RELATIVE RISKS ANALYSIS IN NUTRITIONAL EPIDEMIOLOGY

A Dissertation

by

YANQING WANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Raymond Carroll
Co-Chair of Committee,	Bani Mallick
Committee Members,	Veera Baladandayuthapani
	Ulisses Braga-Neto
Head of Department,	Simon Sheather

August 2014

Major Subject: Statistics

ABSTRACT

Motivated by a logistic regression problem involving diet and cancer, we reconsider the problem of forming a confidence interval for the ratio of two location parameters. We develop a new methodology, which we call the Direct Integral Method for Ratios (DIMER). In simulations, we compare this method to many others, including Wald's method, Fieller's interval, Hayya's method, the nonparametric bootstrap and the parametric bootstrap. These simulations show that, generally, DIMER more closely achieves the nominal confidence level, and in those cases that the other methods achieve the nominal levels, DIMER generally has smaller confidence interval lengths. We also show that DIMER eliminates the probability of infinite length or enormous length confidence intervals, something that can occur in Fieller's interval.

Furthermore, we study the real Healthy Eating Index-2005 (HEI-2005) data set from the NIH-AARP Study of Diet and Health, consider a weighted logistic regression model in which there are multiple subpopulations, and multiple diseases within each subpopulation. Based on this model, we present six different approaches to form the confidence intervals for the relative risks of different diseases in different subpopulations, including DIMER. The asymptotic distributions of the estimates for the log(relative risks) by the maximum likelihood and the nonparametric bootstrap method are provided. Next, the algorithms are presented to perform hypothesis tests and likelihood ratio tests to check there are significant differences between our proposed model and the other three logistic regression models or not. In addition, the adaptive lasso and an estimator with bounded constrains are described for variable selection and a novel algorithm to solve the nonlinear regression model with L_1 norm penalty is proposed. The application of all those methods to the HEI-2005 data are illustrated.

Additionally, we expand the linear function of nutrition components inside the logistic regression model to a nonlinear case. More than that, we consider there are some limitations from the knowledge of biology and nutrition and propose a logistic regression model involving I-spline basis functions and an algorithm to solve it. Application to the real HEI-200d data set and comparison to a logistic model with total HEI scores are also presented.

DEDICATION

To my family for their endless love, support and encouragement.

ACKNOWLEDGEMENTS

Upon the completion of this dissertation, I would like to express my sincere and deepest gratitude to my advisor Dr. Raymond Carroll, who is a world-famous researcher on statistics and a great teacher, not only for his continuous encouragement and guidance throughout my research study, but also his helping into details with great patience. And his generous financial support on research and travel expenses guarantees me to concentrate on my work. Furthermore, during my PhD study instructed by Dr. Carroll, his diligence and methodical way to work, his kindness and willing to help people around, all these virtues have deeply impressed and influenced me, which will benefit me for the whole life.

I would also like to thank my co-advisor Dr. Bani Mallick and my committee members Dr. Veera and Dr.Brago-Neto. They all have outstanding achievements in research and favor my research work with helpful guidance and comments, which greatly expand my view on research.

Finally, I would like to appreciate all my colleagues in the department of Statistics at Texas A&M University, the discussions and cooperation with them gave me a lot of inspirations and knowledge.

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iv

ACKNOWLEDGEMENTS v

TABLE OF CONTENTS vi

LIST OF FIGURES viii

LIST OF TABLES x

1. INTRODUCTION 1

2. THE DIRECT INTEGRAL METHOD FOR CONFIDENCE INTERVALS FOR THE RATIO OF TWO LOCATION PARAMETERS 6

   2.1 Introduction 6
   2.2 Methodology 7
   2.3 Simulations 12
   2.4 Discussion 28

3. RELATIVE RISKS ANALYSIS AND MODEL COMPARISON IN DIETARY INDEX MODELING FOR HEI-2005 29

   3.1 Introduction 29
   3.2 Basic Model 31
   3.3 Relative Risks Analysis 34
   3.4 Results and Discussions 41

4. MODEL COMPARISON AND VARIABLE SELECTION IN DIETARY INDEX MODELING FOR HEI-2005 47

   4.1 Introduction 47
   4.2 Model Comparison in HEI-2005 47
   4.3 Variable Selection by the Bounded Constrains 54
   4.4 The Adaptive Lasso Method 55
   4.5 Result of the Adaptive Lasso 60

5. VARIABLE SELECTION IN DIETARY INDEX MODELING FOR HEI-2005 69

5.1 Introduction 69
5.2 I-spline Basis Function and Regression Model 70
5.3 Results and Discussion 74

6. CONCLUSIONS 79

REFERENCES 81

APPENDIX A. SUPPLEMENTARY MATERIAL FOR CHAPTER 2 87

A.1 Fieller's Method 87
A.2 Proofs of Lemmas 1-2 88
A.3 More Details of Simulation Results in Chapter 2 90
A.4 Bootstrapping Details 95

APPENDIX B. SUPPLEMENTARY MATERIAL FOR CHAPTER 3 96

B.1 Proof of Lemma 3 96
B.2 Plots of Confidence Intervals Lengths 100

APPENDIX C. SUPPLEMENTARY MATERIAL FOR CHAPTER 4 104

C.1 Proof of Lemma 6 104
C.2 Figures of Cross Validation Results for Colorectal Cancer Data 105

APPENDIX D. SUPPLEMENTARY MATERIAL FOR CHAPTER 5 108

D.1 Expression of I-splines basis function with specifications 108
D.2 Figures for I-spline Analysis in HEI-2005 111
D.3 I-Spline Analysis Results for Modified HEI-2005 Components 115

LIST OF FIGURES

FIGURE Page

B.1 Relationship between the 95% confidence interval by using the inverse Fisher matrix and the number of points which involved in computing the derivative $\partial \widehat{S}_{0.90,k\ell}(\Lambda_{k\ell})/\partial \Lambda_{k\ell}^{\mathrm{T}}$ and $\partial \widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})/\partial \Lambda_{k\ell}^{\mathrm{T}}$. 101

B.2 Relationship between the 95% confidence interval by using the sandwich method and the number of points which involved in computing the derivative $\partial \widehat{S}_{0.90,k\ell}(\Lambda_{k\ell})/\partial \Lambda_{k\ell}^{\mathrm{T}}$ and $\partial \widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})/\partial \Lambda_{k\ell}^{\mathrm{T}}$. 102

B.3 Relationship between the 95% confidence interval by using the model transformation method and the number of points which involved in computing the derivative $\partial \widehat{S}_{0.90,k\ell}(\Lambda_{k\ell})/\partial \Lambda_{k\ell}^{\mathrm{T}}$ and $\partial \widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})/\partial \Lambda_{k\ell}^{\mathrm{T}}$. . 103

C.1 Prediction error of cross-validation method in the adaptive lasso for the logistic regression of the colorectal cancer on HEI component scores of men (293615 with 2151 cases) and women together (198245 with 959 cases), where β_{11} = -1, $\widehat{\beta}_{12} = -0.7411$ and $\gamma = 1.5$. 106

C.2 Prediction error of cross-validation method in the adaptive lasso for the logistic regression of the colorectal cancer on HEI component scores of men (293615 with 2151 cases) and women together (198245 with 959 cases), where β_{11} = -1, $\widehat{\beta}_{12} = -0.7411$ and $\gamma = 1.5$. 107

D.1 I-spline analysis results for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$. 112

D.2 I-spline analysis results for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$. 113

D.3 I-spline analysis results for the logistic regression of the colorectal and lung cancers on modified HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$. 114

D.4 I-spline analysis results for the logistic regression of the colorectal and lung cancers on modified HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$. 115

LIST OF TABLES

TABLE Page

2.1 Confidence intervals for β_{21} in a simulation study with 2000 replications and true parameter values $(\beta_{10}, \beta_{20}, \beta_{21}, \omega) = (0.00, 0.00, 1.00, 1.00)$ for the linear regression model $Y_{1i} = \beta_{10} - X_{1i}\omega + \varepsilon_{1i}; Y_{2j} = \beta_{20} + \beta_{21}X_{2j}\omega + \varepsilon_{2j}$. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval and DIMER–Direct Integral Method for Ratios. 19

2.2 Confidence intervals for β_{21} in a simulation study with 2000 replications and true parameter values $(\beta_{10}, \beta_{20}, \beta_{21}, \omega) = (0.00, 0.00, 1.00, 0.75)$ for the linear regression model $Y_{1i} = \beta_{10} - X_{1i}\omega + \varepsilon_{1i}; Y_{2j} = \beta_{20} + \beta_{21}X_{2j}\omega + \varepsilon_{2j}$. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval and DIMER–Direct Integral Method for Ratios. 20

2.3 Confidence intervals for μ in a simulation study with 2000 replications for linear regression model $Y_i = \beta(X_i - \mu) + \epsilon_i$ with Setting I: $(\beta, \mu) = (1.00, 1.00)$. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval and DIMER–Direct Integral Method for Ratios. 26

2.4 Confidence intervals for μ in a simulation study with 2000 replications for linear regression model $Y_i = \beta(X_i - \mu) + \epsilon_i$ with $(\beta, \mu) = (2.00, 1.00)$, where ϵ_i follows a skew normal distribution with mean 0, variance 1 and skewness 0.78. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval, DIMER–Direct Integral Method for Ratios and LR–Likelihood ratio test. 27

3.1 Description of the HEI-2005 scoring system. Except for saturated fat and SoFAAS, density is obtained by multiplying usual intake by 1000 and dividing by usual intake of kilo-calories. For saturated fat, density is 9×100 usual saturated fat (grams) divided by usual calories, i.e., the percentage of usual calories coming from usual saturated fat intake. For SoFAAS, the density is the percentage of usual intake that comes from usual intake of calories, i.e., the division of usual intake of SoFAAS by usual intake of calories. Here, "DOL" is dark green and orange vegetables and legumes. Also, "SoFAAS" is calories from solid fats, alcoholic beverages and added sugars. The total HEI-2005 score is the sum of the individual component scores. 30

3.2 Relative Risks and their 95% Confidence intervals for colorectal cancer on HEI component scores with model $\mathrm{pr}(Y_{ik\ell} = 1|X_{i\ell}, Z_{ik\ell}) = H(\alpha_{k\ell} + \beta_{k\ell}\sum_{j=1}^{J} X_{ij\ell}\omega_j + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell})$, where $\beta_{11} = -1$. IF–the inverse Fisher method, SM–the sandwich method, DIMER–the direct integral method for ratios, MT–the model transformation method, NB–the nonparametric bootstrap method and MNB–the modified nonparametric bootstrap method. The nonparametric bootstrap method has 2500 simulated data sets and the modified bootstrap method has 2320 simulated data sets. 45

3.3 Bootstrap results of Relative Risks for colorectal cancer on HEI component scores. RR–relative risks. Method I–by percentile of relative risk from bootstrap. Method II–the exponent of $\mathcal{V}_{k\ell}^{*} \pm 1.96(\widehat{D}_{k\ell}^{*})^{1/2}$ based on the normal approximation of log(Relative Risk). The nonparametric bootstrap method has 2500 simulated data sets and the modified bootstrap method has 2320 simulated data sets. 46

4.1 Relative risks comparisons between four different models by the modified nonparametric bootstrap method with B=2316. $\Delta_{\mathcal{V}_{k\ell}}$–mean of the difference of log(relative risks). $\text{TS}_{\Delta_{\mathcal{V}_{k\ell}}}$–test statistic of log(relative risks). p-value$_{\Delta_{\mathcal{V}_{k\ell}}}$–p-value of the hypothesis test for log(relative risks). $\Delta_{\exp(\mathcal{V}_{k\ell})}$–mean of the difference of relative risks. $\text{TS}_{\Delta_{\exp(\mathcal{V}_{k\ell})}}$–test statistics of relative risks. p-value$_{\Delta_{\exp(\mathcal{V}_{k\ell})}}$–p-value of the hypothesis test for relative risks. DRR–Difference of relative risks. RRR–Ratio of relative risks. 52

4.2 Likelihood Ratio tests between four different models, where the test statistics follow chi-square distributions. 53

4.3 Variable selection results by the bounded constrain algorithm for the logistic regression of the colorectal cancer on HEI component scores of men and women together, where $\beta_{11} = -1, \widehat{\beta}_{12} = -0.7411$. 63

4.4 Variable selection results by using the adaptive lasso method and positive constrain for the logistic regression of the colorectal cancer on HEI component scores of men and women together, where $\beta_{11} = -1, \widehat{\beta}_{12} = -0.7411$ and $\gamma = 1.5$. MLE–maximum likelihood estimator. 64

4.5 Analysis results of $\widehat{\omega}_{HEI}$ by the adaptive lasso method for the logistic regression of the colorectal cancer on HEI component scores of men (293615 with 2151 cases) and women (198245 with 959 cases) together, where $\beta_{11} = -1, \widehat{\beta}_{12} = -0.7411$ and $\gamma = 1.5$. ω_{HEI}–algorithm is in Section 4.5. MLE–maximum likelihood estimator. 65

4.6 Analysis results for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1, \gamma = 1.5$. $\widehat{\omega}_{HEI}$–algorithm is in Section 4.5. MLE–maximum likelihood estimator. 66

4.7 Analysis results for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1, \gamma = 1.5$. ω_{HEI}–algorithm is in Section 4.5. MLE–maximum likelihood estimator. 67

4.8 Analysis results of $\widehat{\beta}_{k\ell}$ for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1, \gamma = 1.5$. Unweighted Model–pr$(Y_{ik\ell} = 1|X_{ij\ell}, Z_{ik\ell}) = H(\alpha_{k\ell} + \beta_{k\ell}\sum_{j=1}^{J} X_{ij\ell} + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell})$, Weighted Model–pr$(Y_{ik\ell} = 1|X_{ij\ell}, Z_{ik\ell}) = H(\alpha_{k\ell} + \beta_{k\ell}\sum_{j=1}^{J} X_{ij\ell}\widehat{\omega}_j + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell})$. LR–Likelihood ratio test statistic, log-likelihood(weighted model)$-$log-likelihood(unweighted model), which follows a chi-squared distribution with degree freedom 1. 68

5.1 I-spline analysis results for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$. 76

5.2 I-spline analysis results for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$. 77

5.3 I-spine analysis results of $\widehat{\beta}_{k\ell}$ for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$. LR–Likelihood ratio test statistic. 78

A.1 Confidence intervals for β_{21} in a simulation study with 2000 replications and true parameter values $(\beta_{10}, \beta_{20}, \beta_{21}, \omega) = (0.00, 0.00, 1.00, 1.00)$ for the linear regression model $Y_{1i} = \beta_{10} - X_{1i}\omega + \varepsilon_{1i}; Y_{2j} = \beta_{20} + \beta_{21}X_{2j}\omega + \varepsilon_{2j}$. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval and DIMER–Direct Integral Method for Ratios. 91

A.2 Confidence intervals for β_{21} in a simulation study with 2000 replications and true parameter values $(\beta_{10}, \beta_{20}, \beta_{21}, \omega) = (0.00, 0.00, 1.00, 0.75)$ for the linear regression model $Y_{1i} = \beta_{10} - X_{1i}\omega + \varepsilon_{1i}; Y_{2j} = \beta_{20} + \beta_{21}X_{2j}\omega + \varepsilon_{2j}$. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval and DIMER–Direct Integral Method for Ratios. 92

A.3 Confidence intervals for μ in a simulation study with 2000 replications for linear regression model $Y_i = \beta(X_i - \mu) + \epsilon_i$ with Setting I: $(\beta, \mu) = (1.00, 1.00)$. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval and DIMER–Direct Integral Method for Ratios. 93

A.4 Confidence intervals for μ in a simulation study with 2000 replications for linear regression model $Y_i = \beta(X_i - \mu) + \epsilon_i$ with $(\beta, \mu) = (2.00, 1.00)$, where ϵ_i follows a skew normal distribution with mean 0, variance 1 and skewness 0.78. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval, DIMER–Direct Integral Method for Ratios and LR–Likelihood ratio test. 94

D.1 I-spline analysis results for the logistic regression of the colorectal and lung cancers on modified HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$. 116

D.2 I-spline analysis results for the logistic regression of the colorectal and lung cancers on modified HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$. 117

D.3 I-spine analysis results of $\widehat{\beta}_{k\ell}$ for the logistic regression of the colorectal and lung cancers on modified HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1, \gamma = 1.5$. Unweighted Model–pr$(Y_{ik\ell} = 1|X_{ij\ell}, Z_{ik\ell}) = H(\alpha_{k\ell} + \beta_{k\ell}\sum_{j=1}^{J}X_{ij\ell} + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell})$, Weighted Model–pr$(Y_{ik\ell} = 1|X_{i1\ell}, ..., X_{iJ\ell}, Z_{ik\ell}) = H[\alpha_{k\ell} + \beta_{k\ell}\sum_{j=1}^{J}\{\sum_{q=1}^{Q}I_{jq}(X_{ij\ell})\widehat{\omega}_{jq}\} + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell}]$ with constrain $\omega_{jq} \geq 0$. LR–Likelihood ratio test statistic. 118

1. INTRODUCTION

This dissertation focuses on study of the real Healthy Eating Index-2005 (HEI-2005) data set from the NIH-AARP Study of Diet and Health. One of the US Department of Agriculture's (USDA's) strategic objectives is 'to promote healthy diets' and it has developed an associated performance measure, the Healthy Eating Index-2005 (HEI-2005, see Guenther et al., 2008a,b). The HEI-2005 is based on the key recommendations of the 2005 Dietary Guidelines for Americans. The index includes ratios of interrelated dietary components to energy. The HEI-2005 comprises 12 distinct component scores and a total summary score. See Chapter 3 for a list of these components and the standards for scoring, and see Guenther et al. (2008a,b) for details. Intakes of each food or nutrient, represented by one of the 12 components, are expressed as a ratio to energy intake, assessed, and ascribed a score.

The total score defined as the sum of these 12 nutrition components, has been widely used to analyze the relationship between diseases, mortality and individual food intake. Reedy, et al. (2008) show that in a Cox regression for colorectal cancer in the NIH-AARP Study of Diet and Health, with diet assessed by a food frequency questionnaire (FFQ), higher HEI-2005 total scores are statistically significantly associated with lower risk, with a relative risk of 0.72 for men and 0.80 for women.

In my research work, firstly I expand Reedy's work to a weighted logistic regression model. In the other words, we assume each diet component has various weight in regression, although they are the same in Reedy's model. Details for the weighted model are given in Chapter 3. Based on it, one of our key goals to analyze this data set is determining relative risk for each disease in different subpopulations, which is closely related to the lengths of confidence intervals for some parameters in the regression model. Therefore, the question arises to form the confidence intervals of these parameters with reasonable lengths. Two usual approaches, the sandwich method and nonparametric bootstrap, have been performed in simulation study of the data set but both coverages are not favorable when compare with the nominal coverage. In our simulations, estimated values from maximum likelihood estimator (MLE) of these weight parameters have heavily tailed distributions which are not close to normal distributions.

To have more accurate distribution approximation of the estimates, an model transformation is performed. After completion of the transformation, the MLE estimates of these weight parameters, in turn, can be approximately considered as ratios of the two means in some bivariate normal distributions. An usual technique to build these intervals is introduced by Fieller (1932, 1954). In contrast to most other methods, Fieller's interval avoids the distribution approximation of the ratio directly. Instead of it, it uses the distribution character of a new latent variable. This gives widely application area than Hinkley's method (1969) since their approximation needs the probability of positive denominator converge to 1. However, there are several limitations of Fieller's algorithm, which are described detailedly in Appendix A.1. Our simulation results with the real HEI-2005 data set show while Fiellers interval has correct nominal coverage probability in certain cases, it achieves this at of cost of sometimes resulting in confidence intervals of enormous or even infinite length, or even intervals that are the union of disjoint sets. Besides of that, under come circumstances, it is invalid at all.

Consequently, there are many other existing methods in this area and most based on the distribution of the ratio of the estimates of the two location parameters (Geary, 1930; Marsaglia, 1965; Hinkley, 1969; Deaton and Kamerud, 1978; Brody et al., 2002; Cedilnik et al., 2004; Beyene and Moineddin, 2005; Qiao et al., 2006; Pham-Gia et al., 2006; Sherman et al., 2011). Most often, a normal approximation to the distribution is used, with subsequent intervals formed by Wald's method. Hayya et al. (1975) showed that, under certain conditions, the distribution of ratio can be treated as a normal distribution with a second order Taylor expansion. In addition, parametric and nonparametric bootstrap methods are also used.

After we investigating many other existing methods in this area, we came into the conclusion that, for the problem of building a confidence interval for ratio with our data set, except Fieller's interval, the coverages of existing methods all are not sufficiently close to the nominal values. And we have described the shortcomings of the Fieller's interval. Motivated by such a problem, a new methodology named as the *Direct Integral Method for Ratios (DIMER)* is constructed, and details are provided in Chapter 2.

