MODELING DISEASES WITH MULTIPLE DISEASE CHARACTERISTICS:
COMPARISON OF MODELS AND ESTIMATION METHODS

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

BY

MÜNİRE TUĞBA ERDEM

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

JULY 2011

Approval of the thesis

MODELING DISEASES WITH MULTIPLE DISEASE CHARACTERISTICS: COMPARISON OF MODELS AND ESTIMATION METHODS

Submitted by **MÜNİRE TUĞBA ERDEM** in partial fulfillment of the requirements for the degree of **Master of Science in the Department of Statistics, Middle East Technical University** by

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

Prof. Dr. Öztaş Ayhan ________________
Head of Department, **Statistics**

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

Examining Committee Members:

Assist. Prof. Dr. Özlem İlk ________________
Statistics Dept., METU

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

Assist. Prof. Dr. Vilda Purutçuoğlu ________________
Statistics Dept., METU

Assist. Prof. Dr. Ceylan Yozgatlıgil ________________
Statistics Dept., METU

Lütfi Doğan, MD ________________
Surgery Dept.,
Ankara Oncology Research and Education Hospital

Date: 12.07.2011

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: M.Tuğba ERDEM

Signature:

ABSTRACT

MODELING DISEASES WITH MULTIPLE DISEASE CHARACTERISTICS: COMPARISON OF MODELS AND ESTIMATION METHODS

Erdem, Münire Tuğba

M.Sc., Department of Statistics

Supervisor: Assist. Prof. Dr. Zeynep Kalaylıoğlu

July 2011, 121 pages

Epidemiological data with disease characteristic information can be modelled in several ways. One way is taking each disease characteristic as a response and constructing binary or polytomous logistic regression model. Second way is using a new response which consists of disease subtypes created by cross-classification of disease characteristic levels, and then constructing polytomous logistic regression model. The former may be disadvantageous since any possible covariation between disease characteristics is neglected, whereas the latter can capture that covariation behaviour. However, cross-classifying the characteristic levels increases the number of categories of response, so that dimensionality problem in parameter space may occur in classical polytomous logistic regression model. A two staged polytomous logistic regression model overcomes that dimensionality problem. In this thesis, study is progressen in two main directions: simulation study and data analysis parts. In simulation study, models that capture the covariation behaviour are compared in terms of the response model parameter estimators. That is, performances of the maximum likelihood estimation (MLE) approach to classical polytomous logistic regression, Bayesian estimation approach to classical polytomous logistic regression and pseudo-conditional likelihood (PCL) estimation approach to two stage polytomous logistic regression are compared in terms of bias and variation of estimators. Results of the simulation study revealed that for small sized sample and small number of disease subtypes, PCL outperforms in terms of bias and variance. For medium scaled size of total disease subtypes situation when sample size is small, PCL performs better than MLE, however when the sample size gets larger MLE has better performance in terms of standard errors of estimates. In addition, sampling variance of PCL estimators of two stage model converges to asymptotic variance faster than the ML estimators of classical polytomous logistic regression model. In data analysis, etiologic heterogeneity in breast cancer subtypes of Turkish female cancer patients is investigated, and the superiority of the two stage polytomous logistic regression model over the classical polytomous logistic model with disease subtypes is represented in terms of the interpretation of parameters and convenience in hypothesis testing.

Keywords: Two stage polytomous logistic regression model, Etiologic Heterogeneity in Breast Cancer, Pseudo-conditional likehood estimation

ÖZ

HASTALIK KARAKTERİSTİĞİNİN MODELLENMESİ: MODEL VE TAHMİN YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Erdem, Münire Tuğba

Yüksek Lisans, İstatistik Bölümü

Tez Yöneticisi: Yrd. Doç. Dr. Zeynep Kalaylıoğlu

Temmuz 2011, 121 sayfa

Hastalık karakteristiği bilgisi bulunduran epidemiyolojik veri çeşitli şekilde modellenebilir. Bunlardan biri, her bir karakteristiği yanıt değişkeni olarak alıp, iki terimli veya çok terimli lojistik regresyon modeli kurmaktır. Ikinci yol, hastalık karakteristiklerinin kategorilerinin çapraz-sınıflandırılması sonucu elde edilmiş hastalık alt-tiplerinden oluşan yanıt değişkeni üzerine çok terimli lojistik regresyon modeli kurmaktır. Ilk yöntem, karakteristikler arasındaki olası bir ortak değişim davranışını gözardı ettiği için dezavantajlı olabilmektedir. İkinci yöntemde yanıt değişkeni, karakteristiklerin kategorilerinin çapraz-sınıflandırılmasıyla oluşturulmuş hastalık alt-tiplerinden oluştuğu için karakteristikler arası etkileşimi göz önünde bulundurmaktadır. Bununla birlikte, çapraz-sınıflandırma sonucu yanıt değişkeninin kategori sayısı parametre uzayının boyutunu modellemeyi güçleştirecek şekilde artırabilmektedir. İki aşamalı çok terimli lojistik regresyon modeli bu problemi ortadan kaldırmaktadır. Bu tez çalışması, simulasyon ve veri analizi olmak şekilde iki kısımdan oluşmaktadır. Simulasyon kısmında karakteristikler arasındaki ortak değişim durumunu göz önünde bulunduran metodlar parametre tahminleyicilerinin yanlılığı ve varyansı üzerinden karşılaştırılmıştır. Bu metodlar: maksimum olasılık tahminleyicisi (MLE) yaklaşımıyla klasik çok terimli lojistik regresyon modeli, Bayesçi yaklaşımla klasik çok terimli lojistik regresyon modeli, ve sözde-koşullu olabilirlik tahminleyicisi (PCL) yaklaşımıyla iki aşamalı çok terimli lojistik regresyon modelidir. Simulasyon sonuçlarına göre örnek sayısı ve hastalık alt-tipi sayısı az olduğu durumda PCL diğer iki metoda göre daha iyi performans göstermektedir. Orta ölçekli hastalık alt-tipi durumunda örnek sayısı az iken PCL MLE'den daha iyi performansa sahipken, örnek sayısı arttığında ML tahminleyicilerinin standart hataları PCL tahminleyicilerine göre daha düşüktür. Ayrıca, PCL tahminleyicilerinin örneklem varyansları asimtotik varyansa ML tahminleyicilerine göre daha hızlı yakınsar. Tezin veri analizi kısmında Türkiye'deki kadın göğüs kanseri hastaları için göğüs kanserinin etiolojik heterojenliğinin analizi yapılmıştır. Ayrıca, iki aşamalı lojistik regresyon modelinin parametrelerinin yorumlanma ve hipotez testindeki elverişliliği açısından üstünlüğü gösterilmiştir.

Anahtar Kelimeler: İki aşamalı çok terimli lojistik regresyon modeli, Göğüs kanserinde etiolojik heterojenlik, sözde-koşullu-olabilirlik tahmini

ACKNOWLEDGMENTS

I owe my deepest gratitude to my advisor Assist. Prof. Dr. Zeynep Kalaylıoğlu for her guidance, patience and positive attitude during this thesis study. As a student of her, I have always thought that taking benefit from her precious knowledge and extensive academic experience was a great fortune for me. With her friendly manner to share her experiences, self-disciplined stance in academic study and delicate attitude as a person, she has always been a role model for me in academic career.

I would like to present my grateful thanks to thesis committee members of my thesis: Assist. Prof. Dr. Özlem İlk, Assist. Prof. Dr. Vilda Purutçuoğlu, Assist. Prof. Dr. Ceylan Yozgatlıgil and Dr. Lütfi Doğan for their valuable time to review this thesis and their constructive comments and suggestions. I especially would like to thank Dr. Lütfi Doğan for generously sharing with us the dataset used in data analysis part.

Then, my special thanks go to my friends Gül İnan, Nazlı Sarıaslan, Nesrin Mevsim, Göksel Tüker, Neşe Abuk, Özlem Yalçın, Ela Uysal, Ceyda Yazıcı, Serçin Şekkeli, Aysun Yeşilyurt and Akile Sinem Ünlü Engin for their support and sincere friendship. Also, I would like to thank all my instructors, assistant friends and all the members of METU Statistics Department.

Last but not the least, 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 FIGURES

FIGURES

CHAPTER I

INTRODUCTION

In epidemiological studies, one of the main aims is to determine the association between the disease risk factors and the risk of the disease. This association is investigated by statistical modeling of various sorts. It has become common, due to the ever improving medical technology by a recently emerging profession biomedical engineering, that the data about the disease include various different specific characteristics of the disease and not only the disease status represented by absence/presence type of information. As a result, various different subtypes of the diseases, especially the complex diseases such as cancer, are identified by the combination of the important disease characteristics. This type of disease data, with appropriate statistical modeling, makes it possible to understand the association of the disease with factors by going deep into the disease characteristic level. An interesting research question which is actually of crucial interest when such data is available is: whether the effect of the risk factors differs for different disease subtypes. To make things more concrete, let's begin with a motivating example. In this case-control study on breast cancer, a number of characteristics of the breast tumor are recorded such as *tumor size* attaining three levels (T1, T2, T3) and *NA status* with two levels (exist, not exist). In order to assess the effect of the known risk factors such as family history and the number of full term births on the breast cancer risk among Turkish female population by the tumor characteristics specified above, one common and naïve approach to analyze this dataset would be either i) by considering each tumor characteristic as a response variable in a separate/independent logistic regression model, or ii) creating disease subtypes by cross-classifying the disease characteristics, forming a response variable with levels consisting of the resulting subtypes, and running e.g. a generalized logit regression on this response variable with numerous levels. In that sense, the first approach would be having one polytomous logistic regression for tumor size and one binary logistic regression for the NA status and these two separate logistic regression models would be analyzed completely independent from one another. The second approach would be having one large polytomous logistic regression for the response variable consisting of the following categories (T1, NA exist), (T2, NA exist), (T3, NA exist), (T1, NA not exist), (T2, NA not exist), and (T3, NA not exist) as levels. The data is obtained from a hospital-based case-control study on breast cancer which was carried out in Ankara Oncology Research and Education Hospital. Further specifics and the analysis of this data set are given in Chapter 4.

The approaches described above have serious problems. In the first case, the disease characteristics are treated as independent. However this may not be true in reality. For instance most of the time the characteristics of a tumor exhibit covariation. In the second case, cross-classification may lead to response variable to have very large number of categories. This results in dimension problem with the parameter space. Also, in a small or moderate-scale study, estimation problems may occur due to the lack of enough observations corresponding to some categories. To overcome such problems, a two stage polytomous logistic regression model is developed by Chatterjee (2004) in which the first stage models the disease risk given the covariates through the coefficients βs and the second stage models the θs (the details are given in the methods of interest chapter). Notice that, the cross-classification of the two tumor characteristics exemplified in the first paragraph yields 6 disease subtype categories. In a polytomous logistic regression model, this requires 6 regression coefficients for each of the covariates in addition to the 6 intercept parameters resulting in 6(p+1) regression coefficients in total where p is the number of covariates considered in the model. However, with the two staged polytomous logistic regression model, the parameter space of interest is downsized to only 3 dimension in this specific example (the details of the reduction in the dimension of the parameter space in two staged approach is laid out in the methods of interest chapter). It is obvious that, in a small or moderate sample-size study, two staged modelling approach provides advantage in estimation. Also note that, constructing disease subtypes and using the two staged model that considers the multivariate nature in characteristics, we get a chance to examine the etiologic heterogeneity of the disease under investigation. This is one of the main advantages of the two staged modeling approach over constructing separate binary/multinomial logistic regression models. When the two stage modeling of Chatterjee is used all the interesting research questions in an epidemiologic study can be hypothesized based on only the second stage parameters without really needing to express the hypotheses in terms of the main model parameters (first stage parameters). Nevertheless it is not unreasonable to be curious about how the classical method (i.e. (ii) in the first paragraph) and the two staged approach would compare in terms of the efficiency of the main model parameter estimators as the size of the study and the total number of disease subtypes increase. In epidemiologic investigations, all the results are given in terms of the second stage parameters when the two staged approach is used but not in terms of the main model parameters. We hope that our findings in this thesis will ensure the appropriateness of this common practice.

In this study, we will make three major contributions: firstly, using a Monte Carlo simulation experiment, we will compare the performances of the three methods in terms of the estimation of first stage parameters, namely β’s. Methods we compared are: (1) a frequentist approach, wherein maximum likelihood estimation (MLE) to estimate a polytomous logistic regression model is applied when the response variable consists of the disease subtypes obtained by cross-classification of the disease characteristic levels; (2) a Bayesian approach to estimate the model mentioned in (1); (3) a two staged approach to polytomous logistic regression model through pseudo-conditional likelihood estimation (PCL) on multivariate disease characteristics data. As the second contribution, we will demonstrate the practical advantages of the two stage approach by applying the methods of consideration on a breast cancer dataset. In that sense, maximum likelihood estimation of classical polytomous logistic regression with response in the form of disease subtype and pseudo-conditional likelihood estimation (PCL) of two staged logistic regression with response in the form of multivariate disease characteristics are compared in terms of ease in interpretation, standard errors of estimates and parameters used in the hypothesis testing. As the third contribution, we will unveil the underlying association between the breast cancer risk and its known risk factors for Turkish females by employing the two stage method. To sum up, we wish to make the following contributions through this thesis work:

1. A bias and efficiency comparison between the maximum likelihood estimation for classical polytomous logistic regression and conditional likelihood estimation for two stage logistic regression in terms of the main model parameters, namely βs.

2. Illustration about the practical advantage of two stage logistic regression for testing whether the strength of the relationship between a risk factor and a certain tumor characteristic depends on another tumor characteristic.

3. A detailed picture on the heterogeneity in the etiology of breast cancer subtypes for Turkish female breast cancer patients.

The rest of the thesis is organized as follows: having motivated the main problem of the study in this chapter, in chapter 2, methodology of the three approaches are presented; maximum likelihood estimation of classical polytomous logistic regression model is explained in section 2.1, Bayesian estimation of model parameters of classical polytomous logistic regression via WinBUGS is dealt with in section 2.2, and PCL estimation of second stage parameters of two staged logistic regression model is discussed in section 2.3. In chapter 3, comparisons of three methods are done through a simulation experiment designed to cover a different range of realistic sample scenarios. Data generation procedure is explained in 3.1 and functions and procedures to obtain parameter estimates from the above mentioned methods are presented in Section 3.2. Section 3.3 illustrates the results of the simulation study. Analysis of Turkish breast cancer dataset is given in Chapter 4. In section 4.1, disease characteristics are taken independently and binary/multinomial logistic regression models are built on the subjects with disease in the sample, in section 4.2 classical polytomous logistic regression model is built after creating disease subtype information data, and in section 4.3, binary/polytomous logistic regression and two stage logistic regression models are compared in terms of interpretation of parameters and standard errors of estimates, then in section 4.4, conclusions on the data analysis are presented. An overview of the results and possible extensions for future works are discussed in Chapter 5.

CHAPTER II

METHODS OF INTEREST

In this chapter, models and methods that are used throughout the thesis are introduced. Fundamentals of maximum likelihood estimation and Bayesian estimation of classical polytomous logistic regression models where the response variable is the disease subtypes which are created by cross-classification of disease characteristics, and pseudo-conditional likelihood estimation of two stage polytomous logistic regression (Chatterjee, 2004) are presented in the subsequent sections.

2.1. MAXIMUM LIKELIHOOD ESTIMATION FOR POLYTOMOUS LOGISTIC REGRESSION

Regression models are used to explain the association between a response variable and predictors. When the response variable is in nominal or ordinal scale, by using a proper link function (logit, probit etc.), we can regress a categorical variable on covariates which are either numeric or categorical. When the response variable is dichotomous, we can use the binary logistic regression model. McFadden (1974) has introduced a modification of this model for categorical response with more than two levels, and this model is called as polytomous (or multinomial) logistic regression model with *logit* link function (Hosmer and Lemeshow, 2000).

For a categorical response Y with M+1 levels, and p covariates, the polytomous logistic regression model has the following form:

$$\ln \frac{P(Y_i=m|X)}{P(Y_i=0|X)} = \alpha_m + X_i\beta_m \text{ , m=1,..., M}$$

We know that Y has a multinomial distribution with each level having the probability $P(Y=m)$, where $P(Y=m)=\frac{e^{\alpha_m+X_i\beta_m}}{1+\sum_{m=1}^{M}e^{\alpha_m+X_i\beta_m}}$ and $P(Y=0)=\frac{1}{1+\sum_{m=1}^{M}e^{\alpha_m+X_i\beta_m}}$

Then, the likelihood function for a sample of n independent observations is:

$$L(\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})=\prod_{i=1}^{n}[P(Y=0)]^{y_{0i}}\,[P(Y=1)]^{y_{1i}}\ldots\,[P(Y=M)]^{y_{Mi}}$$

$$=\prod_{i=1}^{n}[\frac{1}{1+\sum_{m=1}^{M}e^{\alpha_m+X_i\beta_m}}]^{y_{0i}}[\frac{e^{\alpha_1+X_i\beta_1}}{1+\sum_{m=1}^{M}e^{\alpha_m+X_i\beta_m}}]^{y_{1i}}\ldots\,[\frac{e^{\alpha_M+X_i\beta_M}}{1+\sum_{m=1}^{M}e^{\alpha_m+X_i\beta_m}}]^{y_{Mi}}$$

where $y_{mi}=1$ if $Y_i=m$, and $y_{mi}=0$ otherwise; m=1,..., M.

In order to get the maximum likelihood estimates (MLE's) of model parameters (**α,β**), natural logarithm of likelihood function is written as the following form,

$$\ln L(\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})=\sum_{i=1}^{n}\left\{y_{1i}(\alpha_1+X_i\beta_1)+\cdots+y_{Mi}(\alpha_M+X_i\beta_M)-\ln\left(1+e^{\alpha_1+X_i\beta_1}+\cdots+e^{\alpha_M+X_i\beta_M}\right)\right\}$$

and the score equations are obtained by taking the derivatives of $\ln L(\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})$ with respect to M(p+1) unknown parameters:

$$\frac{\partial\ln L(\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})}{\partial\boldsymbol{\beta}_m}=\sum_{i=1}^{n}y_{mi}X_i-\frac{X_ie^{\alpha_m+X_i\beta_m}}{1+e^{\alpha_1+X_i\beta_1}+\cdots+e^{\alpha_M+X_i\beta_M}}=\sum_{i=1}^{n}X_i(y_{mi}-P(Y_i=m)),$$

m=0,…,M

By finding the roots of the score equations above, we get the MLE's of (α, β). Since it not possible to write $\hat{\alpha}$ and $\hat{\beta}$ in closed forms, the solution is found by using numerical root finding methods such as Newton-Raphson algorithm.

The asymptotic distribution of maximum likelihood estimators of logistic regression models are Normal (Agresti, 2002).

2.2. BAYESIAN ESTIMATION FOR POLYTOMOUS LOGISTIC REGRESSION

Bayesian paradigm is based on letting the parameters have a probabilistic distribution rather than confining them to single values. In that respect Bayesian methods treat the parameters as random variables as opposed to the frequentist statistical approaches that treat them as constants. The advantage of Bayesian point of view is that any kind of idea the researcher has about the parameter prior to the observed data at hand can be employed along with the observed data. The researcher's thought about the parameter may have been driven by either some solid information in similar situations/experiments or by intuition alone. Either way, in Bayesian approach the information that will be used to find the unknown is accrued through the collaboration of the observed data and the prior knowledge regarding the parameter.

In this study, we have the unstructural polytomous logistic regression model parameters to be estimated. We can write the model as follows and explain the basics of Bayesian approach to estimate these parameters through this model:

$\ln \frac{P(Y_i=m|X)}{P(Y_i=0|X)} = \alpha_m + \beta_{1m}X_1 + \cdots + \beta_{pm}X_p$, where m=1,…,M is the different levels of the response variable for diseased subjects, and p is the number of covariates.

Here, the aim is to estimate θ=(α,β) using Bayesian approach. To do this, we should first obtain the distribution of these parameters given the sample we have at hand:

$$f(\theta|X,Y) \propto f(X,Y|\theta)f(\theta) \qquad (2.1)$$

where $f(\theta|X,Y)$ is the posterior distribution of θ, $f(X,Y|\theta)$ is the likelihood and $f(\theta)$ is the prior distribution of the parameter of interest.

2.2.1. Prior and Likelihood

In Bayesian approach, we assume that parameters to be estimated, **θ** have their own distributions which is called as prior distribution and represented by f(**θ**). As it can be seen from (2.1), while getting the posterior distribution of θ, we face with a trade-off between the information coming from data via likelihood function $f(X,Y|\theta)$ and the information coming from the prior knowledge via prior probability density function $f(\theta)$. In some rare cases, the prior distribution of θ may be quite certain that we do not need to know much about the data. However, the opposite direction may be the case as well: when we do not have enough prior information or we want to let the data say all the story, we can use non-informative priors. There are several choices for determining non-informative priors: e.g. uniform priors on a large range, improper priors in conjugate families, Jeffrey's prior (Link and Barker, 2010). In our case, since we know that the approximate distribution of model parameters of logistic regression are Normal, it is reasonable to use Normal distribution to characterize the prior information about the parameters. We used multivariate Normal where diagonals of the variance covariance matrix of this joint prior density implies diffuse prior distribution for each regression coefficient.

For the independent and identically distributed sample with response $Y_1, \dots, Y_n$, and the predictors $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ the likelihood function is:

$$f(X,Y|\theta) = \prod_{i=1}^{n} f(Y_i, X_i|\theta)$$

The likelihood has the information provided by the observed sample for a certain parameter value.

For the multinomial logistic regression model, with response variable Y having M+1 levels, covariate matrix having nxp dimension, and parameters $\boldsymbol{\delta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, the likelihood function is as follows:

$$L(\boldsymbol{\delta}\,|\,\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{n}[P(Y=0)]^{y_{0i}}\,[P(Y=1)]^{y_{1i}}\ldots\,[P(Y=M)]^{y_{Mi}}$$

$$= \prod_{i=1}^{n}\left[\frac{1}{1+\sum_{m=1}^{M} e^{m+X_i\beta_m}}\right]^{y_{0i}}\left[\frac{e^{\alpha_1+X_i\beta_1}}{1+\sum_{m=1}^{M} e^{\alpha_m+X_i\beta_m}}\right]^{y_{1i}}\ldots\left[\frac{e^{\alpha_M+X_i\beta_M}}{1+\sum_{m=1}^{M} e^{\alpha_m+X_i\beta_m}}\right]^{y_{Mi}}$$

where y_{mi} is the disease status of the ith subject, taking value 1 if the subject has disease subtype m, and the other $y_{.i}$'s are equal to 0. m=1,…,M

2.2.2. Markov Chain Monte Carlo and Posterior Calculations

Posterior distribution is the joint density of the parameters when the existence of observed data is taken into account. It is obtained by the product of the likelihood and the joint priors. Then once the joint posterior density of the parameters is obtained, marginal posterior density of each parameter is to be derived to do Bayesian inference for each parameter. This requires integrating out the other variables to get the marginal posterior density of one parameter. That is,

$$f(\boldsymbol{\delta}_j) = \int\int\ldots\int f(\delta_1,\delta_2,\ldots,\delta_M)\,d\delta_1\ldots d\delta_{j-1}d\delta_{j+1}\ldots d\delta_M$$

However, since the analytical derivation of such multiple integrals is obviously cumbersome, some iterative algorithms based on a similar idea for Monte Carlo integration are developed. Through the application of these algorithms one can obtain a sequence of random variables coming from the posterior distribution when the chain satisfies the ergodicity conditions, namely irreducibility, aperiodicity and positive recurrence.

One of these iterative algorithms for posterior distribution estimation is Gibbs sampling and it is first introduced by Geman and Geman (1984). It is derived from the idea of Accept-Reject sampling. It makes use of the full conditional posterior density $f(\delta_j|\delta_{j|},X,Y)$ as the proposal distribution where $\delta_{j|}=(\delta_1,\ldots,\delta_{j-1},\delta_{j+1},\ldots,\delta_M)$. The value generated at each iteration is accepted since those proposal distributions lead to acceptance with probability 1. The advantage of Gibbs sampler is that: since the random values are generated from unidimensional distributions which are the conditional distributions in a known form, it is easy to obtain those values by the help of almost all computational softwares (Ntzoufras, 2009). Since Gibbs sampling does not require to specify the proposal distribution in each step, it is advantegeous together with the ease in computation. However, Gibbs sampler is not effective for the case that parameter space is complicated or parameters have high correlation.

The Gibbs sampling algorithm is as follows (Ntzoufras, 2009):

1) Set initial values for δ: $\delta^{(0)}$

2) For t=1,…,T repeat the following steps:

    a. Set $\delta = \delta^{(t-1)}$

    b. For j=1,…,M, update δ_j by drawing from $f(\delta_{j|}, X, Y)$

    c. Set $\delta^{(t)} = \delta$ and save it as the generated set of values at t+1 iteration of the algorithm.

so, the generated chain for δ in t steps is as the following:

$\delta_1^{(t)}$ from $f(\delta_1 | \delta_2^{(t-1)}, \dots, \delta_M^{(t-1)}, X, Y)$

$\delta_2^{(t)}$ from $f(\delta_2 | \delta_1^{(t)}, \delta_3^{(t-1)} \dots, \delta_M^{(t-1)}, X, Y)$

$\vdots$

$\delta_M^{(t)}$ from $f(\delta_M | \delta_1^{(t)}, \delta_3^{(t)} \dots, \delta_{M-1}^{(t)}, X, Y)$

Here, $f(\delta_{j|}, X, Y)s$ are called the full conditional likelihoods and can be written as $f\left(\delta_{j|}, X, Y\right) \propto f(\delta | X, Y)$ where all the variables in $f(\delta | X, Y)$ are fixed except δ_j.

In our case, we can write the full conditional likelihood for δ_j as follows:

$$f(\delta_j |\, \delta_1, \dots, \delta_{j-1}, \delta_{j+1}, \dots, \delta_M) = \prod_{\{i:y_{ji}=1\}} \frac{e^{\alpha_j + X_i\beta_j}}{1 + \sum_{m=1}^{M} e^{\alpha_m + X_i\beta_m}} \times \prod_{\{i:y_{ji}\neq 1\}} \frac{1}{1 + \sum_{m=1}^{M} e^{\alpha_m + X_i\beta_m}} \times f(\delta_j)$$

i,j=1,…,n; m=1,…,M

All these calculations are carried out in WinBUGS which is a programming language based software that is used in generating random samples from the posterior distribution of the parameters of a model through Gibbs sampling. After specifying the model, data, priors, chain size, burnin period in WinBUGS, sample coming from the posterior density is generated via Markov Chain Monte Carlo algorithms.

2.3. PSEUDO-CONDITIONAL LIKELIHOOD ESTIMATION IN TWO STAGED POLYTOMOUS LOGISTIC REGRESSION

In epidemiological studies, when the disease characteristic information is available and the effect of factors differs according to the different disease subtypes which are constituted by cross-classifying disease characteristics, one may want to do the analysis in the specific disease characteristic level. For illustration, consider that there are two tumor characteristics each having two categories: *tumor size* (*T1, T2*) and *NA status* (*exist, not exist*). One approach to analyze that kind of data is taking characteristics one by one as a response and constructing logistic regression model for each of them seperately, i.e. building binary logistic regression model on *tumor size* where the link is *ln(P(Tumor size=T2)/ P(Tumor size=T1))* and similarly on *NA status* where the link is *ln(P(NA=exist)/ P(NA=not exist)).* Notice that in this approach all the models are constructed independently from each other.This way of modeling is incapable of taking the account for any possible interaction behaviour existing among the tumor characteristics. That is to say, any association between a covariate and the *tumor size* determined this way will be unadjusted for *NA status*. In order to include that interaction behaviour, one can create a response variable consisting of disease subtypes which are obtained by cross-classifying the levels of disease characteristics, i.e. considering our example, M=2×2=4 disease subtypes are

(*T1, NA exist*), (*T1, NA not exis,*), (*T2, NA exist*), (*T2, NA not exist*) captures the inter-relation between *tumor size* and *NA status.* However, in such a case, number of levels of response variable may be very high as the number of characteristics or number of categories of each characteristics gets larger. This may cause estimation problems due to the insufficient number of cases corresponding to each disease subtype. Chatterjee (2004) developed an efficient method for modelling the response data with multivariate disease characteristics information. The method is based on a two staged modelling approach.

At the first stage of the two staged modelling, a polytomous logistic regression model is constructed to investigate the effects of the covariates on disease subtypes which are obtained by cross-classification. Then, at the second stage, new low-dimensional parameters are obtained through a transformation matrix which holds the relation between the first and the second stage parameters. Estimation of model parameters can be done concentrating on the covariate coefficients of the first stage model. In other words, the intercept parameters of first stage models which hold the odds for baseline disease level can be omitted by leaving them unspecified in the estimation procedure. Therefore this method can be thought as semiparametric, and it is advantageous since the number of parameters to be estimated is reduced. Assume that there are K characteristics of the disease, and kth characteristic has M_k levels. Then, in total, there will be $M= M_1 \times M_2 \times \ldots \times M_K$ disease subtypes obtained by cross-classification of levels of the disease characteristics. Let Y_i be the disease subtype status of the i^{th} subject among n subjects. Y_i takes either one of the M+1 values; $Y_i=0$ if the subject is disease-free and $Y_i=m$ if the subject has m^{th} disease subtype where m=1,…,M. And let X_i be the vector of covariates for the i^{th} subject with px1 dimension. Then, at the first stage, we can write the following classical unstructured polytomous logistic regression model:

$$P(Y_i = m \,|X_i) = \frac{e^{\alpha_m + X_i^T \, \beta_m}}{1+\sum_{m=1}^{M} e^{\alpha_m + X_i^T \, \beta_m}} \tag{2.2}$$

where α_m is the intercept, and β_m is the regression parameter for the disease subtype m. Here, e^{β_m} represents the odds ratio which expresses the association between the covariate and the mth disease subtype relative to the disease-free status. Dealing with p covariates, it is clear that the total number of regression coefficients Q=$M_1 \times M_2 \times \ldots \times M_K \times p$ can easily become too large. This can easily result in estimation problems as some of the disease subtype categories may include only very few or no subject. To overcome such problems that are caused by high dimensional parameter space, Chatterjee (2004) developed a novel approach in which the number of parameters are greatly reduced.

For effective illustration of the two staged method we will assume that there is only one covariate that effects the disease outcome. The same idea can then easily be extended to multi-covariate situations. With one covariate in the logistic regression model, the regression coefficients of the first stage model will be an Mx1 vector that is denoted by $\boldsymbol{\beta}$ =(β_1, β_2,…, β_M). We can also represent these parameters in the form of combinations of disease characteristics levels. For instance, $\{\beta_m\}_{m=1}^{M}$ can be represented as $\{\beta_{i_1 i_2 \ldots i_K}\}_{i_1=1, i_2=1, \ldots, i_K=1}^{M_1 M_2 \ldots M_K}$. By the help of this representation, the relationship between the first and second stage parameters can easily be shown as follows:

$$\beta_{i_1 i_2 \ldots i_K} = \theta^{(0)} + \sum_{k_1=1}^{K} \theta_{k_1(i_{k_1})}^{(1)} + \sum_{k_1=1}^{K} \sum_{k_2>k_1}^{K} \theta_{k_1 k_2(i_{k_1} i_{k_2})}^{(2)} + \ldots + \theta_{12\ldots K(i_1 i_2 \ldots i_K)}^{(K)} \qquad (2.3)$$

where $\theta^{(0)}$ is the regression coefficient for the reference disease subtype, and $\theta^{(1)}$'s represent the first order contrasts, and $\theta^{(2)}$'s represent the second order contrasts and so on.

Representation of these relationships can be done through the illustrative example introduced previously in this section:

Let *tumor size=T1* and *NA=not exist* be the reference levels and let each of them is coded as 1. Also let *tumor size=T2* and *NA=exist* are coded as 2. Then the first stage covariate coefficients are written in terms of second stage parameters as follows:

$$\beta_{11} = \theta^{(0)} + \theta_{1(1)}^{(1)} + \theta_{2(1)}^{(1)}$$

$$\beta_{12} = \theta^{(0)} + \theta^{(1)}_{1(1)} + \theta^{(1)}_{2(2)}$$

$$\beta_{21} = \theta^{(0)} + \theta^{(1)}_{1(2)} + \theta^{(1)}_{2(1)}$$

$$\beta_{22} = \theta^{(0)} + \theta^{(1)}_{1(2)} + \theta^{(1)}_{2(2)}$$

where, $\theta^{(1)}_{1(1)}$ is the parameter for the first category of first characteristic: *tumor size*= *T1*, and $\theta^{(1)}_{1(2)}$ is for the second category of first cgaracteristic: *tumor size=T2*; $\theta^{(1)}_{2(1)}$ is the parameter for the first category of second characteristic: *NA=not exist*, and $\theta^{(1)}_{2(2)}$ is for the second category of second characteristic: *NA= exist*.

Reference level disease subtype is formed by i)choosing one level as reference for each disease characteristic, and ii) the disease subtype identified by these reference categories is the reference level disease subtype. For instance, (*T1*, *not exist*) is the reference disease subtype. Note that for identifiability, level of the $\theta^{(k)}$'s that contains the reference level is to be set at zero, except for $\theta^{(0)}$. That is, $\theta^{(1)}_{1(1)}$ and $\theta^{(1)}_{2(1)}$ are set to be zero.

If we set all first and higher order contrasts, i.e. $\theta^{(k)}$'s to be zero, the odds ratio corresponding to reference level disease subtype, exp($\theta^{(0)}$) will give the common covariate odds ratio that the effect of the covariate is indifferent in levels of the characteristic. That is, e.g. for our illustrative example above, the odds of having a tumor with small size or large size is not different for any change in covariate.

Setting second order contrasts to zero, we have the following model:

$$\beta_{i_1 i_2 \dots i_K} = \theta^{(0)} + \sum_{k_1=1}^{K} \theta^{(1)}_{k_1(i_{k_1})} \tag{2.4}$$

This model assumes that the effect of the covariates to one characteristic is independent of the other characteristic. $\theta^{(1)}_{k_1(i_{k_1})}$ is the log-odds ratio of having i_{k_1}th level of the k_1th characteristics to reference level for a one unit change in the covariate. For instance, $\theta_{2(2)}^{(1)}$ corresponds to the following odds ratio for our illustrative example:

$$e^{\theta_{2(2)}^{(1)}} = \frac{P(NA = 2,\ Size|X+1) \Big/ P(NA = 1, Size|X+1)}{P(NA = 2, Size|X) \Big/ P(NA = 1, Size|X)}$$

Allowing interaction between the disease characteristics, we have the following second order contrast model:

$$\beta_{i_1 i_2 \dots i_K} = \theta^{(0)} + \textstyle\sum_{k_1=1}^{K} \theta_{k_1(i_{k_1})}^{(1)} + \sum_{k_1=1}^{K} \sum_{k_2>k_1}^{K} \theta_{k_1 k_2(i_{k_1} i_{k_2})}^{(2)} \qquad (2.5)$$

In this model, e.g. $\theta_{k1k2(i_{k_1} i_{k_2})}^{(2)}$ is the "log-odds ratio" representing the effect of one unit change in the covariate on a certain disease characteristic for different levels of another characteristic. For example, $\theta_{12(22)}^{(2)}$ for our illustrative example corresponds to the following odds ratio:

$$e^{\theta_{12(22)}^{(2)}} = \frac{\dfrac{P(NA = 2, Size = 2|X+1)/P(NA = 1, Size = 2|X+1)}{P(NA = 2, Size = 2|X)/P(NA = 1, Size = 2|X)}}{\dfrac{P(NA = 2, Size = 1|X+1)/P(NA = 1, Size = 1|X+1)}{P(NA = 2, Size = 1|X)/P(NA = 1, Size = 1|X)}}$$

This is the odds ratio for the association between the covariate and *NA status* for cases with *tumor size*=2 versus the odds ratio for the association between the covariate and *NA status* for cases with *tumor size*=1. If this is equal to 1, then the effect of the covariate on *NA status* does not change with respect to the *tumor size* and the second order contrast model (2.5) reduces to the first order contrast model (2.4).

Estimation of the second stage parameters are accomplished by a novel maximum likelihood procedure called pseudo-conditional likelihood estimation (PCL) method (Chatterjee, 2004). Notice that the large number of disease subtypes results in large number of intercept parameters. In this case the joint maximum likelihood estimation of the intercept parameters and the second stage parameters is likely to be numerically difficult. Since these intercept parameters of the first stage model parameters, namely α_m, are not of scientific interest, Chatterjee considered a conditional likelihood in which the nuisance intercept parameters are vanished. The PCL estimation takes only the covariate coefficient parameters of unstructured polytomous logistic regression into account.

