one can gain information regarding the characteristics of the score test statistics by permuting the data in hand and recalculating the test statistic for each permutation. It has been observed through an extensive Monte Carlo simulation experiment that when the required minimum subtype frequency is not satisfied, the success of score test statistic in attaining the nominal significance level is fully dependent on the use of permutation test-based critical value.

The findings of this thesis study should be discussed with the point in mind that the covariate studied here is continuous. However, the case in which the covariate is dichotomous should also be investigated to ascertain whether the same conclusions apply. Therefore, what happens in the presence of binary covariate may be considered as a future study. The other important point should be underlined is the extension of the findings to diseases with number of subtypes greater than 16. As discussed before, it was observed that not only the total sample size but also the minimum subtype frequency should be attained for the asymptotic properties of the score test to hold. Recall that for a certain sample size, say 1000, if there are 8 ($2\times2\times2$) disease subtypes to be studied, it is likely that the required minimum subtype frequency can be attained. However, if the number of disease subtypes is doubled (i.e. 16; $4 \times 2 \times 2$), it becomes less likely to satisfy the criterion regarding the minimum subtype size and so, the permutation test-based approach comes into the picture as a remedy to handle such a case. All these reveal that even for a sample of size 1000, which may not be regarded as cost-efficient, the chance of using the asymptotic approach gets less probable as the number disease subtypes gets far from 16. Although increasing the sample size in such a case may be seemed to be a solution at first glance, the crucial problem arises from not having enough number of subjects with a certain disease subtype in the population of interest. It is obvious that such a case is more likely when the number of disease subtypes is large and the problem cannot be overcome by increasing the sample size. Therefore, whatever the total sample size is, use of the asymptotic properties may most probably become risky due to the fact that the minimum subtype frequency will not be attained. All in all, based upon the scenarios studied in this thesis study, permutation test-based approach is also proposed to be appropriate for diseases with number of subtypes greater than 16.

Results of this work will provide the researches of different fields with a powerful statistical testing tool for important research questions. For example, a cancer researcher will be able to test (i) if having a cancer in the family is more associated with the patient's tumor being large, (ii) if sufficient fiber intake has bigger effect on positive nodal status (i.e. risk of spread of cancer to other lymph nodes) than negative nodal status.

REFERENCES

[1] A.K. Bera and Y. Bilias. Rao's score, Neyman's C(a) and Silvey's LM tests: an essay on historical developments and some new results. *Journal of Statistical Planning and Inference,* 97:9-44, 2001.

[2] D.D. Boos and L.A. Stefanski. *Essential Statistical Inference.* Springer, 2013.

[3] D. McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics,* Edited by Zarembka, Academic Press, New York, 1974.

[4] D.V. Hosmer and S. Lemeshow. *Applied Logistic Regression.* Wiley Series in Probability and Statistics, 2000.

[5] D.G. Kleinbaum and M. Klein. *Logistic Regression: A Self-Learning Text.* Springer, 2010.

[6] R.A. Fisher. *The Design of Experiments.* Oliver and Boyd, Edinburgh, 1935.

[7] M.D. Ernst. Permutation methods: a basis for exact inference. *Statistical Science,* 19(4):676-685, 2004.

[8] M.E. Sherman, D.L. Rimm, X. Yang, N. Chatterjee, C. Brinton, J. Lissowska, B. Peplonska, D. Szeszenia, A.B. Mikolajczak, W. Zatonski, R. Cartun, D. Mandich, G. Rymkiewicz, D.M. Sikor, S. Lukaszek, B. Kordek, Z. Kalaylioglu, M. Harigopal, L. Charrette, R.T. Falk, D. Richesson, W.F. Anderson, S.M. Hewitt, M.G. Closas. Variation in breast cancer hormone receptor and HER2 levels by etiologic factors: a population-based analysis. *International Journal of Cancer,* 121(5):1079-85, 2007.

[9] M. Garcia-Closas, L.A. Brinton, N. Chatterjee, B. Peplonska, N. Szeszenia-Dabrowska, A. Bardin-Mikolajczak, W. Zatonski, A. Blair, W.F. Anderson, G. Rymkiewicz, D. Mazepa-Sikora, R. Kordek, S. Lukaszek, Z. Kalaylioglu, M. Sherman. Established breast cancer risk factors by clinically important tumor characteristics. *British Journal of Cancer,* 95(1):123-129, 2006.

[10] M.T. Erdem. Modeling diseases with multiple disease characteristics: comparison of models and estimation methods. Master's thesis, Middle East Technical University, 2011.

[11] N. Chatterjee. A two-stage regression model for epidemiological studies with multivariate disease classification data. *Journal of the American Statistical Association,* 99(465):127-138, 1994.

[12] R.F. Engle. Wald, likelihood ratio, and lagrange multiplier tests in econometrics. *Handbook of Econometrics II,* Edited by Z. Griliches and M.D. Intriligator, Elsevier, 796-801, 1984.

[13] W.J. Welch. Construction of Permutation Tests. *Journal of the American Statistical Association,* 85(411):693-698, 1990.