After solving the problem of how to accurately estimate the confidence interval for the ratio, based on the weighted model previously mentioned, I turned to analysis the relative risks for different diseases in different subpopulations. For HEI-2005 data, besides Reedy's work, George, et al. (2010) illustrate that higher HEI-2005 total scores are associated with lower levels of chronic inammation among breast cancer survivors. Chiuve, et al. (2012) show that the HEI-2005 total score and the Alternative Healthy Eating Index (AHEI) are significant predictors of chronic diseases such as coronary heart disease, diabetes, stroke and cancer, and that closer adherence to the 2005 Dietary Guidelines may lower the risk of major chronic diseases. The AHEI is also associated with all cause mortality (Akbaraly, et al., 2011). Additionally, there are some other works related to the HEI-2005 data sets (Fungwe et al., 2009; Kipnis et al., 2009; Kipnis et al., 2009; Kott et al., 2009; Sinha et al., 2010; Tooze et al., 2002; Tooze et al., 2006; Zhang et al., 2011).

As mentioned previously, for the HEI-2005 data, one of our main purposes is to study the relative risk of different diseases in various subpopulations. In this dissertation, I propose six different algorithm to analyze it. Details for the various approaches and asymptotic distributions are provided in Section 3. Additionally, hypothesis tests and likelihood ratio tests are performed to compare our proposed weighted model with the other three, including the one used in Reedy's work.

Furthermore, I propose two different methods, the adaptive lasso (Zou, 2006) and an estimator with bounded constrains, for variable selection upon the weighted regression model. One of the most famous methods to solve the lasso problem is the Least Angle Regression (LARS) which was presented by Efron et al., (2004). An efficient package *lars* in R has been widely used. By the coordinate descent algorithm, Friedman et al., (2007, 2010) proposed solutions for regressions with L_1 norm penalty, which resulting in significant time saving when compared to solutions by LARS. Additionally, Wang and Leng (2007) introduced a method of least squares approximation (LSA) for unified lasso estimation method. The basis of LSA was to approximate a nonlinear regression model with L_1 norm penalty to a least squares minimization problem with the same penalty, while there are numerous efficient solutions for the latter one. Additionally, Wang and Leng (2007) suggested to use the R package *lars* directly after obtaining the approximation least square expression.

Since there are quadratic terms for parameters in our nonlinear regression model, even after the least square approximation, we could not use the *lars* package directly. Therefore, we propose a novel algorithm to solve the nonlinear regression model with L_1 norm penalty and apply it to the real HEI-2005 data set with the weighted model.

Next, in order to study the influences of food intake amount on disease, I expand the weighted model to a logistic regression model containing a nonlinear equation about the diet intake amount. Some constrains from the nutriology and biology are also involved here. Details of these constrains are described in Chapter 5. Combining all these factors, I apply I-spline basis function (Ramsay, 1988) for the nonlinear equation fitting in the logistic regression since it is always monotone increasing and non-negative. In Ramsay's work, the exact expressions of the I-spline basis functions for the second order were provided and I expand them to the third order. And then the application and analysis results in the HEI-2005 data are illustrated.

The arrangement for this dissertation is described as follows: In Chapter 2, I propose a new method named as the *Direct Integral Method for Ratios (DIMER)*, which has been used to calculate the confidence intervals for the two location parameters, and the comparisons to some existing methods have been carried out through simulations. In Chapter 3, the structure of the Healthy Eating Index-2005 (HEI-2005) is firstly described. Then a weighted logistic regression model is built and various methods are proposed to calculate the relative risks for different diseases in HEI-2005 data set, including DIMER. Next, the applications in the nutrition data are illustrated. In order to compare the different models for the relationship between diseases and nutrition diets, I apple Hypothesis test and likelihood ratio test in the Chapter 4 to compare four different logistic regression models, and propose two different methods for variable selection: positive bounded constrains and the adaptive lasso method. Furthermore, I develop a novel algorithm to solve the nonlinear regression model with L_1 norm penalty. Finally in Chapter 5, the weighted logistic model presented in Chapter 3 is expanded to nonlinear equations for nutrition components in a logistic regression, which combines some constrains from nutrition and biology. Results of applications to real HEI-2005 data set are also illustrated. Conclusions are summarized in Chapter 6.

2. THE DIRECT INTEGRAL METHOD FOR CONFIDENCE INTERVALS FOR THE RATIO OF TWO LOCATION PARAMETERS

2.1 Introduction

The work in this Chapter is partially motivated by an analysis of the Healthy Eating Index-2005 (HEI-2005, see Guenther et al., 2008a,b) data set from the NIH-AARP Study of Diet and Health (Reedy et al., 2008). In that study, there are two independent subpopulations for different multiple diseases, and we wish to estimate and form confidence intervals for ratios of their relative risks. As shown in Chapter 3, after a model transformation method, this problem reduces to the well-known problem of computing a confidence interval for the ratio of two location parameters.

As described in Chapter 1, performances of existing methods to form the confidence interval of ratio are not favorable. Motivated by such kind of problem, we develope a new methodology named as the *Direct Integral Method for Ratios (DIMER)*. This methodology is also based on the distribution of the ratio of the estimates of the two location parameters, which we show can be computed easily by numerical integration, in contrast to many other methods, followed by simulation to compute a $100(1-\alpha)\%$ confidence interval. In our simulation studies, we show that DIMER closely achieves nominal coverage, unlike the Wald methods and the method of Hayya et al. (1975). DIMER is also much faster computationally than the bootstrap methods, which is important in examples such as ours, where the model is a nonlinear logistic regression based on samples of huge sizes (in tens of thousands or more).

In Section 2.2 we describe the methodology, while Section 2.3 compares various methods via simulation studies. Simulations based on the actual data reinforce the conclusions of the simulations in Section 2.3. Technical details and additional results are given in the Appendix A.

2.2 Methodology

2.2.1 Outline

Consider two random variables T_1 and T_2 which have density functions $f_1\{(t_1 - u_1)/v_1\}$ and $f_2\{(t_2 - u_2)/v_2\}$, respectively, with means μ_1 and μ_2 and standard deviations v_1 and v_2. In other words, f_1 and f_2 are the density functions of the standardized version T_1 and T_2, respectively. Let $F_1(\cdot)$ and $F_2(\cdot)$ denote the corresponding distribution functions. We are interested in making inference for the ratio μ_1/μ_2. We will outline a series of cases where it is possible to compute easily the cumulative distribution function of $\widehat{r} = T_1/T_2$.

2.2.2 Independent Case

Suppose that T_1 and T_2 are independent. We show the following result in Appendix A.2.

Lemma 1 *Define*

$$g(z|x, \mu_1, \mu_2, v_1, v_2) = \begin{cases} (1 - F_1[\{x(\mu_2 + v_2 z) - \mu_1\}/v_1])f_2(z)exp(z^2) & \text{if } z \leq -\mu_2/v_2, \\ F_1[\{x(\mu_2 + v_2 z) - \mu_1\}/v_1]f_2(z)exp(z^2) & \text{if } z > -\mu_2/v_2. \end{cases}$$

Then the cumulative distribution function of $\widehat{r} = T_1/T_2$ *is given by*

$$pr(\widehat{r} \leq x) = \int_{-\infty}^{\infty} g(z|x, \mu_1, \mu_2, v_1, v_2)\,exp(-z^2)dz,$$

a quantity that is easily computed by Gauss-Hermite quadrature.

Variable x at here is defined as a possible value of $\widehat{r}$, and then we define a partial part inside the integral as $g(z)$ for simplicity, which is a function of x and parameters

(μ_1, μ_2, v_1, v_2). In Sections 2.2.3 and 2.2.4, the definitions for all x and $g(z)$ are similar as here.

<u>Remark</u> 1 If the parameters v_1 and v_2 are unknown, we can apply Lemma 3 using their estimated values. However, we have found that a much more efficient approximation can be developed in the case of normally distributed T_1 and T_2. Suppose their estimated variances are $\widehat{v}_1^2$ and $\widehat{v}_2^2$ which are independent of each other, and independent of T_1 and T_2, and have degrees of freedom d_1 and d_2, respectively. Then, both $(T_1 - \mu_1)/\widehat{v}_1$ and $(T_2 - \mu_2)/\widehat{v}_2$ follow the t-distribution with d_1 and d_2 degrees of freedom, respectively. *As an approximation*, from these t-distributions, we fix $\widehat{v}_1^2$ and $\widehat{v}_2^2$, and by making a change of variables, we get an approximation to the distribution of (T_1, T_2) which better reflects the estimation of $(\widehat{v}_1^2, \widehat{v}_2^2)$. We then apply Lemma 3. Thus, $g(z|x, \mu_1, \mu_2, \widehat{v}_1^2, \widehat{v}_2^2)$ is approximated by

$$g(z|x, \mu_1, \mu_2, \widehat{v}_1^2, \widehat{v}_2^2) \approx \begin{cases} (1 - F_{t,d_1}[\{x(\mu_2 + \widehat{v}_2 z) - \mu_1\}/\widehat{v}_1])f_{t,d_2}(z)\exp(z^2) & \text{if } z \leq -\mu_2/\widehat{v}_2, \\ F_{t,d_1}[\{x(\mu_2 + \widehat{v}_2 z) - \mu_1\}/\widehat{v}_1]f_{t,d_2}(z)\exp(z^2) & \text{if } z > -\mu_2/\widehat{v}_2, \end{cases}$$

where $f_{t,d}(\cdot)$ and $F_{t,d}(\cdot)$ are the t-density with d degrees of freedom and the corresponding cumulative distribution function, respectively.

2.2.3 Dependent Case of Two Normally Distributed Variables with Known Covariance Matrix

Suppose now that (T_1, T_2) are jointly normally distributed with means (μ_1, μ_2), variances (v_1^2, v_2^2), covariance v_{12} and that (v_1^2, v_2^2, v_{12}) are known. Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and distribution function. We show the following result in Appendix A.2.

Lemma 2 *Define* $g(z|x, \mu_1, \mu_2, v_1^2, v_2^2, v_{12})$ *as follows. If* $z \leq -\mu_2/v_2$, *then*

$$g(z|x, \mu_1, \mu_2, v_1^2, v_2^2, v_{12}) = (2\pi)^{-1/2}(1 - \Phi[\{x(\mu_2 + v_2 z) \\ -(\mu_1 + zv_{12}/v_2)\}v_2/\sqrt{v_1^2 v_2^2 - v_{12}^2}]) \exp(z^2/2).$$

If $z > -\mu_2/v_2$, *then*

$$g(z|x, \mu_1, \mu_2, v_1^2, v_2^2, v_{12}) = (2\pi)^{-1/2}\Phi[\{x(\mu_2 + v_2 z) \\ -(\mu_1 + zv_{12}/v_2)\}v_2/\sqrt{v_1^2 v_2^2 - v_{12}^2}] \exp(z^2/2).$$

Then the distribution function of $\widehat{r}$ *is*

$$pr(\widehat{r} \leq x) = \int_{-\infty}^{\infty} g(z|x, \mu_1, \mu_2, v_1^2, v_2^2, v_{12}) exp(-z^2) dz,$$

which again can be computed by Gauss-Hermite quadrature.

Of course, when $v_{12} = 0$, Lemma 2 is a special case of Lemma 1.

2.2.4 Dependent Case of Two Normally Distributed Variables with Estimated Covariance Matrix

In this section, we discuss the cumulative distribution of the ratio $\widehat{r} = T_1/T_2$ when T_1 and T_2 are jointly normally distributed with jointly estimated variance and covariance which have the same number of degrees of freedom d, and these estimates are independent of T_1 and T_2. These are the same assumptions noted in Fieller (1954). Define the estimates of the variances and covariance of T_1 and T_2 as $\widehat{v}_1^2, \widehat{v}_2^2$ and $\widehat{v}_{12}$. Let $\eta = v_{12}/v_2^2$. For fixed η, write $W = T_1 - \eta T_2$, Then W and T_2 are independent. In addition, if $\widehat{v}_1^2, \widehat{v}_2^2$ and $\widehat{v}_{12}$ are computed from the sample covariance matrix of normal random variables from a sample of size $d + 1$, then we also have that $T_1 - \eta T_2$ and T_2 are independent of their estimated variances $\widehat{v}_1^2 - 2\eta\widehat{v}_{12} + \eta^2\widehat{v}_2^2$ and $\widehat{v}^2$, which are independent of each other and also have d degrees of freedom.

We use the following algorithm, based on the approximation used in Section 2.2.2. Under our assumptions, the variables $Z_1 = \{(T_1 - \eta T_2) - (\mu_1 - \eta\mu_2)\}/\sqrt{\widehat{v}_1^2 - 2\eta\widehat{v}_{12} + \eta^2\widehat{v}_2^2}$ and $Z_2 = (T_2 - \mu_2)/\widehat{v}_2$ are independent and both have t-distributions with d degrees of freedom As in Remark 1, we then make the approximation that the density of (T_1, T_2), having fixed the estimated covariance matrix, is approximately

$$\widehat{v}_2^{-1}(\widehat{v}_1^2 - 2\eta\widehat{v}_{12} + \eta^2\widehat{v}_2^2)^{-1/2} f_{t,d}[\{(t_1 - \eta t_2) - (\mu_1 - \eta\mu_2)\}/\sqrt{\widehat{v}_1^2 - 2\eta\widehat{v}_{12} + \eta^2\widehat{v}_2^2}] f_{t,d}\{(t_2 - \mu_2)/\widehat{v}_2\}.$$

If $z \leq -\mu_2/\widehat{v}_2$, define

$$g(z|x, \mu_1, \mu_2, \widehat{v}_1^2, \widehat{v}_2^2, \widehat{v}_{12}, \eta)$$
$$= \left(1 - F_{t,d}\left[\{(x - \eta)(\mu_2 + v_2 z) - (\mu_1 - \eta\mu_2)\}/\sqrt{\widehat{v}_1^2 - 2\eta\widehat{v}_{12} + \eta^2\widehat{v}_2^2}\right]\right) f_{t,d}(z)\exp(z^2),$$

while if $z > -\mu_2/\widehat{v}_2$, define

$$g(z|x, \mu_1, \mu_2, \widehat{v}_1^2, \widehat{v}_2^2, \widehat{v}_{12}, \eta)$$
$$= F_{t,d}\left[\{(x - \eta)(\mu_2 + v_2 z) - (\mu_1 - \eta\mu_2)\}/\sqrt{\widehat{v}_1^2 - 2\eta\widehat{v}_{12} + \eta^2\widehat{v}_2^2}\right] f_{t,d}(z)\exp(z^2).$$

Then, using the same devise as in Remark 1 we have that

$$\text{pr}(\widehat{r} \leq x) \approx \int_{-\infty}^{\infty} g(z|x, \mu_1, \mu_2, \widehat{v}_1^2, \widehat{v}_2^2, \widehat{v}_{12}, \eta)\exp(-z^2)dz. \tag{2.1}$$

In practice, η is unknown, so we use $\widehat{\eta} = \widehat{v}_{12}/\widehat{v}_2^2$ to estimate it.

2.2.5 Algorithm for Computing the Confidence Interval of Ratios

In the cases in Sections 2.2.2-2.2.4, the distribution function of $\widehat{r}$ is expressed as $F(x; r) = \text{pr}(\widehat{r} \leq x; r = \mu_1/\mu_2)$ when $\mu_2 \neq 0$. The ratio $\widehat{\mu}_1/\widehat{\mu}_2$ is an estimate of $r = \mu_1/\mu_2$, so that we can view $F(x; \widehat{\mu}_1/\widehat{\mu}_2)$ as an estimate of the population distribution function $F(x; r)$. Efron (1981) and Benton and Krishnamoorthy (2002) pointed out that if we generate values $\widehat{r}_i$, $i = 1, ..., m$, from $F(x; \widehat{\mu}_1/\widehat{\mu}_2)$, we can make inference about the parameter r using the distribution of the generated $\widehat{r}_i$'s.

The main difference between our approach and that of Benton and Krishnamoorthy is that instead of generating a larger number of $\widehat{r}_i$'s and then obtaining its percentiles, we compute the percentile of $\widehat{r}_i$ directly. Consequently, our method is much faster computationally. Specifically, our simulation results indicate that DIMER usually needs less than 30 iteration steps to obtain the quantile of a distribution, but in Benton and Krishnamoorthy (2002), they used $m = 100,000$ $\widehat{r}_i$'s to get the quantiles.

Define the $\alpha/2$ quantile for $F(x; \widehat{\mu}_1/\widehat{\mu}_2)$ as $\widehat{r}_{\alpha/2|\widehat{\mu}_1/\widehat{\mu}_2}$. Then an approximate $100(1-\alpha)\%$ confidence interval for r is $(\widehat{r}_{\alpha/2|\widehat{\mu}_1/\widehat{\mu}_2}, \widehat{r}_{1-\alpha/2|\widehat{\mu}_1/\widehat{\mu}_2})$. Here we give the steps of our iterative algorithm to obtain the quantiles.

- Step 1. Give two initial values of $\widehat{r}_{\alpha/2|\widehat{\mu}_1/\widehat{\mu}_2}$ as $\widehat{r}_{\alpha_1} < 0 < \widehat{r}_{\alpha_2}$ and both have sufficiently large absolute values to make sure that $\widehat{r}_{\alpha/2|\widehat{\mu}_1/\widehat{\mu}_2}$ is inside the interval $(\widehat{r}_{\alpha_1}, \widehat{r}_{\alpha_2})$.

- Step 2. Apply Gauss-Hermit quadrature to the cumulative distribution function of $\widehat{r}$ to obtain $c_{\alpha/2} = \text{pr}\{\widehat{r} \leq (\widehat{r}_{\alpha_1} + \widehat{r}_{\alpha_2})/2\}$. If $c_{\alpha/2} < \alpha/2$, let $\widehat{r}_{\alpha_1} = (\widehat{r}_{\alpha_1} + \widehat{r}_{\alpha_2})/2$; if $c_{\alpha/2} > \alpha/2$, let $\widehat{r}_{\alpha_2} = (\widehat{r}_{\alpha_1} + \widehat{r}_{\alpha_2})/2$; if $c_{\alpha/2} = \alpha/2$, stop the iteration and let $\widehat{r}_{\alpha/2|\widehat{\mu}_1/\widehat{\mu}_2} = (\widehat{r}_{\alpha_1} + \widehat{r}_{\alpha_2})/2$.

- Step 3. Repeat Step 2 until $c_{\alpha/2}$ is close to $\alpha/2$ and/or the difference $|\widehat{r}_{\alpha_2} - \widehat{r}_{\alpha_1}|$ is sufficiently small. Then we have $\widehat{r}_{\alpha/2|\widehat{\mu}_1/\widehat{\mu}_2} = (\widehat{r}_{\alpha_1} + \widehat{r}_{\alpha_2})/2$.

- Step 4. Repeat Steps 1–3 to obtain $\widehat{r}_{1-\alpha/2|\widehat{\mu}_1/\widehat{\mu}_2}$.

In summary, two different original points are given for the estimate of $r_{\alpha/2|\widehat{\mu}_1/\widehat{\mu}_2}$, then with repeating Steps (1∼2), the distance between these two points gradually becomes smaller and smaller until converges to a single point, which is our expected result. After obtaining the lower limit $r_{\alpha/2|\widehat{\mu}_1/\widehat{\mu}_2}$ of the confidence interval, then we repeat steps 1∼3 to obtain the upper limit estimate $\widehat{r}_{1-\alpha/2|\widehat{\mu}_1/\widehat{\mu}_2}$. Furthermore, since this is a bisection method, it is absolutely non-sensitive to the starting values, and the true value of $r_{\alpha/2|\widehat{\mu}_1/\widehat{\mu}_2}$ or $r_{1-\alpha/2|\widehat{\mu}_1/\widehat{\mu}_2}$ is certainly be included as long as the range between the two original points are set large enough.

2.3 Simulations

In this section, we report simulation results on two simple linear regression models. The first part (Section 2.3.1) is to illustrate an application of the formulas in Section 2.2.2 where the two variables are independent. The second part (Section 2.3.2) is an example to demonstrate the performance of our method developed in Section 2.2.4 when the two variables are dependent with estimated variance and covariance which are independent of the two variables and have same degrees of freedom. In both simulations, some other possible methods are outlined and compared with our method. Since dependence case relies on normality assumption, it would be important to evaluate how DIMER would perform if such an assumption is violated. Therefore, in the second simulation, we also consider a case when T_1 and T_2 do not have normal distributions. Furthermore, the confidence intervals' coverage, which is corresponding to the hypothesis test by the likelihood ratio test, is also compared at here. More details of the simulations are available in the Appendix A.

2.3.1 Linear Model When the Two Estimates are Independent

2.3.1.1 Setup

Consider a linear regression model as,

$$Y_{1i} = \beta_{10} + X_{1i}\beta_{11} + \varepsilon_{1i}, i = 1, ..., n_1;$$
$$Y_{2j} = \beta_{20} + X_{2j}\beta_{21} + \varepsilon_{2j}, j = 1, ..., n_2,$$

where in group 1, Y_{1i} denotes i^{th} response and X_{1i} denotes the i^{th} predictor; in group 2, Y_{2j} denotes the j^{th} response and X_{2j} denotes the j^{th} predictor. And ε_{1i} and ε_{2j} are independently normally distributed with mean zero and variance v_1^2 and v_2^2, respectively. Our interest is in the ratio of the two slopes, which is β_{21}/β_{11}.