PCL of a case-control data is as follows:

$$L_{PCL} = \prod_{i\epsilon C_1} \frac{e^{X_i^T \beta_{m_i}}}{e^{X_i^T \beta_{m_i}} + \sum_{j\epsilon Co} e^{X_j^T \beta_{m_i}}}$$

where i,j=1,…,n; i≠j; C_0 is the set of nondiseased subjects, C_1 is the set of diseased subjects, d_i is the observed disease subtype of the i^{th} diseased subject. Note that, this likelihood is derived from the model 2.2 and does not include the intercept parameters.

PCL score equations corresponding to the second stage parameters θ are $\frac{\partial L_{PCL}}{\partial \beta} \frac{\partial \beta}{\partial \theta} = 0$. Using the relation between first and second stage parameters $\beta = Z\theta$ where Z is the transformation matrix representing the relation in model 2.3, score equations can be written as $Z^T T_\beta = 0$

where $T_\beta = (T_{\beta_1}^T, \dots, T_{\beta_m}^T)^T$

and $T_{\beta_m} = \sum_{i\epsilon C_1} I(Y_i = m) \times \left\{X_i - \frac{X_i \exp(X_i^T \beta_m) + \sum_{j\epsilon C_o} X_j \exp(X_j^T \beta_m)}{\exp(X_i^T \beta_m) + \sum_{j\epsilon C_o} X_j \exp(X_j^T \beta_m)}\right\}$

Solving the score equations for θ, we obtain the maximum likelihood estimates of second stage parameters, $\hat{\theta}$. Asymptotic Normality of $\hat{\theta}$ is proven by Chatterjee (2004).

CHAPTER III

SIMULATION STUDY

A simulation experiment is designed to compare the performances of the three different approaches, namely MLE, Bayesian estimation and the two stage approach for polytomous logistic regression analysis for disease outcome with subtype information under various different scenarios. More specifically, we designed different case-control studies based on different number of disease characteristics with different number of levels and different sample sizes. We compared the three methods through important measures of reliability of statistical methods, namely bias and mean squared error of the estimators. Also, relative efficiency and asymptotic relative efficiencies of the parameter estimates produced from three methods are compared. The simulation experiments are programmed in MATLAB 7.8.

3.1. DATA GENERATION

Data is generated according to the following procedure: First, we decided that the number of characteristics to be 3 in all of the scenarios. Then, three different cases are considered for the disease characteristics: (1) with levels $M_1=2$, $M_2=2$, $M_3=2$; (2) with levels $M_1=4$, $M_2=4$, $M_3=4$; (3) with levels $M_1=6$, $M_2=6$, $M_3=4$. This way, we had small, middle and large scale disease characteristic scenarios. For each of these scenarios for disease characteristics, we generated samples of size 500, 1000 and 2000 where the number of diseased subjects are equal to the number of disease-free subjects in each of these samples. For ease in illustration we considered a single risk factor. The results of this comparative experiment will carry over to multi risk factor polytomous logistic regression nevertheless. One covariate from standard normal distribution is generated in each of the scenarios. To illustrate the data generation process in full detail, let’s consider the first case: M_1=2, M_2=2, M_3=2.

As the first step of the data generation process, we set the true values of the second stage model parameters, θ’s at the same values in Chatterjee (2004) which provides that the percentage of diseased people in the population is about 10%. These are $\theta^{(0)}$=0.35, $\theta^{(1)}_{2(1)}$=0.15, $\theta^{(1)}_{2(2)}$=0, $\theta^{(1)}_{3(2)}$=0.5. Then, true values of first stage parameters are computed as β_1=0.35, β_2=0.85, β_3=0.35, β_4=0.85, β_5=0.5, β_6=1, β_7=0.5, β_8=1 by the relation shown in model 2.4. To obtain the true values of the intercept parameters, α_1,…, α_M, second order contrasts model is used, whereas the first order contrasts model is used to obtain the true values of the coefficients, β_1,…, β_M. Having the values of polytomous logistic regression parameters and a covariate of size N'x1, we got the probabilities of each of the M=2×2×2=8 disease subtypes and probability of being disease-free for the ith of the N' subject by:

$$p_{mi} = P(Y_i = m \mid X_i) = \frac{e^{\alpha m + X_i^T \beta_m}}{1 + \sum_{m=1}^{M} e^{\alpha m + X_i^T \beta_m}}$$

and $p_{0i} = P(Y_i = 0 \mid X_i) = \frac{1}{1 + \sum_{m=1}^{M} e^{\alpha m + X_i^T \beta_m}}$, m=1,…,8

Disease status has multinomial distribution with the probabilities as specified above. That is to say, i^{th} subject has the disease subtype m with probability p_{mi} and is disease-free with probability p_{oi}. Therefore, by using these probabilities we randomly generated disease subtype status for each of the N' subjects from the multinomial distribution. Now, we have a sample of N' =7000 subjects with response as disease status Y_i, where Y_i=0,…,8, and a continuous covariate generated from N(0,1) where i=1,…, N'. Suppose that n_{case} is the number of diseased people (cases) in the whole sample. We selected disease-free people (controls) as many as the number of the cases, n_{case}, from the sample of size N'. At the end, we have a new sample of size n= n_{case} + $n_{control}$ (where n_{case} = $n_{control}$) with disease status and corresponding covariate information. n_{case} was set as 250, 500 and 1000 to investigate the effect of sample size on estimation. Generating the N' subjects in the first stage and then random selection of cases and controls in the second stage as described above has mainly two advantages: i) this process mimics the real life situation in which there are N' subjects in the population of interest 10% of whom have the disease and n of them are selected for the study, ii) this process enables us to have a control over the percentage of the cases in the sample for the simulation study. The second point is especially important as it allows us to include sufficient number of cases in the study sample when the disease has a low prevalence.

The same procedure is applied for the scenario where characteristic has levels M_1=4, M_2=4, M_3=4 and for the scenario where characteristic has levels M_1=6, M_2=6, M_3=4. In these cases, response variable Y has (4×4×4)+1=65 levels and (6×6×4)+1=145 levels respectively. Note that 1's are added for the category representing the disease-free state. The procedure which is explained in the previous paragraph is the same for these scenarios as well except for the levels of characteristics. True values for the first stage parameters are obtained through model (2.4) by using the true values of the second stage parameters which are set under the idea of keeping disease prevalance at around 50% for both of M=4×4×4 and M=6×6×4 situations.

A summary of the scenarios considered in simulation experiment is presented in Figure 3.1:

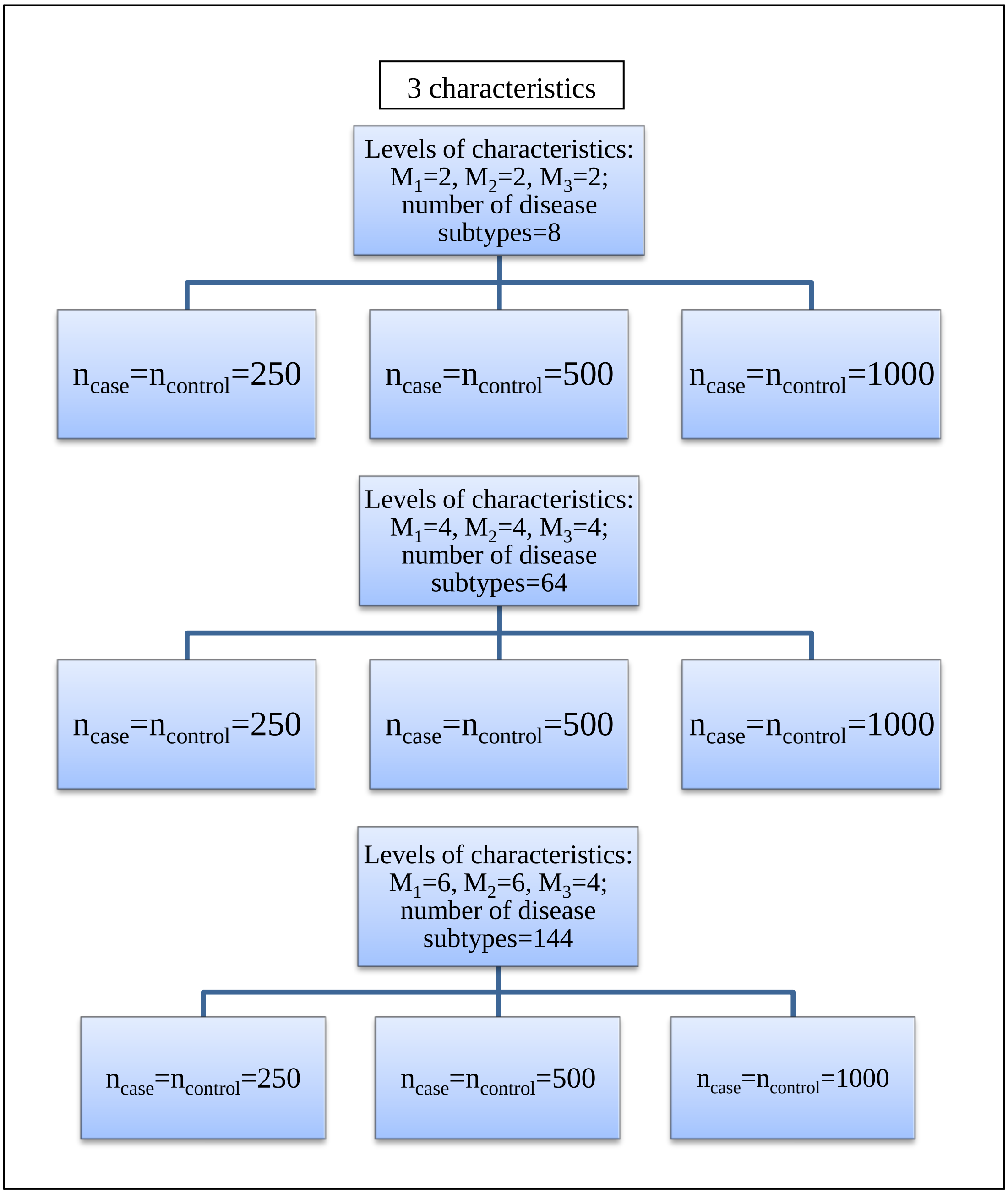

3 characteristics
Levels of characteristics: M_1=2, M_2=2, M_3=2; number of disease subtypes=8
n_{case}=$n_{control}$=250
n_{case}=$n_{control}$=500
n_{case}=$n_{control}$=1000
Levels of characteristics: M_1=4, M_2=4, M_3=4; number of disease subtypes=64
n_{case}=$n_{control}$=250
n_{case}=$n_{control}$=500
n_{case}=$n_{control}$=1000
Levels of characteristics: M_1=6, M_2=6, M_3=4; number of disease subtypes=144
n_{case}=$n_{control}$=250
n_{case}=$n_{control}$=500
n_{case}=$n_{control}$=1000

Figure 3.1: Summary of data generation scenarios.

3.2. PARAMETER ESTIMATION

In the next subsections, the disease subtype scenario with characteristic levels M_1=2, M_2=2, M_3=2 will be used for clear explanation of the details when needed. The parameter estimates are obtained from the three methods and simulation specific aspects of these methods are stated below.

3.2.1. Maximum Likelihood Estimation

For maximum likelihood estimation, *mnrfit* function in MATLAB is used. *mnrfit(X,Y)* is the function to carry out polytomous logistic regression in MATLAB in frequentist aspect. It takes the covariates and the categorical response as input, and gives parameter estimates and some statistics corresponding to that estimates such as standard error estimates, t-test statistics to test significance of parameters, p-values, estimate of variance-covariance matrix of regression coefficient estimators etc. For our sample case, since the response variable takes values 0 to 8 and we consider one single covariate, 8 intercept and 8 covariate coefficient estimates are obtained from the models built on samples N times where N is the number of simulated data sets, i.e. number of Monte Carlo simulation iterations. However, since the intercept parameters are not of any specific interest, the parameters of our focus are the coefficients of the covariates.

3.2.2. Bayesian Estimation

Bayesian estimation of the model parameters are carried out in WinBUGS.As said earlier we programmed the simulation in MATLAB. To perform the Bayesian estimation for each simulated data set, we called WinBUGS from within MATLAB through the use of *mat2bugs* function. As it is stated in chapter 2, WinBUGS generates Markov chain for each parameter in the model using Gibbs sampling. Markov chain is a sequence of random variables which has serial correlation within. In a Markov chain, the state at time (iteration) t is only dependent on the previous observation that is at time (iteration) t-1 and conditionally independent of the earlier iterations given the one at time t-1. After an adequate number of iterations B, distributions of the values of the chain will converge to the equilibrium distribution (in our case the posterior distribution) by Ergodic Theorem as long as the certain regularity conditions are satisfied by the chain (see e.g. Gilks et al. (1996) for further details on these conditions). After discarding the first B values from the chain, we get the convergent values, that is the values which are random draws from the equilibrium distribution. The beauty of the Markov chains is that, once the convergence is attained at time B, the chain from that point on is oblivious to the starting values. For our scenario with one single covariate and M=2×2×2=8 disease subtypes we have $\boldsymbol{\alpha}$ as a 1x8 vector of intercepts and $\boldsymbol{\beta}$ as 1x8 vector of covariate coefficients polytomous logistic regression model. Initial values for the two chains for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are set to different values at the beginning so that for each model parameter two different chains are constructed. For the first chain, we set $\boldsymbol{\alpha} = [0,0,0,0,0,0,0,0]$ and $\boldsymbol{\beta} = [0,0,0,0,0,0,0,0]$, whereas for the second chain $\boldsymbol{\alpha} = [0.3, 1.0, 2.2, 0.6, 1.4, 1.6, 2, 3.0]$ and $\boldsymbol{\beta} = [3, 3.2, 5, 2, 6, 2.2, 1.7, 1.0]$. To decide on the time point at which the convergence is attained, or in other words, the burnin period that is the chain from the initial point to the beginning of the convergence, we can look at the trace plots obtained in WinBUGS. In order to determine the sufficient chain size and burnin point, we used a pilot sample which is generated according to the scenario with character levels M_1=2, M_2=2, M_3=2; n_{case}=250; one standard normal covariate. Trace plots showed that for covariate coefficients, the convergence is attained at about 500th iteration for all of the 8 parameters: $\boldsymbol{\beta} = \beta_1, \dots, \beta_8$ (Figure B.1). Hence the burnin period in WinBUGS is decided to be the first 500 iterations; that is, first 500 iterations are discarded and not used for the posterior inference. We also conducted an exploratory analysis to determine the length of the chain. Monte Carlo (MC) error in WinBUGS, a measure of the variability of the estimate that takes the correlation within the chain into the account, is a criterion for determining the chain size. Larger MC error implies a need for having more Gibbs sampling iterations, i.e. a longer chain. Though this method is subjective, it is still a big help to the practitioner to determine the length of the chain (Table B.1 and Table B.2). Brooks-Gelman-Rubin plots are used for determining the chain size in MCMC processes. For each of the parameters, plots are stabilized at 1000th iteration, so that in our simulation design, chain size is entered as 2500. After extracting first 500 values, we had a chain of size 2000.

3.2.3. Pseudo-Conditional Likelihood Estimation

Second stage regression parameters are obtained by the MATLAB codes originally written by Chatterjee (2004). Recall that, we have disease subtype and one covariate information in the datasets generated by the simulations. For pseudo-conditional likelihood estimation of second stage parameters, we need to specify which subtype is corresponding to which character levels. As an example, again consider that there are three characteristics where each has 2 levels resulting in M=2×2×2=8 disease subtypes. In the datasets, response variable Y takes values from 0 to 8 where 0 indicates the disease-free state (i.e. our controls) . These values correspond to the following levels:

Table 3.1: Disease subtype categories for 3 characteristics with levels M_1=2, M_2=2, M_3=2

(Disease Subtype)	**Characteristic 1**	**Characteristic 2**	**Characteristic 3**
0	-	-	-
1	0	0	0
2	0	0	1
3	0	1	0
4	0	1	1
5	1	0	0
6	1	0	1
7	1	1	0
8	1	1	1

For each of the N=1000 Monte Carlo simulations, we first obtained the second stage parameters **θ**'s, then via model 2.4 first stage coefficients are obtained in order to make the results comparable with the previous two approaches. True **θ**'s and the corresponding true **β**'s are given in section 3.1 for M_1=2, M_2=2, M_3=2 case. And for the other characteristic scenarios, i.e. for M_1=4, M_2=4, M_3=4 and M_1=6, M_2=6, M_3=4, true **β**'s are in tables 3.5 and 3.8 respectively.

3.3. RESULTS

Data generation and coefficient estimation procedures are explained in sections 3.1 and 3.2. Following measures to compare the performances of three methods through the coefficient parameter estimates (**β**'s) are obtained for each of the scenarios (Figure 3.1) and displayed in simulation summary tables (Table 3.2 to Table 3.10).

Monte Carlo Averages: $\hat{E}(\hat{\beta}) = \bar{\hat{\beta}}_j = \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_{ij}$, j=1,…,M

Bias: $\widehat{Bias}\,(\hat{\beta}_j) = \left[\bar{\hat{\beta}}_j - \beta_j\right]$

Mean Square Errors: $\widehat{MSE}(\hat{\beta}_j) = (\bar{\hat{\beta}}_j - \beta_j)^2 + \frac{1}{N-1}\sum_{i=1}^{N}(\hat{\beta}_{ij} - \bar{\hat{\beta}}_j)^2$

Monte Carlo Standard Errors: $SE(\hat{\beta}_j) = \frac{1}{\sqrt{N-1}}\sqrt{\sum_{i=1}^{N}(\hat{\beta}_{ij} - \bar{\hat{\beta}}_j)^2}$

Asymptotic standard errors (est(se)): $\overline{SE(\hat{\beta}_j)} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{Var(\hat{\beta}_{ij})}$

where $Var(\hat{\beta}_j)$ is the asymptotic variance of $\hat{\beta}_j$ which corresponds to the jth diagonal element of the estimated variance-covariance matrix which is the inverse of the Fisher's information matrix.

In the simulation study, we considered 3 scenarios for disease subtypes and 3 different sample sizes. This results in 9 different scenarios in total. There are 9 tables in this subsection displaying the finite sample properties of the estimators under each of these scenarios. We summarize the findings from our simulation based investigation in the following three aspects:

1. The relative performance of the methods as the sample size increases for a fixed number of disease subtypes,
2. The relative performance of the methods as the number of disease subtypes increases for a fixed sample size,

3. The comparison of the estimators in terms of the sample sizes at which the asymptotics hold.

Aspect 1: The simulation results from the point of Aspect 1 view are as follows

- Tables 3.2-3.4 correspond to the disease with 3 characteristics with 2 levels in each, namely 2×2×2. This represents a disease with small number of characteristics and small number of levels for each characteristic. When the disease under investigation is of this type, Pseudo Conditional Likelihood (PCL) estimation of two stage logistic regression performs the best for all types of sample sizes considered in terms of estimating the β parameters. Bias, MSE, Monte Carlo standard errors and averages of standard error estimates of coefficients are the smallest in PCL results. Comparing MLE and Bayesian methods, we observe that MLE performs slightly better in terms of MSE. When the total number of disease subtypes is small, the efficiency of PCL estimators remain superior than the other estimator types for all sample sizes from small to large.

- Tables 3.5-3.7 correspond to the disease with 3 characteristics with 4 levels in each, namely 4×4×4. This represents a disease with a small number of characteristics and moderate number of levels for each characteristic. In this case, the number of disease response categories in the first-stage model is 64 and the total number of first-stage parameters with a single covarate in the model is 64 (for regression coefficients) + 64 (for intercepts) = 128. When this is the case, it is not efficient to obtain Bayesian estimates since these estimates are computed iteratively in WinBUGS and we observed that the computations took a very long time. For this reason, it is not practical to run Bayesian analysis in the simulation study. However one should note that it is still suitable to consider Bayesian method for the analysis of a single data set in real life applications. From this point on, we will frame our comparison focus to PCL and MLE. When the sample size is small, e.g. n=500 (n_{case}=250), PCL estimators have smaller MSE then MLE estimators. As the sample size increases (i.e. as the sample size becomes sufficiently large), MLE is unbeatable in general. Overall, we can say that, when the total number of disease subtypes is large with respect to the sample size, PCL gives better estimators. However, as the sample size gets larger, MLE outperforms in terms of accuracy and efficiency of the estimators. In other words, when the proportion of number of parameters to be estimated in polytomous logistic regression model to the sample size is high, PCL performs better.

- Tables 3.8-3.10 correspond to the disease with 3 characteristics with 6, 6, and 4 levels respectively in each, namely 6×6×4. This represents a disease with a small number of characteristics but a large number of levels for each characteristic resulting in a very large number of disease subtypes, namely 6×6×4=144. In this case, both MLE and PCL estimators have a large bias and large standard error. The bias and the standard error decrease with increasing sample size. It seems that one would need quite a large sample for the estimators to have tolerable bias when the total number of disease subtypes is so large. Another observation possibly derived from these tables is that for all the sample size situations, PCL estimators have smaller bias, whereas ML estimates have better results in terms of measures related to the variation. That is, MC standard errors and average standard error estimates of coefficients are smaller with MLE procedure. These results imply that PCL makes more accurate point estimation with large variation, whereas expected ML estimates are far from the true value yet with small variation. However, it should be noted that for a few number of β_j's in large sample size scenarios, MC standard errors of ML estimators are higher than the MC standard errors of PCL estimators. Investigating the reason behind this situation, we were only able to justify that as the number of observations corresponding to disease category increases then the MC standard error of corresponding estimator decreases for PCL and MLE both.

Aspect 2: The simulation results from the point of Aspect 2 view are as follows

- For a relatively small sample size, e.g. n=500 (n_{case}=250), PCL estimation performs better for small to moderate size of total disease subtypes. When the total number of disease subtypes is large both PCL and MLE performs inefficiently in this case.

- Similar results are also observed for other sample sizes investigated, namely n=1000 (n_{case}=500) and n=2000 (n_{case}=1000).

Aspect 3: The simulation results from the point of Aspect 3 view are as follows

- In the tables, the est(se) and MC error being close implies that the variability of the estimator, i.e. $Var(\hat{\beta})$ can be approximately estimated by the asymptotic variance formula. That is the sampling variance of the estimator attains to its asymptotic variance for the given sample size. Note that the values of est(se) and MC error are a little closer for PCL than they are for MLE for all the sample sizes considered here. That means, the sampling variance of the β estimators based on PCL in two stage logistic regression converges to the asymptotic variances a little faster than the β estimators based on MLE in classical polytomous logistic regression

In addition, efficiency analysis is conducted to see the relative variation of estimators obtained by PCL estimation to MLE. To do this, relative efficiency and asymptotic relative efficiency for each of the parameter estimator is computed. Relative efficiency is a measure for the variability of one estimator to the another, and can be estimated in a Monte Carlo simulation study by using emprical variances of estimators (Li et al, 2001):

$$RE = \frac{Var(\hat{\beta}_j)_{PCL}}{Var(\hat{\beta}_j)_{MLE}}$$

where $Var(\hat{\beta}_j) = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{\beta}_{ij} - \bar{\hat{\beta}}_j)^2$ is the emprical variance of the estimator $\hat{\beta}_j$. i=1…N, j=1…M

Asymptotic relative efficiency also measures the relative variation between two estimator, comparing the asymptotic variances of the estimators :

$$ARE = \frac{AVar(\hat{\beta}_j)_{PCL}}{AVar(\hat{\beta}_j)_{MLE}}$$

where $AVar(\hat{\beta}_j) = \frac{1}{N}\sum_{i=1}^{N} Var(\hat{\beta}_{ij})$ is the asymptotic variance of the estimators $\hat{\beta}_j$, i=1…N, j=1…M

If relative efficiency or asymptotic relative efficiency is smaller than 1, then PCL estimator is more efficient, since the variance of PCL estimator is smaller than the ML estimator. Table A.1 gives the relative and asymptotic relative efficiencies for the M=2×2×2 case. For all sample size scenarios, PCL is more efficient than the MLE. For M=4×4×4 and M=6×6×4, as the sample size increases, MLE becomes more efficient in general. However, for large sample size scenarios (n=1000, n=2000) for M=6×6×4, we observed that comparison between PCL and MLE in terms of RE is not uniform over all the β's. For some β's PCL is superior than MLE.

Table 3.2: Simulation summary for the scenario: 3 disease characteristics with levels M_1=2, M_2=2, M_3=2; n_{case}=$n_{control}$=250

Characteristic level: 2x2x2, ncase=250											
	MLE						**Bayesian**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
1	0.35	0.3581	0.1639	0.1819	0.0081	0.0331	0.3539	0.1663	0.1851	0.0039	0.0343
2	0.85	0.8816	0.2083	0.2367	0.0316	0.057	0.9097	0.2053	0.2449	0.0597	0.0635
3	0.35	0.36	0.2252	0.2616	0.01	0.0685	0.3653	0.228	0.2652	0.0153	0.0705
4	0.85	0.8731	0.2268	0.2647	0.0231	0.0706	0.89	0.2184	0.2714	0.04	0.0753
5	0.5	0.5205	0.2223	0.2642	0.0205	0.0702	0.5188	0.2254	0.2654	0.0188	0.0708
6	1	1.039	0.2208	0.248	0.039	0.063	1.0528	0.2147	0.253	0.0528	0.0668
7	0.5	0.5337	0.244	0.2826	0.0337	0.081	0.5377	0.2398	0.2885	0.0377	0.0847
8	1	1.0244	0.1939	0.2294	0.0244	0.0532	1.0477	0.1915	0.2366	0.0477	0.0583
	PCL										
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**					
1	0.35	0.3552	0.1383	0.1483	0.0052	0.022					
2	0.85	0.8631	0.1747	0.1902	0.0131	0.0363					
3	0.35	0.3537	0.1604	0.1753	0.0037	0.0308					
4	0.85	0.8615	0.1799	0.1983	0.0115	0.0395					
5	0.5	0.5148	0.1636	0.1881	0.0148	0.0356					
6	1	1.0226	0.1877	0.2161	0.0226	0.0472					
7	0.5	0.5132	0.1691	0.1951	0.0132	0.0383					
8	1	1.0211	0.1796	0.2093	0.0211	0.0443					

Table 3.3: Simulation summary for the scenario: 3 disease characteristics with levels M_1=2, M_2=2, M_3=2; n_{case}=$n_{control}$=500

Characteristic level: 2x2x2, ncase=500										
	MLE						**Bayesian**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
1	0.35	0.3563	0.115	0.147	0.0063	0.0216	0.3515	0.1153	0.1481	0.0015	0.0219
2	0.85	0.8707	0.1454	0.1897	0.0207	0.0364	0.8857	0.1404	0.194	0.0357	0.0389
3	0.35	0.362	0.1565	0.1991	0.012	0.0398	0.3654	0.1598	0.2008	0.0154	0.0406
4	0.85	0.8574	0.1576	0.2008	0.0074	0.0404	0.8593	0.1522	0.2021	0.0093	0.0409
5	0.5	0.5112	0.1547	0.1953	0.0112	0.0383	0.5079	0.1558	0.195	0.0079	0.0381
6	1	1.0203	0.1536	0.1921	0.0203	0.0373	1.0244	0.1506	0.1936	0.0244	0.0381
7	0.5	0.5151	0.1688	0.2189	0.0151	0.0481	0.516	0.1645	0.2214	0.016	0.0493
8	1	1.0182	0.136	0.1763	0.0182	0.0314	1.0304	0.1339	0.1793	0.0304	0.0331
	PCL										
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**					
1	0.35	0.3542	0.0972	0.1201	0.0042	0.0145					
2	0.85	0.8586	0.1243	0.1435	0.0086	0.0207					
3	0.35	0.3537	0.1122	0.1375	0.0037	0.0189					
4	0.85	0.8582	0.1278	0.1492	0.0082	0.0223					
5	0.5	0.509	0.1147	0.1415	0.009	0.0201					
6	1	1.0135	0.1341	0.1584	0.0135	0.0253					
7	0.5	0.5086	0.1185	0.1482	0.0086	0.022					
8	1	1.013	0.1288	0.1557	0.013	0.0244					

Table 3.4: Simulation summary for the scenario: 3 disease characteristics with levels M_1=2, M_2=2, M_3=2; n_{case}=$n_{control}$=1000

Characteristic level: 2x2x2, ncase=1000											
	MLE						**Bayesian**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
1	0.35	0.3509	0.0998	0.0807	0.0009	0.01	0.3489	0.1001	0.0801	-0.0011	0.01
2	0.85	0.8598	0.1203	0.1019	0.0098	0.0146	0.8681	0.1219	0.0983	0.0181	0.0152
3	0.35	0.3511	0.1333	0.1095	0.0011	0.0178	0.352	0.1326	0.1114	0.002	0.0176
4	0.85	0.8668	0.1367	0.111	0.0168	0.019	0.8689	0.137	0.1069	0.0189	0.0191
5	0.5	0.5066	0.1265	0.1083	0.0066	0.016	0.5034	0.1257	0.1097	0.0034	0.0158
6	1	1.0075	0.1265	0.1069	0.0075	0.0161	1.0084	0.1265	0.1026	0.0084	0.0161
7	0.5	0.5024	0.1463	0.1182	0.0024	0.0214	0.5006	0.1468	0.1153	0.0006	0.0215
8	1	1.0065	0.1082	0.0949	0.0065	0.0117	1.0122	0.1085	0.0925	0.0122	0.0119
	PCL										
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**					
1	0.35	0.3511	0.0833	0.0683	0.0011	0.0069					
2	0.85	0.8584	0.101	0.0882	0.0084	0.0103					
3	0.35	0.3514	0.0955	0.0788	0.0014	0.0091					
4	0.85	0.8587	0.1049	0.0908	0.0087	0.0111					
5	0.5	0.5002	0.094	0.0805	0.0002	0.0088					
6	1	1.0076	0.1053	0.0951	0.0076	0.0111					
7	0.5	0.5006	0.0992	0.0832	0.0006	0.0098					
8	1	1.0079	0.1035	0.0916	0.0079	0.0108					

Table 3.5: Simulation summary for the scenario: 3 disease characteristics with levels M_1=4, M_2=4, M_3=4; n_{case}=$n_{control}$=250

Characteristic level: 4x4x4, ncase=250												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
1	-3.5	-3.9262	1.2424	1.0804	-0.4262	1.3489		-3.8685	0.7481	0.9626	-0.3685	1.0624
2	-3.49	-3.9784	1.2167	1.0565	-0.4884	1.3548		-3.8417	0.7486	0.9535	-0.3517	1.0328
3	-3.48	-3.9614	1.3229	1.0737	-0.4814	1.3847		-3.8519	0.7477	0.9574	-0.3719	1.0549
4	-3.9	-4.5226	2.6227	1.0812	-0.6226	1.5567		-4.3264	0.9045	1.1025	-0.4264	1.3972
5	-3.48	-3.9462	1.1472	1.0392	-0.4662	1.2972		-3.833	0.741	0.9541	-0.353	1.0349
6	-3.47	-3.9017	0.9521	1.0721	-0.4317	1.3359		-3.8063	0.7386	0.945	-0.3363	1.006
7	-3.46	-3.9059	1.1229	1.0219	-0.4459	1.2432		-3.8164	0.737	0.9473	-0.3564	1.0244
8	-3.88	-4.4967	0.9829	1.0583	-0.6167	1.5003		-4.2909	0.8939	1.0891	-0.4109	1.355
9	-3.5	-3.9851	0.9337	1.0949	-0.4851	1.434		-3.8754	0.7535	0.9943	-0.3754	1.1295
10	-3.49	-3.9056	1.4885	1.1427	-0.4156	1.4785		-3.8487	0.7581	0.9934	-0.3587	1.1155
11	-3.48	-3.9747	1.4437	1.126	-0.4947	1.5126		-3.8588	0.7557	0.9807	-0.3788	1.1053
12	-3.9	-4.5584	1.2887	1.2263	-0.6584	1.9371		-4.3333	0.9121	1.1333	-0.4333	1.4722
13	-3.4	-3.8389	1.4169	1.0768	-0.4389	1.3522		-3.7425	0.7173	0.92	-0.3425	0.9637
14	-3.39	-3.8315	1.0346	1.1295	-0.4415	1.4708		-3.7157	0.722	0.9232	-0.3257	0.9583
15	-3.38	-3.7936	1.0101	1.1003	-0.4136	1.3818		-3.7259	0.7203	0.92	-0.3459	0.9661
16	-3.8	-4.3496	5.9692	1.3276	-0.5496	2.0646		-4.2004	0.8865	1.0663	-0.4004	1.2974
17	-3.5	-3.9205	1.0744	1.0762	-0.4205	1.3351		-3.8656	0.7707	0.9832	-0.3656	1.1003
18	-3.49	-3.9372	3.421	1.3015	-0.4472	1.894		-3.8388	0.7912	0.996	-0.3488	1.1136
19	-3.48	-3.9812	5.1201	1.3622	-0.5012	2.1068		-3.8489	0.7896	1.0054	-0.3689	1.1469
20	-3.9	-4.4756	4.7158	1.3727	-0.5756	2.2156		-4.3235	0.9393	1.1519	-0.4235	1.5061
21	-3.48	-3.918	0.9281	0.9629	-0.438	1.1189		-3.8301	0.7562	0.9804	-0.3501	1.0837
22	-3.47	-3.8744	3.2026	1.2567	-0.4044	1.7428		-3.8033	0.7743	0.9933	-0.3333	1.0977

Continuation of Table 3.5.

Characteristic level: 4x4x4, ncase=250												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
23	-3.46	-3.9897	2.0748	1.2064	-0.5297	1.736		-3.8135	0.7716	1.0012	-0.3535	1.1273
24	-3.88	-4.5098	3.2775	1.2067	-0.6298	1.8526		-4.288	0.9224	1.1438	-0.408	1.4747
25	-3.5	-3.9723	7.9519	1.2669	-0.4723	1.828		-3.8725	0.792	1.0096	-0.3725	1.158
26	-3.49	-4.1948	15.8982	1.3125	-0.7048	2.2193		-3.8457	0.8157	1.0297	-0.3557	1.1869
27	-3.48	-4.203	14.9343	1.4234	-0.723	2.5488		-3.8559	0.8125	1.0231	-0.3759	1.188
28	-3.9	-4.5695	13.8667	1.2674	-0.6695	2.0545		-4.3304	0.9596	1.1775	-0.4304	1.5717
29	-3.4	-3.8127	1.2894	1.1422	-0.4127	1.475		-3.7396	0.7434	0.9395	-0.3396	0.998
30	-3.39	-3.8437	4.3344	1.3811	-0.4537	2.1133		-3.7128	0.7689	0.965	-0.3228	1.0355
31	-3.38	-3.9701	4.7018	1.2878	-0.5901	2.0065		-3.723	0.766	0.9679	-0.343	1.0545
32	-3.8	-4.3805	17.2649	1.3053	-0.5805	2.0408		-4.1975	0.9239	1.1156	-0.3975	1.4026
33	-3.4	-3.8348	0.8676	1.0891	-0.4348	1.3751		-3.7429	0.7232	0.928	-0.3429	0.9788
34	-3.39	-3.8685	3.7409	1.2627	-0.4785	1.8233		-3.7161	0.741	0.9383	-0.3261	0.9868
35	-3.38	-3.8278	0.866	1.0038	-0.4478	1.2081		-3.7263	0.7194	0.9185	-0.3463	0.9636
36	-3.8	-4.3695	0.9472	1.0811	-0.5695	1.4931		-4.2008	0.8736	1.0556	-0.4008	1.2748
37	-3.38	-3.8624	7.9333	1.3134	-0.4824	1.9579		-3.7074	0.7366	0.9436	-0.3274	0.9976
38	-3.37	-3.9957	9.7011	1.3962	-0.6257	2.3409		-3.6806	0.7515	0.9539	-0.3106	1.0063
39	-3.36	-3.7811	2.5228	1.2751	-0.4211	1.8032		-3.6908	0.7292	0.9328	-0.3308	0.9795
40	-3.78	-4.2887	6.4995	1.3062	-0.5087	1.9649		-4.1653	0.8798	1.0632	-0.3853	1.2789
41	-3.4	-3.8069	1.2492	1.0231	-0.4069	1.2122		-3.7498	0.7272	0.9511	-0.3498	1.0269
42	-3.39	-3.926	4.6842	1.2812	-0.536	1.9286		-3.7231	0.7493	0.9693	-0.3331	1.0505
43	-3.38	-3.7847	1.0239	1.0312	-0.4047	1.2272		-3.7332	0.7261	0.9329	-0.3532	0.9951
44	-3.8	-4.3391	1.5605	1.1351	-0.5391	1.5791		-4.2077	0.8803	1.0791	-0.4077	1.3308

Continuation of Table 3.5.