The model could be rewritten as follows for simple expression of the ratio

$$\begin{aligned} Y_{1i} &= \beta_{10} + \beta_{11}X_{1i}\omega + \varepsilon_{1i}, i = 1, ..., n_1; \\ Y_{2j} &= \beta_{20} + \beta_{21}X_{2j}\omega + \varepsilon_{2j}, j = 1, ..., n_2, \end{aligned} \tag{2.2}$$

where we set $\beta_{11} = 1$ for identifiability, and then the ratio of slopes now is β_{21} and where ω could be considered as the slope in the regression model for the first group data or the interaction between two slopes when β_{11} set to 1.

Now our interest is to construct a confidence interval for β_{21}. The loglikelihood function of the data is

$$\begin{aligned} \mathcal{L} \propto & -n_1\log(v_1) - (2v_1^2)^{-1/2}\textstyle\sum_{i=1}^{n_1}(Y_{1i} - \beta_{10} + X_{1i}\omega)^2 \\ & -n_2\log(v_2) - (2v_2^2)^{-1/2}\textstyle\sum_{j=1}^{n_2}(Y_{2j} - \beta_{20} - \beta_{21}X_{2j}\omega)^2. \end{aligned}$$

The maximum likelihood estimates are

$$\begin{aligned} \widehat{\omega} &= -\{\textstyle\sum_{i=1}^{n_1} Y_{1i}(X_{1i} - \overline{X_1})\}/\{(\textstyle\sum_{i=1}^{n_1} X_{1i}^2 - n_1\overline{X_1}^2)\}, \\ \widehat{\beta}_{21} &= \{\textstyle\sum_{j=1}^{n_2} Y_{2j}(X_{2j} - \overline{X_2})\}/\{\widehat{\omega}(\textstyle\sum_{j=1}^{n_2} X_{2j}^2 - n_2\overline{X_2}^2)\}. \end{aligned}$$

First, define $\lambda = \beta_{21}\omega$ and its estimate $\widehat{\lambda} = \{\sum_{j=1}^{n_2} Y_{2j}(X_{2j} - \overline{X_2})\}/\{(\sum_{j=1}^{n_2} X_{2j}^2 - n_2\overline{X_2}^2)\}$. Both $(\widehat{\lambda} - \lambda)/\widehat{v}_\lambda$ and $(\widehat{\omega} - \omega)/\widehat{v}_\omega$ follow independent standard t distributions with degrees of freedom $n_2 - 2$ and $n_1 - 2$, respectively, where $\widehat{v}_\lambda^2 = \{(n_2 - 2)^{-1}\sum_{j=1}^{n_2}(Y_{2j} - \widehat{\beta}_{20} - X_{2j}\widehat{\lambda})^2\}/(\sum_{j=1}^{n_2} X_{2j}^2 - n_2\overline{X_2}^2)$ and $\widehat{v}_\omega^2 = \{(n_1 - 2)^{-1}\sum_{i=1}^{n_1}(Y_{1i} - \widehat{\beta}_{10} + X_{1i}\widehat{\omega})^2\}/(\sum_{i=1}^{n_1} X_{1i}^2 - n_1\overline{X_1}^2)$.

By the development in Section 2.2.2, the estimated cumulative distribution function of $\widehat{\beta}_{21}$ is

$$\mathrm{pr}(\widehat{\beta}_{21} \leq x) \approx \int_{-\infty}^{\infty} g(z|x, \omega, \lambda, \widehat{v}_\lambda^2, \widehat{v}_\omega^2)\exp(-z^2)dz,$$

where

$$g(z|x, \omega, \lambda, \widehat{v}_\lambda^2, \widehat{v}_\omega^2) = \begin{cases} (1 - F_{t,n_2-2}[\{x(\omega + \widehat{\sigma}_\omega z) - \lambda\}/\widehat{\sigma}_\lambda])f_{t,n_1-2}(z)\exp(z^2) & \text{if } z \leq -\omega/\widehat{\sigma}_\omega, \\ F_{t,n_2-2}[\{x(\omega + \widehat{\sigma}_\omega z) - \lambda\}/\widehat{\sigma}_\lambda]f_{t,n_1-2}(z)\exp(z^2) & \text{if } z > -\omega/\widehat{\sigma}_\omega. \end{cases}$$

Applying the algorithm in Section 2.2.5, we obtain a confidence interval by DIMER. To compare DIMER with other possible methods, in Section 2.3.1.2, we outline an application of the Wald interval by inverting the Fisher score matrix, Fieller's interval and Hayya's method in our linear regression model.

2.3.1.2 Comparison with Other Possible Methods

To form a confidence interval for $\widehat{\beta}_{21}$, one common method in practice for estimating the variance of the estimates is the inverse Fisher score information matrix, which is estimated as

$$\begin{pmatrix} n_1/\widehat{v}_1^2 & 0 & 0 & -\sum_{i=1}^{n_1} X_{1i}/\widehat{v}_1^2 \\ 0 & n_2/\widehat{v}_2^2 & \widehat{\omega}\sum_{j=1}^{J} X_{2j}/\widehat{v}_2^2 & \widehat{\beta}_{21}\sum_{j=1}^{J} X_{2j}/\widehat{v}_2^2 \\ 0 & \widehat{\omega}\sum_{j=1}^{J} X_{2j}/\widehat{v}_2^2 & \widehat{\omega}^2\sum_{j=1}^{J} X_{2j}^2/\widehat{v}_2^2 & \widehat{\omega}\widehat{\beta}_{21}\sum_{j=1}^{J} X_{2j}^2/\widehat{v}_2^2 \\ -\sum_{i=1}^{n_1} X_{1i}/\widehat{v}_1^2 & \widehat{\beta}_{21}\sum_{j=1}^{J} X_{2j}^2/\widehat{v}_2^2 & \widehat{\omega}\widehat{\beta}_{21}\sum_{j=1}^{J} X_{2j}^2/\widehat{v}_2^2 & \sum_{i=1}^{n_1} X_{1i}^2/\widehat{v}_1^2 + \widehat{\beta}_{21}^2\sum_{j=1}^{J} X_{2j}^2/\widehat{v}_2^2 \end{pmatrix}.$$

Denote the standard error of $\widehat{\beta}_{21}$ by this method as $se_{\widehat{\beta}_{12},\text{Fisher}}$, so that a $(1-\alpha)100\%$ confidence interval for β_{21} is $(\widehat{\beta}_{21} - z_{\alpha/2} se_{\widehat{\beta}_{12},\text{Fisher}}, \widehat{\beta}_{21} + z_{\alpha/2} se_{\widehat{\beta}_{12},\text{Fisher}})$, where $z_{\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution.

In this linear regression setting, Fieller's interval cannot be applied directly since $\widehat{v}^2_{\widehat{\omega}}$ and $\widehat{v}^2_{\widehat{\lambda}}$ are estimated independently. In this case, by the Welch-Satterthwaite equation (Satterthwaite, 1946; Welch, 1947), the degrees of freedom of $(\widehat{v}^2_{\widehat{\lambda}} + \beta_{21}^2 \widehat{v}^2_{\widehat{\omega}})$ is approximately given by $d_F = (\widehat{v}^2_{\widehat{\lambda}} + \beta_{21}^2 \widehat{v}^2_{\widehat{\omega}})^2 / \{(\widehat{v}^2_{\widehat{\lambda}})^2/(n_2-2) + (\beta_{21}^2 \widehat{v}^2_{\widehat{\omega}})^2/(n_1-2)\}$. We may use $\widehat{\beta}_{21}$ instead of β_{21} in the expression to obtain the estimated degrees of freedom $d_F^* = (\widehat{v}^2_{\widehat{\lambda}} + \widehat{\beta}_{21}^2 \widehat{v}^2_{\widehat{\omega}})^2 / \{(\widehat{v}^2_{\widehat{\lambda}})^2/(n_2-2) + (\widehat{\beta}_{21}^2 \widehat{v}^2_{\widehat{\omega}})^2/(n_1-2)\}$. Then we have $a = \widehat{\omega}^2 - t^2_{d_F^*,\alpha/2}\widehat{\sigma}^2_{\widehat{\omega}}, b = -2\widehat{\omega}\widehat{\lambda}$ and $c = \widehat{\lambda}^2 - t^2_{d_F^*,\alpha/2}\widehat{\sigma}^2_{\widehat{\lambda}}$ used in Appendix A.1. Here $\rho = 0$ since $\widehat{\omega}$ and $\widehat{\lambda}$ are independent.

Remark 2 Fieller's interval has a peculiarity in that sometimes it leads to an imaginary interval. Looking at the detailed description of Fieller's method given in Appendix A.1, we see that if $b^2 - 4ac < 0$, then there is no real solution to Fieller's method. Since this actually occurs in our simulations, we will say that when it does, Fieller's interval is “*invalid*”.

Another method was proposed by Hayya et al. (1975) in a not very well-known article. They suggested a normal approximation to the true cumulative distribution function of the ratio $\widehat{r} = T_1/T_2$ obtained by a second order Taylor expansion. By Monte Carlo simulations, they concluded that if the absolute value of the correlation between T_1 and T_2 is less or equal to 0.5, the coefficient of variation of T_2 is less or equal to 0.09 and the coefficient of variation of T_1 is larger than 0.19, the ratio $\widehat{r} = T_1/T_2$ is approximately normally distributed with

$$\begin{aligned} E(\widehat{r}) &\approx (\mu_{T_1}/\mu_{T_2}) + v_{T_2}^2\mu_{T_1}/\mu_{T_2}^3 - \rho v_{T_2}v_{T_1}/\mu_{T_2}^2, \\ \mathrm{var}(\widehat{r}) &\approx v_{T_2}^2\mu_{T_1}^2/\mu_{T_2}^4 + v_{T_1}^2/\mu_{T_2}^2 - 2\rho v_{T_2}v_{T_1}\mu_{T_1}/\mu_{T_2}^3, \end{aligned}$$

where ρ is the correlation between T_1 and T_2 and $\widehat{r}$ is corresponding to $\widehat{\beta}_{21}$ in the model (2.2).

In our context, $\rho = 0$ since $\widehat{\omega}$ and $\widehat{\lambda}$ are independent. The conditions of Hayya et al. (1975) thus reduce to only two: $\mathrm{cv}(\widehat{\omega}) \leq 0.09$ and $\mathrm{cv}(\widehat{\lambda}) > 0.19$. This can be thought of as

$$\widehat{v}_1/[\widehat{\omega}\sqrt{\{n_1(\textstyle\sum_{i=1}^{n_\ell}X_{1i}^2 - n_1\overline{X_1}^2)\}}] \leq 0.09,\ \ \widehat{v}_2/[\widehat{\beta}_{21}\widehat{\omega}\sqrt{\{n_2(\textstyle\sum_{j=1}^{J}X_{2j}^2 - n_2\overline{X_2}^2)\}}] > 0.19,$$

where $\widehat{v}_1^2 = (n_1-2)^{-1}\sum_{i=1}^{n_1}(Y_{1i} - \widehat{\beta}_{10} + X_{1i}\widehat{\omega})^2$ and $\widehat{v}_2^2 = (n_2-2)^{-1}\sum_{j=1}^{n_2}(Y_{2j} - \widehat{\beta}_{20} - X_{2j}\widehat{\lambda})^2$.

Assuming that the two conditions are satisfied, the distribution of $\widehat{\beta}_{21}$ can be approximated as a normal distribution with mean and variance

$$\begin{aligned} \widehat{\mu}_{\widehat{\beta}_{21},\mathrm{Hayya}} &\approx [1 + \widehat{v}_1^2/\{\widehat{\omega}^2(\textstyle\sum_{i=1}^{n_\ell}X_{1i}^2 - n_1\overline{X_1}^2)\}]\widehat{\beta}_{21}, \\ \widehat{\sigma}^2_{\widehat{\beta}_{21},\mathrm{Hayya}} &\approx \widehat{v}_2^2/\{\widehat{\omega}^2(\textstyle\sum_{j=1}^{J}X_{2j}^2 - n_2\overline{X_2}^2)\} + \widehat{v}_1^2\widehat{\beta}_{21}^2/\{\widehat{\omega}^2(\textstyle\sum_{i=1}^{n_\ell}X_{1i}^2 - n_1\overline{X_1}^2)\}. \end{aligned}$$

A confidence interval with coverage probability $1-\alpha$ is constructed as $\widehat{\mu}_{\widehat{\beta}_{21},\text{Hayya}} \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{\beta}_{21},\text{Hayya}}$. In addition, we have applied the nonparametric bootstrap and the parametric bootstrap; see the details in Appendix A.4.

2.3.1.3 Simulation Results

We conducted simulation studies to assess the performance of the six algorithms in the linear regression model (2.2): the inverse Fisher score, Hayya's method, the nonparametric bootstrap, the parametric bootstrap, Fieller's interval and our proposed DIMER. For simplicity, in all settings, we fixed the variance of ε_{1i} and ε_{2j} to be 1, and without loss of generality, let the intercepts β_{10} and β_{20} be 0. We generated X_{1i} and X_{2j} independently from the standard normal distribution.

We considered two parameter configurations: $(\beta_{10}, \beta_{20}, \beta_{21}, \omega)$=$(0, 0, 1, 1)$ and $(0, 0, 1, 0.75)$. For each parameter setting, we performed simulations for $(n_1, n_2) = (18, 18)$, $(25, 25)$, $(50, 50)$. In each case, we generated 2000 data sets. Depends on Efron and Tibshirani (1994), $B = 400$ was applied as the number of replications for both nonparametric bootstrap and parametric bootstrap methods, and for the rest part of this article, all bootstrap computations were adopt this value for B.

The results for the first parameter configuration $(\beta_{10}, \beta_{20}, \beta_{21}, \omega) = (0, 0, -1, -1)$ with $(n_1, n_2) = (18, 18)$, $(25, 25)$, $(50, 50)$ are given in Table 2.1. Table 2.2 presents the results for setting $(\beta_{10}, \beta_{20}, \beta_{21}, \omega) = (0, 0, 1, 0.75)$ with $(n_1, n_2) = (18, 18)$, $(25, 25)$, $(50, 50)$. QQ plots (not shown here) comparing the quantiles of $\widehat{\beta}_{21}$ to the quantiles of the standard normal distribution in the two parameter configurations with $n_1 = n_2 = 18$ clearly show that for small to moderate sample sizes, normal approximations are not appropriate.

In Table 2.2, when $n_1 = n_2 = 18$, the averaged estimation for β_{21} is -2.85, while the true value is -1.00. The reason for this difference is because beta follows a

Cauchy likely distribution, and one of characteristics for this distribution is that it has severely heavy tails. For example, the maximum estimation for the absolute value of $\widehat{\beta}_{21}$ has reached 3138 in this case, compared to its true value 1.00. Therefore, the outlier is dramatically large. But even in such circumstance, the median estimation for it is still 1.00, which is the same as the true value.

The inverse Fisher information matrix algorithm has the lowest coverage probabilities. Hayya's method has behavior somewhat intermediate between the inverse Fisher score and the other methods, and it also has very low coverage probabilities when the sample sizes are small.

The performance of two bootstrap methods is acceptable when the sample sizes are relatively large. When the sample sizes are small to moderate, the coverage rate of the bootstrap methods for the 90% confidence intervals are higher than the nominal coverage probability but the coverage rate of the 99% confidence intervals are lower than the nominal values.

Fieller's interval has good performance overall in coverage. However, when the sample sizes are small and moderate ($n_1 = n_2 = 18$ and $n_1 = n_2 = 25$), Fieller's interval can be invalid in the sense described in Remark 2. Even if it is valid, it also has substantial probability to produce infinite confidence interval lengths. The inverse Fisher information method produced the shortest confidence interval lengths, but it is not a good method to apply here since the coverage rates are far below the nominal values. Hayya's method remains stable but has a low coverage when the sample sizes are small. Compared with the two bootstrap methods, our method obviously has markedly shorter lengths in the 90% and 95% confidence intervals when the sample sizes are small and moderate, especially when $(n_1, n_2) = (18, 18)$.

	Mean of Coverage			Mean of Length			Median of Length			90% Quantile of Length		
Method	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI
$n_1 = n_2 = 18, \mathrm{cv}(\widehat{\omega}) = 0.260, \mathrm{cv}(\widehat{\lambda}) = 0.265.$												
$\mathrm{mean}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.01, 0.01, 1.10, 1.00), \mathrm{median}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.01, 0.01, 1.00, 1.00)$												
IF	84.05	89.40	94.60	1.63	1.95	2.56	1.09	1.30	1.71	2.83	3.38	4.44
HM	88.50	92.90	96.70	1.74	2.08	2.73	1.15	1.37	1.80	2.31	2.75	3.61
NB	92.15	94.50	97.75	20.66	24.62	32.35	1.67	1.98	2.61	31.39	37.40	49.15
PB	92.00	94.20	97.35	38.84	46.28	60.83	1.49	1.78	2.34	22.75	27.10	35.62
FI	89.85	95.05	99.35	∞	∞	∞	1.39	1.80	3.08	4.28	8.25	∞
DIMER	91.45	95.90	99.50	2.69	4.92	63.53	1.43	1.88	3.35	3.74	6.12	37.32
$b^2 - 4ac < 0$	0.00	0.05	0.45									
$a < 0$	2.90	5.65	14.47									
$n_1 = n_2 = 25, (\widehat{\beta}_{10}, \mathrm{cv}(\widehat{\omega}) = 0.211, \mathrm{cv}(\widehat{\lambda}) = 0.210.$												
$\mathrm{mean}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.00, 1.05, 1.00), \mathrm{median}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.00, 1.00, 1.00).$												
IF	86.15	92.15	96.50	1.35	1.60	2.11	0.95	1.13	1.49	2.17	2.59	3.41
HM	90.15	94.75	98.20	1.12	1.33	1.75	0.97	1.16	1.52	1.65	1.97	2.59
NB	92.55	95.30	98.45	10.25	12.21	16.05	1.17	1.40	1.83	7.55	8.99	11.82
PB	92.55	95.45	98.40	7.23	8.61	11.31	1.13	1.35	1.77	3.84	4.57	6.01
FI	90.15	95.90	99.60	∞	∞	∞	1.10	1.38	2.12	2.17	3.02	6.92
DIMER	91.15	96.30	99.70	1.78	2.75	10.96	1.12	1.42	2.23	2.20	3.06	6.74
$b^2 - 4ac < 0$	0.00	0.00	0.05									
$a < 0$	0.50	1.00	4.45									
$n_1 = n_2 = 50, \mathrm{cv}(\widehat{\omega}) = 0.144, \mathrm{cv}(\widehat{\lambda}) = 0.148.$												
$\mathrm{mean}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.00, 1.02, 1.00), \mathrm{median}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.00, 1.00, 1.00).$												
IF	90.00	93.10	97.25	0.94	1.12	1.47	0.67	0.79	1.04	1.38	1.64	2.15
HM	90.65	95.50	98.65	0.71	0.84	1.11	0.67	0.79	1.04	0.95	1.13	1.48
NB	92.15	95.40	98.55	0.84	1.00	1.32	0.70	0.84	1.10	1.10	1.31	1.72
PB	92.15	96.10	98.65	0.80	0.95	1.25	0.71	0.84	1.11	1.09	1.29	1.70
FI	91.20	95.75	99.00	0.76	0.93	1.35	0.70	0.86	1.19	1.04	1.29	1.90
DIMER	91.50	95.80	99.10	0.77	0.94	1.36	0.71	0.87	1.21	1.05	1.31	1.94
$b^2 - 4ac < 0$	0.00	0.00	0.00									
$a < 0$	0.00	0.00	0.00									

Table 2.1: Confidence intervals for β_{21} in a simulation study with 2000 replications and true parameter values $(\beta_{10}, \beta_{20}, \beta_{21}, \omega) = (0.00, 0.00, 1.00, 1.00)$ for the linear regression model $Y_{1i} = \beta_{10} - X_{1i}\omega + \varepsilon_{1i}; Y_{2j} = \beta_{20} + \beta_{21}X_{2j}\omega + \varepsilon_{2j}$. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval and DIMER–Direct Integral Method for Ratios.

	Mean of Coverage			Mean of Length			Median of Length			90% Quantile of Length		
Method	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI
$n_1 = n_2 = 18, (\widehat{\beta}_{10}, \mathrm{cv}(\widehat{\omega}) = 0.346, \mathrm{cv}(\widehat{\lambda}) = 0.353.$												
$\mathrm{mean}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.01, 0.01, 2.85, 0.75), \mathrm{median}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.01, 0.01, 1.00, 0.75).$												
IF	83.60	88.60	94.15	4.51	5.38	7.07	1.46	1.74	2.29	4.59	5.47	7.19
HM	86.45	91.65	95.55	2974.62	3544.48	4658.24	1.54	1.83	2.41	4.06	4.84	6.36
NB	93.35	95.10	97.75	74.05	88.24	115.97	4.54	5.41	7.11	105.09	125.22	164.57
PB	93.05	94.55	97.35	1634.42	1947.53	2559.48	3.55	4.23	5.56	94.25	112.31	147.60
FI	91.44	96.04	99.35	∞	∞	∞	2.13	2.97	7.75	∞	∞	∞
DIMER	92.80	96.55	99.55	7.57	15.87	56.16	2.15	3.05	8.28	10.03	25.88	105.14
$b^2 - 4ac < 0$	0.65	1.60	7.05									
$a < 0$	11.17	16.92	34.64									
$n_1 = n_2 = 25, \mathrm{cv}(\widehat{\omega}) = 0.281, \mathrm{cv}(\widehat{\lambda}) = 0.280.$												
$\mathrm{mean}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.00, 1.17, 0.75), \mathrm{median}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.00, 1.00, 0.75).$												
IF	85.65	91.40	96.30	2.02	2.40	3.16	1.27	1.51	1.99	3.20	3.81	5.01
HM	89.45	94.15	97.75	5.06	6.03	7.92	1.30	1.55	2.03	2.69	3.20	4.21
NB	93.40	95.55	98.25	42.36	50.47	66.33	2.07	2.47	3.24	45.18	53.84	70.75
PB	93.10	95.50	98.30	53.39	63.62	83.61	1.91	2.28	2.99	35.70	42.54	55.90
FI	91.05	96.49	99.65	∞	∞	∞	1.59	2.11	3.90	5.72	15.03	∞
DIMER	92.40	96.95	99.75	4.53	9.82	35.54	1.62	2.16	4.15	4.64	7.96	61.76
$b^2 - 4ac < 0$	0.05	0.15	1.10									
$a < 0$	4.20	6.96	19.26									
$n_1 = n_2 = 50, \mathrm{cv}(\widehat{\omega}) = 0.192, \mathrm{cv}(\widehat{\lambda}) = 0.197.$												
$\mathrm{mean}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.00, 1.04, 0.75), \mathrm{median}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.01, 1.00, 0.75)$												
IF	89.20	93.00	97.25	1.19	1.41	1.86	0.89	1.07	1.40	2.00	2.38	3.13
HM	90.80	95.30	98.45	0.99	1.17	1.54	0.89	1.06	1.39	1.43	1.70	2.23
NB	93.00	95.60	98.35	2.57	3.06	4.02	1.01	1.20	1.58	2.33	2.77	3.64
PB	92.75	96.10	98.65	3.79	4.52	5.93	1.00	1.19	1.57	2.19	2.61	3.43
FI	91.30	95.80	99.10	∞	∞	∞	0.97	1.21	1.77	1.73	2.28	4.13
DIMER	91.55	96.05	99.10	1.16	1.52	3.70	0.98	1.22	1.81	1.75	2.31	4.23
$b^2 - 4ac < 0$	0.00	0.00	0.00									
$a < 0$	0.05	0.25	1.30									

Table 2.2: Confidence intervals for β_{21} in a simulation study with 2000 replications and true parameter values $(\beta_{10}, \beta_{20}, \beta_{21}, \omega) = (0.00, 0.00, 1.00, 0.75)$ for the linear regression model $Y_{1i} = \beta_{10} - X_{1i}\omega + \varepsilon_{1i}; Y_{2j} = \beta_{20} + \beta_{21}X_{2j}\omega + \varepsilon_{2j}$. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval and DIMER–Direct Integral Method for Ratios.

This is true whether length is measured by mean length, medial length, the interquartile range of length, or the 90^{th} percentile of length, the interquartile range of length shown in Tables A.1 and A.2 in the Appendix A. In the length comparison for the mean, median, interquartile range and 90% quantile of the 99% confidence intervals, the results from our method are occasionally higher than those of the nonparametric bootstrap and parametric bootstrap, because the interval coverage rates of the latter two methods are somewhat lower than the nominal coverage probability. When the sample sizes are small, DIMER and Fieller's interval have similar median and interquartile ranges of lengths, but our method is much shorter in terms of mean length and the 90^{th} percentile of length.

2.3.2 Linear Model with Two Dependent Estimates

2.3.2.1 Setup

Consider another simple linear regression model:

$$Y_i = \beta(X_i - \mu) + \epsilon_i, i = 1, ..., n,$$

where the parameter of interest μ is -1 multiplied by the ratio between the intercept $-\beta\mu$ and the slope β, and ϵ_i is independent and identically normally distributed with mean zero. If one wants to obtain the confidence interval for the intercept/slope ratio, they can simply calculate the inverse value for limits the β's confidence intervals and multiply by -1.

Let $\lambda = \beta\mu$ and define $\mathbf{X}_{\text{new}} = (-\mathbf{1}, \mathbf{X})$, where $X = (X_1, \ldots, X_n)^{\text{T}}$. Then the maximum likelihood estimates are $(\widehat{\lambda}, \widehat{\beta})^{\text{T}} = (\mathbf{X}_{\text{new}}^{\text{T}}\mathbf{X}_{\text{new}})^{-}\mathbf{X}_{\text{new}}^{\text{T}}\mathbf{Y}$, where $(\mathbf{X}_{\text{new}}^{\text{T}}\mathbf{X}_{\text{new}})^{-}$ is a generalized inverse of $\mathbf{X}_{\text{new}}^{\text{T}}\mathbf{X}_{\text{new}}$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\text{T}}$. The estimated covariance matrix of $(\widehat{\lambda}, \widehat{\beta})^{\text{T}}$ is $\widehat{\sigma}^2(\mathbf{X}_{\text{new}}^{\text{T}}\mathbf{X}_{\text{new}})^{-}$, where $\widehat{\sigma}^2 = (n-2)^{-1}\sum_{i=1}^{n}(Y_i - \widehat{\lambda} - \widehat{\beta}X_i)^2$. Write the estimated variances as $\widehat{v}_{\widehat{\lambda}}^2$ and $\widehat{v}_{\widehat{\beta}}^2$, and write the estimated covariance as $\widehat{v}_{\widehat{\lambda},\widehat{\beta}}$. Then $\widehat{v}_{\widehat{\lambda}}^2, \widehat{v}_{\widehat{\beta}}^2$ and $\widehat{v}_{\widehat{\lambda},\widehat{\beta}}$ are independent of $\widehat{\lambda}$ and $\widehat{\beta}$ and jointly estimated with the same degrees of freedom $n - 2$.

Under these conditions, this case is particularly suitable for the application of Fieller's interval. Our intention here is to illustrate that a confidence interval constructed by our DIMER performs at least equally or even better than Fieller's interval in terms of coverage rates, but without Fieller's method's limitations on confidence interval length.

Using the results in Section 2.2.4, the estimated cumulative distribution of $\widehat{\mu}$ is

$$\mathrm{pr}(\widehat{\mu} \leq x) \approx \int_{-\infty}^{\infty} g(z|x, \mu_\lambda, \mu_\beta, \widehat{v}_\lambda^2, \widehat{v}_\beta^2, \widehat{v}_{\lambda,\beta}, \widehat{\eta}) \mathrm{exp}(-z^2) dz,$$

where $\widehat{\eta} = \widehat{v}_{\lambda,\beta}/\widehat{v}_\beta^2$, and $g(z|x, \mu_\lambda, \mu_\beta, \widehat{v}_\lambda^2, \widehat{v}_\beta^2, \widehat{v}_{\lambda,\beta}, \widehat{\eta})$ is defined as follows.

If $z \leq -\mu_\beta/\widehat{v}_\beta$, define

$$\begin{aligned} & g(z|x, \mu_\lambda, \mu_\beta, \widehat{v}_\lambda^2, \widehat{v}_\beta^2, \widehat{v}_{\lambda,\beta}, \widehat{\eta}) \\ & \quad = \left(1 - F_{t,n-2}\left[\{(x - \widehat{\eta})(\mu_\beta + v_\beta z) - (\mu_\lambda - \widehat{\eta}\mu_\beta)\}/\sqrt{\widehat{v}_\lambda^2 - 2\widehat{\eta}\widehat{v}_{\lambda,\beta} + \widehat{\eta}^2\widehat{v}_\beta^2}\right]\right) f_{t,n-2}(z)\mathrm{exp}(z^2), \end{aligned}$$

while if $z > -\mu_\omega/\widehat{v}_\omega$, define

$$\begin{aligned} & g(z|x, \mu_1, \mu_2, \widehat{v}_1^2, \widehat{v}_2^2, \widehat{v}_{12}, \eta) \\ & \quad = F_{t,n-2}\left[\{(x - \widehat{\eta})(\mu_\beta + v_\beta z) - (\mu_\lambda - \widehat{\eta}\mu_\beta)\}/\sqrt{\widehat{v}_\lambda^2 - 2\widehat{\eta}\widehat{v}_{\lambda,\beta} + \widehat{\eta}^2\widehat{v}_\beta^2}\right] f_{t,n-2}(z)\mathrm{exp}(z^2). \end{aligned}$$

Accordingly, in order to compare the results, the other five methodologies, the inverse Fisher score, Hayya's method, the nonparametric bootstrap, the parametric bootstrap and Fieller's interval are also performed and the corresponding results are presented in Section 2.3.2.2.

2.3.2.2 Simulation Results

We performed simulations on two test cases to compare the performance of the six methods of forming confidence intervals: the inverse Fisher score, Hayya's method, the nonparametric bootstrap, the parametric bootstrap, Fieller's interval and DIMER.

We generated ϵ_i and X_i from independently standard normal distribution. The number of simulations was 2000 and there were 400 bootstrap replications for each simulation. Two cases were set as follows: Case 1, $(\beta, \mu) = (1.00, 1.00)$. That is to say, in the circumstance of intercept equals to 1.00, we established the confidence intervals for the ratio between intercept multiplied by -1 and slope. In Case 2, $(\beta, \mu) = (2.00, 1.00)$. In the settings, we defined $\beta = 2$, that is ratio$= 2$. In order to investigate DIMER's performance under the non-normal condition, at here we defined ϵ_i in response follows a skew normal distribution with zero mean, variance of 1 and skewness of 0.78. Under such circumstance, both $\widehat{\beta}\widehat{\mu}$ and $\widehat{\mu}$ are not normally distributed.And we also compared the coverage from likelihood ratio tests in this case, although we did not obtain the lower and upper limits for the confidence intervals from this method. Both cases were performed with sample sizes $n = 10, 25, 50$.

In the first case (Table 2.3), DIMER is always competitive with Fieller's interval in coverage and gives reasonable lengths of confidence intervals. More importantly, it will never be invalid. In contrast, Fieller's interval has a positive probability to be invalid, especially if the sample sizes are small ($n = 10$). When $n = 10$, DIMER has shorter lengths for the mean of 90% and 95% confidence intervals while it has a higher mean value of 99% confidence intervals than the Hayya's method since the coverage of the latter approach is much lower than the nominal value. The median, interquartile range and 90% percentile of the confidence intervals by the inverse Fisher information and Hayya's method are lower than DIMER, but they behave poorly in coverage, where the interquartile range of length is shown in Table A.3 in the Appendix A. The nonparametric bootstrap and the parametric bootstrap have much longer lengths than the inverse Fisher score and Hayya's method. However, their coverage rates are still not very favorable. Fieller's interval and our DIMER have good behavior in coverage overall. A QQ plot (not given here) showing the quantiles of $\widehat{\mu}$ when $n = 10$ over the quantiles of the standard normal distribution indicates that the distribution of $\widehat{\mu}$ has much longer tails than the normal distribution.

When the sample size is small ($n = 10$), DIMER and Fieller'e interval perform better than the other four methods in coverage while when the sample size is large

($n = 50$), they are still the best in coverage. The values of the mean, median, interquartile range and 90% quantile by DIMER are rather stable. It has longer lengths of 99% confidence intervals than the other methods except Fieller's interval because their coverage rates are lower than the nominal value. Performances of the inverse Fisher information and Hayya's method are improved when the sample size increases to 50, and DIMER and Fieller's interval still perform the best. Interval lengths by all method are quite close. We changed parameter values to $(\beta, \mu) = (2.00, 1.00)$ and ϵ_i had a skew normal distribution with skewness of 0.78 in Case 2. The simulation results are given in Table 2.4. Theoretically, DIMER relies on normality assumption in case of dependent. However, even when this normal assumption was not satisfied, the simulation results show that the performance of all the methods are fairly close to that in Case 1 and DIMER still has the best performance compared to all other method, especially when the sample size was small (n=10). The coverage from the likelihood ratio test is the best among all methods except DIMER and the Fieller's interval, but in practical applications, generally speaking, it is not easy and straightforward to compute the confidence interval for parameters by using the fact that twice the difference in these log-likelihoods follows a chi-square distribution. Details of interquartile range of lengths are shown in Tables A.2 in the Appendix A.

Combining all these factors together, along with the much longer computational time for the bootstrap methods, this simulation suggests that our DIMER is at least competitive with and often superior to the other methods proposed in the literature. Based on these simulations in this section, we recommend DIMER, as it is easy to compute and it performs stably and reliably. Overall, it behaves the best in terms of both coverage probability and confidence interval length.

	Mean of Coverage			Mean of Length			Median of Length			90% Quantile of Length		
Method	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI
$n = 10, (\beta, \mu) = (1.00, 1.00), \text{cv}(\widehat{\beta}) = 0.385, \text{cv}(\widehat{\beta}\widehat{\mu}) = 0.343, \rho(\widehat{\beta}, \widehat{\beta}\widehat{\mu}) = -0.008.$ $\text{mean}(\widehat{\beta}, \widehat{\mu}) = (0.99, 1.06), \text{median}(\widehat{\beta}, \widehat{\mu}) = (1.01, 0.98).$												
IF	89.70	93.05	97.05	77.28	92.09	121.02	1.51	1.80	2.37	5.13	6.11	8.03
HM	78.55	84.90	90.70	51.90	61.84	81.27	1.45	1.73	2.27	4.63	5.52	7.25
NB	93.00	94.55	96.70	236.37	281.65	370.15	6.96	8.29	10.89	132.71	158.14	207.83
PB	90.95	93.15	95.90	193.31	230.35	302.73	3.80	4.53	5.95	129.34	154.12	202.55
FI	91.39	95.74	99.30	∞	∞	∞	2.35	3.76	57.35	∞	∞	∞
DIMER	92.70	95.95	99.15	11.29	23.58	96.24	2.25	3.36	12.71	15.18	36.78	124.01
$b^2 - 4ac < 0$	2.40	5.00	21.15									
$a < 0$	18.44	27.37	48.57									
$n = 25, (\beta, \mu) = (1.00, 1.00), \text{cv}(\widehat{\beta}) = 0.214, \text{cv}(\widehat{\beta}\widehat{\mu}) = 0.205, \rho(\widehat{\beta}, \widehat{\beta}\widehat{\mu}) = 0.005.$ $\text{mean}(\widehat{\beta}, \widehat{\mu}) = (0.99, 1.08), \text{median} = (\widehat{\beta}, \widehat{\mu}) = (0.99, 1.01).$												
IF	91.70	94.95	97.90	1.72	2.05	2.70	0.95	1.13	1.49	1.68	2.00	2.63
HM	87.70	93.10	97.40	1.53	1.82	2.40	0.94	1.12	1.47	1.60	1.91	2.50
NB	91.80	94.85	97.85	9.57	11.40	14.98	1.09	1.30	1.71	7.37	8.78	11.53
PB	91.75	95.00	98.20	8.63	10.28	13.51	1.09	1.30	1.71	4.34	5.17	6.80
FI	89.90	95.05	99.20	∞	∞	∞	1.08	1.38	2.19	2.28	3.29	10.03
DIMER	90.35	94.95	99.10	1.92	2.71	17.66	1.08	1.37	2.15	2.24	3.17	7.57
$b^2 - 4ac < 0$	0.00	0.00	0.10									
$a < 0$	0.60	1.20	5.41									
$n = 50, (\beta, \mu) = (1.00, 1.00), \text{cv}(\widehat{\beta}) = 0.141, \text{cv}(\widehat{\beta}\widehat{\mu}) = 0.144, \rho(\widehat{\beta}, \widehat{\beta}\widehat{\mu}) = -0.011.$ $\text{mean}(\widehat{\beta}, \widehat{\mu}) = (1.00, 1.02), \text{median}(\widehat{\beta}, \widehat{\mu}) = (0.99, 1.00).$												
IF	91.90	95.60	98.25	0.70	0.84	1.10	0.66	0.79	1.04	0.93	1.11	1.46
HM	90.45	95.10	98.15	0.70	0.83	1.09	0.66	0.79	1.03	0.95	1.13	1.49
NB	92.10	95.30	98.10	0.91	1.08	1.42	0.69	0.82	1.08	1.14	1.35	1.78
PB	93.00	95.55	98.30	0.78	0.93	1.22	0.70	0.83	1.10	1.09	1.30	1.71
FI	90.60	95.05	99.50	0.76	0.94	1.38	0.70	0.86	1.21	1.07	1.33	1.98
DIMER	90.65	95.00	99.40	0.76	0.94	1.36	0.70	0.85	1.20	1.07	1.33	1.97
$b^2 - 4ac < 0$	0.00	0.00	0.00									
$a < 0$	0.00	0.00	0.00									

Table 2.3: Confidence intervals for μ in a simulation study with 2000 replications for linear regression model $Y_i = \beta(X_i - \mu) + \epsilon_i$ with Setting I: $(\beta, \mu) = (1.00, 1.00)$. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval and DIMER–Direct Integral Method for Ratios.

	Mean of Coverage			Mean of Length			Median of Length			90% Quantile of Length		
Method	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI
$n = 10$, (β, μ)=(2.00,1.00), cv($\widehat{\omega}$)=0.190, cv($\widehat{\lambda}$)=0.174, $\rho(\widehat{\omega}, \widehat{\lambda})$=0.021.												
mean($\widehat{\beta}, \widehat{\mu}$)=(2.01,1.03), median($\widehat{\beta}, \widehat{\mu}$)=(2.00,0.99).												
IF	91.05	93.60	97.65	0.90	1.07	1.41	0.77	0.92	1.21	1.35	1.61	2.12
HM	81.75	88.60	94.75	0.80	0.96	1.26	0.71	0.84	1.11	1.26	1.50	1.97
NB	91.00	94.15	97.35	8.18	9.75	12.82	0.94	1.12	1.48	8.81	10.50	13.80
PB	88.30	93.15	97.00	5.10	6.08	7.99	0.79	0.94	1.24	2.39	2.85	3.74
FI	89.64	94.98	98.84	∞	∞	∞	0.89	1.15	1.94	2.05	3.08	19.86
DIMER	90.15	94.95	98.80	1.54	2.36	7.16	0.89	1.14	1.86	1.98	2.87	8.15
LR	91.95	95.95	99.35									
$b^2 - 4ac < 0$	0.05	0.40	1.05									
$a < 0$	0.90	1.86	7.93									
$n = 25$, (β, μ)=(2.00,1.00), cv($\widehat{\omega}$)=0.107, cv($\widehat{\lambda}$)=0.105, $\rho(\widehat{\omega}, \widehat{\lambda})$=-0.013.												
mean($\widehat{\beta}, \widehat{\mu}$)=(2.01,1.01), median($\widehat{\beta}, \widehat{\mu}$)=(2.01,1.00).												
IF	90.20	94.75	97.85	0.50	0.59	0.78	0.47	0.56	0.74	0.63	0.76	0.99
HM	87.80	92.35	97.45	0.48	0.57	0.75	0.46	0.54	0.72	0.63	0.75	0.99
NB	88.20	93.35	97.75	0.51	0.60	0.79	0.47	0.56	0.74	0.70	0.83	1.09
PB	89.65	93.95	97.75	0.51	0.61	0.80	0.48	0.57	0.74	0.70	0.83	1.09
FI	89.75	94.30	98.90	0.52	0.64	0.91	0.49	0.60	0.84	0.72	0.89	1.28
DIMER	89.80	94.30	98.80	0.53	0.64	0.90	0.49	0.60	0.83	0.72	0.88	1.26
LR	90.70	95.80	99.20									
$b^2 - 4ac < 0$	0.00	0.00	0.00									
$a < 0$	0.00	0.00	0.00									
$n = 50$, (β, μ)=(2.00,1.00), cv($\widehat{\omega}$)=0.072, cv($\widehat{\lambda}$)=0.072, $\rho(\widehat{\omega}, \widehat{\lambda})$=-0.021.												
mean($\widehat{\beta}, \widehat{\mu}$)=(2.00,1.01), median($\widehat{\beta}, \widehat{\mu}$)=(2.00,1.00).												
IF	89.90	94.35	98.85	0.34	0.40	0.53	0.33	0.40	0.52	0.41	0.48	0.64
HM	88.90	93.95	98.65	0.33	0.40	0.52	0.33	0.39	0.51	0.41	0.49	0.64
NB	88.35	93.40	98.25	0.33	0.40	0.52	0.33	0.39	0.51	0.41	0.49	0.65
PB	89.75	94.40	98.90	0.34	0.41	0.53	0.33	0.40	0.52	0.43	0.51	0.67
FI	89.50	94.60	98.75	0.35	0.42	0.56	0.34	0.41	0.55	0.43	0.52	0.71
DIMER	89.55	94.55	98.65	0.35	0.42	0.56	0.34	0.41	0.55	0.43	0.52	0.70
LR	89.90	95.15	99.30									
$b^2 - 4ac < 0$	0.00	0.00	0.00									
$a < 0$	0.00	0.00	0.00									

Table 2.4: Confidence intervals for μ in a simulation study with 2000 replications for linear regression model $Y_i = \beta(X_i - \mu) + \epsilon_i$ with $(\beta, \mu) = (2.00, 1.00)$, where ϵ_i follows a skew normal distribution with mean 0, variance 1 and skewness 0.78. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval, DIMER–Direct Integral Method for Ratios and LR–Likelihood ratio test.