Characteristic level: 4x4x4, ncase=250											
	MLE							**PCL**			
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
45	-3.3	-3.7469	0.8448	1.0079	-0.4469	1.2157		-3.6169	0.6929	0.8871	-0.3169	0.8874
46	-3.29	-3.7102	5.7931	1.3022	-0.4202	1.8723		-3.5901	0.7165	0.9108	-0.3001	0.9196
47	-3.28	-3.6994	0.8443	1.0301	-0.4194	1.237		-3.6003	0.6927	0.8829	-0.3203	0.8821
48	-3.7	-4.2529	3.8059	1.317	-0.5529	2.0403		-4.0748	0.8563	1.0207	-0.3748	1.1822
49	-3.47	-3.9514	0.9416	1.0601	-0.4814	1.3555		-3.8229	0.7308	0.9321	-0.3529	0.9934
50	-3.46	-3.8913	0.9514	1.0477	-0.4313	1.2836		-3.7961	0.7287	0.9243	-0.3361	0.9673
51	-3.45	-3.9252	0.9293	1.0137	-0.4752	1.2535		-3.8063	0.7277	0.9228	-0.3563	0.9784
52	-3.87	-4.4447	1.5641	1.1167	-0.5747	1.5774		-4.2808	0.8845	1.0776	-0.4108	1.3301
53	-3.45	-3.8795	3.8054	1.0548	-0.4295	1.2971		-3.7874	0.724	0.9272	-0.3374	0.9736
54	-3.44	-3.8513	0.9475	0.974	-0.4113	1.1178		-3.7607	0.719	0.9194	-0.3207	0.9482
55	-3.43	-3.9261	0.8761	0.9958	-0.4961	1.2376		-3.7708	0.7171	0.9163	-0.3408	0.9557
56	-3.85	-4.3468	1.0074	1.0591	-0.4968	1.3684		-4.2453	0.8742	1.0674	-0.3953	1.2955
57	-3.47	-3.9071	0.8919	1.0072	-0.4371	1.2055		-3.8298	0.7326	0.9462	-0.3598	1.0247
58	-3.46	-4.006	0.9083	1.0608	-0.546	1.4233		-3.8031	0.7349	0.9469	-0.3431	1.0142
59	-3.45	-3.9094	1.0693	1.0271	-0.4594	1.266		-3.8132	0.7321	0.928	-0.3632	0.9931
60	-3.87	-4.396	1.688	1.0955	-0.526	1.4768		-4.2877	0.8895	1.093	-0.4177	1.3692
61	-3.37	-3.7637	0.8283	1.0234	-0.3937	1.2023		-3.6969	0.6976	0.8863	-0.3269	0.8923
62	-3.36	-3.7383	0.9998	1.0204	-0.3783	1.1844		-3.6701	0.6998	0.8912	-0.3101	0.8905
63	-3.35	-3.7516	0.9425	0.9812	-0.4016	1.124		-3.6803	0.6977	0.8821	-0.3303	0.8873
64	-3.77	-4.3143	3.1867	1.263	-0.5443	1.8915		-4.1548	0.8648	1.0391	-0.3848	1.2279

Table 3.6: Simulation summary for the scenario: 3 disease characteristics with levels M_1=4, M_2=4, M_3=4; n_{case}=$n_{control}$=500

Characteristic level: 4x4x4, ncase=500												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
1	-3.5	-3.7329	0.5704	0.6606	-0.2329	0.4906		-3.8089	0.5417	0.764	-0.3089	0.6792
2	-3.49	-3.7068	0.5723	0.7273	-0.2168	0.576		-3.8097	0.5437	0.7712	-0.3197	0.6969
3	-3.48	-3.7174	0.5768	0.6843	-0.2374	0.5246		-3.7962	0.5392	0.751	-0.3162	0.664
4	-3.9	-4.2104	0.6224	0.6927	-0.3104	0.5762		-4.2929	0.6589	0.8898	-0.3929	0.9461
5	-3.48	-3.7218	0.566	0.6949	-0.2418	0.5414		-3.7865	0.5372	0.756	-0.3065	0.6655
6	-3.47	-3.7327	0.5377	0.6514	-0.2627	0.4933		-3.7873	0.5377	0.7684	-0.3173	0.6912
7	-3.46	-3.7017	0.5446	0.6513	-0.2417	0.4826		-3.7738	0.5331	0.746	-0.3138	0.655
8	-3.88	-4.1607	0.5639	0.6555	-0.2807	0.5084		-4.2705	0.6531	0.8856	-0.3905	0.9368
9	-3.5	-3.7167	0.576	0.7161	-0.2167	0.5597		-3.812	0.5436	0.7718	-0.312	0.693
10	-3.49	-3.6952	0.6538	0.7283	-0.2052	0.5725		-3.8128	0.5475	0.7862	-0.3228	0.7223
11	-3.48	-3.735	0.6017	0.719	-0.255	0.582		-3.7993	0.543	0.7643	-0.3193	0.6861
12	-3.9	-4.1997	0.6836	0.7536	-0.2997	0.6577		-4.296	0.6621	0.9019	-0.396	0.9702
13	-3.4	-3.6353	0.5728	0.6785	-0.2353	0.5157		-3.6939	0.5205	0.7359	-0.2939	0.6279
14	-3.39	-3.5933	0.5951	0.7176	-0.2033	0.5563		-3.6947	0.5248	0.746	-0.3047	0.6494
15	-3.38	-3.5865	0.5803	0.685	-0.2065	0.5119		-3.6812	0.5201	0.7228	-0.3012	0.6132
16	-3.8	-4.0558	1.038	0.9664	-0.2558	0.9994		-4.1779	0.6448	0.8703	-0.3779	0.9003
17	-3.5	-3.7486	0.5782	0.6748	-0.2486	0.5172		-3.8184	0.5558	0.7932	-0.3184	0.7306
18	-3.49	-3.7115	1.0394	0.889	-0.2215	0.8393		-3.8193	0.5688	0.8046	-0.3293	0.7558
19	-3.48	-3.6892	0.9967	0.8937	-0.2092	0.8425		-3.8057	0.5648	0.7925	-0.3257	0.7342
20	-3.9	-4.2426	1.1312	0.9632	-0.3426	1.0452		-4.3024	0.6801	0.9243	-0.4024	1.0163
21	-3.48	-3.7167	0.5456	0.6383	-0.2367	0.4635		-3.796	0.5471	0.7868	-0.316	0.7189
22	-3.47	-3.677	0.733	0.7645	-0.207	0.6274		-3.7968	0.5589	0.8033	-0.3268	0.7521

Continuation of Table 3.6.

Characteristic level: 4x4x4, ncase=500												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
23	-3.46	-3.6852	0.769	0.8188	-0.2252	0.7212		-3.7833	0.5548	0.7891	-0.3233	0.7272
24	-3.88	-4.2105	0.7497	0.8091	-0.3305	0.7639		-4.28	0.671	0.9214	-0.4	1.0089
25	-3.5	-3.7011	1.0905	0.9948	-0.2011	1.03		-3.8215	0.5672	0.8084	-0.3215	0.7569
26	-3.49	-3.8205	7.4079	1.1931	-0.3305	1.5327		-3.8224	0.5817	0.8265	-0.3324	0.7936
27	-3.48	-3.7091	7.5432	1.2472	-0.2291	1.6079		-3.8088	0.5776	0.8128	-0.3288	0.7688
28	-3.9	-4.2122	5.1124	1.2458	-0.3122	1.6494		-4.3055	0.6909	0.9425	-0.4055	1.0528
29	-3.4	-3.5979	0.5914	0.6955	-0.1979	0.5229		-3.7034	0.5365	0.7721	-0.3034	0.6883
30	-3.39	-3.6195	1.1106	0.9709	-0.2295	0.9952		-3.7043	0.552	0.7864	-0.3143	0.7172
31	-3.38	-3.6308	1.7328	0.9525	-0.2508	0.9701		-3.6907	0.5478	0.7719	-0.3107	0.6923
32	-3.8	-4.0671	3.8338	1.1659	-0.2671	1.4306		-4.1875	0.6675	0.9107	-0.3875	0.9794
33	-3.4	-3.5786	0.5665	0.6775	-0.1786	0.4908		-3.6841	0.5203	0.7226	-0.2841	0.6029
34	-3.39	-3.5869	0.817	0.8973	-0.1969	0.8439		-3.6849	0.5322	0.7384	-0.2949	0.6322
35	-3.38	-3.6057	0.5421	0.6517	-0.2257	0.4756		-3.6714	0.5163	0.7151	-0.2914	0.5962
36	-3.8	-4.0957	0.5566	0.6245	-0.2957	0.4775		-4.1681	0.6341	0.8477	-0.3681	0.8541
37	-3.38	-3.5835	1.2693	0.9224	-0.2035	0.8922		-3.6617	0.5278	0.7273	-0.2817	0.6083
38	-3.37	-3.569	3.0653	1.1867	-0.199	1.4478		-3.6625	0.5381	0.7483	-0.2925	0.6455
39	-3.36	-3.5771	0.9475	0.8842	-0.2171	0.829		-3.649	0.5224	0.7231	-0.289	0.6064
40	-3.78	-4.0358	0.8413	0.9226	-0.2558	0.9167		-4.1457	0.638	0.8545	-0.3657	0.8638
41	-3.4	-3.601	0.5672	0.6447	-0.201	0.4561		-3.6872	0.522	0.7276	-0.2872	0.6119
42	-3.39	-3.5839	1.2046	0.9193	-0.1939	0.8826		-3.688	0.5357	0.7509	-0.298	0.6527
43	-3.38	-3.6047	0.6087	0.6615	-0.2247	0.488		-3.6745	0.5199	0.7259	-0.2945	0.6136
44	-3.8	-4.0972	0.5956	0.7004	-0.2972	0.5789		-4.1712	0.6372	0.8577	-0.3712	0.8734
45	-3.3	-3.5131	0.5472	0.6749	-0.2131	0.5008		-3.5691	0.5	0.7024	-0.2691	0.5658

Continuation of Table 3.6.

Characteristic level: 4x4x4, ncase=500												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
46	-3.29	-3.4701	1.0456	0.907	-0.1801	0.855		-3.5699	0.5147	0.7214	-0.2799	0.5988
47	-3.28	-3.4679	0.5557	0.6624	-0.1879	0.474		-3.5564	0.4981	0.6952	-0.2764	0.5597
48	-3.7	-3.9705	0.8203	0.9174	-0.2705	0.9148		-4.0531	0.6208	0.8353	-0.3531	0.8225
49	-3.47	-3.7495	0.5738	0.6897	-0.2795	0.5537		-3.7751	0.5328	0.7501	-0.3051	0.6557
50	-3.46	-3.6865	0.5388	0.6728	-0.2265	0.5039		-3.7759	0.5331	0.7574	-0.3159	0.6734
51	-3.45	-3.675	0.5506	0.675	-0.225	0.5062		-3.7623	0.5292	0.7421	-0.3123	0.6483
52	-3.87	-4.1452	0.5588	0.6753	-0.2752	0.5317		-4.2591	0.6487	0.8772	-0.3891	0.9208
53	-3.45	-3.6728	0.5707	0.6874	-0.2228	0.5221		-3.7526	0.5288	0.7409	-0.3026	0.6405
54	-3.44	-3.6816	0.5151	0.6459	-0.2416	0.4755		-3.7534	0.5276	0.7535	-0.3134	0.666
55	-3.43	-3.6735	0.5173	0.6109	-0.2435	0.4325		-3.7399	0.5235	0.736	-0.3099	0.6377
56	-3.85	-4.1665	0.536	0.6259	-0.3165	0.4919		-4.2366	0.6433	0.872	-0.3866	0.9098
57	-3.47	-3.6657	0.5459	0.6633	-0.1957	0.4783		-3.7782	0.5331	0.7536	-0.3082	0.6629
58	-3.46	-3.6863	0.5431	0.6541	-0.2263	0.479		-3.779	0.5353	0.7683	-0.319	0.692
59	-3.45	-3.7112	0.5437	0.6585	-0.2612	0.5018		-3.7654	0.5313	0.7512	-0.3154	0.6638
60	-3.87	-4.1931	0.5649	0.7036	-0.3231	0.5994		-4.2622	0.6505	0.8857	-0.3922	0.9382
61	-3.37	-3.5692	0.5366	0.6616	-0.1992	0.4774		-3.6601	0.5101	0.7206	-0.2901	0.6034
62	-3.36	-3.6005	0.5395	0.6072	-0.2405	0.4265		-3.6609	0.5127	0.7309	-0.3009	0.6247
63	-3.35	-3.5431	0.5383	0.6586	-0.1931	0.4711		-3.6473	0.5085	0.7127	-0.2973	0.5964
64	-3.77	-3.9965	0.8209	0.84	-0.2265	0.7569		-4.1441	0.6335	0.8567	-0.3741	0.8738

Table 3.7: Simulation summary for the scenario: 3 disease characteristics with levels M_1=4, M_2=4, M_3=4; n_{case}=$n_{control}$=1000

Characteristic level: 4x4x4, ncase=1000												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
1	-3.5	-3.6289	0.3875	0.4974	-0.1289	0.264		-3.7161	0.4131	0.6162	-0.2161	0.4264
2	-3.49	-3.6291	0.3925	0.4711	-0.1391	0.2413		-3.7018	0.4127	0.6188	-0.2118	0.4277
3	-3.48	-3.5923	0.3904	0.4984	-0.1123	0.261		-3.693	0.4109	0.6129	-0.213	0.421
4	-3.9	-4.0822	0.3928	0.4831	-0.1822	0.2666		-4.1715	0.5021	0.722	-0.2715	0.5949
5	-3.48	-3.6045	0.3877	0.4872	-0.1245	0.2529		-3.6959	0.4097	0.6073	-0.2159	0.4155
6	-3.47	-3.5838	0.3723	0.4566	-0.1138	0.2215		-3.6816	0.4085	0.6122	-0.2116	0.4195
7	-3.46	-3.5889	0.3706	0.4745	-0.1289	0.2418		-3.6728	0.4064	0.6044	-0.2128	0.4106
8	-3.88	-4.0558	0.3812	0.4761	-0.1758	0.2576		-4.1513	0.4977	0.7158	-0.2713	0.5859
9	-3.5	-3.6376	0.3915	0.4866	-0.1376	0.2557		-3.7217	0.4155	0.6122	-0.2217	0.4239
10	-3.49	-3.6449	0.4075	0.4935	-0.1549	0.2675		-3.7074	0.4162	0.6169	-0.2174	0.4278
11	-3.48	-3.6396	0.4033	0.5006	-0.1596	0.276		-3.6986	0.4141	0.6061	-0.2186	0.4152
12	-3.9	-4.0822	0.4122	0.5088	-0.1822	0.292		-4.1771	0.5052	0.7202	-0.2771	0.5954
13	-3.4	-3.5163	0.3855	0.4709	-0.1163	0.2353		-3.6024	0.396	0.5945	-0.2024	0.3944
14	-3.39	-3.5119	0.4049	0.5028	-0.1219	0.2677		-3.5881	0.3968	0.5995	-0.1981	0.3987
15	-3.38	-3.4841	0.4058	0.4835	-0.1041	0.2446		-3.5794	0.3949	0.5928	-0.1994	0.3912
16	-3.8	-3.9343	0.5677	0.6477	-0.1343	0.4376		-4.0578	0.4883	0.7033	-0.2578	0.5611
17	-3.5	-3.6528	0.3905	0.4998	-0.1528	0.2731		-3.7296	0.4211	0.6318	-0.2296	0.4518
18	-3.49	-3.6116	0.5146	0.6262	-0.1216	0.4069		-3.7153	0.4267	0.6386	-0.2253	0.4586
19	-3.48	-3.6129	0.5257	0.662	-0.1329	0.456		-3.7066	0.4248	0.6329	-0.2266	0.4519
20	-3.9	-4.0941	0.5288	0.6503	-0.1941	0.4606		-4.185	0.514	0.7433	-0.285	0.6337
21	-3.48	-3.6234	0.3554	0.4542	-0.1434	0.2269		-3.7094	0.4151	0.6188	-0.2294	0.4356
22	-3.47	-3.59	0.4494	0.5475	-0.12	0.3142		-3.6951	0.42	0.628	-0.2251	0.4451

Continuation of Table 3.7.

Characteristic level: 4x4x4, ncase=1000												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
23	-3.46	-3.5831	0.4425	0.5548	-0.1231	0.323		-3.6864	0.4178	0.6204	-0.2264	0.4362
24	-3.88	-4.0438	0.4516	0.5596	-0.1638	0.3399		-4.1648	0.5076	0.7337	-0.2848	0.6194
25	-3.5	-3.6159	0.5708	0.6515	-0.1159	0.4378		-3.7352	0.4286	0.6314	-0.2352	0.454
26	-3.49	-3.5678	1.2169	0.9296	-0.0778	0.8702		-3.7209	0.4351	0.6403	-0.2309	0.4633
27	-3.48	-3.6148	1.1629	0.8869	-0.1348	0.8048		-3.7122	0.4329	0.6299	-0.2322	0.4507
28	-3.9	-4.0426	1.3998	0.9448	-0.1426	0.913		-4.1906	0.5212	0.7446	-0.2906	0.6389
29	-3.4	-3.5413	0.3948	0.4999	-0.1413	0.2699		-3.616	0.4049	0.6081	-0.216	0.4164
30	-3.39	-3.4745	0.5819	0.6755	-0.0845	0.4635		-3.6017	0.4119	0.6175	-0.2117	0.4262
31	-3.38	-3.4922	0.5752	0.6852	-0.1122	0.4821		-3.5929	0.4099	0.611	-0.2129	0.4186
32	-3.8	-3.8714	1.3646	0.9229	-0.0714	0.8569		-4.0714	0.5009	0.7231	-0.2714	0.5965
33	-3.4	-3.5336	0.3857	0.4935	-0.1336	0.2614		-3.6076	0.3971	0.5962	-0.2076	0.3986
34	-3.39	-3.5224	0.5104	0.6337	-0.1324	0.4191		-3.5933	0.4018	0.6016	-0.2033	0.4032
35	-3.38	-3.4878	0.3685	0.4783	-0.1078	0.2404		-3.5846	0.3937	0.592	-0.2046	0.3924
36	-3.8	-3.9585	0.3778	0.4597	-0.1585	0.2365		-4.063	0.484	0.7068	-0.263	0.5688
37	-3.38	-3.4738	0.5475	0.6722	-0.0938	0.4607		-3.5874	0.3998	0.5927	-0.2074	0.3943
38	-3.37	-3.4591	1.0565	0.8756	-0.0891	0.7746		-3.5731	0.4038	0.6004	-0.2031	0.4017
39	-3.36	-3.4637	0.4851	0.6042	-0.1037	0.3759		-3.5644	0.3954	0.5889	-0.2044	0.3886
40	-3.78	-3.923	0.5032	0.6153	-0.143	0.399		-4.0428	0.4846	0.7053	-0.2628	0.5665
41	-3.4	-3.5435	0.3838	0.4734	-0.1435	0.2447		-3.6132	0.3994	0.5953	-0.2132	0.3998
42	-3.39	-3.4899	0.5357	0.6067	-0.0999	0.378		-3.5989	0.4053	0.6029	-0.2089	0.4071
43	-3.38	-3.5192	0.3829	0.4826	-0.1392	0.2523		-3.5902	0.3968	0.5883	-0.2102	0.3903

Continuation of Table 3.7.

Characteristic level: 4x4x4, ncase=1000											
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
44	-3.8	-3.9769	0.3952	0.5025	-0.1769	0.2838		-4.0686	0.487	0.7077	-0.2686	0.573
45	-3.3	-3.4135	0.3784	0.4707	-0.1135	0.2345		-3.4939	0.3804	0.5742	-0.1939	0.3673
46	-3.29	-3.4222	0.5471	0.6524	-0.1322	0.4431		-3.4796	0.3867	0.5822	-0.1896	0.3749
47	-3.28	-3.4032	0.3766	0.4772	-0.1232	0.243		-3.4709	0.3781	0.5717	-0.1909	0.3632
48	-3.7	-3.868	0.4995	0.5681	-0.168	0.351		-3.9494	0.4706	0.6881	-0.2494	0.5357
49	-3.47	-3.6171	0.3878	0.4846	-0.1471	0.2564		-3.697	0.4088	0.6111	-0.227	0.425
50	-3.46	-3.5929	0.3688	0.4718	-0.1329	0.2403		-3.6827	0.4073	0.6124	-0.2227	0.4246
51	-3.45	-3.565	0.3718	0.477	-0.115	0.2408		-3.674	0.4057	0.6079	-0.224	0.4197
52	-3.87	-4.0525	0.3789	0.4611	-0.1825	0.2459		-4.1524	0.4971	0.7202	-0.2824	0.5984
53	-3.45	-3.623	0.3856	0.4708	-0.173	0.2516		-3.6768	0.4053	0.5983	-0.2268	0.4094
54	-3.44	-3.5727	0.3544	0.4515	-0.1327	0.2215		-3.6625	0.403	0.6019	-0.2225	0.4118
55	-3.43	-3.5681	0.3542	0.4396	-0.1381	0.2123		-3.6538	0.4012	0.5954	-0.2238	0.4046
56	-3.85	-4.0426	0.3611	0.4445	-0.1926	0.2347		-4.1323	0.4927	0.7107	-0.2823	0.5848
57	-3.47	-3.6071	0.3669	0.4661	-0.1371	0.2361		-3.7026	0.4102	0.6068	-0.2326	0.4223
58	-3.46	-3.596	0.3716	0.4777	-0.136	0.2467		-3.6883	0.4098	0.6102	-0.2283	0.4245
59	-3.45	-3.5922	0.3682	0.4648	-0.1422	0.2363		-3.6796	0.4079	0.6007	-0.2296	0.4136
60	-3.87	-4.0669	0.3763	0.4697	-0.1969	0.2594		-4.1581	0.4994	0.7181	-0.2881	0.5987
61	-3.37	-3.4897	0.3675	0.4516	-0.1197	0.2183		-3.5834	0.3907	0.5859	-0.2134	0.3888
62	-3.36	-3.5201	0.3639	0.4349	-0.1601	0.2147		-3.5691	0.3904	0.5897	-0.2091	0.3914
63	-3.35	-3.4779	0.3674	0.4671	-0.1279	0.2345		-3.5603	0.3888	0.5843	-0.2103	0.3856
64	-3.77	-3.9225	0.4795	0.606	-0.1525	0.3905		-4.0388	0.4826	0.6987	-0.2688	0.5604

Table 3.8: Simulation summary for the scenario: 3 disease characteristics with levels M_1=6, M_2=6, M_3=4; n_{case}=$n_{control}$=250

Characteristic level: 6x6x4, ncase=250												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
1	-6.7	-5.073	2.3326	2.2241	1.627	7.5938		-7.9901	3.0159	2.7916	-1.2901	9.4574
2	-6.69	-5.1658	2.3018	2.3099	1.5242	7.6588		-7.8841	2.9577	2.6629	-1.1941	8.5167
3	-6.68	-5.0436	2.361	2.3594	1.6364	8.2442		-7.9116	2.964	2.708	-1.2316	8.8503
4	-7.1	-5.5813	2.2284	2.5204	1.5187	8.659		-8.4358	3.3455	2.9324	-1.3358	10.3833
5	-6.68	-5.1721	2.2997	2.3322	1.5079	7.7126		-8.0007	3.1366	2.8632	-1.3207	9.9421
6	-6.67	-5.2498	2.2055	2.3052	1.4202	7.3307		-7.8948	3.0769	2.7295	-1.2248	8.9505
7	-6.66	-5.2512	2.2091	2.2905	1.4088	7.2311		-7.9222	3.0825	2.7765	-1.2622	9.3024
8	-7.08	-5.9598	2.0372	2.4342	1.1202	7.1803		-8.4465	3.4643	2.9919	-1.3665	10.8187
9	-6.7	-5.1954	2.2699	2.306	1.5046	7.5813		-7.9466	3.1092	2.8274	-1.2466	9.5482
10	-6.69	-5.2963	2.1635	2.2753	1.3937	7.1195		-7.8406	3.0474	2.6735	-1.1506	8.4714
11	-6.68	-5.1094	2.1963	2.2332	1.5706	7.4539		-7.8681	3.0544	2.7384	-1.1881	8.9104
12	-7.1	-5.0605	2.4875	2.4656	2.0395	10.2387		-8.3923	3.4416	2.9634	-1.2923	10.452
13	-6.6	-4.9287	2.334	2.2377	1.6713	7.8002		-7.8603	3.0731	2.831	-1.2603	9.6029
14	-6.59	-4.7684	2.4641	2.2485	1.8216	8.374		-7.7544	3.0228	2.7199	-1.1644	8.7535
15	-6.58	-4.7291	2.515	2.2111	1.8509	8.3146		-7.7818	3.0299	2.7813	-1.2018	9.1802
16	-7	-5.1463	2.4864	2.4936	1.8537	9.6541		-8.3061	3.403	2.9956	-1.3061	10.6796
17	-5.4	-4.1186	2.0606	1.6575	1.2814	4.3894		-6.3372	2.2237	2.3222	-0.9372	6.2709
18	-5.39	-4.1331	2.2771	1.7966	1.2569	4.8077		-6.2313	2.1727	2.1866	-0.8413	5.489
19	-5.38	-4.2025	1.8992	1.6421	1.1775	4.0829		-6.2588	2.1605	2.2129	-0.8788	5.669
20	-5.8	-4.4562	2.1022	1.8733	1.3438	5.3151		-6.783	2.5182	2.4626	-0.983	7.0305
21	-6.3	-4.751	2.2968	2.055	1.549	6.6223		-7.4599	2.6994	2.6932	-1.1599	8.5986
22	-6.29	-4.954	2.123	2.0819	1.336	6.1189		-7.3539	2.6333	2.5409	-1.0639	7.5882

Continuation of Table 3.8.

Characteristic level: 6x6x4, ncase=250												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
23	-6.28	-4.9412	2.1158	2.1037	1.3388	6.2179		-7.3814	2.6396	2.5961	-1.1014	7.9529
24	-6.7	-5.2982	2.1108	2.2522	1.4018	7.0374		-7.9057	3.0133	2.8277	-1.2057	9.4493
25	-6.7	-5.0906	2.2875	2.2073	1.6094	7.4624		-7.9198	2.8349	2.6012	-1.2198	8.2541
26	-6.69	-5.3346	2.1116	2.2358	1.3554	6.8358		-7.8138	2.7708	2.559	-1.1238	7.8114
27	-6.68	-5.3187	2.0634	2.2752	1.3613	7.0297		-7.8413	2.7746	2.5908	-1.1613	8.0607
28	-7.1	-5.6675	2.0882	2.4301	1.4325	7.9572		-8.3656	3.1582	2.769	-1.2656	9.2691
29	-6.68	-4.9997	2.3643	2.3342	1.6803	8.272		-7.9304	2.9614	2.6978	-1.2504	8.8416
30	-6.67	-5.3403	2.1298	2.2639	1.3297	6.8932		-7.8245	2.8961	2.6486	-1.1545	8.3479
31	-6.66	-5.3909	2.0563	2.2378	1.2691	6.6182		-7.852	2.8996	2.6823	-1.192	8.6154
32	-7.08	-5.9888	2.0088	2.4087	1.0912	6.9925		-8.3762	3.2821	2.8508	-1.2962	9.8072
33	-6.7	-4.3249	2.848	2.1819	2.3751	10.4016		-7.8763	2.9492	2.6776	-1.1763	8.5534
34	-6.69	-4.546	2.676	2.1915	2.144	9.3995		-7.7703	2.8821	2.6091	-1.0803	7.9745
35	-6.68	-4.4497	2.7015	2.1639	2.2303	9.657		-7.7978	2.8867	2.6607	-1.1178	8.3291
36	-7.1	-4.5397	2.8257	2.3711	2.5603	12.1768		-8.3221	3.2726	2.8377	-1.2221	9.5462
37	-6.6	-4.8781	2.3794	2.2475	1.7219	8.016		-7.79	2.8922	2.6592	-1.19	8.4874
38	-6.59	-5.0836	2.2759	2.2907	1.5064	7.5166		-7.6841	2.8361	2.6342	-1.0941	8.1359
39	-6.58	-5.0354	2.2385	2.2909	1.5446	7.6341		-7.7116	2.8412	2.6829	-1.1316	8.4783
40	-7	-5.4114	2.2784	2.3802	1.5886	8.189		-8.2358	3.2158	2.8506	-1.2358	9.6531
41	-5.4	-4.2194	2.0453	1.7632	1.1806	4.5027		-6.267	2.0404	2.1257	-0.867	5.2701
42	-5.39	-4.149	2.0756	1.7326	1.241	4.5419		-6.161	1.9828	2.0956	-0.771	4.9861
43	-5.38	-4.4149	1.6285	1.6838	0.9651	3.7667		-6.1885	1.9658	2.1043	-0.8085	5.0817
44	-5.8	-4.8262	1.8233	1.8457	0.9738	4.3551		-6.7127	2.3273	2.299	-0.9127	6.1187

Continuation of Table 3.8.

Characteristic level: 6x6x4, ncase=250												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
45	-6.3	-4.6851	2.3325	2.0869	1.6149	6.9632		-7.3896	2.5175	2.5073	-1.0896	7.4738
46	-6.29	-5.1779	1.8668	2.0237	1.1121	5.3319		-7.2837	2.4448	2.4442	-0.9937	6.9614
47	-6.28	-5.0467	1.9371	2.0308	1.2333	5.6452		-7.3112	2.4499	2.4856	-1.0312	7.2417
48	-6.7	-5.477	1.9612	2.1845	1.223	6.2677		-7.8354	2.8257	2.6692	-1.1354	8.4136
49	-6.6	-5.0257	2.2654	2.2289	1.5743	7.4467		-7.7982	2.7194	2.5829	-1.1982	8.107
50	-6.59	-5.2549	2.119	2.2479	1.3351	6.8357		-7.6923	2.6511	2.5311	-1.1023	7.6215
51	-6.58	-5.19	2.1664	2.1918	1.39	6.7362		-7.7198	2.6574	2.5555	-1.1398	7.8298
52	-7	-6.6515	1.2837	1.8772	0.3485	3.6454		-8.244	3.0177	2.7114	-1.244	8.8992
53	-6.58	-5.0091	2.2556	2.2278	1.5709	7.4306		-7.8089	2.8415	2.6585	-1.2289	8.5778
54	-6.57	-5.4942	1.8915	2.1714	1.0758	5.8724		-7.703	2.7719	2.5995	-1.133	8.0412
55	-6.56	-5.4828	1.898	2.1758	1.0772	5.8945		-7.7304	2.777	2.6264	-1.1704	8.2677
56	-6.98	-6.7821	1.1503	1.6047	0.1979	2.6144		-8.2547	3.137	2.7741	-1.2747	9.3206
57	-6.6	-5.189	2.1724	2.2301	1.411	6.9641		-7.7548	2.8139	2.6586	-1.1548	8.4017
58	-6.59	-5.4388	1.9507	2.1297	1.1512	5.8606		-7.6488	2.7451	2.5805	-1.0588	7.78
59	-6.58	-5.5458	1.8717	2.0989	1.0342	5.4752		-7.6763	2.7497	2.6252	-1.0963	8.0935
60	-7	-6.5204	1.3843	1.8997	0.4796	3.839		-8.2005	3.1147	2.7804	-1.2005	9.1717
61	-6.5	-5.0478	2.1921	2.2144	1.4522	7.0124		-7.6685	2.773	2.5609	-1.1685	7.9237
62	-6.49	-5.0029	2.2243	2.1434	1.4871	6.8059		-7.5626	2.7137	2.5257	-1.0726	7.5294
63	-6.48	-4.9867	2.2333	2.1677	1.4933	6.9287		-7.59	2.7206	2.5688	-1.11	7.8308
64	-6.9	-6.4469	1.4044	1.88	0.4531	3.7396		-8.1143	3.0722	2.7188	-1.2143	8.8666
65	-5.3	-4.2028	1.9421	1.6867	1.0972	4.0488		-6.1454	1.9034	2.0973	-0.8454	5.1133
66	-5.29	-4.1176	1.9266	1.7516	1.1724	4.4427		-6.0395	1.8404	2.0554	-0.7495	4.7865

Continuation of Table 3.8.

Characteristic level: 6x6x4, ncase=250												
	MLE							PCL				
j	true beta (B_j)	MC average	est(se)	MC standard error	bias	MSE		MC average	est(se)	MC standard error	bias	MSE
67	-5.28	-4.3583	1.5505	1.5068	0.9217	3.12		-6.067	1.8238	2.0547	-0.787	4.8411
68	-5.7	-5.0546	1.049	1.4129	0.6454	2.4129		-6.5912	2.1523	2.2237	-0.8912	5.7391
69	-6.2	-4.7698	2.1789	2.0351	1.4302	6.1869		-7.2681	2.3917	2.4856	-1.0681	7.319
70	-6.19	-5.1982	1.7819	1.9236	0.9918	4.6838		-7.1621	2.3141	2.4122	-0.9721	6.7638
71	-6.18	-5.1555	1.7922	1.9508	1.0245	4.8553		-7.1896	2.3215	2.4462	-1.0096	7.0031
72	-6.6	-6.2411	1.0902	1.5751	0.3589	2.6097		-7.7139	2.6705	2.6068	-1.1139	8.036
73	-6.67	-5.0563	2.3365	2.2531	1.6137	7.6807		-7.8216	2.6374	2.4176	-1.1516	7.171
74	-6.66	-6.0575	1.352	1.7375	0.6025	3.382		-7.7157	2.5476	2.3214	-1.0557	6.5035
75	-6.65	-6.1754	1.3223	1.7157	0.4746	3.1688		-7.7432	2.5519	2.3292	-1.0932	6.6201
76	-7.07	-6.7533	1.276	1.8443	0.3167	3.5016		-8.2674	2.9481	2.5189	-1.1974	7.7785
77	-6.65	-4.9778	2.3642	2.3051	1.6722	8.1096		-7.8323	2.7636	2.4755	-1.1823	7.5261
78	-6.64	-6.1561	1.3353	1.7362	0.4839	3.2485		-7.7263	2.6731	2.3722	-1.0863	6.8077
79	-6.63	-6.062	1.3108	1.7814	0.568	3.4962		-7.7538	2.6769	2.3832	-1.1238	6.9425
80	-7.05	-6.7024	1.2669	1.7939	0.3476	3.3387		-8.2781	3.0711	2.5644	-1.2281	8.0843
81	-6.67	-4.8826	2.3915	2.2814	1.7874	8.3996		-7.7781	2.7373	2.4452	-1.1081	7.2069
82	-6.66	-6.1847	1.315	1.7237	0.4753	3.197		-7.6722	2.6459	2.3192	-1.0122	6.4034
83	-6.65	-6.1886	1.3401	1.8034	0.4614	3.4652		-7.6997	2.6507	2.3502	-1.0497	6.6253
84	-7.07	-6.4672	1.5508	2.09	0.6028	4.7313		-8.2239	3.0498	2.5418	-1.1539	7.7924
85	-6.57	-4.7964	2.4478	2.2018	1.7736	7.9933		-7.6919	2.6975	2.4263	-1.1219	7.1454
86	-6.56	-5.8096	1.5803	1.9583	0.7504	4.3979		-7.5859	2.6155	2.3487	-1.0259	6.5691
87	-6.55	-5.7588	1.6252	1.9695	0.7912	4.505		-7.6134	2.6206	2.3765	-1.0634	6.7788
88	-6.97	-6.3846	1.5659	2.1096	0.5854	4.7931		-8.1377	3.0077	2.5574	-1.1677	7.9036

Continuation of Table 3.8.

Characteristic level: 6x6x4, ncase=250												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
89	-5.37	-4.0852	2.188	1.7088	1.2848	4.5705		-6.1688	1.815	1.8976	-0.7988	4.2388
90	-5.36	-4.506	1.424	1.5541	0.854	3.1446		-6.0629	1.7168	1.7989	-0.7029	3.7301
91	-5.35	-4.6346	1.0862	1.3241	0.7154	2.265		-6.0904	1.6982	1.7734	-0.7404	3.693
92	-5.77	-5.0393	1.2024	1.4732	0.7307	2.7041		-6.6146	2.0826	1.9917	-0.8446	4.68
93	-6.27	-4.6328	2.3753	2.0561	1.6372	6.908		-7.2915	2.3105	2.2509	-1.0215	6.1099
94	-6.26	-5.6527	1.2751	1.6351	0.6073	3.0425		-7.1855	2.2086	2.1251	-0.9255	5.3727
95	-6.25	-5.6162	1.2494	1.5985	0.6338	2.9568		-7.213	2.2127	2.1431	-0.963	5.5203
96	-6.67	-6.2589	1.2527	1.7177	0.4111	3.1196		-7.7372	2.6042	2.3459	-1.0672	6.6422
97	-5.2	-4.0441	2.0594	1.6699	1.1559	4.1246		-5.9548	1.6404	1.8672	-0.7548	4.0562
98	-5.19	-4.3875	1.1099	1.3431	0.8025	2.4478		-5.8489	1.513	1.735	-0.6589	3.4443
99	-5.18	-4.0331	1.933	1.6461	1.1469	4.025		-5.8764	1.5794	1.8207	-0.6964	3.7998
100	-5.6	-4.3979	2.0618	1.8306	1.2021	4.796		-6.4006	1.9333	2.0022	-0.8006	4.6497
101	-5.18	-3.7539	2.5848	1.643	1.4261	4.7333		-5.9655	1.8029	1.9849	-0.7855	4.5568
102	-5.17	-4.2134	1.5004	1.5175	0.9566	3.218		-5.8595	1.6775	1.8489	-0.6895	3.894
103	-5.16	-3.8214	2.4112	1.6467	1.3386	4.5035		-5.887	1.7384	1.9337	-0.727	4.2677
104	-5.58	-3.9776	2.5168	1.7502	1.6024	5.6308		-6.4113	2.0838	2.1	-0.8313	5.1009
105	-5.2	-3.7762	2.5444	1.6158	1.4238	4.6381		-5.9113	1.7886	1.9671	-0.7113	4.3756
106	-5.19	-4.152	1.5337	1.5267	1.038	3.4082		-5.8054	1.6664	1.8025	-0.6154	3.6277
107	-5.18	-3.7781	2.4741	1.6154	1.4019	4.5749		-5.8329	1.7267	1.9137	-0.6529	4.0885
108	-5.6	-3.711	2.7842	1.598	1.889	6.1218		-6.3571	2.078	2.0914	-0.7571	4.9471
109	-5.1	-3.7548	2.4558	1.6037	1.3452	4.3814		-5.8251	1.7643	1.9355	-0.7251	4.272
110	-5.09	-4.01	1.7811	1.6271	1.08	3.8138		-5.7191	1.6593	1.8318	-0.6291	3.7513

Continuation of Table 3.8.

Characteristic level: 6x6x4, ncase=250												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
111	-5.08	-3.5673	2.6406	1.475	1.5127	4.4639		-5.7466	1.7221	1.9379	-0.6666	4.2
112	-5.5	-3.7778	2.706	1.6777	1.7222	5.7807		-6.2709	2.0462	2.1028	-0.7709	5.0162
113	-3.9	-3.1584	1.8089	1.3208	0.7416	2.2945		-4.302	1.0457	1.397	-0.402	2.1133
114	-3.89	-3.0625	1.073	1.1956	0.8275	2.1141		-4.1961	0.8998	1.2516	-0.3061	1.6602
115	-3.88	-3.0935	1.5414	1.323	0.7865	2.3689		-4.2235	0.963	1.3206	-0.3435	1.862
116	-4.3	-3.4016	1.8516	1.4731	0.8984	2.9771		-4.7478	1.2145	1.5297	-0.4478	2.5404
117	-4.8	-3.6153	2.392	1.5564	1.1847	3.8258		-5.4247	1.41	1.7888	-0.6247	3.5901
118	-4.79	-3.8864	1.2591	1.3529	0.9036	2.6468		-5.3187	1.2547	1.6214	-0.5287	2.9086
119	-4.78	-3.5974	2.1528	1.5089	1.1826	3.6755		-5.3462	1.3346	1.7247	-0.5662	3.2952
120	-5.2	-3.8525	2.3576	1.6284	1.3475	4.4676		-5.8704	1.6554	1.9128	-0.6704	4.1085
121	-4.7	-3.8561	1.7162	1.573	0.8439	3.1867		-5.3302	1.4763	1.6732	-0.6302	3.1968
122	-4.69	-3.8155	1.702	1.519	0.8745	3.0722		-5.2242	1.4099	1.572	-0.5342	2.7567
123	-4.68	-3.81	1.6983	1.4951	0.87	2.9921		-5.2517	1.4153	1.6021	-0.5717	2.8936
124	-5.1	-3.9049	1.9873	1.586	1.1951	3.9436		-5.7759	1.7616	1.7912	-0.6759	3.6651
125	-4.68	-3.5446	2.2571	1.4584	1.1354	3.4163		-5.3408	1.6442	1.7972	-0.6608	3.6666
126	-4.67	-3.5128	2.251	1.4431	1.1572	3.4214		-5.2349	1.5776	1.6901	-0.5649	3.1754
127	-4.66	-3.5585	2.1697	1.4524	1.1015	3.3229		-5.2624	1.5831	1.7227	-0.6024	3.3306
128	-5.08	-3.8329	2.3308	1.6332	1.2471	4.2226		-5.7866	1.9164	1.8937	-0.7066	4.0855
129	-4.7	-3.4496	2.3204	1.5051	1.2504	3.8287		-5.2867	1.637	1.8003	-0.5867	3.5852
130	-4.69	-3.6021	2.1772	1.4943	1.0879	3.4165		-5.1807	1.5716	1.6638	-0.4907	3.009
131	-4.68	-3.5543	2.1588	1.5311	1.1257	3.6117		-5.2082	1.5733	1.724	-0.5282	3.2512
132	-5.1	-3.5998	2.6669	1.5268	1.5002	4.5818		-5.7324	1.9154	1.9056	-0.6324	4.0314

Continuation of Table 3.8.

Characteristic level: 6x6x4, ncase=250												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
133	-4.6	-3.4318	2.3734	1.4242	1.1682	3.393		-5.2004	1.6168	1.7793	-0.6004	3.5263
134	-4.59	-3.4154	2.4429	1.4704	1.1746	3.5419		-5.0945	1.5704	1.7096	-0.5045	3.1773
135	-4.58	-3.443	2.404	1.4815	1.137	3.4876		-5.122	1.5786	1.7645	-0.542	3.4074
136	-5	-3.4916	2.6231	1.5223	1.5084	4.5924		-5.6462	1.8896	1.9307	-0.6462	4.1453
137	-3.4	-2.6947	1.394	1.163	0.7053	1.85		-3.6773	0.9441	1.2382	-0.2773	1.61
138	-3.39	-2.754	1.5985	1.3113	0.636	2.1239		-3.5714	0.9037	1.1383	-0.1814	1.3286
139	-3.38	-2.6963	1.1786	1.1587	0.6837	1.81		-3.5989	0.8626	1.1242	-0.2189	1.3118
140	-3.8	-3.0191	1.4799	1.3228	0.7809	2.3596		-4.1231	1.0767	1.3444	-0.3231	1.9119
141	-4.3	-3.4769	2.0817	1.4806	0.8231	2.8697		-4.8	1.2628	1.6009	-0.5	2.813
142	-4.29	-3.3971	1.8168	1.4343	0.8929	2.8544		-4.6941	1.1811	1.4628	-0.4041	2.3032
143	-4.28	-3.3904	1.8058	1.4421	0.8896	2.871		-4.7215	1.1883	1.5087	-0.4415	2.4711
144	-4.7	-3.6365	2.08	1.5338	1.0635	3.4837		-5.2458	1.4914	1.7053	-0.5458	3.2061

Table 3.9: Simulation summary for the scenario: 3 disease characteristics with levels M_1=6, M_2=6, M_3=4; n_{case}=$n_{control}$=500

Characteristic level: 6x6x4, ncase=500												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
1	-6.7	-5.5187	1.4144	1.7514	1.1813	4.4629		-7.6286	1.9078	1.9178	-0.9286	4.5404
2	-6.69	-5.5108	1.407	1.7268	1.1792	4.3725		-7.6067	1.8888	1.8541	-0.9167	4.2782
3	-6.68	-5.4915	1.4181	1.7173	1.1885	4.3615		-7.6219	1.8931	1.9007	-0.9419	4.4998
4	-7.1	-5.9487	1.3505	1.7734	1.1513	4.4704		-8.1175	2.1625	2.0404	-1.0175	5.1985
5	-6.68	-5.4794	1.4159	1.7052	1.2006	4.3492		-7.6234	1.9567	1.9346	-0.9434	4.6328
6	-6.67	-5.4193	1.335	1.6729	1.2507	4.3629		-7.6015	1.9364	1.8671	-0.9315	4.3537
7	-6.66	-5.5798	1.347	1.6379	1.0802	3.8496		-7.6167	1.9413	1.9213	-0.9567	4.6066
8	-7.08	-5.9892	1.3357	1.7361	1.0908	4.204		-8.1122	2.21	2.0447	-1.0322	5.2463
9	-6.7	-5.415	1.4089	1.6824	1.285	4.4815		-7.6267	1.9644	1.9563	-0.9267	4.686
10	-6.69	-5.5614	1.335	1.6878	1.1286	4.1224		-7.6048	1.9442	1.8952	-0.9148	4.4286
11	-6.68	-5.6249	1.3224	1.6151	1.0551	3.7217		-7.62	1.9489	1.9487	-0.94	4.681
12	-7.1	-5.894	1.489	1.9073	1.206	5.0922		-8.1156	2.2211	2.077	-1.0156	5.3452
13	-6.6	-5.4056	1.4278	1.6716	1.1944	4.2207		-7.5115	1.9088	1.9174	-0.9115	4.5071
14	-6.59	-5.1588	1.5497	1.7524	1.4312	5.1191		-7.4896	1.8944	1.8726	-0.8996	4.3159
15	-6.58	-5.1351	1.5832	1.7836	1.4449	5.2687		-7.5048	1.8988	1.9199	-0.9248	4.5411
16	-7	-5.5436	1.5525	1.897	1.4564	5.7199		-8.0004	2.1641	2.0416	-1.0004	5.169
17	-5.4	-4.3273	1.2691	1.3422	1.0727	2.9522		-6.0572	1.3223	1.5294	-0.6572	2.7711
18	-5.39	-4.2327	1.4476	1.4148	1.1573	3.341		-6.0354	1.3054	1.4678	-0.6454	2.571
19	-5.38	-4.4095	1.1322	1.3114	0.9705	2.6616		-6.0505	1.2996	1.5029	-0.6705	2.7084
20	-5.8	-4.839	1.215	1.4406	0.961	2.9989		-6.5461	1.5508	1.6162	-0.7461	3.1689
21	-6.3	-5.0735	1.4305	1.6463	1.2265	4.2146		-7.149	1.6789	1.8058	-0.849	3.9819
22	-6.29	-5.2104	1.2706	1.5314	1.0796	3.5108		-7.1271	1.6555	1.7488	-0.8371	3.7589

Continuation of Table 3.9.

Characteristic level: 6x6x4, ncase=500											
	MLE							**PCL**			
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
23	-6.28	-5.1592	1.2974	1.5703	1.1208	3.7221		-7.1423	1.6599	1.7907	-0.8623	3.95
24	-6.7	-5.6874	1.2522	1.5855	1.0126	3.539		-7.6379	1.9241	1.9049	-0.9379	4.5084
25	-6.7	-5.542	1.3903	1.7058	1.158	4.2508		-7.6202	1.8752	1.8844	-0.9202	4.3977
26	-6.69	-5.5979	1.2796	1.6318	1.0921	3.8553		-7.5984	1.8523	1.825	-0.9084	4.1557
27	-6.68	-5.6456	1.2716	1.61	1.0344	3.6622		-7.6135	1.857	1.8729	-0.9335	4.3794
28	-7.1	-6.294	1.1875	1.6704	0.806	3.4399		-8.1091	2.1277	2.0178	-1.0091	5.0899
29	-6.68	-5.4234	1.4645	1.7419	1.2566	4.6134		-7.615	1.9258	1.8985	-0.935	4.4786
30	-6.67	-5.6742	1.2462	1.6107	0.9958	3.5858		-7.5931	1.9012	1.8351	-0.9231	4.2196
31	-6.66	-5.6556	1.2411	1.5264	1.0044	3.3385		-7.6083	1.9065	1.8909	-0.9483	4.4746
32	-7.08	-6.1824	1.1708	1.5633	0.8976	3.2497		-8.1039	2.176	2.0194	-1.0239	5.1262
33	-6.7	-4.7875	1.9307	1.8727	1.9125	7.1645		-7.6183	1.9428	1.933	-0.9183	4.5797
34	-6.69	-5.1231	1.6814	1.8947	1.5669	6.0451		-7.5965	1.9192	1.8764	-0.9065	4.3424
35	-6.68	-5.0398	1.7671	1.8885	1.6402	6.2567		-7.6116	1.9242	1.9311	-0.9316	4.597
36	-7.1	-5.0414	1.9096	2.0016	2.0586	8.2441		-8.1072	2.1957	2.0636	-1.0072	5.2731
37	-6.6	-5.3631	1.4558	1.6832	1.2369	4.3631		-7.5031	1.877	1.8958	-0.9031	4.4097
38	-6.59	-5.3656	1.4162	1.7058	1.2244	4.4089		-7.4813	1.8586	1.8558	-0.8913	4.2385
39	-6.58	-5.3139	1.462	1.7515	1.2661	4.671		-7.4964	1.8636	1.9042	-0.9164	4.4659
40	-7	-5.9578	1.3457	1.8037	1.0422	4.3396		-7.992	2.13	2.0302	-0.992	5.1057
41	-5.4	-4.2693	1.2714	1.3244	1.1307	3.0324		-6.0489	1.2864	1.4671	-0.6489	2.5735
42	-5.39	-4.336	1.2575	1.3179	1.054	2.8477		-6.027	1.264	1.4099	-0.637	2.3936
43	-5.38	-4.3996	1.0181	1.1889	0.9804	2.3747		-6.0422	1.2585	1.4473	-0.6622	2.5331
44	-5.8	-4.8751	1.0907	1.3628	0.9249	2.7126		-6.5378	1.5125	1.5688	-0.7378	3.0055

Continuation of Table 3.9.

Characteristic level: 6x6x4, ncase=500												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
45	-6.3	-5.1312	1.4106	1.5889	1.1688	3.8908		-7.1406	1.6449	1.7666	-0.8406	3.8275
46	-6.29	-5.2762	1.1328	1.4513	1.0138	3.1342		-7.1188	1.617	1.714	-0.8288	3.6246
47	-6.28	-5.3	1.1361	1.4406	0.98	3.0357		-7.1339	1.6218	1.7575	-0.8539	3.8179
48	-6.7	-5.7563	1.1262	1.4914	0.9437	3.115		-7.6295	1.8875	1.8773	-0.9295	4.3882
49	-6.6	-5.3253	1.4487	1.6993	1.2747	4.5125		-7.4466	1.7642	1.7661	-0.8466	3.836
50	-6.59	-5.6356	1.2715	1.5655	0.9544	3.3619		-7.4248	1.7393	1.6925	-0.8348	3.5612
51	-6.58	-5.5368	1.2946	1.6116	1.0432	3.6854		-7.4399	1.7452	1.7575	-0.8599	3.8282
52	-7	-6.4391	0.7681	1.0867	0.5609	1.4955		-7.9355	2.0018	1.872	-0.9355	4.3794
53	-6.58	-5.3802	1.3699	1.6169	1.1998	4.0537		-7.4414	1.8141	1.7972	-0.8614	3.972
54	-6.57	-5.6478	1.1566	1.5224	0.9222	3.1679		-7.4195	1.7875	1.72	-0.8495	3.6802
55	-6.56	-5.6576	1.154	1.5148	0.9024	3.1089		-7.4347	1.794	1.7926	-0.8747	3.9785
56	-6.98	-6.443	0.6819	0.9747	0.537	1.2383		-7.9303	2.0501	1.8888	-0.9503	4.4707
57	-6.6	-5.3972	1.3295	1.6252	1.2028	4.0881		-7.4447	1.8218	1.8274	-0.8447	4.0529
58	-6.59	-5.6192	1.1426	1.476	0.9708	3.1211		-7.4228	1.7957	1.7576	-0.8328	3.7829
59	-6.58	-5.6533	1.1232	1.4889	0.9267	3.0756		-7.438	1.8018	1.8288	-0.858	4.0808
60	-7	-6.4603	0.7971	1.1383	0.5397	1.5871		-7.9336	2.0616	1.9302	-0.9336	4.5974
61	-6.5	-5.395	1.3103	1.6236	1.105	3.8571		-7.3295	1.7655	1.7775	-0.8295	3.8477
62	-6.49	-5.3573	1.3551	1.6655	1.1327	4.057		-7.3076	1.7452	1.7249	-0.8176	3.6438
63	-6.48	-5.3362	1.355	1.6244	1.1438	3.947		-7.3228	1.7515	1.79	-0.8428	3.9145
64	-6.9	-6.2744	0.8331	1.131	0.6256	1.6705		-7.8184	2.0036	1.8845	-0.9184	4.3949
65	-5.3	-4.2374	1.1953	1.2945	1.0626	2.8048		-5.8753	1.1674	1.3627	-0.5753	2.1879
66	-5.29	-4.2432	1.1467	1.3556	1.0468	2.9333		-5.8534	1.1409	1.2875	-0.5634	1.9752

Continuation of Table 3.9.

Characteristic level: 6x6x4, ncase=500												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
67	-5.28	-4.3103	0.938	1.1313	0.9697	2.2203		-5.8686	1.1369	1.346	-0.5886	2.158
68	-5.7	-4.8332	0.644	0.9553	0.8668	1.664		-6.3642	1.3707	1.4248	-0.6642	2.471
69	-6.2	-5.0527	1.3364	1.5415	1.1473	3.6924		-6.967	1.5314	1.6839	-0.767	3.424
70	-6.19	-5.2813	1.0537	1.3432	0.9087	2.6299		-6.9452	1.5007	1.6181	-0.7552	3.1884
71	-6.18	-5.3035	1.0659	1.3677	0.8765	2.6389		-6.9603	1.5068	1.6781	-0.7803	3.4249
72	-6.6	-6.0234	0.6697	0.9588	0.5766	1.2518		-7.4559	1.7569	1.7616	-0.8559	3.8359
73	-6.67	-5.4479	1.416	1.6841	1.2221	4.3298		-7.5511	1.8006	1.7931	-0.8811	3.9914
74	-6.66	-5.9516	0.8117	1.0953	0.7084	1.7015		-7.5292	1.7647	1.7095	-0.8692	3.6777
75	-6.65	-5.9019	0.8104	1.0999	0.7481	1.7695		-7.5444	1.7688	1.7731	-0.8944	3.9436
76	-7.07	-6.5593	0.7607	1.078	0.5107	1.4228		-8.0399	2.0461	1.9049	-0.9699	4.5693
77	-6.65	-5.2751	1.5138	1.7596	1.3749	4.9867		-7.5458	1.853	1.8173	-0.8958	4.1051
78	-6.64	-5.9883	0.8014	1.1416	0.6517	1.7279		-7.524	1.8161	1.7301	-0.884	3.7746
79	-6.63	-5.93	0.7979	1.1311	0.7	1.7693		-7.5391	1.8208	1.8014	-0.9091	4.0717
80	-7.05	-6.4858	0.767	1.0573	0.5642	1.4362		-8.0347	2.0967	1.9154	-0.9847	4.6384
81	-6.67	-5.3891	1.4485	1.7224	1.2809	4.6074		-7.5492	1.86	1.8298	-0.8792	4.121
82	-6.66	-5.9828	0.7957	1.0547	0.6772	1.571		-7.5273	1.8234	1.7493	-0.8673	3.8122
83	-6.65	-5.9479	0.797	1.0991	0.7021	1.7009		-7.5424	1.8278	1.82	-0.8924	4.1089
84	-7.07	-6.3932	0.9156	1.2692	0.6768	2.069		-8.038	2.1074	1.9398	-0.968	4.7001
85	-6.57	-5.2614	1.5107	1.7244	1.3086	4.6858		-7.4339	1.8037	1.8018	-0.8639	3.9927
86	-6.56	-5.7977	0.932	1.2339	0.7623	2.1036		-7.4121	1.7725	1.739	-0.8521	3.75
87	-6.55	-5.7212	0.9577	1.2749	0.8288	2.3123		-7.4272	1.7769	1.8028	-0.8772	4.0196
88	-6.97	-6.3199	0.9284	1.255	0.6501	1.9977		-7.9228	2.0495	1.9149	-0.9528	4.5745

Continuation of Table 3.9.

Characteristic level: 6x6x4, ncase=500												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
89	-5.37	-4.3284	1.3405	1.3737	1.0416	2.9719		-5.9797	1.1974	1.3762	-0.6097	2.2658
90	-5.36	-4.4167	0.84	1.0393	0.9433	1.97		-5.9578	1.1544	1.2872	-0.5978	2.0142
91	-5.35	-4.4654	0.6815	0.8624	0.8846	1.5263		-5.973	1.1478	1.3445	-0.623	2.1959
92	-5.77	-4.8897	0.7348	0.9805	0.8803	1.7363		-6.4686	1.4145	1.4476	-0.6986	2.5835
93	-6.27	-5.0257	1.504	1.667	1.2443	4.327		-7.0715	1.5675	1.6717	-0.8015	3.4371
94	-6.26	-5.5257	0.7564	1.0541	0.7343	1.6503		-7.0496	1.5243	1.5935	-0.7896	3.1626
95	-6.25	-5.5096	0.7505	1.0327	0.7404	1.6148		-7.0648	1.5287	1.6535	-0.8148	3.398
96	-6.67	-6.0141	0.7596	1.0234	0.6559	1.4775		-7.5603	1.8025	1.7581	-0.8903	3.8835
97	-5.2	-4.206	1.2078	1.3066	0.994	2.6953		-5.7493	1.0971	1.3344	-0.5493	2.0824
98	-5.19	-4.3061	0.6924	0.8883	0.8839	1.5704		-5.7274	1.0375	1.2297	-0.5374	1.801
99	-5.18	-4.1678	1.1513	1.2623	1.0122	2.6178		-5.7426	1.074	1.3167	-0.5626	2.0502
100	-5.6	-4.5925	1.255	1.4124	1.0075	3.0101		-6.2382	1.3233	1.4544	-0.6382	2.5226
101	-5.18	-3.9094	1.6366	1.4472	1.2706	3.7089		-5.7441	1.1696	1.4057	-0.5641	2.2942
102	-5.17	-4.2124	0.9064	1.1063	0.9576	2.1407		-5.7222	1.1096	1.3003	-0.5522	1.9958
103	-5.16	-3.972	1.5585	1.3935	1.188	3.3532		-5.7374	1.1462	1.3939	-0.5774	2.2764
104	-5.58	-4.2811	1.6486	1.496	1.2989	3.9251		-6.233	1.3891	1.5045	-0.653	2.6898
105	-5.2	-4.0196	1.6071	1.4524	1.1804	3.5028		-5.7474	1.1815	1.4251	-0.5474	2.3306
106	-5.19	-4.2374	0.9363	1.1328	0.9526	2.1908		-5.7255	1.1234	1.3294	-0.5355	2.054
107	-5.18	-3.9649	1.5221	1.3939	1.2151	3.4194		-5.7407	1.1589	1.4212	-0.5607	2.3341
108	-5.6	-3.9874	1.9146	1.4575	1.6126	4.725		-6.2363	1.4066	1.5386	-0.6363	2.772
109	-5.1	-3.8196	1.6307	1.3483	1.2804	3.4574		-5.6322	1.1354	1.3683	-0.5322	2.1555
110	-5.09	-4.1073	1.0493	1.1694	0.9827	2.3332		-5.6103	1.0861	1.2939	-0.5203	1.945

Continuation of Table 3.9.

Characteristic level: 6x6x4, ncase=500												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
111	-5.08	-3.9026	1.6979	1.3452	1.1774	3.1958		-5.6255	1.1226	1.3786	-0.5455	2.198
112	-5.5	-4.0865	1.8377	1.5256	1.4135	4.3256		-6.1211	1.3561	1.4879	-0.6211	2.5996
113	-3.9	-3.0558	1.0807	1.1333	0.8442	1.9971		-4.178	0.6769	0.9793	-0.278	1.0363
114	-3.89	-3.0425	0.6787	0.7715	0.8475	1.3135		-4.1561	0.5965	0.8638	-0.2661	0.817
115	-3.88	-3.0889	0.9384	0.9727	0.7911	1.5721		-4.1712	0.6374	0.9472	-0.2912	0.9821
116	-4.3	-3.4672	1.1098	1.1903	0.8328	2.1102		-4.6668	0.7904	1.0401	-0.3668	1.2163
117	-4.8	-3.715	1.5346	1.3383	1.085	2.9683		-5.2697	0.9219	1.2569	-0.4697	1.8003
118	-4.79	-3.8488	0.7739	0.9195	0.9412	1.7314		-5.2478	0.8467	1.1612	-0.4578	1.5581
119	-4.78	-3.729	1.2905	1.2733	1.051	2.7259		-5.263	0.8917	1.2424	-0.483	1.7769
120	-5.2	-4.0313	1.4833	1.3539	1.1687	3.1989		-5.7586	1.1144	1.3405	-0.5586	2.1091
121	-4.7	-3.7589	1.0513	1.1871	0.9411	2.295		-5.149	0.9793	1.1741	-0.449	1.5801
122	-4.69	-3.7723	1.0165	1.1149	0.9177	2.0853		-5.1271	0.9517	1.0965	-0.4371	1.3934
123	-4.68	-3.7	1.0074	1.1065	0.98	2.1847		-5.1423	0.9574	1.1499	-0.4623	1.5359
124	-5.1	-4.1119	1.1728	1.3136	0.9881	2.7017		-5.6379	1.1999	1.2923	-0.5379	1.9592
125	-4.68	-3.659	1.4921	1.3208	1.021	2.787		-5.1438	1.0535	1.242	-0.4638	1.7576
126	-4.67	-3.6919	1.3619	1.2926	0.9781	2.6275		-5.1219	1.0237	1.1618	-0.4519	1.5539
127	-4.66	-3.6713	1.3823	1.2804	0.9887	2.617		-5.137	1.0316	1.2248	-0.477	1.7277
128	-5.08	-3.9713	1.518	1.3769	1.1087	3.1251		-5.6326	1.267	1.3367	-0.5526	2.0921
129	-4.7	-3.6304	1.4853	1.3032	1.0696	2.8424		-5.1471	1.0665	1.2717	-0.4471	1.8171
130	-4.69	-3.6647	1.3433	1.2709	1.0253	2.6664		-5.1252	1.0386	1.2024	-0.4352	1.6351
131	-4.68	-3.6754	1.3439	1.2887	1.0046	2.6698		-5.1404	1.0452	1.2635	-0.4604	1.8085
132	-5.1	-3.8691	1.7481	1.3856	1.2309	3.4349		-5.636	1.2862	1.3821	-0.536	2.1973

Continuation of Table 3.9.

Characteristic level: 6x6x4, ncase=500												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
133	-4.6	-3.5723	1.4089	1.2638	1.0277	2.6532		-5.0319	1.0268	1.2305	-0.4319	1.7007
134	-4.59	-3.5029	1.5168	1.2951	1.0871	2.8592		-5.01	1.0107	1.1868	-0.42	1.5849
135	-4.58	-3.5393	1.545	1.2391	1.0407	2.6184		-5.0252	1.0165	1.2381	-0.4452	1.7311
136	-5	-3.7815	1.7557	1.3503	1.2185	3.3082		-5.5208	1.2391	1.3463	-0.5208	2.0837
137	-3.4	-2.7258	0.8613	0.9887	0.6742	1.432		-3.5776	0.6191	0.8549	-0.1776	0.7625
138	-3.39	-2.7426	0.9339	0.982	0.6474	1.3835		-3.5558	0.5985	0.7812	-0.1658	0.6378
139	-3.38	-2.6795	0.7367	0.8368	0.7005	1.1908		-3.5709	0.582	0.8122	-0.1909	0.6961
140	-3.8	-3.003	0.9107	0.9758	0.797	1.5874		-4.0665	0.6952	0.9002	-0.2665	0.8814
141	-4.3	-3.4426	1.2732	1.216	0.8574	2.2139		-4.6694	0.819	1.1124	-0.3694	1.3739
142	-4.29	-3.4322	1.0845	1.1708	0.8578	2.1065		-4.6475	0.783	1.0481	-0.3575	1.2263
143	-4.28	-3.404	1.1243	1.1194	0.876	2.0205		-4.6627	0.7896	1.0917	-0.3827	1.3384
144	-4.7	-3.6718	1.2791	1.2518	1.0282	2.6243		-5.1583	0.9972	1.1881	-0.4583	1.6215

Table 3.10: Simulation summary for the scenario: 3 disease characteristics with levels M_1=6, M_2=6, M_3=4; n_{case}=$n_{control}$=1000

Characteristic level: 6x6x4, ncase=1000												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
1	-6.7	-5.7143	0.8938	1.1944	0.9857	2.3981		-7.5005	1.4055	1.7092	-0.8005	3.5619
2	-6.69	-5.719	0.8595	1.1867	0.971	2.3512		-7.468	1.3879	1.6821	-0.778	3.4347
3	-6.68	-5.6516	0.8893	1.1883	1.0284	2.4697		-7.4598	1.3847	1.6765	-0.7798	3.4186
4	-7.1	-6.2029	0.8161	1.1837	0.8971	2.2059		-7.9396	1.5909	1.7968	-0.8396	3.9333
5	-6.68	-5.6969	0.8827	1.2069	0.9831	2.4231		-7.5	1.4409	1.7179	-0.82	3.6238
6	-6.67	-5.762	0.8479	1.2012	0.908	2.2673		-7.4676	1.4226	1.694	-0.7976	3.5057
7	-6.66	-5.5997	0.8523	1.184	1.0603	2.5261		-7.4593	1.4195	1.6842	-0.7993	3.4753
8	-7.08	-6.171	0.7948	1.1275	0.909	2.0975		-7.9392	1.6253	1.8061	-0.8592	4.0004
9	-6.7	-5.7032	0.8722	1.2409	0.9968	2.5335		-7.4795	1.4314	1.6998	-0.7795	3.497
10	-6.69	-5.7091	0.8365	1.1965	0.9809	2.3939		-7.4471	1.4131	1.6715	-0.7571	3.3671
11	-6.68	-5.6614	0.8444	1.1653	1.0186	2.3955		-7.4388	1.41	1.6619	-0.7588	3.3376
12	-7.1	-6.0839	0.9385	1.2963	1.0161	2.713		-7.9187	1.6175	1.7907	-0.8187	3.8768
13	-6.6	-5.6259	0.8814	1.1823	0.9741	2.3467		-7.3917	1.395	1.6789	-0.7917	3.4454
14	-6.59	-5.4589	0.9883	1.3292	1.1311	3.0461		-7.3592	1.3795	1.6542	-0.7692	3.3282
15	-6.58	-5.4616	1.0091	1.3014	1.1184	2.9445		-7.351	1.3765	1.6458	-0.771	3.3031
16	-7	-5.9472	0.95	1.3438	1.0528	2.9142		-7.8308	1.5802	1.7627	-0.8308	3.7974
17	-5.4	-4.4119	0.8201	1.1174	0.9881	2.225		-5.963	0.9518	1.2841	-0.563	1.9659
18	-5.39	-4.315	0.9051	1.1407	1.075	2.4568		-5.9306	0.9358	1.2629	-0.5406	1.8871
19	-5.38	-4.4068	0.7237	0.9793	0.9732	1.9062		-5.9223	0.9278	1.2535	-0.5423	1.8653
20	-5.8	-4.8323	0.7843	1.0275	0.9677	1.9924		-6.4022	1.1221	1.3697	-0.6022	2.2388
21	-6.3	-5.2792	0.8998	1.1846	1.0208	2.4452		-7.0125	1.2245	1.5647	-0.7125	2.9558
22	-6.29	-5.2858	0.8045	1.073	1.0042	2.1597		-6.9801	1.2046	1.5418	-0.6901	2.8534

Continuation of Table 3.10.

Characteristic level: 6x6x4, ncase=1000												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
23	-6.28	-5.2619	0.8029	1.1299	1.0181	2.3132		-6.9718	1.2015	1.5326	-0.6918	2.8274
24	-6.7	-5.7412	0.7864	1.1207	0.9588	2.1752		-7.4517	1.4043	1.6534	-0.7517	3.2987
25	-6.7	-5.7164	0.8782	1.1873	0.9836	2.3773		-7.4436	1.3718	1.6111	-0.7436	3.1487
26	-6.69	-5.7782	0.7702	1.0319	0.9118	1.8963		-7.4112	1.352	1.5841	-0.7212	3.0295
27	-6.68	-5.7664	0.7898	1.0868	0.9136	2.0158		-7.4029	1.3487	1.583	-0.7229	3.0285
28	-7.1	-6.226	0.7415	1.032	0.874	1.8289		-7.8828	1.5555	1.7111	-0.7828	3.5406
29	-6.68	-5.6601	0.9338	1.2343	1.0199	2.5636		-7.4432	1.408	1.6204	-0.7632	3.2081
30	-6.67	-5.7576	0.7641	1.1137	0.9124	2.0729		-7.4108	1.3875	1.5966	-0.7408	3.098
31	-6.66	-5.7383	0.7717	1.1005	0.9217	2.0606		-7.4025	1.3843	1.591	-0.7425	3.0827
32	-7.08	-6.24	0.7313	1.0284	0.84	1.7631		-7.8824	1.5905	1.7208	-0.8024	3.6051
33	-6.7	-5.2152	1.2643	1.5547	1.4848	4.6217		-7.4227	1.4035	1.6084	-0.7227	3.1092
34	-6.69	-5.3445	1.1081	1.4839	1.3455	4.0125		-7.3903	1.3831	1.5802	-0.7003	2.9873
35	-6.68	-5.3366	1.1253	1.4473	1.3434	3.8994		-7.382	1.3799	1.5748	-0.702	2.9729
36	-7.1	-5.5605	1.223	1.6212	1.5395	4.9985		-7.8619	1.5872	1.7115	-0.7619	3.5095
37	-6.6	-5.5589	0.8881	1.2479	1.0411	2.6411		-7.3348	1.3615	1.5816	-0.7348	3.0413
38	-6.59	-5.5675	0.9001	1.2106	1.0225	2.5112		-7.3024	1.3439	1.5571	-0.7124	2.9321
39	-6.58	-5.4888	0.9039	1.2094	1.0912	2.6533		-7.2941	1.3409	1.5531	-0.7141	2.9222
40	-7	-6.0418	0.8441	1.2062	0.9582	2.373		-7.774	1.5451	1.6777	-0.774	3.4139
41	-5.4	-4.3874	0.8148	1.0182	1.0126	2.062		-5.9062	0.9178	1.1821	-0.5062	1.6535
42	-5.39	-4.3277	0.8169	1.0293	1.0623	2.1881		-5.8738	0.8984	1.1614	-0.4838	1.5828
43	-5.38	-4.378	0.6502	0.8572	1.002	1.7389		-5.8655	0.8903	1.1577	-0.4855	1.5761
44	-5.8	-4.8439	0.6945	0.9595	0.9561	1.8348		-6.3454	1.0854	1.2843	-0.5454	1.9469

Continuation of Table 3.10.

Characteristic level: 6x6x4, ncase=1000												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
45	-6.3	-5.2642	0.9051	1.2313	1.0358	2.5892		-6.9557	1.1907	1.4633	-0.6557	2.571
46	-6.29	-5.3707	0.718	0.9825	0.9193	1.8103		-6.9232	1.1681	1.4407	-0.6332	2.4766
47	-6.28	-5.3014	0.709	1.0156	0.9786	1.9891		-6.915	1.165	1.4362	-0.635	2.4657
48	-6.7	-5.8137	0.7173	1.0434	0.8863	1.8742		-7.3948	1.3683	1.5657	-0.6948	2.9344
49	-6.6	-5.5641	0.8974	1.2473	1.0359	2.6289		-7.3386	1.3089	1.5879	-0.7386	3.0668
50	-6.59	-5.6031	0.8078	1.1187	0.9869	2.2254		-7.3061	1.2883	1.5542	-0.7161	2.9284
51	-6.58	-5.5555	0.787	1.1298	1.0245	2.3259		-7.2979	1.2854	1.5532	-0.7179	2.9277
52	-7	-6.2851	0.5021	0.7285	0.7149	1.0418		-7.7777	1.4849	1.6587	-0.7777	3.3561
53	-6.58	-5.5947	0.8228	1.1638	0.9853	2.3252		-7.3382	1.344	1.5902	-0.7582	3.1034
54	-6.57	-5.7094	0.716	1.0462	0.8606	1.8352		-7.3057	1.3228	1.5597	-0.7357	2.9741
55	-6.56	-5.6613	0.7027	1.0208	0.8987	1.8497		-7.2975	1.32	1.5541	-0.7375	2.9592
56	-6.98	-6.2781	0.4538	0.6803	0.7019	0.9555		-7.7773	1.5192	1.662	-0.7973	3.3979
57	-6.6	-5.6456	0.8409	1.1788	0.9544	2.3004		-7.3176	1.3347	1.5611	-0.7176	2.9521
58	-6.59	-5.6515	0.7155	0.9649	0.9385	1.8117		-7.2852	1.3134	1.5257	-0.6952	2.8109
59	-6.58	-5.7454	0.7054	1.0123	0.8346	1.7212		-7.2769	1.3105	1.5202	-0.6969	2.7968
60	-7	-6.2506	0.5303	0.7502	0.7494	1.1243		-7.7568	1.5115	1.6362	-0.7568	3.2498
61	-6.5	-5.5045	0.8454	1.1809	0.9955	2.3856		-7.2298	1.2984	1.5566	-0.7298	2.9555
62	-6.49	-5.5168	0.8548	1.1922	0.9732	2.3684		-7.1974	1.2802	1.5254	-0.7074	2.8271
63	-6.48	-5.4921	0.841	1.1312	0.9879	2.2556		-7.1891	1.2775	1.5214	-0.7091	2.8175
64	-6.9	-6.1438	0.5364	0.7812	0.7562	1.1822		-7.6689	1.4741	1.623	-0.7689	3.2255
65	-5.3	-4.3208	0.7599	0.9814	0.9792	1.9219		-5.8011	0.8499	1.1486	-0.5011	1.5703
66	-5.29	-4.3204	0.7511	0.9469	0.9696	1.8368		-5.7687	0.8293	1.1185	-0.4787	1.4803

Continuation of Table 3.10.