2.4 Discussion

I have developed the Direct Integral Method for Ratios (DIMER) for forming confidence intervals for the ratio of two means. The method, based on analytical results and further approximations to account for nuisance parameters, is computationally efficient. Compared to other methods in the literature, our simulations indicated that DIMER more nearly achieves nominal coverage levels while at the same time resulting in shorter confidence interval lengths. The most important reason why our DIMER method is better than the other compared methods is that there are severely heavy tail in the distribution of the ratio, our DIMER method avoid this by direct probability computation, while other methods are badly hindered at this part, especially for those methods which based on the assumption that use the normal distribution to approximate the Cauchy likely distribution.

Due to the this reason, the performances of the nonparametric bootstrap method and the parametric bootstrap method are not favorable, although they are usually used as benchmarks to compare with other methods. Firstly, they are still based on the assumption that the ratio approximately follows a normal distribution. Secondly, when calculating the estimated standard deviation for this normal distribution, few outliers due to heavy tails will severely affect the estimation for the standard deviation by bootstraps, and consequently influence on the coverage and lengths of the confidence intervals.

3. RELATIVE RISKS ANALYSIS AND MODEL COMPARISON IN DIETARY INDEX MODELING FOR HEI-2005

3.1 Introduction

Our goal is to expand the Reedy's model to a weighted regression model, which is described in Chapter 1 and then apply it to the relative risk computation for diseases. Therefore, the structure of this Chapter is outlined as follows. In Section 3.2 we describe details the data structure, a weighted logistic model and methodology to obtain estimates and their estimated variance. Section 3.3 compares various methods to form the confidence intervals for relative risks of different diseases in different subpopulations and provides the asymptotic theories for the estimates for the log(relative risks) by the maximum likelihood and the nonparametric bootstrap method. Discussion is shown in 3.4.

Details of the 12 nutrition components are listed in Table 3.1.

Component	Units	HEI-2005 score calculation
Total Fruit	cups	$\min(5, 5 \times (\text{density}/.8))$
Whole Fruit	cups	$\min(5, 5 \times (\text{density}/.4))$
Total Vegetables	cups	$\min(5, 5 \times (\text{density}/1.1))$
DOL	cups	$\min(5, 5 \times (\text{density}/.4))$
Total Grains	ounces	$\min(5, 5 \times (\text{density}/3))$
Whole Grains	ounces	$\min(5, 5 \times (\text{density}/1.5))$
Milk	cups	$\min(10, 10 \times (\text{density}/1.3))$
Meat and Beans	ounces	$\min(10, 10 \times (\text{density}/2.5))$
Oil	grams	$\min(10, 10 \times (\text{density}/12))$
Saturated Fat	% of	if density ≥ 15 score $= 0$
	energy	else if density ≤ 7 score $= 10$
		else if density > 10 score $= 8 - (8 \times (\text{density} - 10)/5)$
		else, score $= 10 - (2 \times (\text{density} - 7)/3)$
Sodium	milligrams	if density ≥ 2000 score$=0$
		else if density ≤ 700 score$=10$
		else if density ≥ 1100
		score $= 8 - \{8 \times (\text{density} - 1100)/(2000 - 1100)\}$
		else score $= 10 - \{2 \times (\text{density} - 700)/(1100 - 700)\}$
SoFAAS	% of	if density ≥ 50 score $= 0$
	energy	else if density ≤ 20 score$=20$
		else score $= 20 - \{20 \times (\text{density} - 20)/(50 - 20)\}$

Table 3.1: Description of the HEI-2005 scoring system. Except for saturated fat and SoFAAS, density is obtained by multiplying usual intake by 1000 and dividing by usual intake of kilo-calories. For saturated fat, density is 9×100 usual saturated fat (grams) divided by usual calories, i.e., the percentage of usual calories coming from usual saturated fat intake. For SoFAAS, the density is the percentage of usual intake that comes from usual intake of calories, i.e., the division of usual intake of SoFAAS by usual intake of calories. Here, "DOL" is dark green and orange vegetables and legumes. Also, "SoFAAS" is calories from solid fats, alcoholic beverages and added sugars. The total HEI-2005 score is the sum of the individual component scores.

3.2 Basic Model

3.2.1 Data Structure and Setting

Although in this Chapter we only apply our methodologies on the experimental data of cohort cancer on two subpopulations: male and female, all our notations and formulas are for the general case which has multiple subpopulations with different multiple diseases. Therefore, all equations and algorithms in this work allow the general data sets enter into the models directly.

In the HEI-2005 data set, let $j = 1, ..., J$ denote the dietary component, where $J = 12$. Let there be $k = 1, ..., K_\ell$ types of disease in subpopulation ℓ, where $\ell = 1, ..., L$ denotes different subpopulation and there are $i = 1, ..., n_{k\ell}$ individuals with available data on disease k and gender ℓ. In practice, we have $n_{k\ell} = n_\ell$.

The data observed are as follows.

- Let $Y_{ik\ell}$ denote a health binary outcome for person i, disease k and gender ℓ.
- Let $(X_{i1\ell}, ..., X_{iJ\ell})$ be the FFQ values for person i either of density for the j^{th} dietary component or the HEI score for that component, $j = 1, ..., J = 12$.
- For each disease/gender, there may be different covariates/confounders, which always include the FFQ for energy, and other possible terms like age, ethnicity, education, body mass index, smoking, physical activity and etc. These covariates/confounders are denoted as $Z_{ik\ell}$.

To weight the component scores in a way that better captures disease risk, we assume the following model

$$\mathrm{pr}(Y_{ik\ell} = 1 | X_{ij\ell}, Z_{ik\ell}) = H(\alpha_{k\ell} + \beta_{k\ell} \textstyle\sum_{j=1}^{J} X_{ij\ell}\omega_j + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell}). \tag{3.1}$$

where $\beta_{11} = -1$ for identifiability.

3.2.2 Scoring Method

Define $\Theta = (\alpha_{11}, \theta_{11}, \alpha_{12}, \ldots, \omega)$ and based on the model (3.1), the the loglikelihood scores functions are computed as follows.

For $(\alpha_{11}, \theta_{11})$ with $i = 1, ..., n_1$

$$f_{i1}(\Theta) = (1, Z_{i11}^{\mathrm{T}})^{\mathrm{T}}\{Y_{i11} - H(\alpha_{11} + \beta_{11}\textstyle\sum_{j=1}^{J} X_{ij1}\omega_j + Z_{i11}^{\mathrm{T}}\theta_{11})\}.$$

For $(\alpha_{k\ell}, \beta_{k\ell}, \theta_{k\ell})$ when $(k, \ell) \neq (1, 1)$ with $i = 1, ..., n_\ell; k = 1, ..., K_\ell; \ell = 1, ..., L$

$$f_{ik\ell 2}(\Theta) = (1, \textstyle\sum_{j=1}^{J} X_{ij\ell}\omega_j, Z_{ik\ell}^{\mathrm{T}})^{\mathrm{T}}\{Y_{ik\ell} - H(\alpha_{k\ell} + \beta_{k\ell}\sum_{j=1}^{J} X_{ij\ell}\omega_j + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell})\}.$$

For ω_j

$$f_{ik\ell 3j}(\Theta) = \beta_{k\ell}X_{ij\ell}\{Y_{ik\ell} - H(\alpha_{k\ell} + \beta_{k\ell}\textstyle\sum_{j=1}^{J} X_{ij\ell}\omega_j + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell})\}.$$

Suppose that there are totally Q parameters in Θ. Define the $Q \times 1$ vector of scores as $S(\Theta)$, and the Fisher scoring Hessian $Q \times Q$ as $F(\Theta)$, which is defined as the expectation of the derivatives of the loglikelihood scores. Then, if Θ_{curr} is the current value of Θ in the iteration, the update is

$$\Theta_{\mathrm{new}} = \Theta_{\mathrm{curr}} - F^{-1}(\Theta_{\mathrm{curr}})S(\Theta_{\mathrm{curr}}).$$

Therefore, after updating Θ_{curr} gradually, all unknown parameters in the regression model (3.1) can be solved.

3.2.3 Sandwich Method

Rewrite the estimating functions in a more convenient as follows. Let $\Psi_{ik\ell} = \left(\Psi_{ik\ell 1}{}^{\mathrm{T}}, \Psi_{ik\ell 2}{}^{\mathrm{T}}, \Psi_{ik\ell 3}{}^{\mathrm{T}}\right)^{\mathrm{T}}$. Here define $\Psi_{ik\ell 1} = f_{i1}$ if $(k,\ell) = (1,1)$, and is equal to 0 otherwise. Define $\Psi_{ik\ell 2} = \left(\Psi_{ik\ell 2.1}{}^{\mathrm{T}}, \ldots, \Psi_{ik\ell 2.L}{}^{\mathrm{T}}\right)^{\mathrm{T}}$, where $\Psi_{ik\ell 2.\ell'} = (\Psi_{ik\ell 22\ell'}{}^{\mathrm{T}}, \ldots, \Psi_{ik\ell 2K_{\ell'}\ell'}{}^{\mathrm{T}})^{\mathrm{T}}$ for $\ell' = 1$ and $\Psi_{ik\ell 2.\ell'} = (\Psi_{ik\ell 21\ell'}{}^{\mathrm{T}}, \ldots, \Psi_{ik\ell 2K_{\ell'}\ell'}{}^{\mathrm{T}})^{\mathrm{T}}$ for $\ell' \neq 1$; $\Psi_{ik\ell 2k'\ell'} = f_{ik\ell 2}$ if $(k,\ell) = (k',\ell')$, and 0 otherwise. Also we have $\Psi_{ik\ell 3} = (f_{ik\ell 31}, f_{ik\ell 32}, \ldots, f_{ik\ell 3J})^{\mathrm{T}}$.

Asymptotically

$$\begin{aligned} N^{-1/2}\textstyle\sum_{\ell=1}^{2}\sum_{i=1}^{n_\ell}\sum_{k=1}^{K_\ell}\Psi_{ik\ell}(\Theta) &\sim \mathrm{Normal}(\mathbf{0}, V_\Psi(\Theta)), \\ V_\Psi(\Theta) &= N^{-1}\textstyle\sum_{\ell=1}^{2} n_\ell \mathrm{cov}\{\sum_{k=1}^{K_\ell}\Psi_{ik\ell}(\Theta)\}, \end{aligned}$$

if $n_1, ..., n_L \to \infty$ and $\max(n_1, ..., n_L)/\min(n_1, ..., n_L) \to c < \infty$, and where $N = \sum_{\ell=1}^{2} K_\ell n_\ell$.

Define $\widehat{\mu}_\ell = n_\ell^{-1}\sum_{i=1}^{n_\ell}\sum_{k=1}^{K_\ell}\Psi_{ik\ell}(\widehat{\Theta})$, so that an estimate for $V_\Psi(\Theta)$ is given by $\widehat{V}_\Psi(\widehat{\Theta}) = N^{-1}\sum_{\ell=1}^{2}\sum_{i=1}^{n_\ell}\{\sum_{k=1}^{K_\ell}\Psi_{ik\ell}(\widehat{\Theta}) - \widehat{\mu}_\ell\}\{\sum_{k=1}^{K_\ell}\Psi_{ik\ell}(\widehat{\Theta}) - \widehat{\mu}_\ell\}^{\mathrm{T}} = V_\Psi(\Theta) + o_p(1)$.

The asymptotic limit distribution of $\widehat{\Theta}$ by the sandwich method (see Carroll, Ruppert and Stefanski, 2006) is as follow.

$$N^{1/2}(\widehat{\Theta} - \Theta) \backsim \mathrm{Normal}\{0, A^{-1}(\Theta)V_\Psi(\Theta)A^{-\mathrm{T}}(\Theta)\},$$

where $A(\Theta) = -N^{-1}\sum_{\ell=1}^{2}\sum_{k=1}^{K_\ell}\sum_{i=1}^{n_\ell}\mathrm{E}\{\partial\Psi_{ik\ell}(\Theta)/\partial\Theta^{\mathrm{T}}\}$ and a consistent estimate of it is $A(\widehat{\Theta}) = -N^{-1}\sum_{\ell=1}^{2}\sum_{k=1}^{K_\ell}\sum_{i=1}^{n_\ell}\mathrm{E}\{\partial\Psi_{ik\ell}(\widehat{\Theta})/\partial\widehat{\Theta}^{\mathrm{T}}\}$.

We use $A^{-1}(\widehat{\Theta})\widehat{V}_\Psi(\widehat{\Theta})A^{-\mathrm{T}}(\widehat{\Theta})$ to estimate $A^{-1}(\Theta)V_\Psi(\Theta)A^{-\mathrm{T}}(\Theta)$ and obtain the estimated variance of MLE $\widehat{\Theta}$.

3.3 Relative Risks Analysis

To be notice here, the asymptotic distributions of the relative risks are directly related to the structure of the covariates $X_{i\ell}$, where $X_{i\ell} = (X_{i1\ell}, \cdots, X_{iJ\ell})^{\mathrm{T}}$. Suppose $X_{i\ell}$'s are regarded as random variables following some parametric model which may be unknown, we define $\Lambda_{k\ell} = (\beta_{k\ell}, \omega^{\mathrm{T}})^{\mathrm{T}}$, its estimate $\widehat{\Lambda}_{k\ell} = (\widehat{\beta}_{k\ell}, \widehat{\omega}^{\mathrm{T}})^{\mathrm{T}}$, and the random variable $S_{ik\ell}(\beta_{k\ell}, \omega) = S_{ik\ell}(\Lambda_{k\ell}) = \beta_{k\ell} X_{i\ell}^{\mathrm{T}} \omega$. Let $S_{\alpha,k\ell}(\beta_{k\ell}, \omega)$ be the α^{th} population percentile of the $S_{ik\ell}(\Lambda_{k\ell})$, i.e., $\alpha = \mathrm{pr}\{S_{ik\ell}(\Lambda_{k\ell}) \le S_{\alpha,k\ell}(\Lambda_{k\ell})\}$.

We are interested in estimating the relative risk for moving from the 10^{th} to the 90^{th} population percentile of the $S_{ik\ell}(\Lambda_{k\ell})$, i.e., we wish to estimate $\mathcal{R}_{k\ell} = \exp\{S_{0.90,k\ell}(\Lambda_{k\ell}) - S_{0.10,k\ell}(\Lambda_{k\ell})\}$, and form a confidence interval for it. This problem can be reduced to construct a confidence interval for $\mathcal{V}_{k\ell} = S_{0.90,k\ell}(\Lambda_{k\ell}) - S_{0.10,k\ell}(\Lambda_{k\ell})$, which we would then exponentiate.

If we assume that the observed $X_{i\ell}$'s are regarded as a sequence of known fixed constants, then this question transfers to estimate the relative risk for moving from the 10^{th} to the 90^{th} sample percentile of the $S_{ik\ell}(\Lambda_{k\ell})$. Let $\widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})$ be the α^{th} sample percentile of the $S_{ik\ell}(\Lambda_{k\ell})$, i.e., $\alpha = n_\ell^{-1} \sum_{i=1}^{n_\ell} \mathrm{I}\{S_{ik\ell}(\Lambda_{k\ell}) \le \widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})\}$. In the other words, the interested term changes to $\mathcal{R}_{k\ell} = \exp\{\widehat{S}_{0.90,k\ell}(\Lambda_{k\ell}) - \widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})\}$ and again we need to form a confidence interval for it. Similarly, this question can be reduced to construct a confidence interval for $\mathcal{V}_{k\ell} = \widehat{S}_{0.90,k\ell}(\Lambda_{k\ell}) - \widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})$.

For simplicity, we use the second assumption in this article because if we assume the $X_{i\ell}$'s are regarded as random variables, the parametric models of $X_{i\ell}$'s need to be built and estimated unknown parameters and might involve misspecification problem.

In Section 3.3.1, we present the asymptotic theories of MLE for the log(relative risks) and estimate by using the nonparametric bootstrap method.

3.3.1 Asymptotic Distributions

3.3.1.1 Asymptotic Distribution of $\widehat{\mathcal{V}}_{k\ell}$

Our estimate for $\mathcal{V}_{k\ell}$ is $\widehat{\mathcal{V}}_{k\ell} = \widehat{S}_{0.90,k\ell}(\widehat{\Lambda}_{k\ell}) - \widehat{S}_{0.10,k\ell}(\widehat{\Lambda}_{k\ell})$. We show the following result in the Appendix B.1.

<u>Lemma</u> 3 *Define*

$$\begin{aligned} D_{k\ell} = & \{\partial\widehat{S}_{0.90,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}^{\mathrm{T}}\}var(\widehat{\Lambda}_{k\ell})\{\partial\widehat{S}_{0.90,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}\} \\ & +\{\partial\widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}^{\mathrm{T}}\}var(\widehat{\Lambda}_{k\ell})\{\partial\widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}\} \\ & -2\{\partial\widehat{S}_{0.90,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}^{\mathrm{T}}\}var(\widehat{\Lambda}_{k\ell})\{\partial\widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}\}. \end{aligned}$$

The asymptotic limit distribution of $\widehat{\mathcal{V}}_{k\ell}$ is given by

$$N^{1/2}(\widehat{\mathcal{V}}_{k\ell} - \mathcal{V}_{k\ell}) \sim Normal(0, D_{k\ell})$$

3.3.1.2 Asymptotic Distribution of $\widehat{\mathcal{V}}^*_{k\ell}$

For the given paired data $(Y_{ik\ell}, X_{i\ell}, Z_{ik\ell})$, we resample them with replacements to a new data set named as $(Y^b_{ik\ell}, X^b_{i\ell}, Z^b_{ik\ell})$ with $b = 1, ..., B$, and then compute the $\widehat{\mathcal{V}}^b_{k\ell}$ based on this sampled data set $(Y^b_{ik\ell}, X^b_{i\ell}, Z^b_{ik\ell})$. In Appendix B.1.1, we prove the following result.

<u>Lemma</u> 4 *Define $\widehat{\mathcal{V}}^*_{k\ell} = B^{-1}\sum_{b=1}^{B}\widehat{\mathcal{V}}^b_{k\ell}$ and $\widehat{D}^*_{k\ell} = (B-1)^{-1}\sum_{b=1}^{B}\left(\widehat{\mathcal{V}}^b_{k\ell} - \widehat{\mathcal{V}}^*_{k\ell}\right)^2$.*

The asymptotic limit distribution of $\widehat{\mathcal{V}}^*_{k\ell}$ is given by

$$B^{1/2}(\widehat{D}^*_{k\ell})^{-1/2}(\widehat{\mathcal{V}}^*_{k\ell} - \mathcal{V}_{k\ell}) \backsim Normal(0,1).$$

In further work, we not only use this asymptotic distribution to construct the confidence intervals for $\mathcal{V}_{k\ell}$, but also run the hypothesis test to verify whether the relative risks are statistically significantly different in four models.

In Section 3.3.2, we describe six methods to form the confidence intervals of $\mathcal{V}_{k\ell}$ so that the corresponding confidence intervals for the relative risks $\mathcal{R}_{k\ell}$ could be easily constructed.

3.3.2 Confidence Interval Construction

3.3.2.1 The Sandwich Method and the Inverse Fisher Score Method

As a benchmark, a first way to form the confidence intervals of $\mathcal{V}_{k\ell}$ is using the asymptotic distribution of $\widehat{\mathcal{V}}_{k\ell}$ in Section 3.3.1.1, where the estimated variance of $\widehat{\mathcal{V}}_{k\ell}$ is achieved by the partial $\{\partial S_{\alpha,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}\}$ and the estimated variance of $\widehat{\Lambda}_{k\ell}$ from the sandwich method. Accordingly, the confidence intervals of $\mathcal{R}_{k\ell}$ are formed.

Instead of the sandwich method, another common method to estimate the variance of $\widehat{\Lambda}_{k\ell}$ is the inverse Fisher score information matrix. After obtaining the estimated variance, similarly, one can construct the confidence intervals of $\mathcal{R}_{k\ell}$ with combining the partial $\{\partial S_{\alpha,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}\}$.

Depends on the results presented in Chapter 2, for the estimated variance of $\widehat{\beta}_{k\ell}$, if calculated by the sandwich method, then the results' accuracy is far from satisfactory. And in that paper, we pointed out that $\widehat{\beta}_{k\ell}$ in model (3.1) can be approximately written as a ratio of two normally distributed variables, which would follow a Cauchy-like distribution. So that the normal distribution approximation for $\widehat{\beta}_{k\ell}$ by the sandwich method is not appropriate for the data set used in our study. Furthermore, this will inevitably influence the accuracy of the estimated variance for the estimates of the other parameters, since they are jointly estimated by using the sandwich method or the inverse Fisher score matrix. For example, refers to other results from our study which is not presented here, we found the estimated variance for $\widehat{\alpha}_{k\ell}$ is not accurate if the sandwich method or the inverse Fisher matrix was applied, even though it is an intercept parameter instead of the weight parameter such as $\widehat{\beta}_{k\ell}$.

However, if the targeted estimate is $\widehat{\omega}$,the estimated variance from the sandwich method or the inverse Fisher matrix is very reliable and stable. This had been proved by the results from the nonparametric bootstrap method. Therefore, we present the process in Section 3.3.2.2 to estimate the confidence intervals of the relative risks.

3.3.2.2 the Direct Integral Method for Ratios

In that paper, for given similar logistic regression model, we proposed an algorithm to compute the confidence interval of $\beta_{k\ell}$ reliably and stably which can be summarized as follows.