Characteristic level: 6x6x4, ncase=1000												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
67	-5.28	-4.3353	0.6109	0.8706	0.9447	1.6505		-5.7604	0.8213	1.115	-0.4804	1.4739
68	-5.7	-4.7748	0.4362	0.6616	0.9252	1.2937		-6.2403	1.006	1.2121	-0.5403	1.7611
69	-6.2	-5.1437	0.8573	1.1366	1.0563	2.4077		-6.8506	1.1259	1.4396	-0.6506	2.4956
70	-6.19	-5.2501	0.6854	0.9351	0.9399	1.7578		-6.8182	1.1022	1.4097	-0.6282	2.3819
71	-6.18	-5.2679	0.6749	0.9428	0.9121	1.7207		-6.8099	1.0994	1.4052	-0.6299	2.3714
72	-6.6	-5.7988	0.4542	0.6986	0.8012	1.13		-7.2898	1.2947	1.5102	-0.6898	2.7564
73	-6.67	-5.6837	0.8859	1.19	0.9863	2.3888		-7.433	1.3428	1.6259	-0.763	3.2258
74	-6.66	-5.8248	0.5286	0.744	0.8352	1.2511		-7.4006	1.3162	1.5879	-0.7406	3.0698
75	-6.65	-5.8305	0.5233	0.767	0.8195	1.2599		-7.3923	1.3125	1.5868	-0.7423	3.0689
76	-7.07	-6.3832	0.4985	0.734	0.6868	1.0104		-7.8722	1.5229	1.707	-0.8022	3.5575
77	-6.65	-5.6897	0.9263	1.2032	0.9603	2.3698		-7.4326	1.3795	1.6341	-0.7826	3.2826
78	-6.64	-5.8368	0.5273	0.7711	0.8032	1.2397		-7.4002	1.3524	1.5993	-0.7602	3.1357
79	-6.63	-5.8163	0.5248	0.7849	0.8137	1.2782		-7.3919	1.3488	1.5938	-0.7619	3.1206
80	-7.05	-6.3266	0.498	0.7529	0.7234	1.0901		-7.8718	1.5585	1.7158	-0.8218	3.6194
81	-6.67	-5.6283	0.947	1.2664	1.0417	2.6888		-7.4121	1.3697	1.6066	-0.7421	3.1319
82	-6.66	-5.8533	0.5228	0.7722	0.8067	1.2471		-7.3796	1.3425	1.5669	-0.7196	2.9732
83	-6.65	-5.8223	0.5198	0.7533	0.8277	1.2526		-7.3714	1.339	1.5616	-0.7214	2.9589
84	-7.07	-6.2904	0.5763	0.833	0.7796	1.3016		-7.8512	1.5504	1.6916	-0.7812	3.472
85	-6.57	-5.5071	0.9501	1.2362	1.0629	2.6578		-7.3242	1.333	1.5865	-0.7542	3.0859
86	-6.56	-5.6979	0.6109	0.8958	0.8621	1.5458		-7.2918	1.3086	1.5506	-0.7318	2.9399
87	-6.55	-5.704	0.6104	0.8951	0.846	1.5169		-7.2835	1.3052	1.5466	-0.7335	2.9301
88	-6.97	-6.1567	0.5914	0.8421	0.8133	1.3705		-7.7634	1.5129	1.664	-0.7934	3.3982

Continuation of Table 3.10.

Characteristic level: 6x6x4, ncase=1000												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
89	-5.37	-4.3831	0.8658	1.0665	0.9869	2.1115		-5.8956	0.8794	1.1818	-0.5256	1.6728
90	-5.36	-4.4444	0.5448	0.7911	0.9156	1.4641		-5.8631	0.8494	1.1455	-0.5031	1.5653
91	-5.35	-4.4096	0.4529	0.6888	0.9404	1.3589		-5.8549	0.8403	1.1419	-0.5049	1.5587
92	-5.77	-4.8553	0.479	0.6905	0.9147	1.3134		-6.3347	1.0431	1.2598	-0.5647	1.906
93	-6.27	-5.246	0.9487	1.2476	1.024	2.605		-6.945	1.1596	1.4752	-0.675	2.632
94	-6.26	-5.3557	0.5054	0.7596	0.9043	1.3948		-6.9126	1.1289	1.4405	-0.6526	2.5008
95	-6.25	-5.4063	0.5028	0.7229	0.8437	1.2344		-6.9043	1.1254	1.4359	-0.6543	2.4901
96	-6.67	-5.8525	0.499	0.7393	0.8175	1.2148		-7.3842	1.3333	1.5573	-0.7142	2.9352
97	-5.2	-4.2008	0.7869	1.0196	0.9992	2.038		-5.67	0.8026	1.1264	-0.47	1.4897
98	-5.19	-4.2499	0.4538	0.6426	0.9401	1.2967		-5.6376	0.7634	1.0722	-0.4476	1.35
99	-5.18	-4.2119	0.7576	0.9753	0.9681	1.8884		-5.6293	0.7791	1.097	-0.4493	1.4054
100	-5.6	-4.5892	0.8315	1.0651	1.0108	2.1562		-6.1092	0.9701	1.2061	-0.5092	1.714
101	-5.18	-4.1406	1.0738	1.1813	1.0394	2.4757		-5.6696	0.8483	1.1431	-0.4896	1.5463
102	-5.17	-4.2324	0.5906	0.81	0.9376	1.5351		-5.6371	0.8093	1.0943	-0.4671	1.4157
103	-5.16	-4.095	0.9908	1.1827	1.065	2.5331		-5.6289	0.8237	1.1122	-0.4689	1.4569
104	-5.58	-4.4087	1.0723	1.236	1.1713	2.8996		-6.1087	1.0128	1.2232	-0.5287	1.7758
105	-5.2	-4.0681	1.0824	1.2246	1.1319	2.7808		-5.649	0.8437	1.1271	-0.449	1.472
106	-5.19	-4.2007	0.5921	0.8069	0.9893	1.6299		-5.6166	0.8045	1.0713	-0.4266	1.3296
107	-5.18	-4.1275	0.9921	1.2076	1.0525	2.5662		-5.6083	0.8193	1.09	-0.4283	1.3716
108	-5.6	-4.2379	1.2659	1.2783	1.3621	3.4895		-6.0882	1.0102	1.211	-0.4882	1.7048
109	-5.1	-3.9923	1.0969	1.1523	1.1077	2.5547		-5.5612	0.8138	1.1047	-0.4612	1.4331
110	-5.09	-4.1471	0.681	0.8797	0.9429	1.663		-5.5288	0.779	1.054	-0.4388	1.3035

Continuation of Table 3.10.

Characteristic level: 6x6x4, ncase=1000												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
111	-5.08	-3.8997	1.1573	1.2032	1.1803	2.8408		-5.5205	0.7947	1.0751	-0.4405	1.3499
112	-5.5	-4.3102	1.2332	1.3246	1.1898	3.1703		-6.0004	0.9767	1.178	-0.5004	1.6382
113	-3.9	-3.1398	0.7088	0.8825	0.7602	1.3566		-4.1325	0.4719	0.7531	-0.2325	0.6213
114	-3.89	-3.0835	0.4548	0.6072	0.8065	1.0191		-4.1001	0.4236	0.6971	-0.2101	0.53
115	-3.88	-3.0369	0.6083	0.7541	0.8431	1.2795		-4.0918	0.446	0.7313	-0.2118	0.5797
116	-4.3	-3.3914	0.7231	0.8914	0.9086	1.6202		-4.5717	0.5604	0.8206	-0.2717	0.7471
117	-4.8	-3.7767	0.9896	1.1357	1.0233	2.3369		-5.182	0.6589	0.9982	-0.382	1.1422
118	-4.79	-3.8619	0.5067	0.7311	0.9281	1.3959		-5.1496	0.6107	0.9477	-0.3596	1.0274
119	-4.78	-3.8331	0.8446	0.9879	0.9469	1.8725		-5.1413	0.6313	0.9707	-0.3613	1.0728
120	-5.2	-4.1237	0.9538	1.1192	1.0763	2.4111		-5.6212	0.8042	1.0773	-0.4212	1.3381
121	-4.7	-3.7374	0.6803	0.8383	0.9626	1.6294		-5.0815	0.7078	0.989	-0.3815	1.1238
122	-4.69	-3.7051	0.6658	0.8967	0.9849	1.774		-5.0491	0.6871	0.9587	-0.3591	1.048
123	-4.68	-3.7284	0.6443	0.8663	0.9516	1.656		-5.0408	0.6848	0.9513	-0.3608	1.0352
124	-5.1	-4.1281	0.7468	0.9694	0.9719	1.8844		-5.5207	0.8706	1.0616	-0.4207	1.3039
125	-4.68	-3.7542	0.9372	1.0511	0.9258	1.9619		-5.0811	0.7549	1.0214	-0.4011	1.2041
126	-4.67	-3.6983	0.8708	1.0034	0.9717	1.951		-5.0487	0.7334	0.997	-0.3787	1.1374
127	-4.66	-3.6597	0.8628	1.0146	1.0003	2.0299		-5.0404	0.7306	0.9827	-0.3804	1.1104
128	-5.08	-4.1046	0.9593	1.0887	0.9754	2.1366		-5.5203	0.9139	1.0934	-0.4403	1.3895
129	-4.7	-3.7402	0.9528	1.0428	0.9598	2.0085		-5.0606	0.7514	1.0053	-0.3606	1.1407
130	-4.69	-3.6853	0.8847	1.0528	1.0047	2.1177		-5.0282	0.7299	0.9736	-0.3382	1.0622
131	-4.68	-3.7773	0.8706	1.0742	0.9027	1.9688		-5.0199	0.7274	0.9594	-0.3399	1.036
132	-5.1	-3.952	1.1492	1.2125	1.148	2.7881		-5.4998	0.913	1.0814	-0.3998	1.3293

Continuation of Table 3.10.

Characteristic level: 6x6x4, ncase=1000												
	MLE							**PCL**				
j	**true beta (B_j)**	**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**		**MC average**	**est(se)**	**MC standard error**	**bias**	**MSE**
133	-4.6	-3.6756	0.9508	1.0681	0.9244	1.9955		-4.9727	0.7238	0.9911	-0.3727	1.1211
134	-4.59	-3.6779	0.9784	1.1243	0.9121	2.0961		-4.9403	0.7086	0.9657	-0.3503	1.0554
135	-4.58	-3.6152	1.0132	1.0911	0.9648	2.1213		-4.932	0.7063	0.9538	-0.352	1.0336
136	-5	-3.9047	1.1194	1.1535	1.0953	2.5302		-5.4119	0.8805	1.0547	-0.4119	1.282
137	-3.4	-2.6858	0.5518	0.7027	0.7142	1.0038		-3.5441	0.4271	0.6484	-0.1441	0.4412
138	-3.39	-2.712	0.6039	0.7307	0.678	0.9936		-3.5117	0.4165	0.6317	-0.1217	0.4139
139	-3.38	-2.7167	0.482	0.6496	0.6633	0.8619		-3.5034	0.4044	0.6166	-0.1234	0.3954
140	-3.8	-2.9669	0.5797	0.7232	0.8331	1.2171		-3.9833	0.4838	0.6995	-0.1833	0.5229
141	-4.3	-3.415	0.8453	0.9685	0.885	1.7211		-4.5936	0.5762	0.8824	-0.2936	0.8649
142	-4.29	-3.4023	0.7149	0.8782	0.8877	1.5593		-4.5611	0.5518	0.8605	-0.2711	0.8139
143	-4.28	-3.4148	0.7153	0.8601	0.8652	1.4884		-4.5529	0.5492	0.8466	-0.2729	0.7912
144	-4.7	-3.709	0.8177	1.0164	0.991	2.015		-5.0327	0.7103	0.9517	-0.3327	1.0165

CHAPTER IV

DATA ANALYSIS

In this chapter, we focus on a data set obtained from a case-control study conducted in Ankara Oncology Research and Education Hospital. There are 500 female subjects in the study 249 of whom are cases (women with breast cancer) and 251 are controls (women without breast cancer). The information about breast cancer characteristics are ascertained from each case. These are the size, type, and the grade of the tumor, and NA, NB, ER, PR, and C-erb-B2 status. The information on the health related risk factors, namely age at first birth, hormone replacement therapy (HRT) status, number of births, and body mass index (BMI), as well as demographical adjusting factors namely age, education level, and region they reside are also collected from both cases and controls. In cancer studies age is usually a confounding factor and thus it should be included in the model. The aim of the study is to investigate the association between the breast cancer related risk factors and the cancer characteristics, and determine the factors significantly affecting the tumor characteristics.

In section 4.1. characteristics of the study sample are shown. In section 4.2. association study is carried out by using both Two Stage Polytomous Logistic Regression with Pseudo-Conditional Likelihood (PCL) approach and unstructured polytomous logistic regression with maximum likelihood estimation (MLE) method. The first method takes all disease characteristics simultaneously into consideration whereas the second method considers each characteristic individually and independently. In that section our main aim is to compare the efficiencies of these two approaches on a typical case-control data set that is typical in studies focusing on cancer research. In section 4.3. we focus on a specific hypothesis and consider two different approaches for testing: two stage polytomous logistic regression with

PCL and unstructured polytomous logistic regression with MLE when response levels are cross-classifications of disease characteristics levels. Our main aim in section 4.3 is to display the convenience provided by the two stage polytomous logistic regression model for testing whether the ORs associated with one disease characteristic changes with respect to another disease characteristics.

4.1. STUDY SAMPLE

As it is stated before, the data are collected in Ankara Oncology Research and Education Hospital from 249 women with breast cancer and 251 women without. Breast cancer is the cancer that occurs in the breast tissue. It generally takes place in the inner lining of milk ducts or the lobules where the milk is produced (Sariego, 2010). It arises from the interaction between a defective gene and environmental conditions. Normally, a cell is divided as many as it is needed and it stops. However, a cancerous cell loses its feature to stop dividing because of the mutations and they no longer stay at the tissue that they belong. Those rapidly growing cells constitute tumors which has certain characteristics like sensitiveness to hormones, tumor type, size, grade, receptor status. In a cell’s cytoplasm and nucleus, receptors function to keep hormones to starting reactions in the cell. Breast cancer cells may have three main receptors on its surface: Estrogen receptor (ER), progesterone receptor (PR) or HER2/neu (C-erb B2) receptor. People without the disease do not have these receptors. For a patient, if ER is positive that means breast cancer is triggered because of the release of estrogen hormone. Likewise, for PR positive, tumors grow in response to progesterone release. C-erb B2 is the “human epidermal growth factor receptor-2” which is a protein that causes aggressiveness in breast cancers (Sotiriou and Pusztai, 2009). Luminal A, luminal B, HER-2 status define the stages of the cancer. Luminal A is an earlier stage in which the cells are similar to nondiseased cells. Those cells have estrogen and progesterone receptors, but do not have HER-2 receptor. Luminal A is known as having low risk of recurrence. Luminal B also has ER and PR but in smaller quantities. That stage of cancer is also in low risk of recurrence. HER-2 positive tumors involve gene mutations related with the human epidermal growth factor. Triple negative tumors have all three receptors as negative but considered as high risk since they are inclined to grow rapidly. Grade of the tumor is determined according to whether cells are differentiated from each other. In an organism, cells take different shapes and functions to form an organ. Cancerous cells become hard to detect that difference. Tumor cells are classified as well differentiated (low grade), moderately differentiated (intermediate grade) and poorly differentiated (high grade), or categorization is grade 1, grade 2, grade 3 respectively. The higher the grade, the more irregular the cell behavior (Harris et al., 2010). In addition to the tumor classification based on grade, tumor type also provides information about the appearance of the cancer. These special types are invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC) and tubular carcinoma. IDC is the cancer which starts at the ducts of the breast and spreads to the duct walls and neighbouring tissue. This is the most common type of the breast cancer. ILC is the cancer that starts at the lobules of the breast and then spreads to the lobule and neighbouring tissue (Nass et al. 2001). Margins of the ILC's are poorly defined so that detecting them by mammography is harder (Miller et al., 2002). Tubular carcinoma is the cancer that starts at the tubules of breast. It is one of the best characterized breast cancer so that it has important prognosis features (Brooks and Harris, 2006). Another classification based on the dimensions of the tumor (tumor size) is as follows: T1: tumor is less than or equal to 20 mm, T2: tumor is between 20 mm and 50 mm, T3: tumor is greater than 50 mm in its greatest dimension, and T4: tumor is directly extended to chest wall or skin with any dimension. Breast cancer may spread to axillary lymph nodes. NA is the status of having the nodes in axillary nodes or not having. In addition, NB shows the number of nodes existing with levels such as N1: Cancer spread to 1-3 axillary lymph nodes, N2: Cancer spread to 4-9 lymph nodes, N3: Cancer has spread to more than 10 lymph nodes. If the number of nodes that cancer is extended increases, then the prognosis is worse (Brooks and Harris, 2006).

A univariate analysis for the association between breast cancer and risk factors are conducted by using chi-square and t-tests (Table 4.1 and Table 4.2). Following tables also show the distribution of the risk factors of interest and adjusting factors for cases and controls.

Table 4.1: Percentages for levels of categorical factors with respect to disease status (case, control), and chi-square test for indepence of disease and factors.

Factor	***Factor levels***	***Case***		***Control***			***P-val.***
		(%)	***(N)***	***(%)***	***(N)***	***Total***	
HRT	*Not taking*	42 %	210	39.4%	197	407	
	Taking	7.8%	39	10.8%	54	93	
	Total		249		251	500	0.092
Family history	*Absent*	38.4%	192	39%	195	387	
	1st order relative	8.4%	42	9.2%	46	88	
	2nd order relative	3%	15	2%	10	25	
	Total		249		251	500	0.549
Mammography	*Never*	31.8%	159	20.8%	104	263	
	Semi-annually	18%	90	29.4%	147	237	
	Total		249		251	500	<.001
Education	*No education*	7.8%	39	3%	15	54	
	Primary school	22.4%	112	22%	110	222	
	Secondary sch.	6.2%	31	5.4%	27	58	
	High school	5.8%	29	10.6%	53	82	
	University	7.6%	38	9.2%	46	84	
	Post graduate	0%	0	0%	0	0	
	Total		249		251	500	<.001
Region	*Central Anatolia*	29.2%	146	33%	165	311	
	East /South-East Anatolia	6.8%	34	5.2%	26	60	
	Blacksea	8.6%	43	7%	35	78	
	Mediterrenean & Aegean&Marm.	5.2%	26	5%	25	51	
	Total		249		251	500	0.383

Frequencies and percentages of the levels of categorical risk factors are given in Table 4.1. P-values correspond to the chi-square test for testing the association between breast cancer and factors. Results revealed that mammography screening status has association with breast cancer status (p<0.001). Also, one of the adjusting factors, education, is not independent of breast cancer, that means, having breast cancer differs for level of the education (p<0.001).

Table 4.2: Sample means for continuous risk factors with respect to disease status, and t-test for the difference of means

Factor	**Case**	**Control**	
	Average *(s.d.)*	**Average *(s.d.)***	**P-value**
Age	51.752 *(10.94)*	46.32 *(9.96)*	<.0001
Age at first menstruation	13.464 *(1.46)*	13.552 *(1.36)*	0.622
Age at first birth	21.946 *(5.17)*	21.611 *(4.80)*	0.614
Age at last birth	29.705 *(5.56)*	27.531 *(5.41)*	0.003
Duration of breast feeding	27.5 *(21.83)*	23.143 *(20.32)*	0.115
Number of births	2.624 *(1.58)*	2.192 *(1.34)*	0.021
Age at menopause	47.713 *(5.29)*	45.711 *(5.20)*	0.043
BMI(pre-menopause)	28.68 *(4.74)*	26.59 *(5.18)*	0.002
BMI(post-menopause)	29.27 *(5.06)*	28.20 *(5.39)*	0.122

Table 4.2 shows the t-test results for the difference of covariates of cases and controls. According to the results, number of births (p=0.02), age at menopause (p=0.04) and BMI before the menopause (p=0.002) differs for diseased and nondiseased women. That is to say, these factors have statistically significant relationship with breast cancer status.

4.2. LOGISTIC REGRESSION ANALYSES FOR ASSOCIATION BETWEEN RISK FACTORS AND TUMOR CHARACTERISTICS

In this section, we will investigate the association between the risk factors of interest and the tumor characteristics both using classical logistic regression and two stage polytomous logistic regression.

4.2.1. Binary/Polytomous/Ordinal Logistic Regression

In order to investigate the association between the tumor characteristics and the risk factors of interest, the health scientist (e.g. the cancer researcher, the cancer epidemiologist, or the biostatistician) tends to perform the statistical analyses on each tumor characteristic separately. That is, they consider only the case data (i.e. the data set that belongs to the breast cancer patients only) and use logistic regression for each tumor characteristic to model the relationship between the probability of the occurrence of a certain characteristic and the risk factors. They apply either dichotomous logistic regression or polytomous logistic regression or ordinal logistic regression depending on the number of levels and scale of the specific tumor characteristic. For instance, the grade of the tumor which can fall into either one of the three classifications, namely low, intermediate, or high, is modeled using ordinal logistic regression whereas the ER status of the tumor which can be either negative or positive is modeled using binary logistic regression. These models are well established and there are many sources today to which one may refer to see the models and methods of analyses (e.g. see Kleinbaum and Klein (2010), Hosmer and Lemeshow (2000)). In this part we employ this approach which is applied by many practitioners as the default procedure.

The logistic regression for each tumor characteristic is fit and the corresponding analyses are performed in SAS software. The odds ratio (OR) and corresponding p-value are obtained for the association between tumor characteristic and factor by considering each tumor characteristic as a response variable. Since the number of cases corresponding to the last level of the response tumor size, the categories *T3* and *T4* are combined. Likewise, *ILC* and *Tubular* are combined for the response *tumor type*, and the categories *N2* (4-9 lymph nodes exist) and *N3* (more than 10 lymph nodes exist) are combined for the response *NB*. Covariates in each model are: age, age at first menstruation, hormone replacement therapy (HRT) status, duration of breast feeding, family history of breast cancer, number of births, age at first birth, mammography history, education level, age at menopause, body-mass index (BMI) for pre-menopause and post-menopause women, geographical region, and menopause status.

Based on these analyses, we find that some of the factors under investigation have statistically significant associations with some tumor characteristics (Table 4.3). Mammography screened woman have lower odds of developing grade 3 tumor versus grade 1 tumor compared to woman not mammography screened (OR=0.207, p=0.0006). Women who take hormone replacement therapy are more likely to have positive C-erb-B2 receptor on breast tumor (OR=2.298, p=0.047) . Women who breast-fed their children in a longer duration have more risk to develop 1-3 enlarged lymph nodes (OR=1.02, p=0.029).

Table 4.3: OR's, CI's, and p-values for polyt/binary/ordinal logistic regr. models.

	Response: Tumor size					
	Tumor size: T2 **(ref= Tumor size: T1)**			**Tumor size: T3,T4** **(ref= Tumor size: T1)**		
	OR	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Age	0.94	0.902 0.980	0.0039	0.961	0.907 1.019	0.1851
Age at first menstruation	1.05	0.818 1.348	0.7018	1.096	0.761 1.578	0.6228
HRT taken (ref=not taken)	1.514	0.587 3.902	0.3905	0.675	0.154 2.960	0.6018
Duration of breast feeding	1.013	0.992 1.035	0.2243	1.012	0.988 1.038	0.3224
Family history (1st degree) (ref=absent)	1.189	0.487 2.899	0.7038	1.084	0.288 4.082	0.9047
Family history (2nd degree) (ref=absent)	2.389	0.468 12.193	0.2949	0.832	0.064 10.760	0.8882
Number of births	0.942	0.660 1.345	0.7433	0.882	0.537 1.449	0.6203
Age at first birth	0.952	0.877 1.033	0.2408	1.023	0.919 1.137	0.6814
Mammography (ref=no)	0.599	0.298 1.204	0.1505	0.394	0.141 1.102	0.0759
Education	0.891	0.649 1.223	0.4740	0.9	0.580 1.397	0.6386
Age at menopause	1.039	0.960 1.124	0.3429	0.967	0.859 1.088	0.5727
BMI (pre-menopause)	1.1086	0.9631 1.2761	0.1508	1.1511	0.9703 1.3658	0.1065
BMI (post menopause)	1.0161	0.9296 1.1108	0.7247	1.0737	0.9474 1.2168	0.2654
Central Anatolia (ref=Other)	1.557	0.532 4.560	0.4189	0.979	0.227 4.219	0.9776
East and South-East Anatolia (ref=Other)	1.035	0.279 3.833	0.9588	1.566	0.286 8.579	0.6052
Black Sea (ref=Other)	2.073	0.575 7.480	0.2653	1.713	0.298 9.842	0.5463
Births status (ref=no birth)	1.452	0.097 21.802	0.7874	0.355	0.009 13.759	0.5789
Menopause status (ref=no)	4.5928	0.0133 1583	0.6091	30.0361	0.0086 10459	0.4135

Continuation of Table 4.3

	Response: Grade					
	Grade 2 **(ref= Grade 1)**			**Grade 3** **(ref= Grade 1)**		
	OR	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Age	0.978	0.93 1.029	0.3886	0.977	0.928 1.03	0.3948
Age at first menstruation	0.982	0.734 1.315	0.9050	1.012	0.742 1.381	0.9407
HRT taken (ref=not taken)	1.083	0.353 3.319	0.8893	1.26	0.398 3.988	0.6940
Duration of breast feeding	1.017	0.99 1.046	0.2155	1.016	0.988 1.045	0.2602
Family history (1st degree) (ref=absent)	1.476	0.491 4.442	0.4882	1.735	0.542 5.556	0.3534
Family history (2nd degree) (ref=absent)	0.756	0.152 3.752	0.7317	0.847	0.161 4.459	0.8451
Number of births	0.944	0.577 1.544	0.8183	1.115	0.674 1.846	0.6713
Age at first birth	0.989	0.888 1.101	0.8333	1.051	0.94 1.175	0.3825
Mammography (ref=no)	0.296	0.126 0.694	0.0051	0.207	0.085 0.508	**0.0006**
Education	0.982	0.673 1.434	0.9258	0.905	0.608 1.347	0.6227
Age at menopause	1.05	0.953 1.158	0.3227	0.982	0.885 1.089	0.7272
BMI (pre-menopause)	1.0867	0.9148 1.2911	0.3439	1.1306	0.7456 1.0491	0.1583
BMI (post menopause)	1.0786	0.9659 1.2045	0.1791	1.0838	0.7538 0.9536	0.1797
Central Anatolia (ref=Other)	1.028	0.251 4.215	0.9699	0.807	0.192 3.395	0.7694
East and South-East Anatolia (ref=Other)	0.817	0.150 4.438	0.8150	0.873	0.156 4.874	0.8770
Black Sea (ref=Other)	1.747	0.337 9.06	0.5067	1.248	0.227 6.855	0.7989
Births status (ref=no birth)	2.296	0.081 65.387	0.6267	0.254	0.008 8.245	0.4399
Menopause status (ref=no)	0.2112	0.00182 243.83	0.6656	6.1564	0.0042 8998.2	0.6250

Continuation of Table 4.3

	Response: Tumor type			**Response: NA**		
	Tumortype: ILC/Tubular (ref=tumor type IDC)			**NA: enlarged lymph nodes exist (ref= no enlarged lymph nodes)**		
	OR	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Age	1.046	0.991 1.103	0.1004	0.958	0.924 0.993	0.0188
Age at first menstruation	1.001	0.741 1.351	0.9955	1.048	0.855 1.283	0.6531
HRT taken (ref=not taken)	0.934	0.272 3.210	0.9133	1.184	0.532 2.634	0.6794
Duration of breast feeding	0.996	0.976 1.018	0.7349	1.017	1.000 1.034	0.05
Family history (1st degree) (ref=absent)	1.616	0.588 4.441	0.3524	1.347	0.620 2.924	0.4518
Family history (2nd degree) (ref=absent)	0.329	0.031 3.525	0.3586	2.189	0.620 7.719	0.2233
Number of births	0.944	0.617 1.443	0.7887	0.711	0.528 0.959	0.0254
Age at first birth	0.96	0.857 1.076	0.4821	1.005	0.936 1.078	0.901
Mammography (ref=no)	1.114	0.470 2.638	0.8065	0.624	0.342 1.139	0.1244
Education	1.166	0.775 1.756	0.4609	0.685	0.517 0.907	0.0082
Age at menopause	1.012	0.921 1.112	0.8065	0.988	0.925 1.056	0.7277
BMI (pre-menopause)	1.1278	0.9632 1.3206	0.1347	0.9111	0.8169 1.0162	0.0946
BMI (post menopause)	0.9452	0.8421 1.0608	0.3384	0.9535	0.8892 1.0233	0.1878
Central Anatolia (ref=Other)	1.976	0.231 16.883	0.5337	1.653	0.638 4.283	0.3007
East and South-East Anatolia(ref=Other)	6.701	0.715 62.772	0.0956	1.313	0.415 4.152	0.6435
Black Sea (ref=Other)	5.782	0.632 52.915	0.1203	1.284	0.424 3.892	0.6586
Births status (ref=no birth)	2.08	0.070 62.148	0.6726	1.23	0.132 11.462	0.8558
Menopause status (ref=No)	89.2106	0.0611 13210	0.227	0.4439	0.0034 58.807	0.7446

Continuation of Table 4.3

	Response: NB					
	NB: 1-3 enlarged lymph nodes (ref=no enlarged lymph nodes)			**NB: more than4-9/ more than 10 enlarged lymph nodes (ref=no enlarged lymph nodes)**		
	OR	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Age	0.96	0.924 0.998	0.0416	0.957	0.913 1.003	0.0684
Age at first menstruation	1.023	0.819 1.278	0.8378	1.069	0.809 1.411	0.6399
HRT taken (ref=not taken)	1.388	0.593 3.245	0.4495	0.815	0.265 2.505	0.7216
Duration of breast feeding	1.02	1.002 1.039	0.0287	1.012	0.992 1.033	0.2527
Family history (1st degree) (ref=absent)	1.554	0.680 3.549	0.296	0.97	0.313 3.003	0.9579
Family history (2nd degree) (ref=absent)	2.395	0.624 9.185	0.2029	1.989	0.392 10.100	0.4069
Number of births	0.627	0.441 0.891	0.0092	0.867	0.598 1.257	0.4511
Age at first birth	0.988	0.913 1.069	0.7582	1.039	0.946 1.141	0.4228
Mammography (ref=no)	0.621	0.321 1.200	0.1565	0.644	0.281 1.473	0.2969
Education	0.7	0.517 0.949	0.0215	0.64	0.437 0.937	0.0219
Age at menopause	0.999	0.930 1.073	0.9769	0.968	0.879 1.066	0.5099
BMI (pre-menopause)	0.9055	0.8036 1.0203	0.103	0.9119	0.8002 1.0393	0.1666
BMI (post menopause)	0.9388	0.8672 1.0164	0.1194	0.9948	0.9025 1.0966	0.9167
Central Anatolia (ref=Other)	1.841	0.647 5.239	0.2528	1.277	0.352 4.631	0.7097
East and South-East Anatolia (ref=Other)	1.429	0.407 5.020	0.5773	1.061	0.219 5.129	0.9414
Black Sea (ref=Other)	0.912	0.258 3.221	0.8857	1.92	0.459 8.029	0.3718
Births status (ref=no birth)	2.195	0.186 25.847	0.5321	0.417	0.021 8.187	0.565
Menopause status (ref=No)	0.4415	0.0021 92.528	0.7643	0.5918	0.0031 149.148	0.6472

Continuation of Table 4.3

	Response: ER			**Response: PR**		
	ER: (+) (ref= ER: (-))			**PR: (+) (ref= PR: (-))**		
	OR	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Age	1.01	0.973 1.047	0.6134	0.987	0.953 1.021	0.4363
Age at first menstruation	1.031	0.829 1.282	0.7861	0.901	0.737 1.101	0.3074
HRT taken (ref=not taken)	0.633	0.289 1.386	0.2525	0.931	0.442 1.963	0.8514
Duration of breast feeding	1.004	0.990 1.019	0.5741	1.013	0.997 1.028	0.1028
Family history (1st degree) (ref=absent)	0.487	0.223 1.062	0.0703	0.554	0.263 1.168	0.1208
Family history (2nd degree) (ref=absent)	1.292	0.353 4.738	0.6988	1.108	0.349 3.519	0.8615
Number of births	1.05	0.770 1.431	0.7579	0.893	0.672 1.186	0.4357
Age at first birth	1.009	0.934 1.089	0.8272	1.04	0.970 1.115	0.2756
Mammography (ref=no)	1.743	0.904 3.364	0.0973	1.742	0.961 3.157	0.0673
Education	1.117	0.837 1.490	0.4523	0.909	0.698 1.182	0.476
Age at menopause	1.047	0.974 1.125	0.2146	1.063	0.995 1.137	0.0706
BMI (pre-menopause)	0.9512	0.8559 1.0548	0.3374	0.9603	0.8713 1.0584	0.4136
BMI (post menopause)	**0.9541**	0.9233 0.9838	0.2117	0.9677	0.9412 0.9949	0.3590
Central Anatolia (ref=Other)	1.173	0.436 3.158	0.7521	1.777	0.695 4.545	0.2304
East and South-East Anatolia (ref=Other)	0.724	0.223 2.351	0.5916	1.036	0.336 3.198	0.9504
Black Sea (ref=Other)	1.041	0.327 3.313	0.9462	2.128	0.716 6.318	0.174
Births status (ref=no birth)	1.421	0.135 14.910	0.7696	0.323	0.037 2.849	0.8972
Menopause status (ref=No)	0.0712	0.000502 10.86	0.2958	0.000301 2.93	0.0296	0.4363

Continuation of Table 4.3

	Response: C-erb-B2		
	C-erb: (+) **(ref=C-erb-B2: (-))**		
	OR	**95% CI**	**p-value**
Age	0.965	0.926 1.006	0.0935
Age at first menstruation	1.03	0.811 1.307	0.8083
HRT taken (ref=not taken)	2.298	1.011 5.223	0.0471
Duration of breast feeding	1.017	1.000 1.034	0.0517
Family history (1st degree) (ref=absent)	1.262	0.525 3.033	0.6033
Family history (2nd degree) (ref=absent)	0.888	0.218 3.623	0.8685
Number of births	0.859	0.606 1.216	0.3913
Age at first birth	0.949	0.862 1.044	0.28
Mammography (ref=no)	0.542	0.256 1.146	0.1087
Education	0.969	0.699 1.343	0.8513
Age at menopause	0.994	0.917 1.078	0.88
BMI (pre-menopause)	0.9732	0.8712 1.0871	0.6304
BMI (post menopause)	1.0112	0.9750 1.0486	0.7899
Central Anatolia (ref=Other)	1.372	0.435 4.326	0.5897
East and South-East Anatolia (ref=Other)	0.856	0.205 3.574	0.8307
Black Sea (ref=Other)	1.269	0.333 4.834	0.7265
Births status (ref=no birth)	2.409	0.153 37.869	0.5318
Menopause status (ref=No)	0.6596	0.000052 2.76	0.8807

4.2.2. Two Stage Polytomous Logistic Regression

In order to investigate the association between the disease characteristics and the covariates in a multivariate way, when the data on subtypes is available, two stage logistic regression by Chatterjee (2004) is used to conveniently estimate these effects. The important risk factors by clinically important tumor characteristics based on a large scale study in Poland using the two stage approach is established (Garcia-Closas et al., 2006; Sherman et al., 2007). We established the risk factors and their effects by important tumor characteristics for Turkish female breast cancer patients. In this approach, the association between a single tumor characteristic and the covariates is adjusted for the remaining tumor characteristics. This way the association between the specific tumor characteristic and the covariates are determined by holding the other tumor characteristics fixed. The results of this analysis is given in Table 4.4. Women who go through mammography seem less likely to develop grade-2 tumor than grade-1 or grade-3 tumor (OR=0.28, p=0.01; OR=0.21, p=0.006). Higher body mass index for women in post-menopausal term has lower association with having enlarged lymph nodes (OR=0.93, p=0.043). Small number of births is associated with PR positive. That is, the odds of PR positive is higher for woman with less number of births compared to a woman with more number of births (p=0.015). Note that, in this approach, the association between risk factor and a tumor characteristic is adjusted for all the remaining tumor characteristics.

Figures 4.1-4.7 represent the confidence intervals of the ORs estimated from two stage model. The significant associations between breast cancer risk factors and tumor characteristics are detected from these figures.

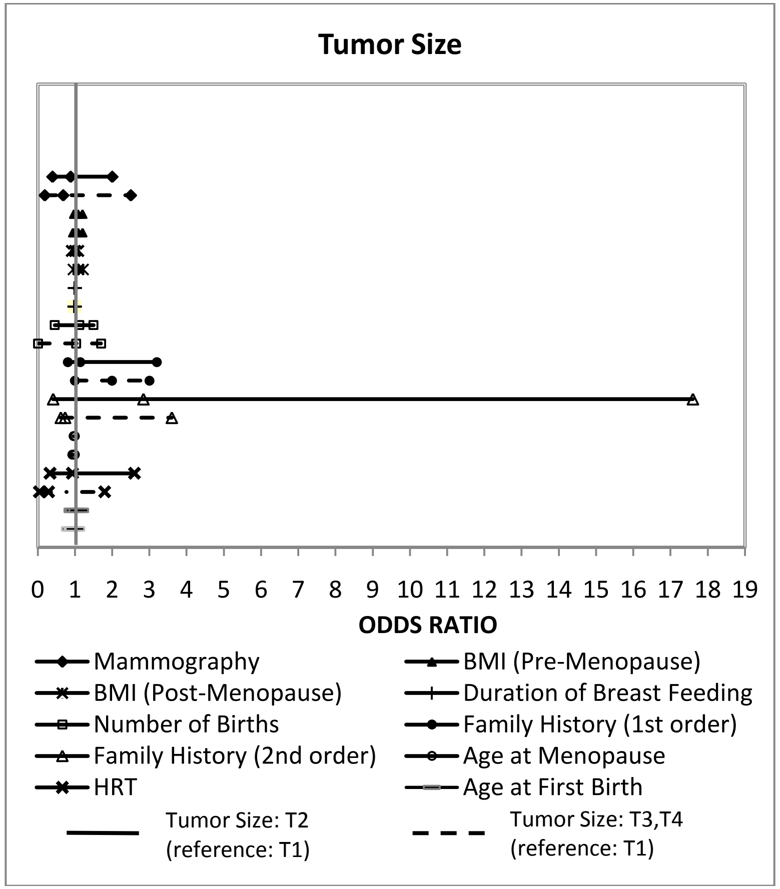


Figure 4.1: 95% Confidence intervals for the OR representing the association between tumor size and risk factors

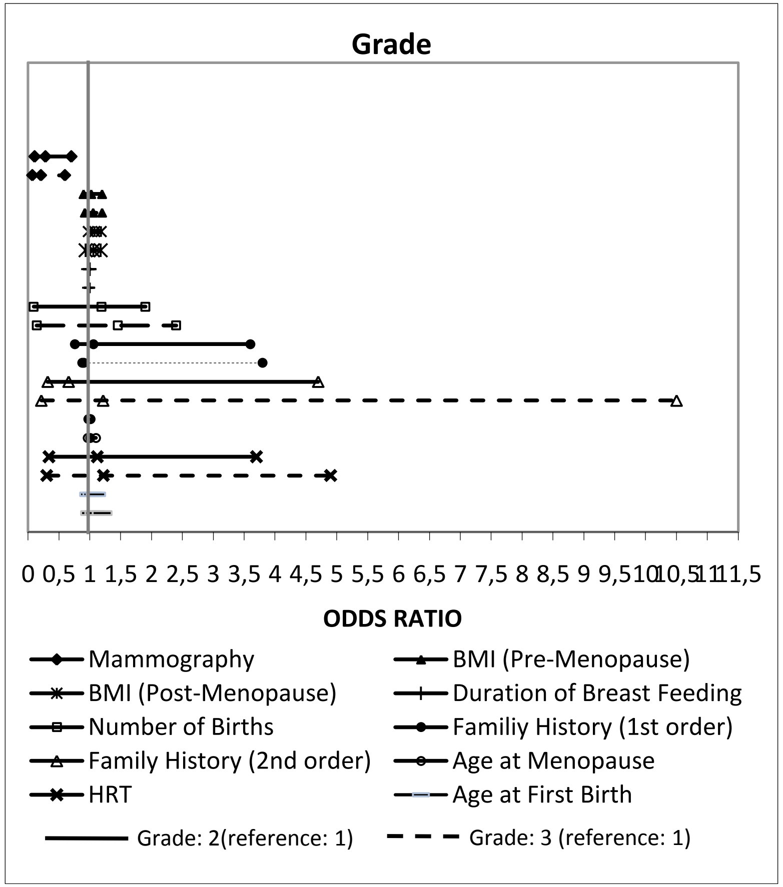


Figure 4.2: 95% Confidence intervals for the OR representing the association between tumor grade and risk factors

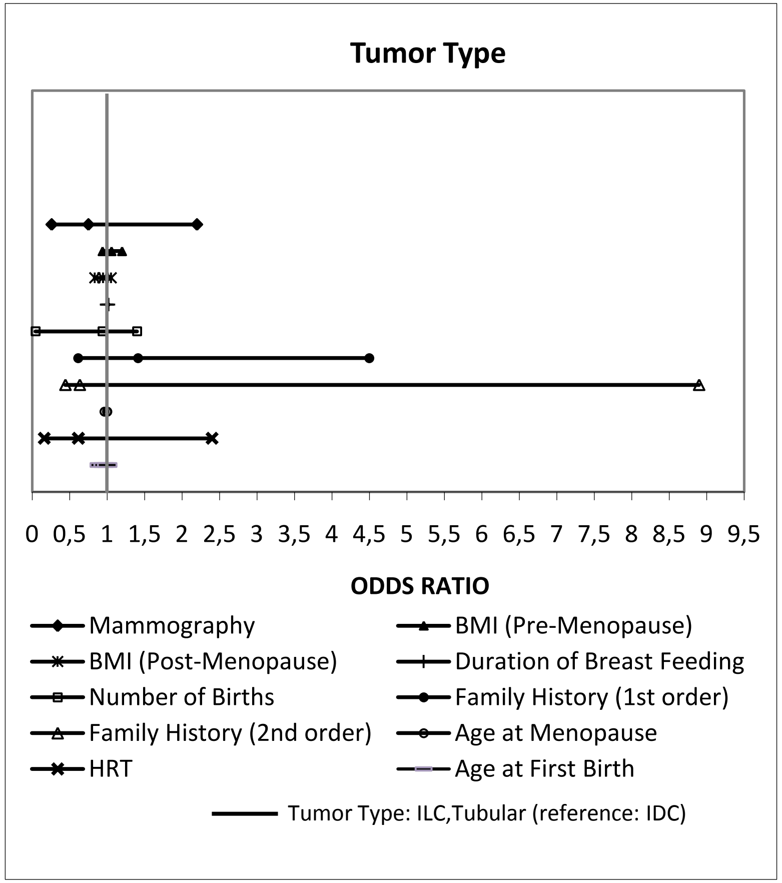


Figure 4.3: 95% Confidence intervals for the ORs representing the association between tumor type and risk factors

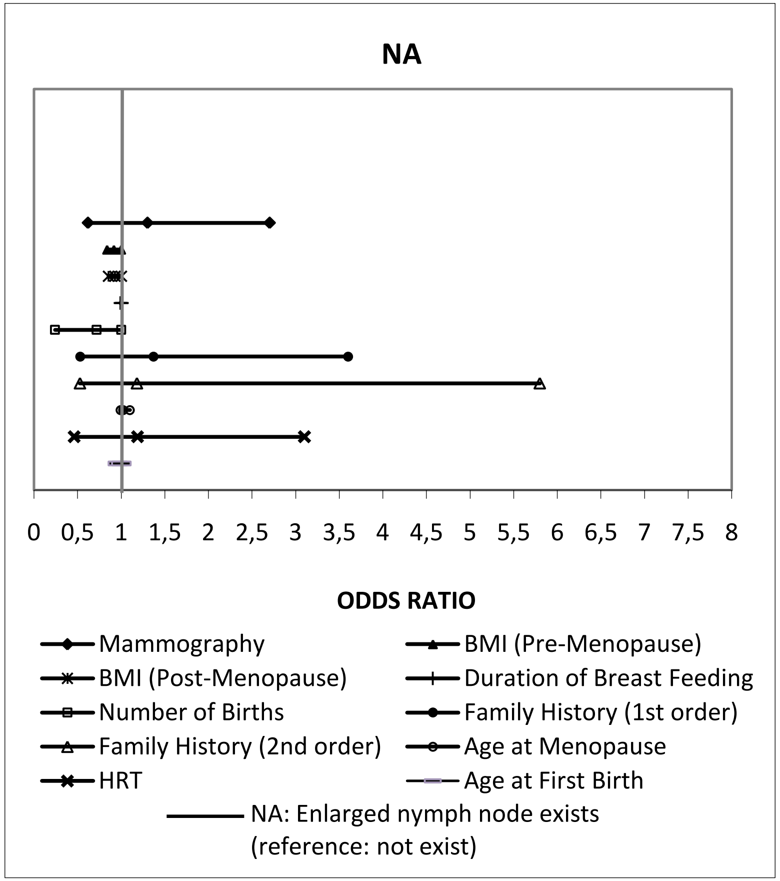


Figure 4.4: 95% Confidence intervals for the ORs representing the association between NA status and risk factors

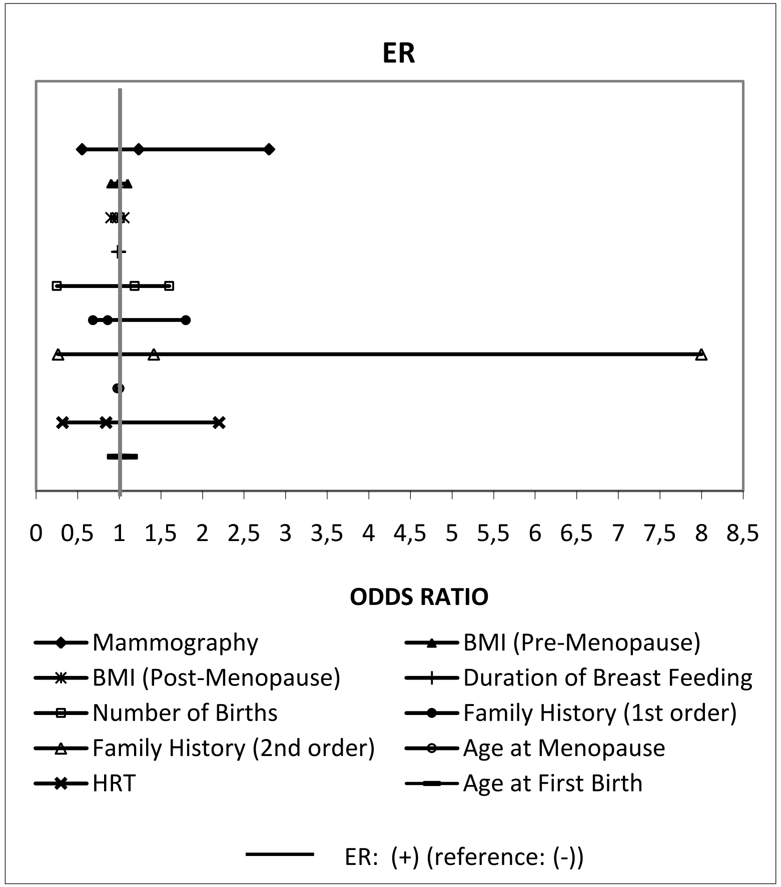


Figure 4.5: 95% Confidence intervals for the ORs representing the association between ER and risk factors

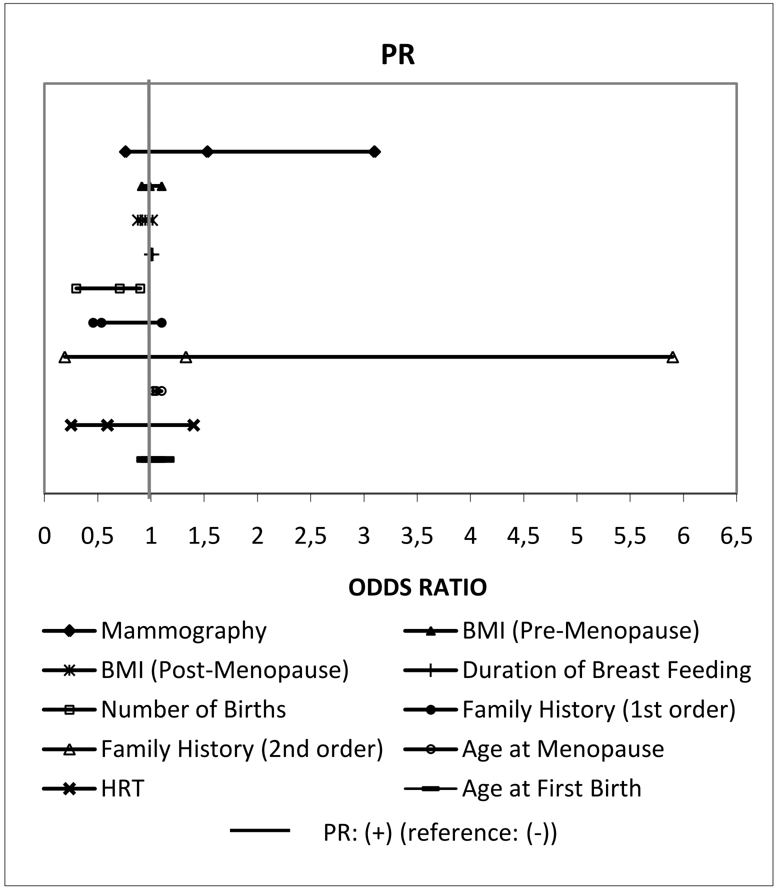


Figure 4.6: 95% Confidence intervals for the ORs representing the association between PR and risk factors

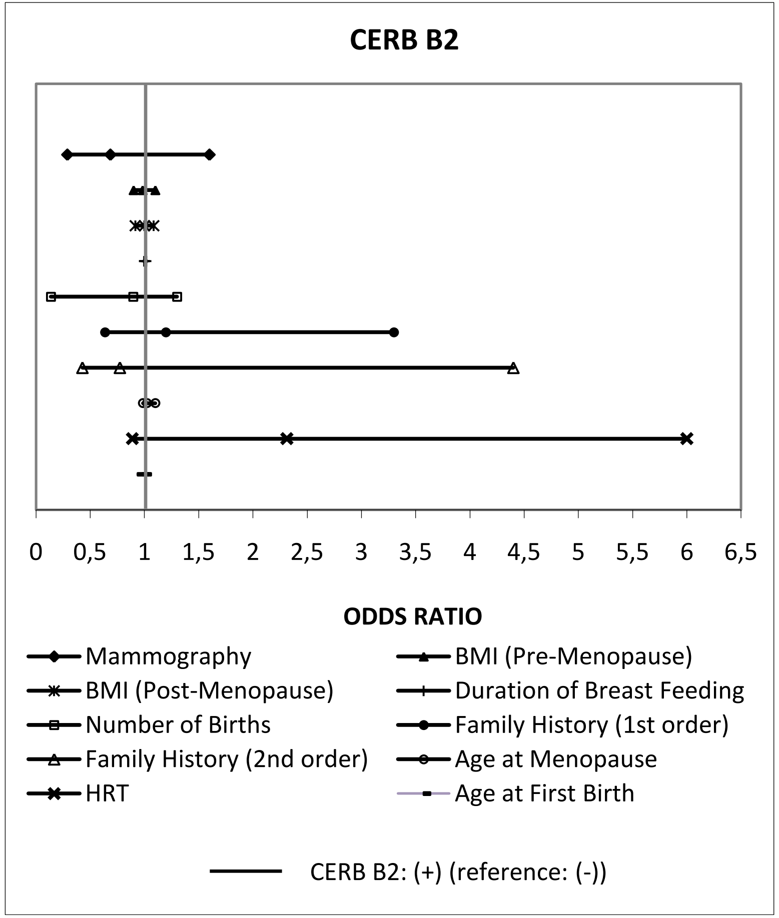


Figure 4.7: 95% Confidence intervals for the ORs representing the association between C-erb-B2 and risk factors

Table 4.4: OR estimates, CIs, and p-values for two stage polyt. logistic regression

	Response: Tumor size					
	Tumor size: T2 (ref= Tumor size: T1)			**Tumor size: T3,T4 (ref= Tumor size: T1)**		
	OR	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Age	0.9657	0.9187 1	0.1713	1.0012	0.9268 1.1	0.975
Age at first menstruation	1.0101	0.7628 1.3	0.944	1.0586	0.6857 1.6	0.7973
HRT taken (ref=not taken)	0.9327	0.3313 2.6	0.8951	0.2899	0.0474 1.8	0.1805
Duration of breast feeding	1.0041	0.9846 1	0.6845	1.0025	0.9727 1	0.8708
Family history (1st degree) (ref=absent)	1.1482	0.8119 3.2	0.792	0.7333	0.6188 3.6	0.7036
Family history (2nd) (ref=absent)	2.8368	0.4111 17.6	0.2622	0.2165	0.1484 5.4	0.3512
Number of births	1.1175	0.4585 1.5	0.4955	1.0321	0.0087 1.7	0.9037
Age at first birth	1.0105	0.9676 1.1	0.6358	0.9683	0.904 1	0.3579
Mammography (ref=no)	0.8862	0.4002 2	0.7657	0.6822	0.1874 2.5	0.5619
Education	1.0516	0.7347 1.5	0.7834	1.3074	0.7516 2.3	0.3426
Age at menopause	0.9972	0.9672 1	0.8567	0.973	0.9258 1	0.2803
BMI (pre-menopause)	1.0784	0.9728 1.2	0.151	1.0814	0.9426 1.2	0.2645
BMI (post menopause)	1.0023	0.9172 1.0873	0.9579	1.0904	0.9584 1.2224	0.1987
Central Anatolia (ref=Other)	1.8892	0.5959 6	0.2799	1.4457	0.2478 8.4	0.6822
East and South-EastAnatolia (ref=Other)	1.0509	0.2355 4.7	0.9481	4.2747	0.4827 37.9	0.1918
Black Sea (ref=Other)	2.1504	0.5183 8.9	0.2916	2.8848	0.3344 24.9	0.3353
Births status (ref=no birth)	0.2358	0.0405 1.4	0.1081	1.1764	0.0857 16.1	0.9033
Menopause status (ref=No)	23.5486	0.4581 1210.4	0.116	2.5773	0.0079 837.7	0.7484

Continuation of Table 4.4.

	Response: Grade					
	Grade 2 (ref= Grade 1)			**Grade 3 (ref= Grade 1)**		
	OR	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Age	1.0175	0.9545 1.1	0.5952	1.0421	0.9688 1.1	0.268
Age at first menstruation	0.8789	0.6269 1.2	0.4541	0.8823	0.5963 1.3	0.531
HRT taken (ref=not taken)	1.1211	0.3393 3.7	0.8514	1.2196	0.3041 4.9	0.7793
Duration of breast feeding	0.9942	0.9689 1	0.6583	0.9905	0.9621 1	0.5229
Family history (1st degree) (ref=absent)	1.0616	0.7604 3.6	0.9228	0.906	0.8793 3.8	0.8924
Family history (2nd) (ref=absent)	0.6559	0.3168 4.7	0.6735	1.2179	0.2168 10.5	0.8576
Number of births	1.1869	0.0923 1.9	0.4506	1.4524	0.1414 2.4	0.145
Age at first birth	1.0468	0.9939 1.1	0.0839	1.0803	1.014 1.2	0.0168
Mammography (ref=no)	0.2826	0.108 0.7	0.01	0.2081	0.0679 0.6	0.006
Education	0.9764	0.6312 1.5	0.9146	0.9335	0.5607 1.6	0.7914
Age at menopause	1.012	0.977 1	0.5074	1.0091	0.9642 1.1	0.6962
BMI (pre-menopause)	1.0252	0.9016 1.2	0.7041	1.0548	0.9201 1.2	0.4442
BMI (post menopause)	1.0801	0.9773 1.1829	0.1418	1.0552	0.933 1.1774	0.3888
Central Anatolia (ref=Other)	0.5335	0.1053 2.7	0.4478	0.3125	0.0496 2	0.2154
East and South-East Anatolia (ref=Other)	0.8802	0.1182 6.6	0.9009	0.7086	0.0741 6.8	0.7649
Black Sea (ref=Other)	0.7409	0.1045 5.3	0.7642	0.2763	0.03 2.5	0.256
Births status (ref=no birth)	0.8762	0.1274 6	0.8931	0.3036	0.0301 3.1	0.3124
Menopause status (ref=No)	0.1519	0.0012 18.6	0.4422	0.1989	0.0008 47.8	0.5638

Continuation of Table 4.4.

	Response: Tumor type			**Response: NA**		
	Tumortype: ILC/Tubular (ref=tumor type IDC)			**NA: enlarged lymph nodes exist (ref= no enlarged lymph nodes)**		
	OR	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Age	1.0442	0.9811 1.1	0.1741	0.945	0.9024 1	0.0163
Age at first menstruation	0.8164	0.5669 1.2	0.2758	1.2124	0.937 1.6	0.143
HRT taken (ref=not taken)	0.617	0.1614 2.4	0.4802	1.1856	0.4608 3.1	0.724
Duration of breast feeding	1.0171	0.9919 1	0.1845	1.0075	0.9896 1	0.4147
Family history (1st degree) (ref=absent)	1.413	0.6147 4.5	0.5592	1.3687	0.5307 3.6	0.5211
Family history (2nd degree) (ref=absent)	0.635	0.4429 8.9	0.7356	1.1784	0.5247 5.8	0.8393
Number of births	0.9398	0.0455 1.4	0.7743	0.7176	0.2411 1	0.0311
Age at first birth	0.9555	0.9053 1	0.0975	1.0028	0.9604 1	0.898
Mammography (ref=no)	0.7513	0.2606 2.2	0.5967	1.2997	0.6171 2.7	0.4903
Education	1.5116	0.9365 2.4	0.0907	0.5699	0.4011 0.8	0.0017
Age at menopause	1.0017	0.9633 1	0.9327	1.0223	0.9922 1.1	0.1487
BMI (pre-menopause)	1.0592	0.9391 1.2	0.3491	0.9161	0.8376 1	0.055
BMI (post menopause)	0.9449	0.8363 1.0536	0.3068	0.9271	0.8537 1.0006	0.0434
Central Anatolia (ref=Other)	4.1114	0.4187 40.4	0.2251	0.9533	0.2986 3	0.9357
East and South-East Anatolia (ref=Other)	11.7268	0.9798 140.4	0.0519	0.577	0.1331 2.5	0.4624
Black Sea (ref=Other)	13.6948	1.2006 156.2	0.0351	0.7735	0.2001 3	0.7097
Births status (ref=no birth)	1.2014	0.1641 8.8	0.8566	1.9011	0.3735 9.7	0.4391
Menopause status (ref=No)	43.5664	0.3391 5597.7	0.1277	0.2583	0.0071 9.4	0.46

Continuation of Table 4.4.

	Response: ER			**Response: PR**		
	ER: (+) **(ref= ER: (-))**			**PR: (+)** **(ref= PR: (-))**		
	OR	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Age	1.0056	0.9592 1.1	0.8181	1.0037	0.9628 1	0.8615
Age at first menstruation	1.0739	0.8183 1.4	0.6073	0.8265	0.6473 1.1	0.1263
HRT taken (ref=not taken)	0.8395	0.3161 2.2	0.7255	0.5912	0.2501 1.4	0.2311
Duration of breast feeding	0.9982	0.9793 1	0.85	1.0159	0.9991 1	0.0643
Family history (1st degree) (ref=absent)	0.6823	0.8612 1.8	0.4336	0.4574	0.5347 1.1	0.0799
Family history (2nd degree) (ref=absent)	1.4134	0.262 8	0.696	1.3285	0.1906 5.9	0.7094
Number of births	1.1842	0.2492 1.6	0.2982	0.7073	0.2982 0.9	0.0152
Age at first birth	1.016	0.9736 1.1	0.4666	1.0204	0.9818 1.1	0.3042
Mammography (ref=no)	1.2323	0.5514 2.8	0.6107	1.5311	0.7616 3.1	0.2319
Education	1.088	0.7593 1.6	0.6457	0.7978	0.5821 1.1	0.1601
Age at menopause	1.0037	0.9742 1	0.8079	1.0506	1.0219 1.1	0.0005
BMI (pre-menopause)	0.9846	0.9004 1.1	0.7338	0.9885	0.913 1.1	0.7761
BMI (post menopause)	0.9772	0.8998 1.0546	0.5595	0.9446	0.8753 1.0139	0.1071
Central Anatolia (ref=Other)	1.0554	0.3289 3.4	0.9278	1.3336	0.4724 3.8	0.5867
East and South-East Anatolia (ref=Other)	0.8861	0.2008 3.9	0.8731	0.8315	0.217 3.2	0.7878
Black Sea (ref=Other)	0.9384	0.2331 3.8	0.9287	2.0501	0.5883 7.1	0.2598
Births status (ref=no birth)	1.0856	0.2132 5.5	0.9213	0.5341	0.1225 2.3	0.4037
Menopause status (ref=No)	0.5448	0.0142 20.9	0.7441	0.2921	0.0119 7.2	0.4516

Continuation of Table 4.4.

	Response c-erb B2		
	C-erb: (+) **(ref=c-erb: (-))**		
	OR	**95% CI**	**p-value**
Age	0.9658	0.9184 1	0.1764
Age at first menstruation	1.0452	0.7923 1.4	0.7543
HRT taken (ref=not taken)	2.3123	0.8885 6	0.0858
Duration of breast feeding	1.0154	0.9972 1	0.0968
Family history (1st degree) (ref=absent)	1.1965	0.6364 3.3	0.7318
Family history (2nd degree) (ref=absent)	0.774	0.4288 4.4	0.772
Number of births	0.897	0.1369 1.3	0.5347
Age at first birth	0.9803	0.9365 1	0.3914
Mammography (ref=no)	0.6864	0.2895 1.6	0.393
Education	1.0224	0.6998 1.5	0.9089
Age at menopause	1.0232	0.9854 1.1	0.2322
BMI (pre-menopause)	0.9833	0.9002 1.1	0.7085
BMI (post menopause)	0.9996	0.9146 1.0847	0.9931
Central Anatolia (ref=Other)	1.4326	0.4165 4.9	0.5684
East and South-East Anatolia (ref=Other)	0.7388	0.1467 3.7	0.7136
Black Sea (ref=Other)	1.4685	0.3361 6.4	0.6095
Births status (ref=no birth)	1.3841	0.2534 7.6	0.7075
Menopause status (ref=No)	0.5243	0.0112 24.6	0.7423

4.3. TESTING THE INTERACTION OF DISEASE CHARACTERISTICS

For illustration, we focus on the disease characteristics tumor type and NA, and risk factors age, BMI, duration of breast feeding and education level. *Tumor type* has two levels (*ILC/Tubular, IDC*) and *NA* has two levels (*exist, not exist*). To test the interaction between disease characteristics levels the following hypotheses are tested:

H_0 : Association between *Type* and *Duration of Breast Feeding* does not change with respect to *NA*

H_1 : Association between *Type* and *Duration of Breast Feeding* changes with respect to *NA*

The testing of these hypotheses can be conducted either by using the parameters of polytomous logistic regression models with subtype information, or by using the parameters of two stage polytomous logistic regression.

4.3.1. Polytomous Logistic Regression for Response Levels Obtained by Cross-classifying the Levels of Characteristics

Disease subtypes which are obtained by cross-classifying the levels of the disease characteristics are modeled by using polytomous logistic regression. Subtypes are obtained for our case as follows:

Table 4.5: Disease subtypes for *Tumor type* and *NA*

Disease Subtype (d)	Tumor Type	NA
0 (Control)	-	-
1	1	0
2	1	1
3	2	0
4	2	1

Subjects without the disease characteristics information, i.e. controls, are labeled with zero.

Polytomous logistic regression models for modeling the log-odds of having a certain disease subtype are as follows:

$$ln\frac{P(D=1|\boldsymbol{X})}{P(D=0|\boldsymbol{X})}= \beta_{01}+\beta_{11}(Age)+\beta_{21}(Duration\ of\ Breast\ Feeding)+\beta_{31}(Education)$$

$$ln\frac{P(D=2|\boldsymbol{X})}{P(D=0|\boldsymbol{X})}= \beta_{02}+\beta_{12}(Age)+\beta_{22}(Duration\ of\ Breast\ Feeding)+\beta_{32}(Education)$$

$$ln\frac{P(D=3|\boldsymbol{X})}{P(D=0|\boldsymbol{X})}= \beta_{03}+\beta_{13}(Age)+\beta_{23}(Duration\ of\ Breast\ Feeding)+\beta_{33}(Education)$$

$$ln\frac{P(D=4|\boldsymbol{X})}{P(D=0|\boldsymbol{X})}= \beta_{04}+\beta_{14}(Age)+\beta_{24}(Duration\ of\ Breast\ Feeding)+\beta_{34}(Education)$$

In these models, *Age* and *Duration of Breast Feeding* are continuous covariates, whereas *education level* is in ordinal scale. We used *mnrfit* function for polytomous logistic regression in MATLAB to obtain the model parameter estimates. Parameter estimates, standard errors and p-value corresponding to test the significance of each parameter is given in Table 4.6:

Table 4.6: Estimates, standard errors and p-values of parameters of polytomous logistic regression model

	Intercept (j=0)			**Age (j=1)**			**Duration of breast feeding (j=2)**			**Education Level (j=3)**		
	Est	**S.E**	**p-val**	**Est.**	**S.E**	**p-val**	**Est**	**S.E**	**p-val**	**Est.**	**S.E.**	**p-val**
$\widehat{\boldsymbol{\beta}}_{j1}$	-4.64	2.39	0.053	0.06	0.04	0.123	-0.001	0.02	0.979	0.77	0.33	0.021
$\widehat{\boldsymbol{\beta}}_{j2}$	3.98	1.60	0.013	-0.04	0.03	0.153	-0.001	0.01	0.899	0.09	0.25	0.723
$\widehat{\boldsymbol{\beta}}_{j3}$	0.51	1.65	0.756	0.02	0.03	0.541	-0.007	0.01	0.505	0.29	0.26	0.249
$\widehat{\boldsymbol{\beta}}_{j4}$	5.43	1.57	0.001	-0.06	0.03	0.020	-0.005	0.01	0.577	0.30	0.25	0.220

where j=0,1,2,3. That is, j is the index that stands for the model parameters.

Odds Ratio related with the null hypothesis is:

$$\text{OR}=\frac{\dfrac{P(Type=2,NA=1|\boldsymbol{Age},\boldsymbol{DBF}=\boldsymbol{25},\boldsymbol{EL})/\,P(Type=1,NA=1|\boldsymbol{Age},\boldsymbol{DBF}=\boldsymbol{25},\boldsymbol{EL})}{P(Type=2,NA=1|\boldsymbol{Age},\boldsymbol{DBF}=\boldsymbol{24},\boldsymbol{EL})/\,P(Type=1,NA=1|\boldsymbol{Age},\boldsymbol{DBF}=\boldsymbol{24},\boldsymbol{EL})}}{\dfrac{P(Type=2,NA=0|\boldsymbol{Age},\boldsymbol{DBF}=\boldsymbol{25},\boldsymbol{EL})/P(Type=1,NA=0|\boldsymbol{Age},\boldsymbol{DBF}=\boldsymbol{25},\boldsymbol{EL})}{P(Type=2,NA=0|\boldsymbol{Age},\boldsymbol{DBF}=\boldsymbol{24},\boldsymbol{EL})/P(Type=1,NA=0|\boldsymbol{Age},\boldsymbol{DBF}=\boldsymbol{24},\boldsymbol{EL})}}$$

where *DBF* is the duration of breast feeding in months. This is in fact the ratio of two odds ratios. In order to obtain this OR, we first write the OR's that can be represented in the form of polytomous logistic regression model:

$$OR_{11}=\frac{\frac{P(D=1|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}{P(D=0|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}}{\frac{P(D=1|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}{P(D=0|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}}=\frac{\frac{P(Type=1,NA=0|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}{P(Control|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}}{\frac{P(Type=1,NA=0|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}{P(Control|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}}$$

$$=\frac{e^{\beta_{01}+\beta_{11}Age+\beta_{21}(25)+\beta_{31}EL}}{e^{\beta_{01}+\beta_{11}Age+\beta_{21}(24)+\beta_{31}EL}}=\frac{\mathrm{e}^{(\mathbf{25})\boldsymbol{\beta}_{\mathbf{21}}}}{\mathrm{e}^{(\mathbf{24})\boldsymbol{\beta}_{\mathbf{21}}}}=\ e^{\boldsymbol{\beta}_{\mathbf{21}}}$$

$$OR_{21}=\frac{\frac{P(D=2|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}{P(D=0|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}}{\frac{P(D=2|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}{P(D=0|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}}=\frac{\frac{P(Type=1,NA=1|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}{P(Control|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}}{\frac{P(Type=1,NA=1|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}{P(Control|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}}$$

$$=\frac{e^{\beta_{02}+\beta_{12}Age+\beta_{22}(25)+\beta_{32}EL}}{e^{\beta_{02}+\beta_{12}Age+\beta_{22}(24)+\beta_{32}EL}}=\frac{\mathrm{e}^{(\mathbf{25})\boldsymbol{\beta}_{\mathbf{22}}}}{\mathrm{e}^{(\mathbf{24})\boldsymbol{\beta}_{\mathbf{22}}}}=\ e^{\boldsymbol{\beta}_{\mathbf{22}}}$$

$$OR_{31}=\frac{\frac{P(D=3|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}{P(D=0|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}}{\frac{P(D=3|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}{P(D=0|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}}=\frac{\frac{P(Type=2,NA=0|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}{P(Control|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}}{\frac{P(Type=2,NA=0|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}{P(Control|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}}$$

$$=\frac{e^{\beta_{03}+\beta_{13}Age+\beta_{23}(25)+\beta_{33}EL}}{e^{\beta_{03}+\beta_{13}Age+\beta_{23}(24)+\beta_{33}EL}}=\frac{\mathrm{e}^{(\mathbf{25})\boldsymbol{\beta}_{\mathbf{23}}}}{\mathrm{e}^{(\mathbf{24})\boldsymbol{\beta}_{\mathbf{23}}}}=\ e^{\boldsymbol{\beta}_{\mathbf{23}}}$$

$$OR_{41}=\frac{\frac{P(D=4|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}{P(D=0|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}}{\frac{P(D=4|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}{P(D=0|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}}=\frac{\frac{P(Type=2,NA=1|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}{P(Control|\boldsymbol{Age},\boldsymbol{DBF=25},\boldsymbol{EL})}}{\frac{P(Type=2,NA=1|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}{P(Control|\boldsymbol{Age},\boldsymbol{DBF=24},\boldsymbol{EL})}}$$

$$=\frac{e^{\beta_{04}+\beta_{14}Age+\beta_{24}(25)+\beta_{34}EL}}{e^{\beta_{04}+\beta_{14}Age+\beta_{24}(25)+\beta_{34}EL}}=\frac{\mathrm{e}^{(\mathbf{25})\boldsymbol{\beta}_{\mathbf{24}}}}{\mathrm{e}^{(\mathbf{24})\boldsymbol{\beta}_{\mathbf{24}}}}=\ e^{\boldsymbol{\beta}_{\mathbf{24}}}$$

So, the odds ratio related with the null hypothesis is represented in terms of the odds ratios coming from the polytomous logistic regression models is as follows,

$$OR = \frac{OR_{41}/\ OR_{21}}{OR_{31}/\ OR_{11}} = \frac{e^{\boldsymbol{\beta}_{24}}/e^{\boldsymbol{\beta}_{22}}}{e^{\boldsymbol{\beta}_{23}}/e^{\boldsymbol{\beta}_{21}}} = e^{\beta_{24}+\beta_{21}-\beta_{22}-\beta_{23}}$$

Therefore, we can now rewrite the null hypothesis in the form of the model parameters

$$\text{H}_0 : \beta_{24} + \beta_{21} - \beta_{22} - \beta_{23}\ = 0 \tag{4.1}$$

$$\text{H}_1 : \beta_{24} + \beta_{21} - \beta_{22} - \beta_{23}\ \neq\ 0$$

To test these hypotheses, we need to obtain Wald’s test statistic which is distributed as $\chi^2_{(r)}$ where r is the number of linear equations in null hypothesis:

$$T_W = \frac{(\hat{\beta}_{24} + \hat{\beta}_{21} - \hat{\beta}_{22} - \hat{\beta}_{23})^2}{\widehat{Var}(\hat{\beta}_{24} + \hat{\beta}_{21} - \hat{\beta}_{22} - \hat{\beta}_{23})} = \frac{(-0.005 - 0.001 + 0.001 + 0.007)^2}{0.0001} = 0.004$$

where

$$\widehat{Var}(\hat{\beta}_{24} + \hat{\beta}_{21} - \hat{\beta}_{22} - \hat{\beta}_{23})$$

$$\begin{aligned} = \ \widehat{Var}(\hat{\beta}_{24}) + \ \widehat{Var}(\hat{\beta}_{21)} + \widehat{Var}(\hat{\beta}_{22})\ + \widehat{Var}(\hat{\beta}_{23})\ + 2\widehat{Cov}(\hat{\beta}_{24}, \hat{\beta}_{21}) - 2\widehat{Cov}(\hat{\beta}_{24}, \hat{\beta}_{22}) \\ - 2\ \widehat{Cov}(\hat{\beta}_{24}, \hat{\beta}_{23})\ - 2\widehat{Cov}(\hat{\beta}_{21}, \hat{\beta}_{22}) - 2\widehat{Cov}(\hat{\beta}_{21}, \hat{\beta}_{23})\ + 2\widehat{Cov}(\hat{\beta}_{22}, \hat{\beta}_{23}) \end{aligned}$$

$$\begin{aligned} = 0.0001 + 0.0002 + 0.0001 + 0.0001 + 2(0.0001) - 2(0.0001) - 2(0.0001) \\ - 2(0.0001) - 2(0.0001) + 2(0.0001) = 0.0001 \end{aligned}$$

Since T_W = 0.004 < 3.84 = $\chi^2_{(1)}$, we do not reject the null hypothesis (p=0.94). Association between *duration of breast feeding* and *tumor type* does not change with respect to the existence of enlarged lymph nodes for the sample we have.