Define a new latent variable $\Gamma_{k\ell} = \beta_{k\ell}\omega$ and rewrite the model (3.1) as

$$\mathrm{pr}(Y_{ipm} = 1|X_{ijm}, Z_{ipm}) = \begin{cases} H(\alpha_{pm} + \beta_{pm}\sum_{j=1}^{J} X_{ij\ell}\Gamma_{k\ell,j}/\beta_{k\ell} + Z_{ipm}^{\mathrm{T}}\theta_{pm}), & \text{if } (p,m) \neq (k,\ell); \\ H(\alpha_{k\ell} + \sum_{j=1}^{J} X_{ij\ell}\Gamma_{k\ell,j} + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell}), & \text{if } (p,m) = (k,\ell), \end{cases} \tag{3.2}$$

with $\beta_{11} = -1$.

Define $\Theta^{k\ell} = (\alpha_{11}, \theta_{11}, \alpha_{12}, \ldots, \Gamma_{k\ell})$. We write the log-likelihood function with latent variable $\Gamma_{k\ell}$ but without ω. Then it is easy to obtain the estimate $\widehat{\Theta}^{k\ell}$ for parameter $\Theta^{k\ell}$ by the scoring method and the estimated variance by the sandwich method. And in that paper, we also provided the algorithm to estimate the covariance of $(\widehat{\omega}, \widehat{\Gamma}_{k\ell})$.

Since we have $\widehat{\Gamma}_{k\ell} = \widehat{\beta}_{k\ell}\widehat{\omega}$, with a $J \times 1$ dimensional constant vector $\mathbf{e}$, $\widehat{\beta}_{k\ell}$ can be thought of as the ratio of two variables $\mathbf{e}^{\mathrm{T}}\widehat{\Gamma}_{k\ell}$ and $\mathbf{e}^{\mathrm{T}}\widehat{\omega}$, which follow a bivariate normal distribution. For simplicity, we set $e = \mathbf{1}$. After setting $\widehat{\beta}_{kl}$ is distributed as ratio of two normal variables, we could apply the direct integral method presented in that paper to form the confidence intervals of $\beta_{k\ell}$, which gives rise to the approaches in Appendix B.1.2.1 to build the confidence intervals for $\mathcal{V}_{k\ell}$ and get the according confidence intervals of the relative risks $\mathcal{R}_{k\ell}$.

In Section 3.3.2.3, we introduce a technology to avoid computing the confidence interval of $\beta_{k\ell}$ for comparison.

3.3.2.3 Model Transformation Method

Since in model (3.3.2.2) we had defined $\Gamma_{k\ell} = \beta_{k\ell}\omega$, and considered the estimated variance of it is as reliable as the estimated variance of $\widehat{\omega}$, one might naturally ask the question: why not to replace the direct integral method by the following one to get the confidence intervals of the relative risks since it is more straightforward after all? For this question, we will answer it detailedly in Section 3.4. First, we describe the algorithm as follows.

Same as the definition $\widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})$, let $\widehat{S}_{\alpha,k\ell}(\Gamma_{k\ell})$ be the α^{th} sample percentile of the $S_{ik\ell}(\Gamma_{k\ell})$, i.e., $\alpha = n_\ell^{-1}\sum_{i=1}^{n_\ell}\mathrm{I}\{S_{ik\ell}(\Gamma_{k\ell}) \le \widehat{S}_{\alpha,k\ell}(\Gamma_{k\ell})\}$.

The relative risk moving from the 10^{th} to the 90^{th} sample percentile of the $\widehat{S}_{ik\ell}(\Gamma_{k\ell})$ is expressed as $\mathcal{R}_{k\ell} = \exp\{\widehat{S}_{0.90,k\ell}(\Gamma_{k\ell}) - \widehat{S}_{0.10,k\ell}(\Gamma_{k\ell})\}$, and oue purpose is to construct a confidence interval for it, which can be reduced to form a confidence interval of $\Omega_{k\ell} = \widehat{S}_{0.90,k\ell}(\Gamma_{k\ell}) - \widehat{S}_{0.10,k\ell}(\Gamma_{k\ell})$.

Define the estimate for $\Omega_{k\ell}$ is $\widehat{\Omega}_{k\ell} = \widehat{S}_{0.90,k\ell}(\widehat{\Gamma}_{k\ell}) - \widehat{S}_{0.10,k\ell}(\widehat{\Gamma}_{k\ell})$. Similarly as Lemma 3, we can write the following result.

<u>Lemma</u> 5 *Define*

$$\begin{aligned} D_{\Omega_{k\ell}} = {} & \{\partial\widehat{S}_{0.90,k\ell}(\Gamma_{k\ell})/\partial\Gamma_{k\ell}^{\mathrm{T}}\}var(\widehat{\Gamma}_{k\ell})\{\partial\widehat{S}_{0.90,k\ell}(\Gamma_{k\ell})/\partial\Gamma_{k\ell}\} \\ & +\{\partial\widehat{S}_{0.10,k\ell}(\Gamma_{k\ell})/\partial\Gamma_{k\ell}^{\mathrm{T}}\}var(\widehat{\Gamma}_{k\ell})\{\partial\widehat{S}_{0.10,k\ell}(\Gamma_{k\ell})/\partial\Gamma_{k\ell}\} \\ & -2\{\partial\widehat{S}_{0.90,k\ell}(\Gamma_{k\ell})/\partial\Gamma_{k\ell}^{\mathrm{T}}\}var(\widehat{\Gamma}_{k\ell})\{\partial\widehat{S}_{0.10,k\ell}(\Gamma_{k\ell})/\partial\Gamma_{k\ell}\}. \end{aligned}$$

The asymptotic limit distribution of $\widehat{\Omega}_{k\ell}$ is given by

$$n_\ell^{1/2}(\widehat{\Omega}_{k\ell} - \Omega_{k\ell}) \sim Normal(0, D_{\Omega_{k\ell}}).$$

If $(k, \ell) = (1, 1)$, $\widehat{\Gamma}_{k\ell}$ is equivalent to $-\widehat{\omega}$. For cases $(k, \ell) \neq (1, 1)$, we use the model (3.3.2.2) to obtain the estimate of $\widehat{\Gamma}_{k\ell}$ and its estimated variance.

Therefore, a confidence interval of $\Omega_{k\ell}$ is formed by the asymptotic distribution of $\widehat{\Omega}_{k\ell}$ and this algorithm avoids to compute the confidence intervals of β_{kl}.

3.3.3 The Nonparametric Bootstrap Method

In Lemma 4, we have already described the asymptotic distribution of the estimate of log(relative risk) by the nonparametric bootstrap method. As a matter of course, the procedure in Appendix B.1.2.2 is applied to calculate the estimation of the log(relative risk) and its distribution.

This procedure is then applied to the male and female cohort cancer 2005-HEI data set. In this data set, we have only one disease, the cohort cancer, in two subpopulations, male and female, where $n_1 = 293616$ (male), $n_2 = 198246$ (female) and there are 3110 incident colorectal cancer cases (2151 in male and 959 in female). For the covariates $Z_{ik\ell}$, there are 24 components for male, for one individual on disease $k = 1$ and gender $\ell = 1$; and for female, there are two more terms, so that 26 components for one individual on disease $k = 1$ when gender $\ell = 2$.

However, refers to the other work of us mentioned previously, the presented simulation results show that, for the logistic regression of HEI-2005 case-control data set (in which the control/case ratio is 3), if computed by the nonparametric bootstrap method, the resulted coverage of confidence intervals for $\beta_{k\ell}$ is not even close to the nominal value. We checked but did report in the paper that when we performed the nonparametric bootstrap method to the whole data set, the coverage is as bad as in the case-control data set. Therefore, a further investigation on why the nonparametric bootstrap method is not feasible in such circumstance is necessary.

The following content summarizes the process: firstly, transfer the model into model (3.3.2.2). This allows us to estimate $(\widehat{\beta}_{k\ell}, \widehat{\Gamma}_{k\ell})$, based on which the corresponding $\widehat{\omega} = \widehat{\Gamma}_{k\ell}/\widehat{\beta}_{k\ell}$ can be back calculated. Usually, the results from this procedure should be the same as the results obtained by applied this data set to model (3.1). However, in some circumstances, the results from model (3.3.2.2) might be diverged while model (3.1)'s results are converged, or different with model (3.1)'s results even if both obtain converged results. Intuitively, we think the reason for this might due to the very low incidence of cohort cancer among females in this data set. In addition, the data applied in the nonparametric bootstrap method requires resampling with replacement, which will cause even in this low incidence, there are still duplicated patients data exists. Also, as Reedy, et al. (2008) pointed out that there might be inherent differences between how men and women complete the AARP food frequency questionnaire, giving rise to increased measurement error. These conclude the possible reasons for the diverged results of model (3.3.2.2). On the other hand, in model (3.1), we directly set the value for β_{11} as -1, and since the male data is more reliable and under this model it mainly dominates the whole simulated data set, so the probability of estimation getting diverged results would be much smaller, although the female data still have large influences on the estimated values.

Based on the above analysis, a modified methodology is put forward. For the same set of simulated nonparametric bootstrap data, both model (3.1) and model (3.3.2.2) were applied to compute the estimation of $\Theta = (\alpha_{11}, \theta_{11}, \alpha_{12}, \ldots, \omega)$. If the two results both converged and are the same, then this simulated data set will be kept. Otherwise, they will be abandoned. We call this process as ' the modified nonparametric bootstrap method'.

3.4 Results and Discussions

The results in the cohort cancer HEI-2005 data set for the six different methodologies are presented and compared in Table 3.2. The six methodologies are listed as below: the inverse Fisher method, the sandwich method, the direct integral method, the model transformation method, the nonparametric bootstrap method and the modified nonparametric bootstrap method.

There are two methods, the inverse Fisher score method and the sandwich method, need to compute the derivative terms $\{\partial\widehat{S}_{0.90,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}\}$ and $\{\partial\widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}\}$. Define $X_{[\alpha],k\ell}$ as $\widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell}) = X^{\mathrm{T}}_{[\alpha],k\ell}\Lambda_{k\ell}$. Since we assumed that $X_{i\ell}$'s are regarded as a sequence of known fixed constants, we have

$$\partial\widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell} = (X^{\mathrm{T}}_{[\alpha],k\ell}\omega, X^{\mathrm{T}}_{[\alpha],k\ell}\beta_{k\ell})^{\mathrm{T}},$$

where $\alpha = (0.10, 0.90)$.

Due to $X_{[\alpha],k\ell}$ is a 12−dimensional vector, one might concern the stability of $(X^{\mathrm{T}}_{[\alpha],k\ell}\omega, X^{\mathrm{T}}_{[\alpha],k\ell}\beta_{k\ell})^{\mathrm{T}}$. An alternative way to compute the term $\partial\widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}$ which in purpose decreases the stability is to involve more points in the computation

$$\partial\widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})/\Lambda_{k\ell} = (2P+1)^{-1}H^{\mathrm{T}}_{\omega,\beta_{k\ell}}\textstyle\sum_{d=-P}^{P}X_{[\alpha]+d,k\ell},$$

where $\mathbf{I}$ is the 12×12 dimensional identity matrix and $H_{\omega,\beta_{k\ell}} = (\omega, \mathbf{I}\beta_{k\ell})$, and $X_{[\alpha]+d,k\ell}$ is defined as follows. Rewrite the sample percentile $\widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})$ as $n_\ell\alpha = \sum_{i=1}^{n_\ell}\mathrm{I}\{S_{ik\ell}(\Lambda_{k\ell}) \le \widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})\}$, and define $n_\ell\alpha+d = \sum_{i=1}^{n_\ell}\mathrm{I}\{S_{ik\ell}(\Lambda_{k\ell}) \le \widehat{S}_{[\alpha]+d,k\ell}(\Lambda_{k\ell})\}$. Then let $\widehat{S}_{[\alpha]+d,k\ell}(\Lambda_{k\ell}) = X^{\mathrm{T}}_{[\alpha]+d,k\ell}\Lambda_{k\ell}$. It is obvious that totally there are $2P+1$ points involved in the derivative computation.

Accordingly, the derivative $\partial\widehat{S}_{\alpha,k\ell}(\Gamma_{k\ell})/\partial\Gamma_{k\ell}$ could also be obtained in this way.

Therefore, for the computation of the relative risk 95% confidence intervals including terms $\partial\widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}$ or $\partial\widehat{S}_{\alpha,k\ell}(\Gamma_{k\ell})/\partial\Gamma_{k\ell}$, sensitivity studies were carried out for the three presented methods' performances, on the influences caused by different number of points used in the derivative calculation. From Figure B.1 to Figure B.3 in the Appendix B.2 we can see, the confidence intervals have the widest range when there is only one point involved in the derivative computation, and it strictly follows the rule that, as the number of points increases, the range's upper limit goes lower while the lower limit rises higher. In other words, as the number of points increases, the range of confidence intervals rapidly converged to narrow-band. Especially, $P = 5$ is a critical value. In the results which is not reported here, we investigated all the cases with various number of points included in derivative calculation, in the range of 1 to 201. As results show, when the number of points increased beyond 11 ($P = 5$), the range becomes steady and the results do not have significant changes. Therefore, we choose 11 points in the derivative computations for all the three methods.

Furthermore, our data set size is rather large ($n_1 = 293616, n_2 = 198246$) comparing to 11, therefore, all 11 points are quite near the α^{th} percentile and it is reasonable to use the average of them to compute the derivatives.

Next, it is clear that both the sandwich method and the model transformation method obtain exactly the same estimation of $\mathcal{R}_{k\ell}$ and it's 95% confidence interval. The reason is explained as following: since ω and $\Gamma_{k\ell}$ are from the same logistic model up to a parameter transformation, there is potential relationship between the estimated variances of $\widehat{\omega}$ and $\widehat{\Gamma}_{k\ell}$. The following formula can be proved by the delta method

$$H_{\omega,\beta_{k\ell}}\mathrm{var}(\widehat{\Lambda}_{k\ell})H^{\mathrm{T}}_{\omega,\beta_{k\ell}} = \mathrm{var}(\widehat{\Gamma}_{k\ell}).$$

Furthermore, in the calculations for the derivative $\partial\widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}$ and $\partial\widehat{S}_{\alpha,k\ell}(\Gamma_{k\ell})/\partial\Gamma_{k\ell}$, both use the same points of $X_{ik\ell}$ for given the same number of points involved in the process. This is because for fixed $X_{i\ell}$, its corresponding $S_{ik\ell}(\Lambda_{k\ell})$ and $S_{ik\ell}(\Gamma_{k\ell})$ have the same order in their sequences. Define $X_{P,[\alpha]+d,k\ell} = (2P+1)^{-1}\sum_{d=-P}^{P} X_{[\alpha]+d,k\ell}$, then we have $\partial\widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell} = H^{\mathrm{T}}_{\omega,\beta_{k\ell}} X_{P,[\alpha]+d,k\ell}$ and $\partial\widehat{S}_{\alpha,k\ell}(\Gamma_{k\ell})/\partial\Gamma_{k\ell} = X_{P,[\alpha]+d,k\ell}$. Therefore, these two methods result the identical estimated variance of $\widehat{\mathcal{V}}_{k\ell}$ and $\widehat{\Omega}_{k\ell}$ through analyses, and this consequently causes their estimations on the confidence interval are exactly the same as each other.

Based on comparisons in Table 3.2 we can easily see: for the male relative risk confidence intervals, the six presented methods obtain rather close results. The reason is that, for the model's identifiability, β_{11} value had been set to be -1, so $S_{\alpha,k\ell} = -X^{\mathrm{T}}_{i\ell}\omega$. Therefore, the only variable considered at here is ω. As we previous stated, we could obtain the reliable estimated variance of its estimate $\widehat{\omega}$ by using the sandwich method or the inverse Fisher score matrix.

However, for the confidence intervals of the female relative risk, the results from the direct integral method and the modified nonparametric bootstrap method are well matched with each other. For the rest of methods, results of the inverse Fisher score method and the sandwich method are lower than the ones from the direct integral method and the modified nonparametric bootstrap method, while the nonparametric bootstrap method's results are higher than them. This observation is just same as the one presented in the other paper of us. In that paper, the confidence interval's coverages for $\beta_{k\ell}$ from the inverse Fisher and the sandwich method are always lower than the nominal value, while the coverage from the usual nonparametric bootstrap method is too high on the contrary.

In Table 3.3, details of the comparison between the results from the nonparametric bootstrap and the modified nonparametric bootstrap methods. Based on our screening rules, there are 2320 effective data sets among the total of 2500 nonparametric bootstrap data sets. Therefore, the B value is set to be equal to 2320 in the modified nonparametric bootstrap. Both of them obtain similar results for male, while the results for female are different. In addition, the two methods have significant difference when estimated the variable $\widehat{\beta}_{k\ell}$, but for the relative risk, their estimations are close to each other.

For the aspect of computation time, due to the extremely large data amount, the time consumption ratio between the modified nonparametric bootstrap method and the direct integral method is around 7000, and this is a special case in which the data set only considers single disease and two subpopulations. In the future, if these formulas were applied to more complicated cases such as several different multiple diseases in multiple subpopulations, the time consumption ratio will consequently become even higher. In conclusion, regardless of reliability, accuracy and computation efficiency, the direct integral method is obviously the best among all the method presented in this paper to compute the relative risks and their confidence intervals.

	Male		Female	
	Relative Risk	95% CI	Relative Risk	95% CI
IF	0.662	(0.585,0.750)	0.752	(0.654,0.866)
SM	0.662	(0.584,0.751)	0.752	(0.654,0.865)
DIMER	0.662	(0.593,0.740)	0.752	(0.584,0.912)
MT	0.662	(0.584,0.751)	0.752	(0.654,0.865)
NB	0.646	(0.573,0.742)	0.753	(0.562,1.046)
MNB	0.646	(0.572,0.740)	0.732	(0.560,0.925)

Table 3.2: Relative Risks and their 95% Confidence intervals for colorectal cancer on HEI component scores with model $\mathrm{pr}(Y_{ik\ell} = 1|X_{i\ell}, Z_{ik\ell}) = H(\alpha_{k\ell} + \beta_{k\ell}\sum_{j=1}^{J} X_{ij\ell}\omega_j + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell})$, where $\beta_{11} = -1$. IF–the inverse Fisher method, SM–the sandwich method, DIMER–the direct integral method for ratios, MT–the model transformation method, NB–the nonparametric bootstrap method and MNB–the modified nonparametric bootstrap method. The nonparametric bootstrap method has 2500 simulated data sets and the modified bootstrap method has 2320 simulated data sets.

	Nonparametric Bootstrap		Modified Nonparametric Bootstrap	
	Male	Female	Male	Female
10th perc T_{il}	0.268	0.340	0.275	0.347
95% CI for 10th perc T_{il}	(-0.195,0.707)	(-0.128,0.785	(-0.170,0.708)	(-0.103,0.787)
90th perc T_{il}	0.707	0.751	0.714	0.759
95% CI for 90th perc T_{il}	(0.211,1.180)	(0.257,1.234)	(0.229,1.180)	(0.271,1.234)
$\widehat{\beta}_{k\ell}$	-0.996	-0.774	-1.000	-0.818
Relative Risk	0.646	0.753	0.646	0.732
95% CI for RR (method I)	(0.573,0.742)	(0.562,1.046)	(0.572,0.740)	(0.560,0.925)
95% CI for RR (Method II)	(0.566,0.735)	(0.546,1.012)	(0.566,0.733)	(0.565,0.932)

Table 3.3: Bootstrap results of Relative Risks for colorectal cancer on HEI component scores. RR–relative risks. Method I–by percentile of relative risk from bootstrap. Method II–the exponent of $\mathcal{V}^{*}_{k\ell} \pm 1.96(\widehat{D}^{*}_{k\ell})^{1/2}$ based on the normal approximation of log(Relative Risk). The nonparametric bootstrap method has 2500 simulated data sets and the modified bootstrap method has 2320 simulated data sets.

4. MODEL COMPARISON AND VARIABLE SELECTION IN DIETARY INDEX MODELING FOR HEI-2005

4.1 Introduction

To verify there are statistical significant difference between our logistic model (3.1) and the Reedy's model or not, as well as other simple models, four different models are compared in Section 4.2 by using the hypothesis test and the likelihood ratio test. Furthermore, a simple technology of bounded constrains estimator in Section 4.3 and the adaptive lasso method in Section 4.4 are interpreted to identify which components in HEI components are more important to cancers. Additionally, a novel solution algorithm for solving the L_1 norm penalty for nonlinear regression model is proposed and we show applications in real HEI-2005 data set.