4.3.2. Two Stage Polytomous Logistic Regression

To test the same hypothesis we can make use of the two stage polytomous regression model parameter estimators.

In the first stage, unstructured polytomous logistic regression model is built same as in section 4.3.1. Then, second stage parameters, given previously in model (2.4) in chapter 2, are estimated through PCL estimation (Table 4.7).

Table 4.7: Estimates, standard errors and p-values of parameters of second stage model

	Age			**Duration of breast feeding (DBF)**			**Education Level (EL)**		
	Est.	**Std. Err.**	**p-val.**	**Est.**	**Std. Err.**	**p-val.**	**Est.**	**Std. Err.**	**p-val.**
$\boldsymbol{\theta}^{\mathbf{0}}$	0.071	0.016	0	-0.003	0.007	0.644	-0.034	0.126	0.785
$\boldsymbol{\theta}^{(\mathbf{1})}_{\mathbf{1(2)}}$	0.022	0.023	0.333	0.01	0.014	0.485	0.495	0.267	0.064
$\boldsymbol{\theta}^{(\mathbf{1})}_{\mathbf{2(2)}}$	-0.051	0.017	0.003	0.007	0.007	0.275	-0.188	0.137	0.170
$\boldsymbol{\theta}^{(\mathbf{2})}_{\mathbf{12(22)}}$	0.015	0.035	0.663	-0.011	0.018	0.546	-0.608	0.382	0.111

Odds ratio related with the null hypothesis can be directly written by second-degree contrasts of parameters, i.e. the parameter corresponds to the interaction between disease characteristics for the covariate *DBF*.

$$OR_{12} = \frac{\dfrac{P(Type = 2, NA = 1|\boldsymbol{Age}, \boldsymbol{DBF = 25}, \boldsymbol{EL})/\ P(Type = 1, NA = 1|\boldsymbol{Age}, \boldsymbol{DBF = 25}, \boldsymbol{EL})}{P(Type = 2, NA = 1|\boldsymbol{Age}, \boldsymbol{DBF = 24}, \boldsymbol{EL})/\ P(Type = 1, NA = 1|\boldsymbol{Age}, \boldsymbol{DBF = 24}, \boldsymbol{EL})}}{\dfrac{P(Type = 2, NA = 0|\boldsymbol{Age}, \boldsymbol{DBF = 25}, \boldsymbol{EL})/P(Type = 2, NA = 0|\boldsymbol{Age}, \boldsymbol{DBF = 25}, \boldsymbol{EL})}{P(Type = 2, NA = 0|\boldsymbol{Age}, \boldsymbol{DBF = 24}, \boldsymbol{EL})/P(Type = 2, NA = 0|\boldsymbol{Age}, \boldsymbol{DBF = 24}, \boldsymbol{EL})}} = e^{\theta_{12(22)}}$$

We can rewrite the null hypothesis in the form of two stage polytomous logistic regression model parameters:

H$_0$: $\boldsymbol{\theta}_{12(22)} = 0$ (4.2)

H$_1$: $\boldsymbol{\theta}_{12(22)} \neq 0$

For testing of the hypotheses (4.2) which are equivalent to hypotheses (4.1) we can again calculate Wald's statistic which is:

$$T_W = \frac{(\hat{\theta}_{12(22)})^2}{\widehat{Var}(\hat{\theta}_{12(22)})} = \left(\frac{-0.011}{0.018}\right)^2 = 0.373$$

Since T_W = 0.373 < 3.81 = $\chi^2_{(1)}$ at 0.05 significance level, we fail to reject H$_0$ (p=0.54) and conclude that association between *type* and *duration of breast feeding* does not change with respect to *NA*. Both of the testing procedures indicate the same decision.

In section 4.3, we tested whether the association between tumor characteristic *type* and covariate *duration of breast feeding* changes with respect to tumor characteristic *NA* through two approaches: i) polytomous logistic regression with response categorized as disease subtypes and ii) two stage polytomous logistic regression with second order contrast parameters estimated for interaction. It is revealed that the odds ratio corresponding to test hypothesis, OR, is equivalent to the odds ratio represented by the interaction parameter $\boldsymbol{\theta}_{12(22)}$. On the other hand, the same odds ratio, OR, can be written as division of $OR_{41}/\ OR_{21}$ and $OR_{31}/\ OR_{11}$. That is, one can obtain the ratio of the odds of having ILC/Tubular type of tumor to IDC type of tumor when NA exists, to the odds of having ILC/Tubular type of tumor when the NA does not exist either directly by obtaining exp $(\theta_{12(22)})$ or $exp((\beta_{24} - \beta_{22}) - (\beta_{23} - \beta_{21}))$. In approach (i) we need to do much more effort to test the hypothesis than approach (ii). It is obvious that using a two staged approach provides us to estimate only one parameter in order to obtain OR, whereas unstructured polytomous logistic regression requires to estimate all first stage parameters for the covariate included in the hypothesis.

CHAPTER V

CONCLUSION

In this thesis work, we have studied the methods based on the polytomous logistic regression to analyze health data with multivariate disease subtype information. We first compared the performances of three different methods through a Monte Carlo simulation experiment and then we implemented two of the methods on a real-life breast cancer dataset and compared these methods in terms of the inference on the model parameters.

In a simulation experiment, we have compared the performances of the three approaches through the accuracy and efficiency of the first stage parameters. We designed sample scenarios with small, moderate and large scaled sample sizes as well as the number of disease subtypes. Results of the simulation experiment is interpreted in three aspects: (1) for small number of disease characteristic case, i.e. M=2×2×2, PCL estimators outperforms the ML and Bayesian estimators of classical polytomous logistic regression parameters in terms of efficiency for small, moderate and large sample sizes. When the disease subtype levels increased to M=4×4×4, because of the computation time Bayesian estimation became difficult for a large number of simulation iterations. Therefore, only MLE and PCL methods are implemented and compared for larger disease subtype scenarios. For M=4×4×4 case, when the sample size is small, PCL estimators had better performance in terms of MSE, however, as the sample size increased efficiency of ML estimators of classical polytomous logistic regression outperformed because of the large sample properties of MLE's. For M=6×6×4 case, standard errors and biases of both ML and PCL estimators are increased compared to the previous disease subtype scenarios, however, for M=6×6×4 case as the sample size increased, standard errors and biases decreased. That means when the number of disease subtypes are increased, sample size should be large enough to have small bias and standard errors. In addition, it is revealed that for all the sample size scenarios, PCL estimators had smaller bias whereas ML estimates had better results in terms of measures related to the variation. This in turn implies that, PCL estimators had better property in terms of accuracy but not in efficiency, while the ML estimators had better performance in terms of efficiency but not in accuracy. (2) for relatively small sample sizes, PCL estimation performed better for small and moderate number of disease subtypes, however when the size of the disease subtypes is large, both PCL and MLE perform inefficiently. (3) the sampling variance of the first stage estimators based on PCL in two stage logistic regression converges to the asymptotic variances slightly faster than the first stage estimators based on MLE in classical polytomous logistic regression.

We investigated the etiologic heterogeneity among breast cancer subtypes for Turkish female breast cancer patients by analyzing a breast cancer dataset with tumor characteristics information collected in Ankara Oncology Research and Education Hospital. First, every tumor characteristic are taken independently and binary/polytomous logistic regression models are constructed. Then, on the same tumor characteristics, a two stage logistic regression model is constructed. The advantage of latter over the former is revealed in the interpretation of the odds ratios that represent the association between covariates and tumor characteristics. Two staged approach provided the advantage of adjusting for the other characteristics in the association of a certain characteristic with covariates, in other words, two stage parameters account for the multivariate nature in the tumor characteristics. We also illustrated the practical advantage of two stage polytomous logistic regression for testing the interaction between the tumor characteristics, i.e. if the association of a characteristic with a covariate differs according to the another tumor characteristic, in terms of hypothesis testing.

In brief, with this thesis work, we mainly made three contributions: (1) classical polytomous logistic regression with ML and Bayesian estimation, and two staged polytomous regression with PCL estimation are compared in terms of bias and efficiency over the main model parameters; (2) a statistical analysis of etiologic heterogeneity in breast cancer subtypes for Turkish female breast cancer patients is conducted; (3) advantage of two stage polytomous logistic regression for investigating the interaction behaviour between tumor characteristics is illustrated.

As a future study, missing covariate and/or missing disease characteristic situation can be considered in two stage polytomous logistic regression approach. Moreover, an efficient testing procedure, possibly based on score test, can be developed for the second stage parameters.

In this thesis, simulation experiment is done by using MATLAB 7.8 and WinBUGS. Analysis of dataset is implemented by using SAS, MATLAB 7.8, WinBUGS and R programs. MATLAB codes of PCL estimation is written by Chatterjee (2004).

REFERENCES

Agresti, A., (2002), *Categorical Data Analysis*, John Wiley Sons Inc. Publication, New Jersey, USA

Brooks, S.A., Harris, A., (2006), *Breast Cancer Research Protocols*, Humana Press Inc., New Jersey, USA

Casella, G., George, E.I., (1992), Explaining the Gibbs Sampler, *The American Statistician*, 46(3), 167-174

Chatterjee, N., (2004), A Two Stage Regression Model for Epidemiological Studies With Multivariate Disease Classification Data, *Journal of the American Statistical Association*, 99(465), 127-138

García-Closas, M, Brinton, L.A., Lissowska, J., Chatterjee, N., Peplonska, B., Anderson, W.F., Szeszenia-Dabrowska, N., Bardin-Mikolajczak, A., Zatonski, W., Blair, A., Kalaylioglu, Z., Rymkiewicz, G., Mazepa-Sikora, D., Kordek, R., Lukaszek, S., Sherman, M.E., (2006), Established Breast Cancer Risk Factors by Clinically Important Tumour Characteristics, *British Journal of Cancer*, 95(1), pp-123-129.

Geman, S., Geman, D., (1984), Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.

Gilks, W.R., Richardson S., Spiegelhalter D.J., (1996), *Markov Chain Monte Carlo in Practice*, Chapman&Hall

Glossary of Statistical Terms, retrieved from http://isi.cbs.nl/glossary/term2031.htm, on June 26, 2011.

Harris, J.R., Lippman, M.E., Morrow, M., Osborne, C.K., (2010), *Diseases of the Breast*, Lippincott Williams & Wilkins

Hosmer D.W., Lemeshow S., (2000), *Applied Logistic Regression*, Wiley Series in Probability and Statistics.

Kleinbaum, D.G., Klein, M., (2010), *Logistic Regression-A Self Learning Text*, Springer.

Li, E., Boos, D., Gumpertz, M., (2001), *Simulation Study in Statistics*, retrieved from http://www4.stat.ncsu.edu/~reich/st810A/ on June 22, 2011

Link, W.A., Barker, R.J., (2010), *Bayesian Inference with Ecological Applications*, Elsevier

McFadden, D., (1974), Conditional Logit Analysis of Qualitative Choice Behavior, *Frontiers in Econometrics*, Edited by Zarembka, Academic Press, New York

Ntzoufras, I., (2009), *Bayesian Inference via WinBUGS*, John Wiley Sons Inc.

Sariego, J., (2010), Breast cancer in the young patient, *The American Surgeon*, 6(12):1397-400

Sherman, M.E., Rimm, D.L, Yang, X.R., Chatterjee, N, Brinton, L.A., Lissowska, J., Peplonska, B., Szeszenia-Dabrowska, N, Zatonski, W., Cartun, R., Mandich, D., Rymkiewicz, G., Ligaj, M., Lukaszek, S., Kordek, R., Kalaylioglu, Z., Harigopal, M., Charrette, L., Falk, R.T., Richesson, D., Anderson, W.F., Hewitt, S.M., García-Closas, M., (2007), Variation in breast cancer hormone receptor and HER2 levels by etiologic factors: a population-based analysis, *Int J. Cancer*, 121(5), 1079-1085.

Sotiriou, C., Pusztai, L., (2009), Gene-Expression Signatures in Breast Cancer, *The New England Journal of Medicine,* 360:790-800

APPENDIX A

TABLES FOR RELATIVE AND ASYMPTOTIC RELATIVE EFFICIENCY

Table A.1: Relative efficiency and Asymptotic relative efficiency for M=2x2x2

Number of Cases=250			Number of Cases=500			Number of Cases=1000		
Index of β	**RE PCL vs MLE**	**ARE PCL vs MLE**	**Index of β**	**RE PCL vs MLE**	**ARE PCL vs MLE**	**Index of β**	**RE PCL vs MLE**	**ARE PCL vs MLE**
1	0.6645	0.7116	1	0.6684	0.7144	1	0.6959	0.7163
2	0.6452	0.7041	2	0.5722	0.7311	2	0.7041	0.7498
3	0.4493	0.5074	3	0.4766	0.514	3	0.5136	0.518
4	0.5615	0.629	4	0.5521	0.6583	4	0.5883	0.6688
5	0.5069	0.5419	5	0.5251	0.5504	5	0.5519	0.5531
6	0.7593	0.7231	6	0.6799	0.7632	6	0.693	0.7914
7	0.4769	0.4804	7	0.4586	0.4927	7	0.4597	0.4955
8	0.8326	0.8579	8	0.7792	0.8977	8	0.9153	0.9312

Table A.2: Relative efficiency and Asymptotic relative efficiency for M=4x4x4

Number of Cases=250			Number of Cases=500			Number of Cases=1000		
Index of β	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**	**Index of β**	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**	**Index of β**	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**
1	0.7938	0.6022	1	1.3378	0.9496	1	1.535	1.0662
2	0.8144	0.6153	2	1.1242	0.9501	2	1.7252	1.0514
3	0.795	0.5652	3	1.2045	0.935	3	1.5123	1.0527
4	1.0397	0.3449	4	1.6501	1.0586	4	2.2333	1.2785
5	0.8429	0.6459	5	1.1835	0.9491	5	1.5539	1.0568
6	0.7768	0.7757	6	1.3916	1.0001	6	1.7975	1.0972
7	0.8593	0.6564	7	1.3119	0.9789	7	1.6224	1.0966
8	1.059	0.9094	8	1.8255	1.1582	8	2.26	1.3056
9	0.8247	0.807	9	1.1616	0.9438	9	1.5829	1.0615
10	0.7557	0.5093	10	1.1653	0.8375	10	1.5627	1.0214
11	0.7586	0.5235	11	1.1299	0.9024	11	1.4662	1.0268
12	0.8542	0.7078	12	1.4323	0.9685	12	2.0038	1.2257
13	0.73	0.5063	13	1.1765	0.9086	13	1.5936	1.0274
14	0.668	0.6978	14	1.0807	0.882	14	1.4219	0.9801
15	0.6991	0.7131	15	1.1134	0.8963	15	1.5034	0.9732
16	0.6451	0.1485	16	0.811	0.6212	16	1.1789	0.8602
17	0.8346	0.7174	17	1.3817	0.9613	17	1.5978	1.0783
18	0.5856	0.2313	18	0.8192	0.5472	18	1.04	0.8291
19	0.5447	0.1542	19	0.7863	0.5666	19	0.914	0.808
20	0.7041	0.1992	20	0.9208	0.6012	20	1.3063	0.972
21	1.0367	0.8148	21	1.5191	1.0027	21	1.8561	1.1679
22	0.6247	0.2418	22	1.1039	0.7625	22	1.3157	0.9347
23	0.6887	0.3719	23	0.9287	0.7215	23	1.2505	0.9443

Continuation of Table A.2.

Number of Cases=250			Number of Cases=500			Number of Cases=1000		
Index of β	**RE PCL vs MLE**	**ARE PCL vs MLE**	**Index of β**	**RE PCL vs MLE**	**ARE PCL vs MLE**	**Index of β**	**RE PCL vs MLE**	**ARE PCL vs MLE**
24	0.8985	0.2814	24	1.2967	0.895	24	1.7192	1.1239
25	0.6351	0.0996	25	0.6604	0.5202	25	0.9393	0.751
26	0.6155	0.0513	26	0.4799	0.0785	26	0.4744	0.3576
27	0.5166	0.0544	27	0.4247	0.0766	27	0.5044	0.3723
28	0.8632	0.0692	28	0.5724	0.1351	28	0.6211	0.3723
29	0.6765	0.5766	29	1.2326	0.9072	29	1.4796	1.0257
30	0.4882	0.1774	30	0.6561	0.497	30	0.8356	0.7079
31	0.5649	0.1629	31	0.6567	0.3161	31	0.7952	0.7125
32	0.7305	0.0535	32	0.6101	0.1741	32	0.6138	0.3671
33	0.7261	0.8335	33	1.1376	0.9185	33	1.4599	1.0295
34	0.5522	0.1981	34	0.6771	0.6514	34	0.9013	0.7873
35	0.8374	0.8307	35	1.2041	0.9524	35	1.532	1.0684
36	0.9534	0.9223	36	1.8423	1.1392	36	2.3637	1.2812
37	0.5162	0.0928	37	0.6217	0.4158	37	0.7774	0.7303
38	0.4667	0.0775	38	0.3976	0.1755	38	0.4702	0.3822
39	0.5351	0.289	39	0.6688	0.5513	39	0.95	0.815
40	0.6626	0.1354	40	0.8577	0.7583	40	1.314	0.963
41	0.8642	0.5821	41	1.2737	0.9204	41	1.581	1.0406
42	0.5724	0.16	42	0.6673	0.4447	42	0.9874	0.7565
43	0.8185	0.7091	43	1.2041	0.8541	43	1.4858	1.0365
44	0.9039	0.5641	44	1.4994	1.0699	44	1.9833	1.2324
45	0.7746	0.8202	45	1.0833	0.9137	45	1.4877	1.0053
46	0.4892	0.1237	46	0.6327	0.4922	46	0.7964	0.7068
47	0.7346	0.8204	47	1.1015	0.8964	47	1.4347	1.0041

Continuation of Table A.2.

Number of Cases=250			Number of Cases=500			Number of Cases=1000		
Index of β	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**	**Index of β**	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**	**Index of β**	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**
48	0.6006	0.225	48	0.8291	0.7569	48	1.4671	0.9422
49	0.7731	0.7761	49	1.1831	0.9286	49	1.5908	1.0541
50	0.7784	0.7659	50	1.2673	0.9894	50	1.6849	1.1043
51	0.8286	0.7831	51	1.2089	0.9611	51	1.624	1.0913
52	0.9313	0.5655	52	1.6874	1.1609	52	2.4395	1.3121
53	0.7727	0.1903	53	1.1617	0.9265	53	1.6148	1.0509
54	0.8912	0.7588	54	1.3611	1.0242	54	1.7772	1.1373
55	0.8467	0.8185	55	1.4512	1.012	55	1.8343	1.1328
56	1.0157	0.8678	56	1.9408	1.2002	56	2.5565	1.3646
57	0.8826	0.8214	57	1.2908	0.9765	57	1.6945	1.1178
58	0.7967	0.8091	58	1.3799	0.9857	58	1.6315	1.1027
59	0.8164	0.6846	59	1.3014	0.9772	59	1.6704	1.1078
60	0.9955	0.527	60	1.5846	1.1517	60	2.3376	1.3271
61	0.75	0.8422	61	1.1863	0.9506	61	1.6831	1.0633
62	0.7629	0.7	62	1.4489	0.9502	62	1.8388	1.0731
63	0.8083	0.7403	63	1.1709	0.9448	63	1.5649	1.0581
64	0.6769	0.2714	64	1.0401	0.7718	64	1.3294	1.0064

Table A.3: Relative efficiency and Asymptotic relative efficiency for M=6x6x4

Number of Cases=250			Number of Cases=500			Number of Cases=1000		
Index of β	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**	**Index of β**	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**	**Index of β**	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**
1	1.5754	1.2929	1	1.1991	1.3489	1	2.0477	1.5724
2	1.3289	1.285	2	1.1529	1.3424	2	2.009	1.6148
3	1.3174	1.2554	3	1.225	1.335	3	1.9903	1.5572
4	1.3536	1.5013	4	1.3238	1.6013	4	2.3041	1.9493
5	1.5072	1.3639	5	1.2872	1.3819	5	2.0262	1.6323
6	1.4021	1.3951	6	1.2457	1.4505	6	1.9887	1.6778
7	1.4694	1.3954	7	1.376	1.4413	7	2.0233	1.6656
8	1.5107	1.7005	8	1.3871	1.6546	8	2.5662	2.045
9	1.5034	1.3698	9	1.3522	1.3943	9	1.8764	1.6413
10	1.3806	1.4085	10	1.2609	1.4563	10	1.9515	1.6894
11	1.5037	1.3908	11	1.4558	1.4738	11	2.0338	1.6697
12	1.4447	1.3835	12	1.1859	1.4917	12	1.9081	1.7236
13	1.6006	1.3167	13	1.3157	1.3368	13	2.0164	1.5826
14	1.4633	1.2267	14	1.1419	1.2224	14	1.5488	1.3958
15	1.5824	1.2047	15	1.1587	1.1993	15	1.5992	1.3641
16	1.4432	1.3686	16	1.1583	1.3939	16	1.7207	1.6634
17	1.9628	1.0792	17	1.2984	1.0419	17	1.3206	1.1605
18	1.4812	0.9541	18	1.0764	0.9018	18	1.2257	1.034
19	1.816	1.1376	19	1.3135	1.1479	19	1.6381	1.2821
20	1.7281	1.1979	20	1.2587	1.2764	20	1.7769	1.4307
21	1.7176	1.1753	21	1.2032	1.1736	21	1.7447	1.3608
22	1.4896	1.2403	22	1.304	1.3029	22	2.0647	1.4973
23	1.5229	1.2476	23	1.3003	1.2794	23	1.8397	1.4963

Continuation of Table A.3.

Number of Cases=250			Number of Cases=500			Number of Cases=1000		
Index of β	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**	**Index of β**	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**	**Index of β**	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**
24	1.5763	1.4275	1	1.4436	1.5365	1	2.1766	1.7857
25	1.3888	1.2393	2	1.2203	1.3488	2	1.8413	1.562
26	1.31	1.3122	3	1.2508	1.4476	3	2.3565	1.7554
27	1.2966	1.3447	4	1.3532	1.4604	4	2.1217	1.7077
28	1.2984	1.5124	5	1.4593	1.7917	5	2.7491	2.0977
29	1.3357	1.2526	6	1.1879	1.315	6	1.7234	1.5077
30	1.3687	1.3598	7	1.298	1.5255	7	2.0551	1.8158
31	1.4367	1.4101	8	1.5346	1.5362	8	2.0902	1.7939
32	1.4007	1.6338	9	1.6685	1.8585	9	2.8001	2.175
33	1.506	1.0355	10	1.0654	1.0062	10	1.0702	1.1101
34	1.4174	1.077	11	0.9807	1.1415	11	1.1339	1.2481
35	1.5119	1.0686	12	1.0456	1.0889	12	1.1839	1.2262
36	1.4324	1.1581	13	1.0629	1.1499	13	1.1144	1.2979
37	1.3999	1.2155	14	1.2685	1.2893	14	1.6063	1.533
38	1.3224	1.2462	15	1.1837	1.3124	15	1.6543	1.4932
39	1.3714	1.2693	16	1.1819	1.2747	16	1.6492	1.4834
40	1.4343	1.4115	17	1.2668	1.5828	17	1.9347	1.8303
41	1.4534	0.9976	18	1.2272	1.0118	18	1.3479	1.1265
42	1.463	0.9553	19	1.1445	1.0052	19	1.273	1.0997
43	1.5617	1.2071	20	1.4818	1.2361	20	1.824	1.3693
44	1.5515	1.2764	21	1.3252	1.3867	21	1.7916	1.5629
45	1.4434	1.0793	22	1.2361	1.1661	22	1.4122	1.3155
46	1.4588	1.3097	23	1.3947	1.4275	23	2.1504	1.6268

Continuation of Table A.3.

Number of Cases=250			Number of Cases=500			Number of Cases=1000		
Index of β	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**	**Index of β**	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**	**Index of β**	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**
47	1.4981	1.2648	47	1.4883	1.4276	47	1.9998	1.643
48	1.4929	1.4408	48	1.5843	1.676	48	2.2518	1.9077
49	1.3428	1.2004	49	1.0802	1.2178	49	1.6206	1.4585
50	1.2678	1.2511	50	1.1687	1.3679	50	1.9302	1.5948
51	1.3594	1.2267	51	1.1893	1.3481	51	1.89	1.6334
52	2.0862	2.3507	52	2.9674	2.6062	52	5.1834	2.9573
53	1.4241	1.2598	53	1.2355	1.3242	53	1.8669	1.6334
54	1.4332	1.4654	54	1.2766	1.5455	54	2.2226	1.8476
55	1.4571	1.4632	55	1.4005	1.5546	55	2.3179	1.8786
56	2.9884	2.7271	56	3.7555	3.0065	56	5.9687	3.3478
57	1.4213	1.2953	57	1.2643	1.3703	57	1.7539	1.5873
58	1.4682	1.4073	58	1.418	1.5716	58	2.5002	1.8356
59	1.5643	1.4691	59	1.5087	1.6041	59	2.2555	1.858
60	2.142	2.25	60	2.8754	2.5862	60	4.7572	2.8502
61	1.3375	1.265	61	1.1986	1.3474	61	1.7375	1.5359
62	1.3884	1.22	62	1.0725	1.2879	62	1.6371	1.4976
63	1.4043	1.2182	63	1.2143	1.2927	63	1.8088	1.5191
64	2.0916	2.1875	64	2.7765	2.4051	64	4.3166	2.7485
65	1.546	0.9801	65	1.1082	0.9766	65	1.3698	1.1183
66	1.377	0.9553	66	0.9021	0.9949	66	1.3955	1.1041
67	1.8594	1.1763	67	1.4154	1.2121	67	1.6402	1.3445
68	2.477	2.0518	68	2.2241	2.1284	68	3.3564	2.3061
69	1.4918	1.0977	69	1.1934	1.1459	69	1.604	1.3132

Continuation of Table A.3.

Number of Cases=250			Number of Cases=500			Number of Cases=1000		
Index of β	**RE PCL vs MLE**	**ARE PCL vs MLE**	**Index of β**	**RE PCL vs MLE**	**ARE PCL vs MLE**	**Index of β**	**RE PCL vs MLE**	**ARE PCL vs MLE**
70	1.5726	1.2987	70	1.4511	1.4241	70	2.2729	1.608
71	1.5723	1.2953	71	1.5054	1.4137	71	2.2215	1.6291
72	2.7391	2.4496	72	3.3756	2.6236	72	4.6728	2.8507
73	1.1513	1.1288	73	1.1336	1.2716	73	1.8668	1.5156
74	1.785	1.8844	74	2.4357	2.1742	74	4.5545	2.4901
75	1.843	1.9298	75	2.5984	2.1827	75	4.2798	2.508
76	1.8654	2.3104	76	3.1226	2.6899	76	5.4093	3.0546
77	1.1534	1.1689	77	1.0666	1.2241	77	1.8445	1.4892
78	1.8669	2.0019	78	2.2968	2.2661	78	4.3017	2.5647
79	1.7897	2.0422	79	2.5367	2.282	79	4.1229	2.5703
80	2.0436	2.4241	80	3.2817	2.7335	80	5.1942	3.1292
81	1.1487	1.1446	81	1.1286	1.2841	81	1.6095	1.4464
82	1.8104	2.012	82	2.7506	2.2916	82	4.118	2.5676
83	1.6983	1.978	83	2.7422	2.2933	83	4.2973	2.5761
84	1.4792	1.9666	84	2.3359	2.3018	84	4.1244	2.6904
85	1.2143	1.102	85	1.0918	1.1939	85	1.6472	1.403
86	1.4385	1.6551	86	1.9862	1.9018	86	2.9962	2.142
87	1.456	1.6125	87	1.9995	1.8555	87	2.9854	2.1384
88	1.4696	1.9208	88	2.3278	2.2075	88	3.9049	2.558
89	1.2332	0.8295	89	1.0037	0.8932	89	1.2278	1.0157
90	1.3398	1.2056	90	1.5339	1.3743	90	2.0968	1.5589
91	1.7938	1.5634	91	2.4308	1.6843	91	2.7478	1.8553
92	1.8278	1.7321	92	2.1796	1.925	92	3.3291	2.1777

Continuation of Table A.3.

Number of Cases=250			Number of Cases=500			Number of Cases=1000		
Index of β	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**	**Index of β**	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**	**Index of β**	**RE** **PCL vs MLE**	**ARE** **PCL vs MLE**
93	1.1984	0.9727	93	1.0057	1.0422	93	1.3983	1.2223
94	1.6891	1.7322	94	2.2853	2.0151	94	3.5957	2.2339
95	1.7976	1.771	95	2.5636	2.0368	95	3.9455	2.2383
96	1.8651	2.0788	96	2.9513	2.3731	96	4.4372	2.6718
97	1.2503	0.7965	97	1.043	0.9083	97	1.2205	1.02
98	1.6687	1.3632	98	1.9165	1.4985	98	2.7841	1.6821
99	1.2234	0.8171	99	1.0881	0.9328	99	1.2653	1.0283
100	1.1963	0.9377	100	1.0603	1.0544	100	1.2824	1.1667
101	1.4595	0.6975	101	0.9435	0.7147	101	0.9364	0.79
102	1.4845	1.118	102	1.3817	1.2242	102	1.8252	1.3702
103	1.379	0.721	103	1.0006	0.7355	103	0.8843	0.8313
104	1.4396	0.828	104	1.0114	0.8426	104	0.9794	0.9445
105	1.4821	0.703	105	0.9628	0.7352	105	0.8471	0.7795
106	1.394	1.0865	106	1.377	1.1999	106	1.7625	1.3587
107	1.4035	0.6979	107	1.0395	0.7614	107	0.8147	0.8258
108	1.7129	0.7464	108	1.1143	0.7347	108	0.8974	0.798
109	1.4566	0.7184	109	1.0299	0.6963	109	0.9192	0.7419
110	1.2675	0.9316	110	1.2243	1.0351	110	1.4355	1.1439
111	1.7261	0.6522	111	1.0503	0.6612	111	0.7984	0.6866
112	1.5711	0.7562	112	0.9512	0.7379	112	0.7909	0.792
113	1.1187	0.5781	113	0.7467	0.6264	113	0.7284	0.6658
114	1.0959	0.8385	114	1.2538	0.8789	114	1.3179	0.9314
115	0.9964	0.6248	115	0.9482	0.6792	115	0.9405	0.7331

Continuation of Table A.3.

Number of Cases=250			Number of Cases=500			Number of Cases=1000		
Index of β	**RE PCL vs MLE**	**ARE PCL vs MLE**	**Index of β**	**RE PCL vs MLE**	**ARE PCL vs MLE**	**Index of β**	**RE PCL vs MLE**	**ARE PCL vs MLE**
116	1.0783	0.6559	116	0.7636	0.7123	116	0.8473	0.775
117	1.321	0.5894	117	0.882	0.6007	117	0.7724	0.6658
118	1.4364	0.9965	118	1.5948	1.094	118	1.6804	1.2053
119	1.3064	0.6199	119	0.952	0.691	119	0.9656	0.7475
120	1.3798	0.7021	120	0.9803	0.7513	120	0.9266	0.8432
121	1.1314	0.8602	121	0.9782	0.9315	121	1.3919	1.0405
122	1.071	0.8283	122	0.9673	0.9362	122	1.1431	1.0321
123	1.1483	0.8334	123	1.0799	0.9504	123	1.2059	1.0629
124	1.2755	0.8864	124	0.9678	1.0231	124	1.1992	1.1658
125	1.5185	0.7284	125	0.8842	0.7061	125	0.9442	0.8055
126	1.3717	0.7008	126	0.8078	0.7517	126	0.9874	0.8423
127	1.4068	0.7296	127	0.9151	0.7463	127	0.9381	0.8468
128	1.3444	0.8222	128	0.9424	0.8347	128	1.0088	0.9527
129	1.4308	0.7055	129	0.9522	0.718	129	0.9295	0.7887
130	1.2396	0.7219	130	0.8951	0.7731	130	0.8552	0.825
131	1.2678	0.7288	131	0.9614	0.7777	131	0.7977	0.8355
132	1.5578	0.7182	132	0.995	0.7358	132	0.7955	0.7944
133	1.5607	0.6813	133	0.9481	0.7288	133	0.8609	0.7613
134	1.3518	0.6428	134	0.8397	0.6664	134	0.7378	0.7242
135	1.4186	0.6567	135	0.9984	0.658	135	0.7642	0.6971
136	1.6087	0.7204	136	0.994	0.7058	136	0.836	0.7866
137	1.1335	0.6772	137	0.7478	0.7188	137	0.8516	0.774
138	0.7536	0.5653	138	0.6329	0.6409	138	0.7474	0.6896

Continuation of Table A.3.

Number of Cases=250			Number of Cases=500			Number of Cases=1000		
Index of β	**RE PCL vs MLE**	**ARE PCL vs MLE**	**Index of β**	**RE PCL vs MLE**	**ARE PCL vs MLE**	**Index of β**	**RE PCL vs MLE**	**ARE PCL vs MLE**
139	0.9414	0.7318	139	0.9422	0.7901	139	0.9009	0.8391
140	1.0329	0.7275	140	0.8511	0.7634	140	0.9356	0.8345
141	1.1691	0.6066	141	0.8369	0.6433	141	0.8302	0.6816
142	1.0402	0.6501	142	0.8014	0.722	142	0.9601	0.7718
143	1.0945	0.658	143	0.9512	0.7023	143	0.9689	0.7679
144	1.2361	0.717	144	0.9007	0.7796	144	0.8769	0.8687

APPENDIX B

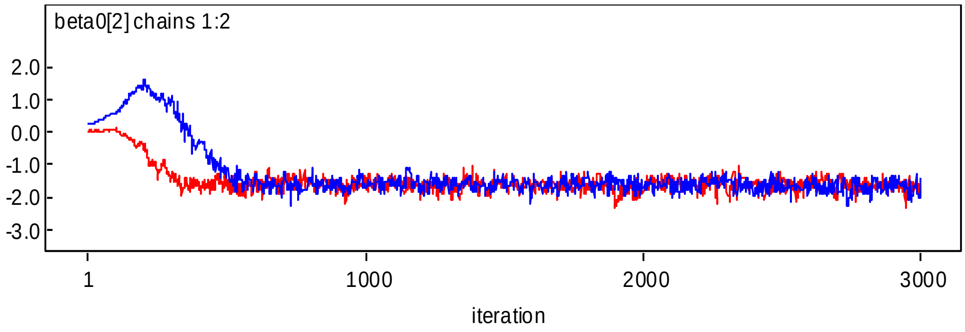
WINBUGS OUTPUT OF CHAIN SIZE AND BURNIN PERIOD DETERMINATION FOR BAYESIAN ESTIMATION IN SIMULATED DATASETS

Table B.1: Posterior summaries for $\alpha_1,\ldots,\alpha_8$ for the pilot dataset

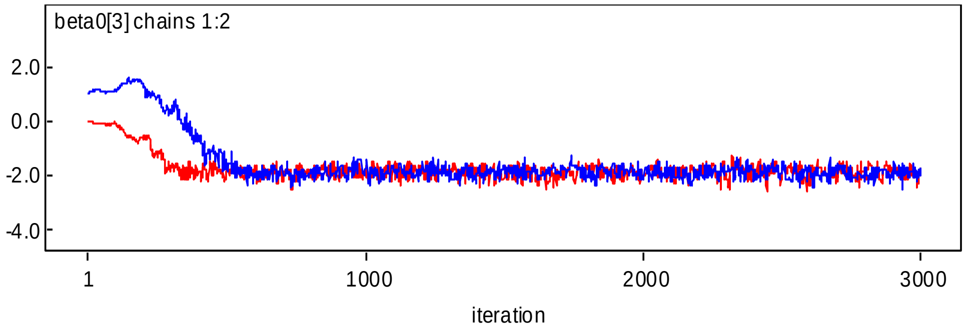
node	mean	sd	MC error	2.50%	median	97.50%	start	sample
beta0[2]	-1.619	0.1817	0.007669	-1.977	-1.611	-1.274	501	5000
beta0[3]	-1.896	0.2021	0.007766	-2.294	-1.892	-1.503	501	5000
beta0[4]	-2.783	0.329	0.02473	-3.437	-2.768	-2.191	501	5000
beta0[5]	-2.617	0.3127	0.01329	-3.247	-2.609	-2.053	501	5000
beta0[6]	-2.633	0.3061	0.01788	-3.311	-2.609	-2.119	501	5000
beta0[7]	-2.275	0.231	0.007698	-2.75	-2.273	-1.83	501	5000
beta0[8]	-2.204	0.2322	0.008757	-2.7	-2.189	-1.769	501	5000
beta0[9]	-2.794	0.3401	0.026	-3.451	-2.784	-2.164	501	5000

Table B.2: Posterior summaries for $\beta_1,\ldots,\beta_8$ for the pilot dataset

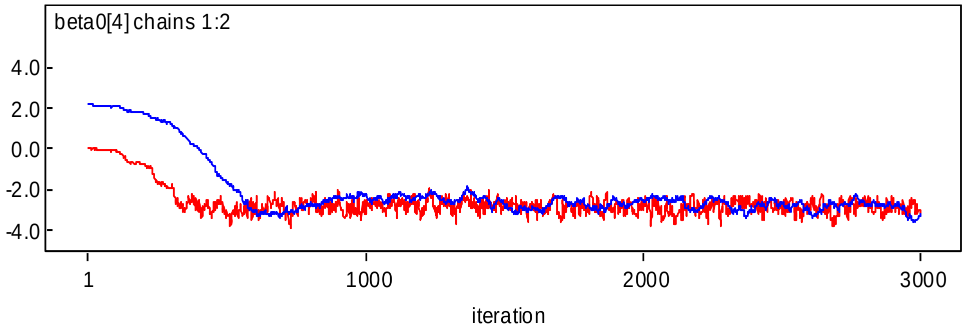
node	mean	sd	MC error	2.50%	median	97.50%	start	sample
beta1[2]	0.6437	0.1778	0.007176	0.3034	0.6413	0.991	501	5000
beta1[3]	0.7261	0.1959	0.007654	0.3593	0.7224	1.122	501	5000
beta1[4]	0.5533	0.2919	0.01304	0.01601	0.5423	1.129	501	5000
beta1[5]	1.018	0.2583	0.01088	0.5315	1.016	1.533	501	5000
beta1[6]	0.7976	0.2811	0.02083	0.2903	0.7894	1.336	501	5000
beta1[7]	0.2882	0.2519	0.008711	-0.2096	0.2895	0.7963	501	5000
beta1[8]	0.2967	0.2388	0.007688	-0.1772	0.296	0.752	501	5000
beta1[9]	1.387	0.2464	0.01619	0.9159	1.376	1.893	501	5000



(a) : Trace plot of two chains for α_1

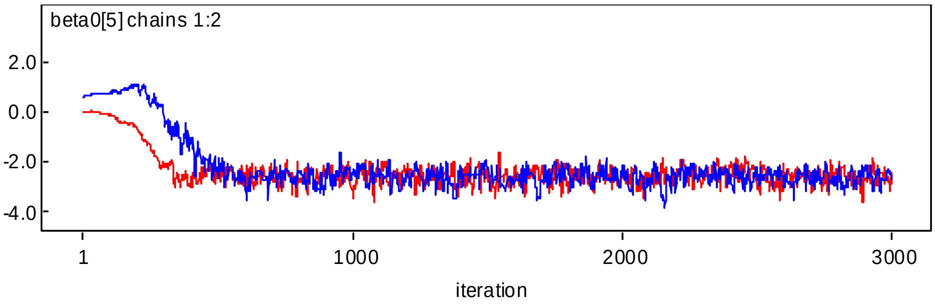


(b) : Trace plot of two chains for α_2

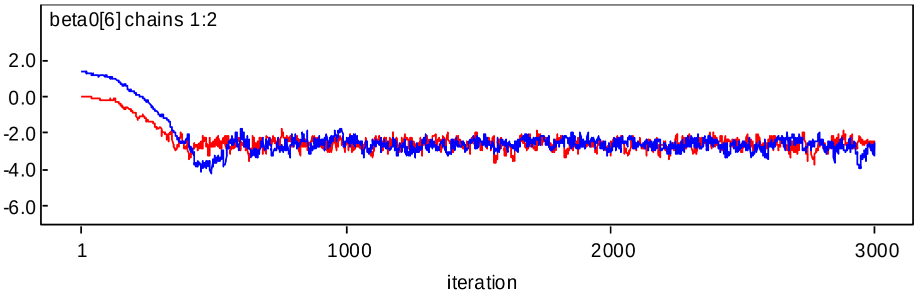


(c) : Trace plot of two chains for α_3

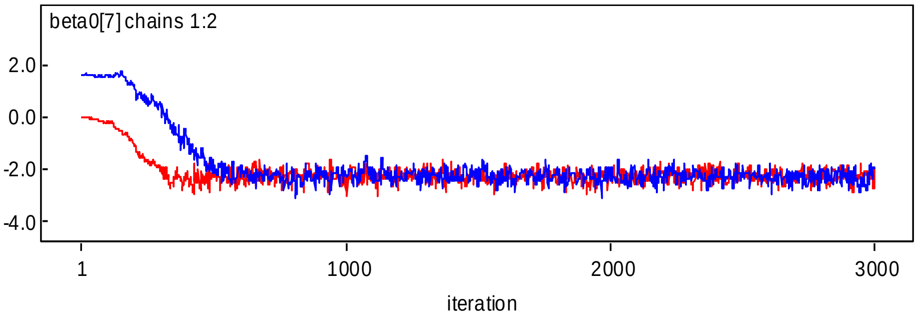
Figure B.1. Trace plots for $\alpha_1,\ldots, \alpha_8$ and $\beta_1,\ldots, \beta_8$



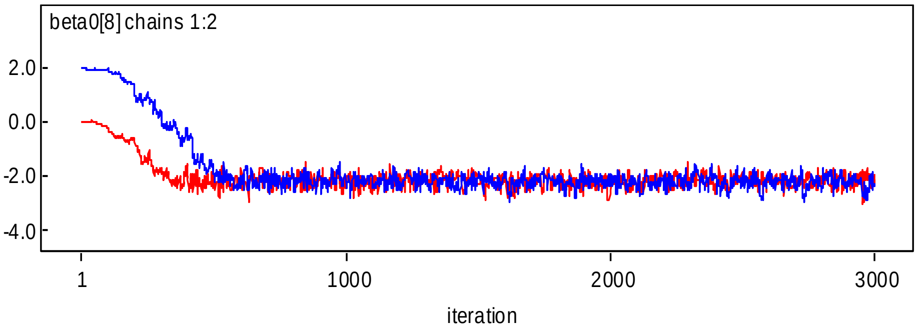
(d) : Trace plot of two chains for α_4



(e): Trace plot of two chains for α_5

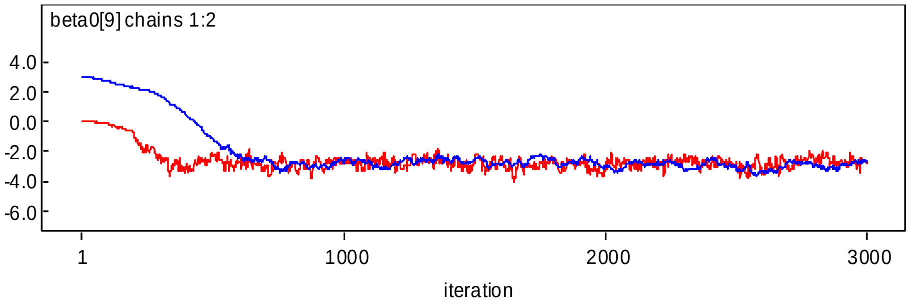


(f) : Trace plot of two chains for α_6

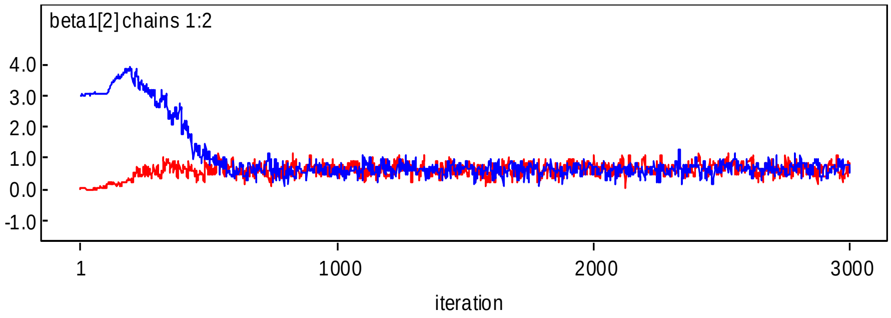


(g): Trace plot of two chains for α_7

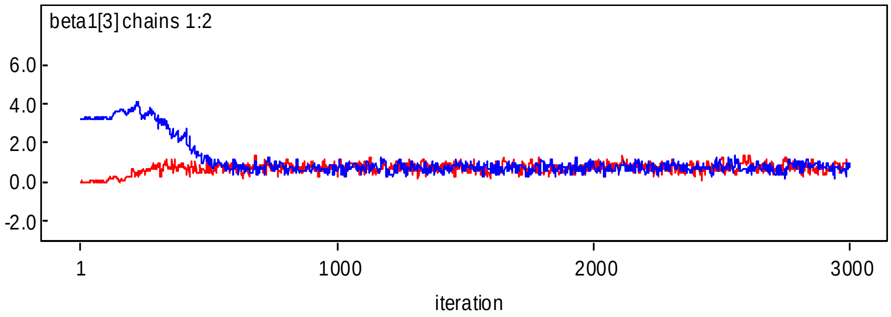
Figure B.1. (continued)



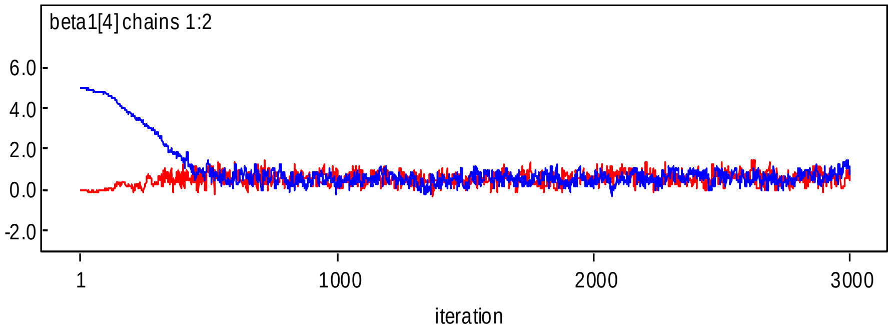
(h) : Trace plot of two chains for α_8



(i): Trace plot of two chains for β_1

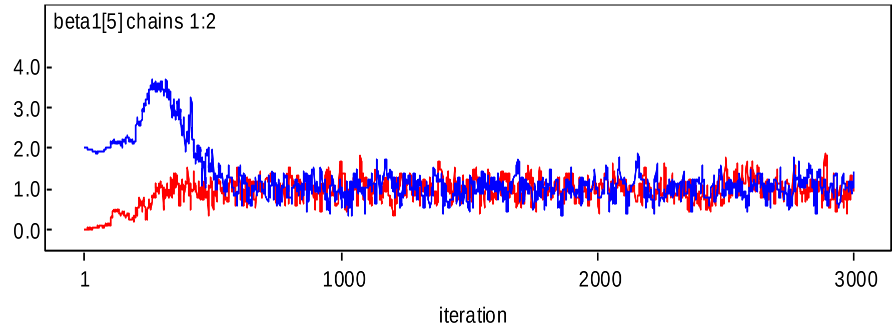


(j) : Trace plot of two chains for β_2

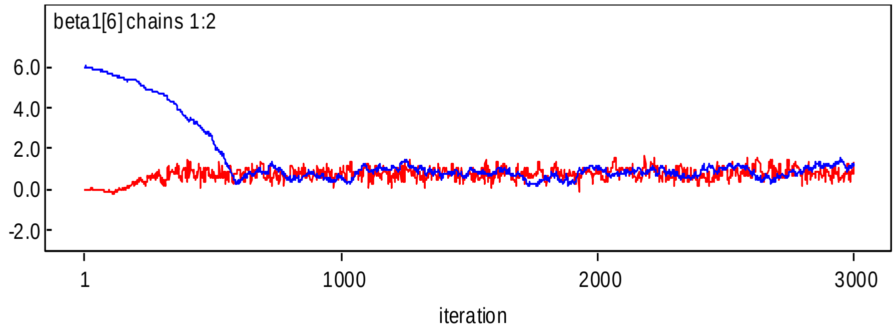


(k) : Trace plot of two chains for β_3

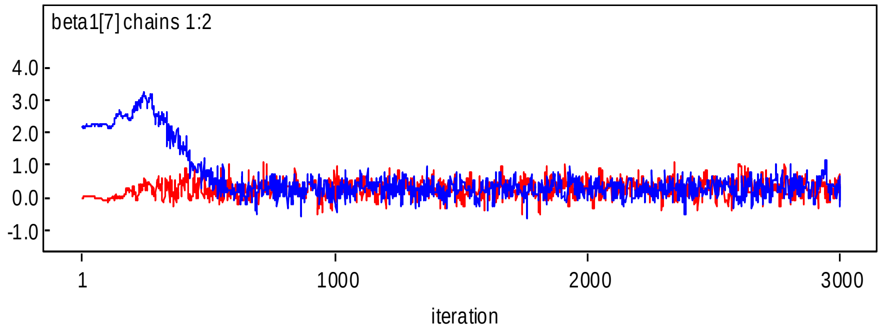
Figure B.1. (continued)



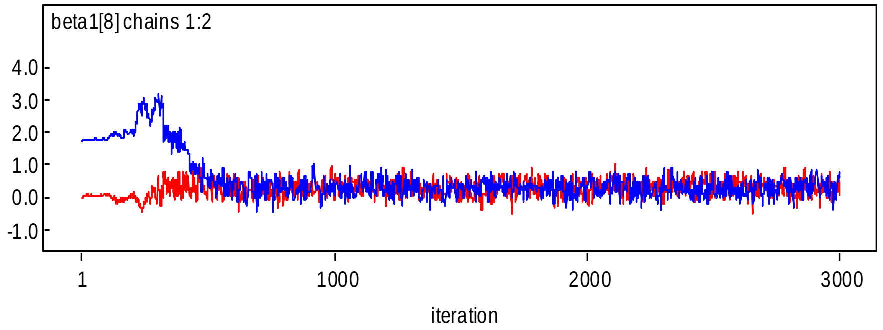
(l) : Trace plot of two chains for β$_4$



(m) : Trace plot of two chains for β$_5$

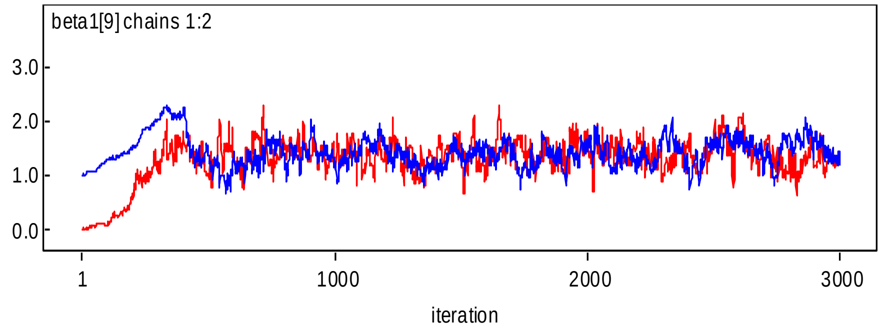


(n) : Trace plot of two chains for β$_6$



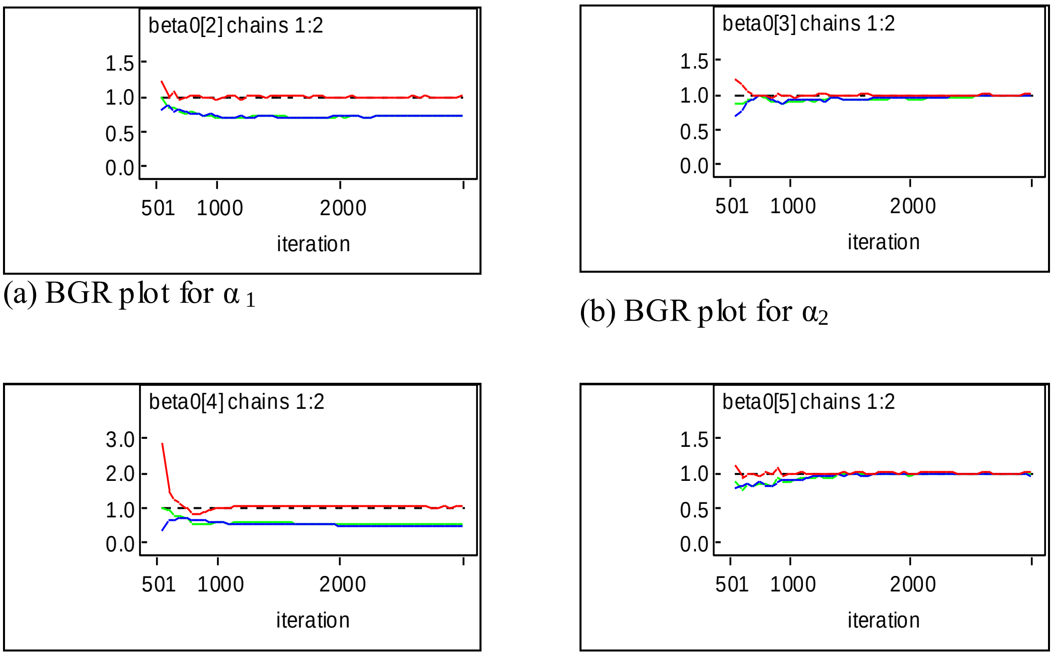
(o) : Trace plot of two chains for β$_7$

Figure B.1. (continued)



(p) : Trace plot of two chains for β_8

Figure B.1. (continued)



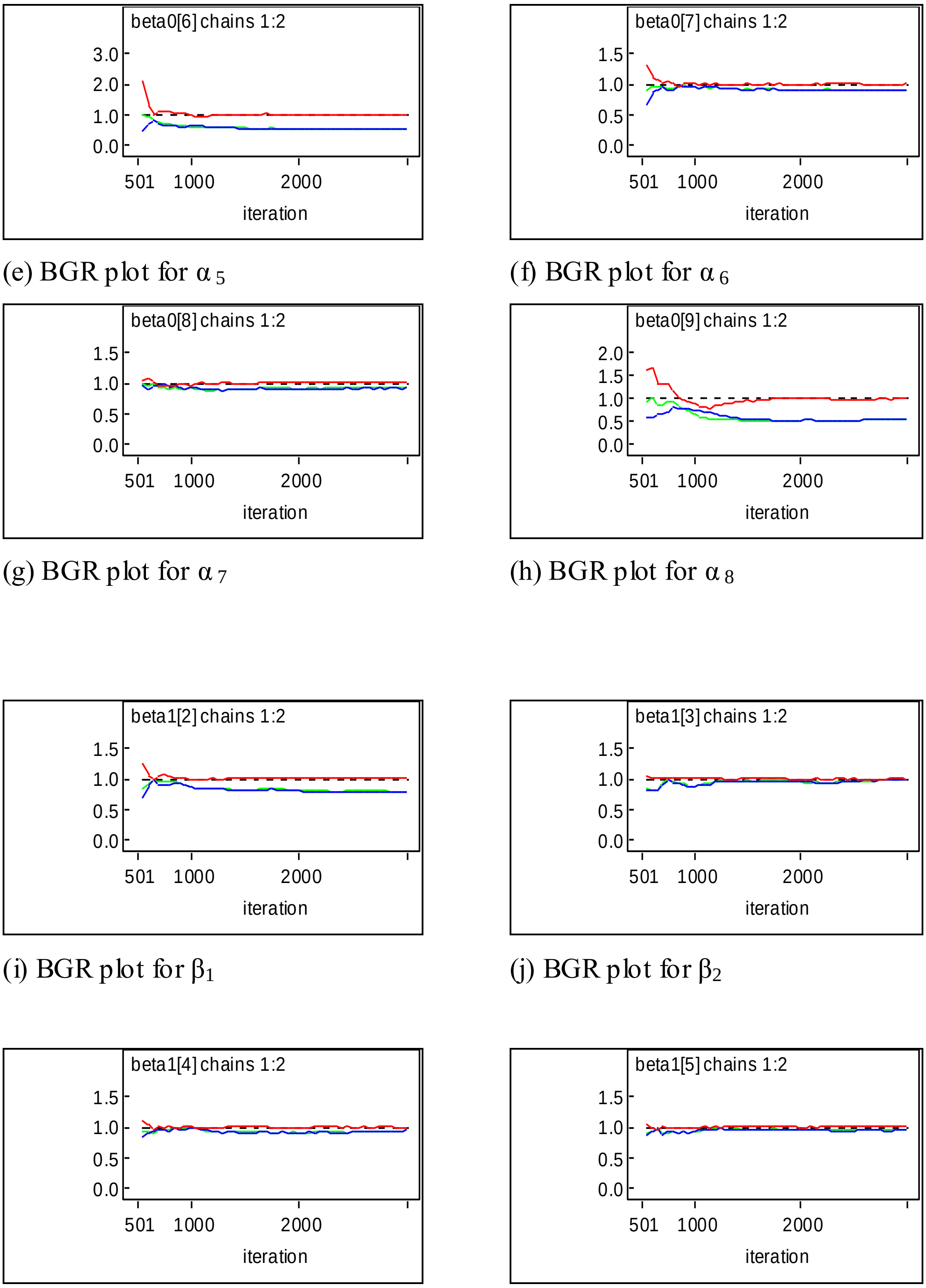
(a) BGR plot for α_1

(b) BGR plot for α_2

(c) BGR plot for α_3

(d) BGR plot for α_4

Figure B.2: Brook-Gelman-Rubin plots for $\alpha_1,\ldots, \alpha_8$ and $\beta_1,\ldots, \beta_8$



(e) BGR plot for α_5

(f) BGR plot for α_6

(g) BGR plot for α_7

(h) BGR plot for α_8

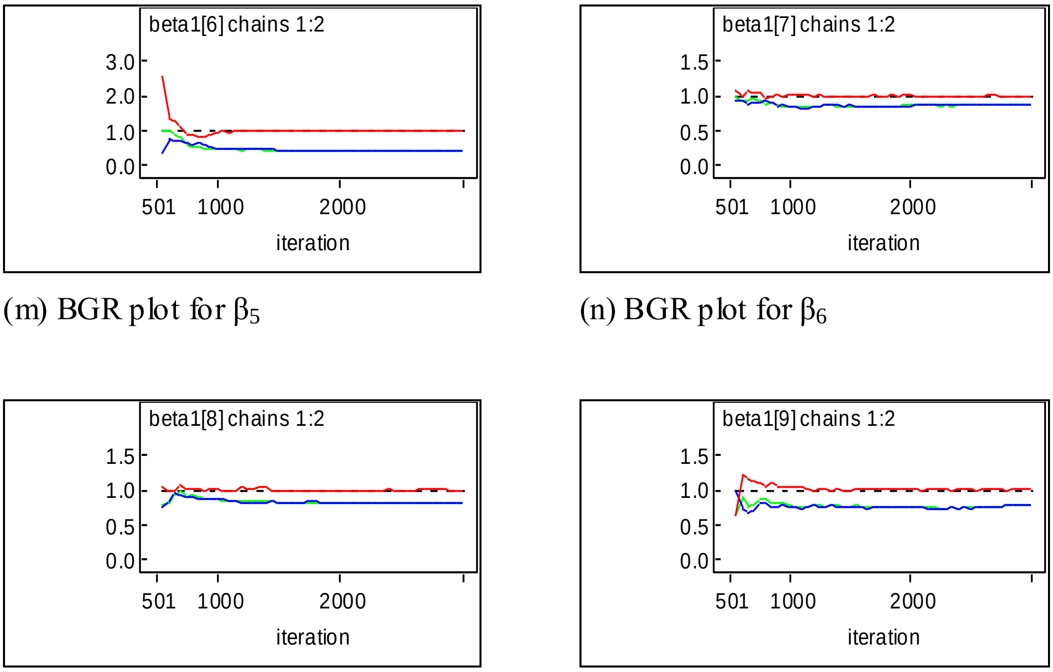
(i) BGR plot for β_1

(j) BGR plot for β_2

(k) BGR plot for β_3

(l) BGR plot for β_4

Figure B.2 (continued)



(m) BGR plot for β_5

(n) BGR plot for β_6

(o) BGR plot for β_7

(p) BGR plot for β_{18}

Figure B.2 (continued)

APPENDIX C

MATLAB CODES FOR SUMULATION DESIGN (M=2x2x2)

```
%simulation design
tic
M = 1000;
ncase = 500;
thetanozero=[.35;.15;0;.5];

N=7000;  %sample that cases will be drawn
dchars=[2 2 2];
theta=[.35;0;.15;0;0;0;.5];
no_disease_categ=prod(dchars);
theta_foralphas = [-3.84; -0.7;-0.7;-0.7;0.5;0.5;0.5];
  % constructing the Z matrix to generate betas
z1 = ones(no_disease_categ,1) ;
z2 = [zeros(4,1) ; ones(4,1)];
z3 = [zeros(2,1) ; ones(2,1); zeros(2,1) ; ones(2,1)];
z4 = [0; 1; 0 ;1 ;0 ;1 ;0 ; 1];
z_betas = [ z1 z2 z3 z4];

   % constructing the Z matrix to generate alphas
  z1 = [zeros(6,1) ; ones(2,1)];
  z2 = [zeros(5,1) ;1 ;0 ; 1];
  z3 = [zeros(3,1) ;1 ;0; 0;0;1];
  z_alphas = [ z_betas z1 z2 z3];
  alpha = z_alphas * theta_foralphas;
%%%%%%

nprm2=sum(dchars)+1-length(dchars); %number of thetas
nsubtype=prod(dchars);

zadd = zeros(nsubtype,nprm2);
     m=1;
     for i=1:2
        for j=1:2
          for k=1:2
               zadd(m,:) = [1  (i==2)  (j==2)  (k==2)];
            m=m+1;
          end
        end
     end


Btrue=zadd*thetanozero;

fidmle = fopen('mcout_222_x1norm_mle.txt','a');
fidbayes = fopen('mcout_222_x1norm_bayes.txt','a');
fidpcl = fopen('mcout_222_x1norm_pcl.txt','a');
```

```
for mc=1:M

    % mnrfit & pcl & winbugs run

    %DATA GENERATION PART%
    %obtain beta's from thetas:
    bet=betas(theta, dchars); %obtain beta's from thetas

    % exposure:
    xs=normrnd(0,1,N,1);

    %**% compute the probabilities P(Di=m|x) as the first step of creating disease
    %subtype category for each individual (i.e. create y)
    p=generateprob(xs,alpha,bet);

    % GENERATE FROM MULTINOMIAL
    R = mnrnd(1,p);
    ds = zeros(N,1);
    for i =1:N
        ds(i)= find(R(i,:));
    end;
    ds = ds -1;

    %Sample from the controls, where sample size of the selected controls will be
    %equal to the number of cases
    [x d]=selectsample2(xs,ds,ncase);  %ncase is the number of cases
    n = length(d);

    % ANALYSIS PART

    %MLE:
     di=d+1;
     nd=newcateg(di);
     [b,s,stats] = mnrfit(x,nd);
     BMLE= (fliplr(b))'; %betas
     BMLEse=(fliplr(stats.se))'; %standard errors of betas
     % 95% confidence intervals for betas:
     BMLElb95=BMLE-1.9599*BMLEse; % lower bound
     BMLEub95=BMLE+1.9599*BMLEse;  % upper bound

    % BAYESIAN ESTIMATION:

    dataStruct = struct('D', di, 'x', x, 'J',9, 'n', n);

    init0 = struct( 'beta0', [nan 0 0 0 0 0 0 0 0], 'beta1', [nan 0 0 0 0 0 0 0 0]);

    [samples, stats, structArray] = matbugs(dataStruct, ...
        fullfile(pwd, 'poly_log_reg_nominal_model.txt'), ...
        'init', init0, ...
        'nChains', 1, ...
        'view', 0, 'nburnin', 500, 'nsamples', 2500, ...
        'thin', 1, 'DICstatus', 0, ...
               'monitorParams', {'beta0', 'beta1'}, ...
        'Bugdir', 'C:/Program Files/WinBUGS14');

    BBYS=[stats.mean.beta0(2:9) ; stats.mean.beta1(2:9)]'; %betas
```

```
  BBYSse=[stats.std.beta0(2:9) ; stats.std.beta1(2:9)]'; %standard errors of betas


   % 95% confidence intervals for betas:
   BBYSlb95=BBYS-1.9599*BBYSse;
   BBYSub95=BBYS+1.9599*BBYSse;

  % PCL ESTIMATION

  %disease subtypes:
  siz=zeros(n,1);
  vil=zeros(n,1);
  mult=zeros(n,1);

  svm= zeros(n,3);
  svm((d==1),:)= repmat([1 1 1],sum((d==1)),1);
  svm((d==2),:)= repmat([1 1 2],sum((d==2)),1);
  svm((d==3),:)= repmat([1 2 1],sum((d==3)),1);
  svm((d==4),:)= repmat([1 2 2],sum((d==4)),1);
  svm((d==5),:)= repmat([2 1 1],sum((d==5)),1);
  svm((d==6),:)= repmat([2 1 2],sum((d==6)),1);
  svm((d==7),:)= repmat([2 2 1],sum((d==7)),1);
  svm((d==8),:)= repmat([2 2 2],sum((d==8)),1);

  siz=svm(:,1);
  vil=svm(:,2);
  mult=svm(:,3);
%
 yy = zeros(n,nsubtype);
   i=0; j=0; k=0;
   mm=0;
   for i=1:2
     for j=1:2
       for k=1:2
         mm=mm+1;
         yy(:,mm)=(siz==i)&(vil==j)&(mult==k);
       end
     end
    end

%one of the inputs of EH_analysis_packed;
status = zeros(n,1);
status = (d>1);

%thetas and std errors from PCL
[est sd vt lb ub]= EH_analysis_packed(yy,zadd,status,x,0,1)
T=est';
BPCL=zadd*T; % betas
BPCLse=sqrt(diag(zadd*vt*zadd')); % standard errors of betas
% 95% confidence intervals for betas:
BPCLlb95=BPCL-1.9599*BPCLse;
BPCLub95=BPCL+1.9599*BPCLse;


estsemle = [BMLE BMLEse BMLElb95 BMLEub95]; %size: 64 by 1
estsebayes = [BBYS BBYSse BBYSlb95 BBYSub95]; %size: 64 by 1
estsepcl = [BPCL BPCLse BPCLlb95 BPCLub95]; %size: 32 by 1
```

```
fprintf(fidmle,  '%g %g %g %g %g %g %g %g %g %g %g %g  %g %g %g %g %g %g %g %g %g
%g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g
%g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g \n', estsemle);
fprintf(fidbayes, '%g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g
%g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g
%g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g \n', estsebayes);
fprintf(fidpcl,  '%g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g %g
%g %g %g %g %g %g %g %g %g %g\n', estsepcl);

mc
end % mc=1:M


fclose(fidmle);
fclose(fidbayes);
fclose(fidpcl);

toc

%%%%%%%%%%FUNCTIONS%%%%%%%%%%%%

% this function will compute the  PROBABILITY OF EACH DISEASE SUBTYPE for
% each subject(xi) and return these probabilities. (returns nx9 matrix )

%n: sample size, x exposure, alpha ve beta are the parameters used in
%diseade subtype creation ; x, alpha, beta column vectors
function[p]=generateprob(x,alpha,bet)
[n c]=size(x); %n is the sample size, c is the number of covariates.
m=length(bet);
 p=zeros(n,m+1); % p: probability of disease subtype

 for i=1:n  % i: subject. there are n subjects in total
       term = zeros(m,1);
       for  j=2:m+1 % j: disease subtype  /there are 8 probs
       term(j-1,1)=exp(alpha(j-1)+bet(j-1)*x(i,1));
       end % j icin

       p(i,2:m+1)=term'/(1+sum(term));
       %p(i,:)=term/ (1+sum(term));

    end % i icin
    p(:,1)=1-sum(p,2);
    sum(p,2);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% theta to beta
%%%dchars is a vector which has the length equal to number of disease characteristics
%and holds the number of levels for each characteristics
function[betas]=betas(theta, dchars)

z=zmatrix(dchars);
betas=z*theta;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%output: x's(cases and controls in equal sizes) and corrwponding disaease
                                                   %subtype status
```

```
function [xfinal dfinal pr]=selectsample(x,d,nc)

n=length(d);
cas=[]; contfull=[]; cont=[]; z=[]; y=[]; dcas=[]; dc=[];
dcont=[]; dco=[]; dcontfull=[]; caslast=[]; dcaslast=[];
 for i=1:n   %case
   if d(i)~=0
       z=x(i,:);
       cas=[cas; z];
       dc=d(i);
       dcas=[dcas; dc];

   else          %control
       y=x(i,:);
       contfull=[contfull;y];
       dco=d(i);
       dcontfull=[dcontfull; dco];

   end

 end

 r= nc;
%r=input('enter number of case..: ')
l=length(cas);
k=length(contfull);

if r>l
   disp('there are not enough number of cases.');
else
ind1=unidrnd(k,r,1);
cont=contfull(ind1);
dcont=dcontfull(ind1);

ind2=unidrnd(l,r,1);
caslast=cas(ind2);
dcaslast=dcas(ind2);
end
xfinal=[caslast; cont];
dfinal=[dcaslast; dcont];

pr=length(dcas)/(length(dcas)+length(dcontfull));

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%obtain beta's from thetas
%ATTENTION: this function for only if there are 3 characteristics each has 2 levels
% dchars is a vector which has the length equal to number of disease characteristics
                         %and holds the number of levels for each characteristics
%consider all levels of theta
function[z]=zmatrix(dchars)
nsubtype=prod(dchars);
nprm2=sum(dchars)+1; %number of thetas

z=zeros(nsubtype,nprm2);
m=1;
for i=1:2
      for j=1:2
        for k=1:2
```

```
              z(m,:) = [1 (i==1) (i==2) (j==1) (j==2) (k==1) (k==2)];
                m=m+1;
        end
      end
   end

%RELATIVE EFFICIENCY
%nc=250
dpcl=load('mcout_222_x1norm_pcl_nc250.txt');
beta_pcl=dpcl(:,1:8);
[N p ]=size(beta_pcl);
mc_var_pcl=var(beta_pcl,1);
dmle=load('mcout_222_x1norm_mle_nc250.txt');
beta1_mle=dmle(:,9:16);
mc_var_mle=var(beta1_mle,1);
   %PCL vs MLE
re_pcl_mle = mc_var_pcl./ mc_var_mle;
re_pcl_mle'

%ASYMPTOTIC RELATIVE EFFICIENCY
avar_pcl = mean(dpcl(:,9:16),1);
avar_mle = mean(dmle(:,25:32),1);
  %PCL vs MLE
  are_pcl_mle = avar_pcl ./ avar_mle;
  are_pcl_mle'
```