4.2 Model Comparison in HEI-2005

The definitions of these four models are listed as below

$$
\begin{aligned}
&\text{Model I:} && \mathrm{pr}(Y_{ik\ell}=1|X_{ij\ell}) = H(\alpha_{k\ell}+\beta_{k\ell}{\textstyle\sum_{j=1}^{J}}X_{ij\ell}),\\
&\text{Model II:} && \mathrm{pr}(Y_{ik\ell}=1|X_{ij\ell}) = H(\alpha_{k\ell}+\beta_{k\ell}{\textstyle\sum_{j=1}^{J}}X_{ij\ell}\omega_j) \text{ with } \beta_{11}=-1,\\
&\text{Model III:} && \mathrm{pr}(Y_{ik\ell}=1|X_{ij\ell},Z_{ik\ell}) = H(\alpha_{k\ell}+\beta_{k\ell}{\textstyle\sum_{j=1}^{J}}X_{ij\ell}+Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell})),\\
&\text{Model IV:} && \mathrm{pr}(Y_{ik\ell}=1|X_{ij\ell},Z_{ik\ell}) = H(\alpha_{k\ell}+\beta_{k\ell}{\textstyle\sum_{j=1}^{J}}X_{ij\ell}\omega_j+Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell}) \text{ with } \beta_{11}=-1.
\end{aligned}
$$

In which, Model III represents Reedy's model, and Model IV refers to the model we presented in Section 3.2.1. Hypothesis tests are carried out for the comparisons between Model I and II, Model II and IV, Model III and IV, respectively. For each pair, differences between the relative risks and the log value are then obtained. The 95% confidence intervals are also presented for the difference and ratio of the corresponding relative risks.

The modified nonparametric bootstrap method, which was introduced in Section 3.3.3, is applied to compute the needed values for the hypothesis test. The details are described in Section 4.2.1 and Section 4.2.2. As for the original data set, since the relatives risk's values from the different models are not independent,therefor, all models are using the same simulated bootstrap data set. In Section 4.2.3, we discuss how to compute the 95% confidence intervals of the difference and ratio for the relative risks in the compared models, as well as how to perform the likelihood ratio test for compared models in Section 4.2.4.

4.2.1 Hypothesis Test for Log Relative Risk

If we define two models required for comparison as Model_{c1} and Model_{c2}. For both compared models Model_{c1} and Model_{c2}, the same b^{th} effective simulated nonparametric bootstrap data set $(Y_{ik\ell,b}, X_{ij\ell,b}, Z_{ik\ell,b})$ for $\ell = 1, \ldots, L$, $k = 1, \ldots, K_\ell$ and $i = 1, \ldots, n_\ell$ is taken into analysis, from which the log(relative risks) for the disease k in the ℓ subpopulation for this data set can be obtained as $\widehat{\mathcal{V}}^b_{k\ell,c1}$ and $\widehat{\mathcal{V}}^b_{k\ell,c2}$, respectively. Based on the asymptotic normal distribution of log(relative risk), the hypothesis test for these two models with equal log (relative risk) can be expressed as

$$\text{H}_0 : \mathcal{V}_{k\ell,c1} = \mathcal{V}_{k\ell,c2},$$

where $\mathcal{V}_{k\ell,c1}$ is the true log(relative risk) in Model_{c1}, and $\mathcal{V}_{k\ell,c2}$ is the true log(relative risk) for Model_{c2}.

As we proved in Appendix B.1.1, the requirement for the modified nonparametric bootstrap method to satisfy asymptotic distribution condition is $n_\ell/\text{B} \to \infty$. It is clear that our data sets met this requirement, since they have $n_1 = 293616$(male),

$n_2 = 198246$(female) and B is in the range $[2000, 2500]$. Therefore, two samples paired t-test can be applied. The test statistic can be written as

$$\mathrm{TS} = (\widehat{\sigma_{d_{k\ell,c1,c2}}})^{-1}\overline{d^b_{k\ell,c1,c2}},$$

where $d^b_{k\ell,c1,c2} = \widehat{\mathcal{V}}^b_{k\ell,c1} - \widehat{\mathcal{V}}^b_{k\ell,c2}$, $\overline{d^b_{k\ell,c1,c2}} = \mathrm{B}^{-1}\sum_{b=1}^{B} d^b_{k\ell,c1,c2}$ and $\widehat{\sigma_{d_{k\ell,c1,c2}}}^2 = (\mathrm{B}-1)^{-1}\sum_{b=1}^{B}(d^b_{k\ell,c1,c2} - \overline{d^b_{k\ell,c1,c2}})^2$.

In the hypothesis test, the test statistic obviously follows the standard student t-distribution which has $B-1$ degree of freedom. Its number of degree of freedom is so large that we can approximately assume the test statistic has a standard normal distribution.

In the following Section 4.2.2, the direct hypothesis test for the relative risks is presented.

4.2.2 Models Comparison for Relative Risk

In the previous Section 4.2.1, the hypothesis test for log(relative risk) had been discussed. Although the conclusion on the log(relative risk) test can be directly transferred to the relationships of the according relative risk, we still present the hypothesis test's procedure which are straightly applied to the relative risk as follows.

According to the asymptotic distribution of log(relative risk), under the same condition $n_\ell/\mathrm{B} \to \infty$, the relative risk can be expressed asymptotically using the delta theorem

$$\mathrm{B}^{1/2}\{\exp(\widehat{\mathcal{V}}^*_{k\ell}) - \exp(\mathcal{V}_{k\ell})\} \backsim \mathrm{Normal}[0, \{\partial\exp(\mathcal{V}_{k\ell})/\partial\mathcal{V}^{\mathrm{T}}_{k\ell}\}\widehat{D}^*_{k\ell}\{\partial\exp(\mathcal{V}_{k\ell})/\partial\mathcal{V}_{k\ell}\}].$$

Furthermore, the null hypothesis which results the equal relative risks for both

Model_{c1} and Model_{c2} can be written as

$$\mathrm{H}_0 : \exp(\mathcal{V}_{k\ell,c1}) = \exp(\mathcal{V}_{k\ell,c2}),$$

where $\exp(\mathcal{V}_{k\ell,c1})$ is the true relative risk in Model_{c1}, and $\exp(\mathcal{V}_{k\ell,c2})$ is the true relative risk for Model_{c2}.

Correspondingly, the statistics for this test is

$$\mathrm{TS} = \mathrm{TS} = (\widehat{\sigma_{d_{k\ell,c1,c2}}})^{-1}\overline{d^b_{k\ell,c1,c2},},$$

where $d^b_{k\ell,c1,c2} = \exp(\widehat{\mathcal{V}}^b_{k\ell,c1}) - \exp(\widehat{\mathcal{V}}^b_{k\ell,c2})$, $\overline{d^b_{k\ell,c1,c2}} = \mathrm{B}^{-1}\sum_{b=1}^{B} d^b_{k\ell,c1,c2}$ and $\widehat{\sigma_{d_{k\ell,c1,c2}}}^2 = (\mathrm{B}-1)^{-1}\sum_{b=1}^{B}(d^b_{k\ell,c1,c2} - \overline{d^b_{k\ell,c1,c2}})^2$.

As same as in Section 4.2.2, the modified nonparametric bootstrap method is utilized to compute the test statistic. Similarly, with the large value of B, the test statistic TS approximately normally distributed.

4.2.3 Confidence Interval of Difference and Ratio of Relative Risk

Another approach to the comparison for the relative risks' statistical difference between two various models, is to construct the confidence intervals for their difference and ratio.

After obtained the relative risks from the modified nonparametric bootstrap method, define their difference and ratio as $\kappa^b_{k\ell} = \exp(\widehat{\mathcal{V}}^b_{k\ell,c1}) - \exp(\widehat{\mathcal{V}}^b_{k\ell,c2})$ and $\gamma^b_{k\ell} = \exp(\widehat{\mathcal{V}}k\ell, c1^b - \widehat{\mathcal{V}}^b_{k\ell,c2})$, the ratio of relative risks, respectively.

Repeat the above procedure for $b = 1, ..., B$, the resulted 95% confidence interval for the difference and ratio of the relative risks are $(\overline{\kappa^b_{k\ell}} \pm 1.96 se_{\kappa^b_{k\ell}})$ and $(\overline{\gamma^b_{k\ell}} \pm 1.96 se_{\gamma^b_{k\ell}})$, respectively.

4.2.4 Likelihood Ratio Test

We also applied the likelihood ratio test to examine whether statistical differences were existing in the four different models. The test statistic equals to two times the negative difference between the log-likelihood functions for the null model and the alternative model. Furthermore, it has a chi-distribution, and its degree of freedom equals to the difference between the numbers of variables for the alternative model and the null model.

4.2.5 Results

The comparisons of relative risks between models are tabulated in Table 4.1, and the results of likelihood ratio tests are presented in Table 4.2. In this study, the modified nonparametric bootstrap method is applied. In total, there are 2500 simulated nonparametric bootstrap data sets being analyzed, in which effect data sets are $B = 2316$. The results in Table 4.1 show that, for the three pairs of models in comparison (model I vs model II, model II vs model IV and model III vs model IV), the relative risks of cohort cancer for female do not have statistical significant difference. A reasonable guess for this, is that, in the questionnaire survey on daily diet, the answers from females are not as accurate as the answers from males (in other word, women usually incline to conceal their true diet information). Turn to the males' relative risks for cohort cancer, when covariates Z are not taken into consideration, the relative risk computed from the overall data set of HEI-2005, has statistical significant difference compared to the relative risk which is obtained from the total score of the HEI-2005 components. If covariates Z were included, in other words, when the FFQ for energy, age, ethnicity, education, body mass index, smoking, physical activity....etc. were taken into consideration, we also come to the same conclusion.

	Model I vs Model II		Model II vs Model IV		Model III vs Model IV	
	Male	Female	Male	Female	Male	Female
$\Delta_{\mathcal{V}_{k\ell}}$	0.09	0.14	0.01	-0.04	0.12	0.15
$\mathrm{TS}_{\Delta_{\mathcal{V}_{k\ell}}}$	2.11	1.50	0.19	-0.67	2.43	1.27
$\text{p-value}_{\Delta_{\mathcal{V}_{k\ell}}}$	0.03	0.13	0.85	0.50	0.02	0.20
$\Delta_{\exp(\mathcal{V}_{k\ell})}$	0.06	0.10	0.00	-0.03	0.08	0.11
$\mathrm{TS}_{\Delta_{\mathcal{V}_{k\ell}}}$	2.10	1.51	0.17	-0.66	2.39	1.27
$\text{p-value}_{\Delta_{\exp(\mathcal{V}_{k\ell})}}$	0.04	0.13	0.86	0.51	0.02	0.20
95 % CI of DRR	(0.00,0.11)	(-0.03,0.23)	(-0.03,0.04)	(-0.11,0.06)	(0.02,0.15)	(-0.06,0.28)
95 % CI of RRR	(1.00,1.18)	(0.94,1.36)	(0.95,1.06)	(0.86,1.07)	(1.02,1.25)	(0.90,1.43)

Table 4.1: Relative risks comparisons between four different models by the modified nonparametric bootstrap method with B=2316. $\Delta_{\mathcal{V}_{k\ell}}$–mean of the difference of log(relative risks). $\mathrm{TS}_{\Delta_{\mathcal{V}_{k\ell}}}$–test statistic of log(relative risks). $\text{p-value}_{\Delta_{\mathcal{V}_{k\ell}}}$–p-value of the hypothesis test for log(relative risks). $\Delta_{\exp(\mathcal{V}_{k\ell})}$–mean of the difference of relative risks. $\mathrm{TS}_{\Delta_{\exp(\mathcal{V}_{k\ell})}}$–test statistics of relative risks. $\text{p-value}_{\Delta_{\exp(\mathcal{V}_{k\ell})}}$–p-value of the hypothesis test for relative risks. DRR–Difference of relative risks. RRR–Ratio of relative risks.

	Model I vs Model II	Model III vs Model IV	Model II vs Model IV
Test Statistics	26.573	33.037	769.691
Degree of freedom	11	11	50
p-value	0.005	0.001	0.000

Table 4.2: Likelihood Ratio tests between four different models, where the test statistics follow chi-square distributions.

On the other hand, when weight factor is assigned to each component of HEI-2005, no matter the computation is based on p-value or on the confidence intervals of difference and ratio, it is clear that the resulted relative risks for male have no statistical significant differences, regardless whether the covariates Z are included or not.

However, if the likelihood ratio tests were performed, all the three pairs of models in comparison (model I vs model II, model II vs model IV and model III vs model IV), show that there are statistical significant differences. The detailed results of the likelihood ratio tests are given in Table 4.2.

4.3 Variable Selection by the Bounded Constrains

The following method is applied to bound all ω_j $(j = 1, \ldots, J)$ as positive.

1. Run the analysis in the original model.

2. Fix α, β, θ, then update ω with constrain that they must be positive.

3. Fix ω, update α, β, θ.

4. Repeat steps 1-3 until it converge

We performed this bounded constrain algorithm by using the Matlab function ‘fmincon’, and results are given in Table 4.3. The estimates from the original data set, their standard errors, and p-values for individual components are also provided for comparison. We see that three components in HEI-2005 were adjusted to zeroes with constrains that $\omega_j > 0$, they are ‘Total Fruit’ with original estimate 0.120 and p-value 0.958, ‘Meats and Beans’ and ‘Saturated Fat’ with original negative estimates -1.646 and -0.706, p-value 0.217 and 0.360, respectively. The other terms are rather near the original estimates and only have trivial changes.

4.4 The Adaptive Lasso Method

We assume that model (3.1) contains both significant and insignificant dietary components and then propose the adaptive lasso for variable selection (Zou, 2006). Let $\mathcal{A} = \{j : \omega_j \neq 0\}$ and further assume that $|\mathcal{A}| = J_0 < J$. Without loss of generality, we assume that $\mathcal{A} = \{1, 2, \ldots, J_0\}$ and a two-step estimating procedure for ω_j is described as follows.

Step I. Let $\alpha = (\alpha_{k\ell} : 1 \leq k \leq K_\ell, 1 \leq \ell \leq L)^{\mathrm{T}}$, $\theta = (\theta_{k\ell}^{\mathrm{T}} : 1 \leq k \leq K_\ell, 1 \leq \ell \leq L)^{\mathrm{T}}$, and $\omega = (\omega_1, \ldots, \omega_J)^{\mathrm{T}}$. By assuming that $\beta_{k\ell}$ are known, denote the negative log-likelihood function as $\mathcal{L}(\alpha, \theta, \omega) = -\mathrm{log}P(Y|X, Z) = -\sum_{\ell=1}^{2}\sum_{k=1}^{K_\ell}\sum_{i=1}^{n_\ell}\{Y_{ik\ell}\mathrm{log}(p_{ik\ell}) +(1 - Y_{ik\ell})\mathrm{log}(1 - p_{ik\ell})\}$, where $p_{ik\ell} = H(\alpha_{k\ell} + \beta_{k\ell}\sum_{j=1}^{J} X_{ij\ell}\omega_j + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell})$. Let $\widehat{\omega}_0 = (\widehat{\omega}_{1,0}, \ldots, \widehat{\omega}_{J,0})^{\mathrm{T}}$ be the unpenalized estimate of ω minimizing the negative likelihood function. Then $\widehat{\omega}_0$ is a root-n consistent estimate of ω (i.e., $\sqrt{N}(\widehat{\omega}_0 - \omega)$ converges to a Normal distribution.). As given in Zou (2006), we pick a $\gamma > 0$ and define the weight vector $\widehat{t} = (\widehat{t}_1, \ldots, \widehat{t}_J)^{\mathrm{T}}$ with $\widehat{t}_j = 1/|\widehat{\omega}_{j,0}|^\gamma$. The estimates of (α, θ, ω) are given by

$$(\widehat{\alpha}, \widehat{\theta}, \widehat{\omega}) = \arg\min_{(\alpha,\theta,\omega)}\{2\mathcal{L}(\alpha, \theta, \omega) + \lambda_N\sum_{j=1}^{J}\widehat{t}_j|\omega_j|\}$$

where λ_N is a regularization parameter controlling the amount of shrinkage, and $\widehat{\omega}$ is the adaptive lasso estimates of ω. Let $\mathcal{A}_N^* = \{j : \widehat{\omega}_j \neq 0\}$. Then $\widehat{\omega} = (\widehat{\omega}_{\mathcal{A}_N^*}^{\mathrm{T}}, \widehat{\omega}_{\mathcal{A}/\mathcal{A}_N^*}^{\mathrm{T}})^{\mathrm{T}}$, where $\widehat{\omega}_{\mathcal{A}_N^*}^{\mathrm{T}} = (\widehat{\omega}_j : j \in \mathcal{A}_N^*)$ and $\widehat{\omega}_{\mathcal{A}/\mathcal{A}_N^*} = (\widehat{\omega}_j : j \in \mathcal{A}/\mathcal{A}_N^*)$.

Step II. We refit the model by using the subset $\mathcal{A}_N^*$ of $(1, \ldots, J)$ selected from Step I. Let $\omega_{\mathcal{A}_N^*}^* = (\omega_j : j \in \mathcal{A}_N^*)$. Then the estimates of $(\alpha, \theta, \omega_{\mathcal{A}_N^*}^*)$ denoted as $(\widetilde{\alpha}, \widetilde{\theta}, \widetilde{\omega}_{\mathcal{A}_N^*}^*)$ are obtained by minimizing $\mathcal{L}(\alpha, \beta, \theta, \omega_{\mathcal{A}_N^*}^*)$ subject to $\omega_j \geq 0$ for all $j \in \mathcal{A}_N^*$, where

$\beta = (\beta_{k\ell} : 1 \leq k \leq K_\ell, 1 \leq \ell \leq L)^{\mathrm{T}}$ and

$$\begin{aligned}
\mathcal{L}(\alpha, \beta, \theta, \omega^*_{\mathcal{A}^*_N}) &= -\textstyle\sum_{\ell=1}^{2}\sum_{k=1}^{K_\ell}\sum_{i=1}^{n_\ell}\{Y_{ik\ell}\log(p^*_{ik\ell}) + (1 - Y_{ik\ell})\log(1 - p^*_{ik\ell})\}, \\
p^*_{ik\ell} &= H(\alpha_{k\ell} + \beta_{k\ell}\textstyle\sum_{\ j\in\mathcal{A}^*_N} X_{ij\ell}\omega_j + Z^{\mathrm{T}}_{ik\ell}\theta_{k\ell}).
\end{aligned}$$

An iteration algorithm to obtain the estimates for the adaptive lasso method is shown in Section 4.4.2.

4.4.1 Oracle Properties of the Adaptive Lasso Estimator

Let $\mathbf{I_{k\ell}}$ be the $\sum_{\ell=1}^{2}K_\ell$-dimensional vector with the $(\sum_{\ell'=1}^{\ell-1}K_{\ell'}+k)^{th}$ element being 1 and others being 0, and $Z^*_{ik\ell} = \{(\mathbf{0}^{\mathrm{T}}_{d_{11}}, \ldots, \mathbf{0}^{\mathrm{T}}_{d_{(k-1)\ell}}), Z^{\mathrm{T}}_{ik\ell}, (\mathbf{0}^{\mathrm{T}}_{d_{(k+1)\ell}}, \ldots, \mathbf{0}^{\mathrm{T}}_{d_{K_L L}})\}^{\mathrm{T}}$, where $d_{k\ell}$ is the dimension of $Z_{ik\ell}$. Let $d = \sum_{\ell=1}^{2}K_\ell + \sum_{k,\ell}d_{k\ell} + J$. Define $(Q_{ikl})_{d\times 1} = (\mathbf{I}^{\mathrm{T}}_{kl}, {Z^*_{ik\ell}}^{\mathrm{T}}, \beta_{k\ell}X^{\mathrm{T}}_{i\ell})^{\mathrm{T}}$, $V_{ikl} = p_{ikl}(1 - p_{ikl})$, $\mathbf{Q}_{N\times d} = \{Q_{ikl} : 1 \leq i \leq n_\ell, 1 \leq k \leq K_\ell, 1 \leq \ell \leq L\}$, and $\mathbf{V}$ is a $N \times N$ diagonal matrix with V_{ikl} as its diagonal elements, where $N = \sum_{\ell=1}^{2}K_\ell n_\ell$. Assume that $\mathbf{Q}^{\mathrm{T}}\mathbf{V}\mathbf{Q}/N \to \mathbf{\Sigma}$, and let $\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}$, where $\mathbf{\Sigma}_{11}$ is a $d_0 \times d_0$ matrix and $d_0 = \sum_{\ell=1}^{2}K_\ell + \sum_{k,\ell}d_{kl} + J_0$.

Lemma 6 *Suppose that $\lambda_N/\sqrt{N} \to 0$ and $\lambda_N N^{(\gamma-1)/2} \to \infty$. Then the adaptive lasso estimates satisfy:*

i) $\lim_N P(\mathcal{A}^*_N = \mathcal{A}) = 1$,

ii) $\sqrt{N}\{(\widehat{\alpha}^{\mathrm{T}}, \widehat{\theta}^{\mathrm{T}}, \widehat{\omega}^{\mathrm{T}}_{\mathcal{A}})^{\mathrm{T}} - (\alpha^{\mathrm{T}}, \theta^{\mathrm{T}}, \omega^{\mathrm{T}}_{\mathcal{A}})^{\mathrm{T}}\} \to Normal(\mathbf{0}, \mathbf{\Sigma}^{-1}_{11})$,

where $\widehat{\omega}_{\mathcal{A}} = (\widehat{\omega}_j : j \in \mathcal{A})$ and $\omega_{\mathcal{A}} = (\omega_j : j \in \mathcal{A})$.

Proof details of Lemma 6 are provided in Appendix C.1.

4.4.2 Algorithm for L_1 Norm Penalty in Non-linear Regression

Generally, assume our target is to minimum the function with L_1 norm as

$$\text{argmin}\{\Phi(\Theta) + \textstyle\sum_{j=1}^{J}\lambda_j|\omega_j|\}, \tag{4.1}$$

where Θ contains all unknown parameters, ω_j denotes parameter with penalty for $j = 1, ..., J$, and λ_j is the tuning parameter for ω_j.

Let $\widehat{\Theta}^*$ be the minimizer to function $\Phi(\Theta)$, as

$$\widehat{\Theta}^* = \text{argmin}_\Theta\{\Phi(\Theta)\}.$$

By Taylor expansion of $\Phi(\Theta)$ at $\widehat{\Theta}^*$, we have

$$\begin{aligned}\Phi(\Theta) &= \Phi(\widehat{\Theta}^*) + \{\partial\Phi(\widehat{\Theta}^*)/\partial\Theta\}^{\mathrm{T}}(\Theta - \widehat{\Theta}^*) + 1/2(\Theta - \widehat{\Theta}^*)^{\mathrm{T}}\{\partial^2\Phi(\widehat{\Theta}^*)/\partial\Theta^2\}(\Theta - \widehat{\Theta}^*)\\ &= C(\widehat{\Theta}^*) + (\Theta - \widehat{\Theta}^*)^{\mathrm{T}}V(\widehat{\Theta}^*)(\Theta - \widehat{\Theta}^*),\end{aligned}$$

where $C(\widehat{\Theta}^*) = \Phi(\widehat{\Theta}^*)$ and $V(\widehat{\Theta}^*) = 1/2\{\partial^2\Phi(\widehat{\Theta}^*)/\partial\Theta^2\}$, both are constants for unknown parameters Θ; and $V(\widehat{\Theta}^*)$ is symmetric and non-negative defined.

Therefore, Eq(4.1) can be written as

$$\text{argmin}\{(\Theta - \widehat{\Theta}^*)^{\mathrm{T}}V(\widehat{\Theta}^*)(\Theta - \widehat{\Theta}^*) + \textstyle\sum_{j=1}^{J}\lambda_j|\omega_j|\}. \tag{4.2}$$

Separate Θ into two parts $\Theta_{-\omega}$ and ω, where $\omega = (\omega_1, ..., \omega_J)^{\mathrm{T}}$ are parameters with penalty and $\Theta_{-\omega}$ contains all other parameters except ω which do not have penalty. Without loss of generality, Eq(4.2) can be re-written as

$$\begin{aligned}
&\operatorname{argmin}(\{(\Theta_{-\omega}-\widehat{\Theta}^*_{-\omega})^{\mathrm{T}},(\omega-\widehat{\omega}^*)^{\mathrm{T}}\}\begin{bmatrix} V_{\Theta_{-\omega}} & V_{\Theta_{-\omega},\omega} \\ V_{\omega,\Theta_{-\omega}} & V_{\omega}\end{bmatrix} \\
&\qquad\{(\Theta_{-\omega}-\widehat{\Theta}^*_{-\omega})^{\mathrm{T}},(\omega-\widehat{\omega}^*)^{\mathrm{T}}\}^{\mathrm{T}}+\textstyle\sum_{j=1}^{J}\lambda_j|\omega_j|) \\
&=\operatorname{argmin}[(\Theta_{-\omega}-\widehat{\Theta}^*_{-\omega})^{\mathrm{T}}V_{\Theta_{-\omega}}(\Theta_{-\omega}-\widehat{\Theta}^*_{-\omega})+2(\Theta_{-\omega}-\widehat{\Theta}^*_{-\omega})^{\mathrm{T}}V_{\Theta_{-\omega},\omega}(\omega-\widehat{\omega}^*) \\
&\qquad+(\omega-\widehat{\omega}^*)^{\mathrm{T}}V_{\omega}(\omega-\widehat{\omega}^*)+\textstyle\sum_{j=1}^{J}\lambda_j|\omega_j|].
\end{aligned}$$

Consequently, based on the current estimate $\tilde{\omega}$ for ω, we update $\Theta_{-\omega}$ by

$$\Theta_{-\omega}=\widehat{\Theta}^*_{-\omega}-(V_{\Theta_{-\omega}})^{-1}V_{\Theta_{-\omega},\omega}(\tilde{\omega}-\widehat{\omega}^*). \tag{4.3}$$

For updating parameter ω_j, assume current estimates for other parameters $\Theta_{-\omega}$ and ω_p with $p\neq j$ are $\tilde{\Theta}_{-\omega}$ and $\tilde{\omega}_p$, respectively. Define ω_j is corresponding to the ${h_j}^{th}$ element in Θ, hence, for fix current estimates of $\Theta_{-\omega}$ and ω_p, we minimize the following equation

$$\operatorname{argmin}\{2(\omega_j-\widehat{\omega}^*_j)\textstyle\sum_{m\neq h_j}V_{h_jm}(\tilde{\Theta}_m-\widehat{\Theta}^*_m)+(\omega_j-\widehat{\omega}^*_j)^2V_{h_jh_j}+\lambda_j|\omega_j|\}.$$

Assume $\omega_j>0$, the partial of above function to ω_j is

$$2\textstyle\sum_{m\neq h_j}V_{h_jm}(\tilde{\Theta}_m-\widehat{\Theta}^*_m)+2(\omega_j-\widehat{\omega}^*_j)V_{h_jh_j}+\lambda_j.$$

Thus, the solution of ω_j with $\omega_j>0$ is

$$\begin{cases}\widehat{\Theta}^*_m-\sum_{m\neq h_j}V_{h_jm}(\tilde{\Theta}_m-\widehat{\Theta}^*_m)/V_{h_jh_j}-\lambda_j/(2V_{h_jh_j}) & , \\ \qquad\text{if } \lambda_j/(2V_{h_jh_j})<\widehat{\Theta}^*_m-\sum_{m\neq h_j}V_{h_jm}(\tilde{\Theta}_m-\widehat{\Theta}^*_m)/V_{h_jh_j}; & \\ 0 & ,\text{otherwise.}\end{cases}$$

Similarly, we obtain the solution of ω_j with $\omega_j < 0$

$$\begin{cases} \widehat{\Theta}^*_m - \sum_{m \neq h_j} V_{h_j m}(\tilde{\Theta}_m - \widehat{\Theta}^*_m)/V_{h_j h_j} + \lambda_j/(2V_{h_j h_j}) & , \\ \qquad \text{if } \lambda_j/(2V_{h_j h_j}) < -(\widehat{\Theta}^*_m - \sum_{m \neq h_j} V_{h_j m}(\tilde{\Theta}_m - \widehat{\Theta}^*_m)/V_{h_j h_j}); & \\ 0 & \text{, otherwise.} \end{cases}$$

In practice since the sign of ω_j is unknown, a reasonable estimate is the sign of $\widehat{\Theta}^*_m - \sum_{m \neq h_j} V_{h_j m}(\tilde{\Theta}_m - \widehat{\Theta}^*_m)/V_{h_j h_j}$ because $\widehat{\Theta}^*_m - \sum_{m \neq h_j} V_{h_j m}(\tilde{\Theta}_m - \widehat{\Theta}^*_m)/V_{h_j h_j}$ is the estimate of ω_j to minimize function $\Phi(\Theta)$ based on current estimates of all other parameters.

We use $\widehat{\Theta}^*$ as the initial guess of Θ, update ω_j one by one for $j = 1, ..., J$, and then update $\Theta_{-\omega}$. Repeat the updating process until it converges.

4.4.3 Details of Procedure

Details of variable selection by the adaptive Lasso are as follows.

1. Run the analysis in the original model.

2. Run the adaptive Lasso only the ω's to update the estimates of the all unknown parameters.

   (a) Separate the data in 10-folds.

   (b) For fixed λ_N, by the modified scoring method, use 9 folds are training data, and 1 fold as test data, to calculate the prediction errors.

   (c) Repeat it until all folds has be treated as test data, add all prediction errors together for this λ_N.

   (d) Repeat all interested values of λ_N, and find out the value of λ_N which minimize the prediction errors, define as λ_0.

3. For λ_0, use all data set to calculate all unknown parameter except $\beta_{k\ell}$, and remove any component score in HEI 2005 whose $\omega_j = 0$.

4. Refit the original model (3.1) subject to $\omega_j \geq 0$ with all parameters except the removed components in ω by using the procedure described in Section 4.3.

4.5 Result of the Adaptive Lasso

During the procedure of cross validation, our calculations for the colorectal cancer data in which 293615 men with 2151 cases and 198245 women with 959 cases, show that this model apparently has the smallest prediction error with tuning parameter $\lambda_N = 0.5$. Figures C.1 and C.2 in Appendix C.2 show the prediction error versus tuning parameter λ_N. For better understanding the relationship between tuning parameter value and performance of the adaptive lasso algorithm, we compared analysis results by using procedure in Section (4.4.3) with six different values of λ_N at $(0.100, 0.250, 0.500, 1.000, 2.000, 5.000)$ with $\gamma = 1.5$.

Table 4.4 gives the results of the colorectal cancer on HEI component scores of men and women together. If one of the terms in ω is removed by the 3^{th} step of the methodology introduced in Section 4.4.3, then its position in the table will be empty. While, if this term is kept through the 3^{th} step, but equals to 0 after the bounded constrain in the 4^{th} step, then its value is written as 0.000 in the corresponding position. As λ_N value increases, more and more terms in ω will be removed or written as 0.000. Furthermore, terms in ω with negative MLE and large p-values usually are kept by the adaptive lasso with small λ_N and adjusted to 0.000 by the positive constrain, and then often been removed by the adaptive lasso as λ_N increases. It is clear that whether one component will be removed or not is depending on the value of λ_N, the individual p-value of this component and the sign of its MLE.

Next, we introduce an algorithm to compute $\widehat{\omega}_{HEI}$ as follows.

- Obtain $\widehat{\omega}$ by the score method.

- Apply the the adaptive lasso method to get $\widehat{\omega}_{ALS}$.

- Remove j^{th} HEI component if $\widehat{\omega}_{ALS,j} = 0$.

- Obtain $\widehat{\omega}_{New}$ by refitting model with residual components and constrain that $\widehat{\omega}_{New} \geq 0$.

- $\widehat{\omega}_{HEI} = \widehat{\omega}_{New} \times /(X_{max}^{\mathrm{T}} \widehat{\omega}_{New})$.

- Use the delta method to get the estimated variance for $\widehat{\omega}_{HEI}$.

Table 4.5 gives the result of $\widehat{\omega}_{HEI}$ for the colorectal cancer data set, in which there are 293615 men with 2151 cases and 198245 women with 959 cases. Table 4.6 shows the result for the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, in which both genders have 38 covariates on smoking except HEI scores. For eliminating the multicollinearity between Total Fruit and Whole Fruit, Total Grains and Whole Total Grains, and Total Vegetables and DOL, Table 4.7 displays the $\widehat{\omega}_{HEI}$ for the later data set on 'Whole Fruit', 'Total Fruit - Whole Fruit', 'Whole Grains', 'Total Grains - Whole Grains', 'DOL' and 'Total Vegetables - DOL', where other HEI score components are kept the same.

It is clear that in the regression of colorectal cancer data, 'Milk', 'Whole Grains' and 'Oil' are three key factors of nutrition components and have significant effects on the disease. Table 4.6 show that in the regression of the colorectal and lung cancers on HEI component scores of men and the breast, colorectal and lung cancers on women together, key factor of nutrition elements are 'Milk', 'Total Grains', 'Sodium', 'Oil', ‘SoFAAS’ and ‘DOL’ and they have significant effects on the diseases. If we change components of ‘Total Fruit’, ‘Total Grains’ and ‘Total Vegetables’ to ‘Total Fruit - Whole Fruit’, ‘Total Grains - Whole Grains’ and ‘Total Vegetables - DOL’, Table 4.7 shows that ‘Whole Grains’, ‘Total Grains - Whole Grains’, ‘Milk’, ‘Sodium’, ‘Oil’, ‘Whole Fruit’ and ‘SoFAAS’ have significant effects in the regression of the latter data set.

Table 4.8 gives the result of $\widehat{\beta}_{k\ell}$, their standard error and p-values for the unweighted model and weighted model, respectively. The Unweighted Model is expressed as $\mathrm{pr}(Y_{ik\ell} = 1|X_{ij\ell}, Z_{ik\ell}) = H(\alpha_{k\ell} + \beta_{k\ell}\sum_{j=1}^{J} X_{ij\ell} + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell})$, and the weighted Model is $\mathrm{pr}(Y_{ik\ell} = 1|X_{ij\ell}, Z_{ik\ell}) = H(\alpha_{k\ell} + \beta_{k\ell}\sum_{j=1}^{J} X_{ij\ell}\widehat{\omega}_j + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell})$, where $\widehat{\omega}_j$ the MLE. The likelihood ratio test is performed for model comparison. Among these five disease, only the colorectal cancer of female has significant difference between the unweighted model and weighted model.

Component	Estimate bounded constrain	Estimate Original	s.e by Sandwich	p-value
Total Fruit	0.000	0.120	2.266	0.958
Whole Fruit	3.072	3.189	2.129	0.134
Total Grains	3.264	3.375	2.529	0.182
Whole Grains	6.069	6.059	2.020	0.003
Total Vegetables	0.074	0.185	2.581	0.943
DOL	2.285	2.458	1.801	0.172
Milk	3.011	2.785	0.766	0.000
Meats and Beans	0.000	-1.646	1.334	0.217
Oil	1.697	1.748	0.797	0.028
Saturated Fat	0.000	-0.706	0.772	0.360
Sodium	0.310	0.007	1.288	0.996
SoFAAS	0.016	0.077	0.482	0.873

Table 4.3: Variable selection results by the bounded constrain algorithm for the logistic regression of the colorectal cancer on HEI component scores of men and women together, where $\beta_{11} = -1, \widehat{\beta}_{12} = -0.7411$.

	$\widehat{\omega}$	$\widehat{\omega}$	$\widehat{\omega}$	$\widehat{\omega}$	$\widehat{\omega}$	$\widehat{\omega}$	MLE	p-value
λ_N	0.100	0.250	0.500	1.000	2.000	5.000		
Total Fruit							0.120	0.958
Whole Fruit	3.182	3.170	3.160	3.174	3.158	3.156	3.189	0.134
Total Grains	3.034	3.005	3.014	3.017	2.999	2.994	3.375	0.182
Whole Grains	6.109	6.106	6.100	6.115	6.112	6.107	6.059	0.003
Total Vegetables							0.185	0.943
DOL	2.187	2.185	2.188	2.187	2.188	2.187	2.458	0.172
Milk	2.988	2.990	2.990	2.990	2.990	2.986	2.785	0.000
Meats and Beans	0.000	0.000	0.000	0.000	0.000		-1.646	0.217
Oil	1.701	1.687	1.682	1.704	1.689	1.682	1.748	0.028
Saturated Fat	0.000	0.000	0.000	0.000			-0.706	0.360
Sodium							0.007	0.996
SoFAAS							0.077	0.873

Table 4.4: Variable selection results by using the adaptive lasso method and positive constrain for the logistic regression of the colorectal cancer on HEI component scores of men and women together, where $\beta_{11} = -1, \widehat{\beta}_{12} = -0.7411$ and $\gamma = 1.5$. MLE–maximum likelihood estimator.

	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	MLE	p-value
λ_N	0.100	0.250	0.500	1.000	2.000	5.000		
Total Fruit							0.120	0.958
Whole Fruit	2.658	2.657	2.656	2.656	2.652	2.654	3.189	0.134
Total Grains	2.536	2.522	2.531	2.523	2.518	2.518	3.375	0.182
Whole Grains	5.123	5.132	5.129	5.125	5.133	5.136	6.059	0.003
Total Vegetables							0.185	0.943
DOL	1.836	1.838	1.837	1.834	1.838	1.839	2.458	0.172
Milk	2.510	2.510	2.508	2.506	2.511	2.512	2.785	0.000
Meats and Beans	0.000	0.000	0.000	0.000	0.000		-1.646	0.217
Oil	1.414	1.415	1.415	1.425	1.418	1.415	1.748	0.028
Saturated Fat	0.000	0.000	0.000	0.000			-0.706	0.360
Sodium							0.007	0.996
SoFAAS							0.077	0.873

Table 4.5: Analysis results of $\widehat{\omega}_{HEI}$ by the adaptive lasso method for the logistic regression of the colorectal cancer on HEI component scores of men (293615 with 2151 cases) and women (198245 with 959 cases) together, where $\beta_{11} = -1, \widehat{\beta}_{12} = -0.7411$ and $\gamma = 1.5$. ω_{HEI}–algorithm is in Section 4.5. MLE–maximum likelihood estimator.

	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	MLE	p-value
λ_N	0.100	0.250	0.500	1.000	2.000	5.000		
Total Fruit	0.314						0.551	0.716
Whole Fruit	1.440	1.613	1.613	1.610	1.657	1.855	2.477	0.081
Total Grains	3.918	3.851	3.854	4.047	4.118	3.822	6.772	0.000
Whole Grains	0.497	0.480	0.483				0.870	0.531
Total Vegetables							0.097	0.956
DOL	1.581	1.545	1.552	1.524	1.555	1.334	2.707	0.034
Milk	1.962	1.938	1.934	1.926	1.937	2.047	3.384	0.000
Meats and Beans	0.504	0.497	0.490	0.445			0.876	0.323
Oil	0.965	0.938	0.939	0.912	0.962	0.967	1.668	0.004
Saturated Fat	0.792	0.804	0.808	0.787	0.654		1.390	0.128
Sodium	0.937	0.939	0.931	0.938	0.906	1.048	1.621	0.003
SoFAAS	0.483	0.493	0.496	0.505	0.515	0.436	0.831	0.015

Table 4.6: Analysis results for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1, \gamma = 1.5$. $\widehat{\omega}_{HEI}$–algorithm is in Section 4.5. MLE–maximum likelihood estimator.

	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	$\widehat{\omega}_{HEI}$	MLE	p-value
λ_N	0.100	0.250	0.500	1.000	2.000	5.000		
Whole Fruit	2.004	2.007	1.991	2.048	2.122	2.130	3.029	0.013
Total Fruit - Whole Fruit	0.000						-0.551	0.716
Whole Grains	1.907	1.902	1.903	1.917	1.886	1.878	7.642	0.000
Total Grains - Whole Grains	0.000	0.000	0.000	0.000	0.000	0.000	-6.772	0.000
DOL	1.342	1.350	1.343	1.349	1.395	1.365	2.804	0.067
Total Vegetables - DOL							-0.097	0.956
Milk	2.130	2.118	2.109	2.166	2.181	2.200	3.384	0.000
Meats and Beans	0.608	0.624	0.648	0.548			0.876	0.323
Oil	1.048	1.045	1.040	1.065	1.129	1.123	1.668	0.004
Saturated Fat	0.251	0.254	0.266	0.211	0.044		1.390	0.128
Sodium	1.106	1.110	1.107	1.114	1.091	1.095	1.621	0.003
SoFAAS	0.704	0.698	0.694	0.708	0.736	0.735	0.831	0.015

Table 4.7: Analysis results for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1, \gamma = 1.5$. ω_{HEI}–algorithm is in Section 4.5. MLE–maximum likelihood estimator.

		Unweighted Model			Weighted Model			Model Comparison	
Gender	Disease	$\widehat{\beta}_{k\ell}$	s.e	p-value	$\widehat{\beta}_{k\ell}$	s.e	p-value	LR	p-value
Male	Colorectal Cancer	-1.413	0.153	0.000	-1.734	0.163	0.000	26.620	0.000
Male	Lung Cancer	-0.862	0.138	0.000	-0.721	0.147	0.000	-14.908	1.000
Female	Breast Cancer	-0.003	0.122	0.979	-0.025	0.128	0.847	0.036	0.850
Female	Colorectal Cancer	-1.034	0.219	0.000	-1.330	0.231	0.000	10.616	0.001
Female	Lung Cancer	-0.523	0.169	0.002	-0.607	0.179	0.001	1.858	0.173

Table 4.8: Analysis results of $\widehat{\beta}_{k\ell}$ for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1, \gamma = 1.5$. Unweighted Model–pr$(Y_{ik\ell} = 1|X_{ij\ell}, Z_{ik\ell}) = H(\alpha_{k\ell} + \beta_{k\ell}\sum_{j=1}^{J} X_{ij\ell} + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell})$, Weighted Model–pr$(Y_{ik\ell} = 1|X_{ij\ell}, Z_{ik\ell}) = H(\alpha_{k\ell}+\beta_{k\ell}\sum_{j=1}^{J} X_{ij\ell}\widehat{\omega}_j+Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell})$. LR–Likelihood ratio test statistic, log-likelihood(weighted model)−log-likelihood(unweighted model), which follows a chi-squared distribution with degree freedom 1.

5. VARIABLE SELECTION IN DIETARY INDEX MODELING FOR HEI-2005

5.1 Introduction

Naturally, the next thing we are interested is what the function would be if the nutrition component related part in the logistic regression turns out to be nonlinear. More than that, one should also note there are limitations from the knowledge of biology and nutrition. In this Chapter, we establish a logistic regression model to satisfy all these requirements, and apply it to the real HEI-2005 data set.

Since our interests focus on the effects of increasing nutrition component values on various diseases across different subpopulations, a logistic regression is built as

$$\mathrm{pr}(Y_{ik\ell} = 1 | X_{i1\ell}, ..., X_{iJ\ell}, Z_{ik\ell}) = H\{\alpha_{k\ell} + \beta_{k\ell} \textstyle\sum_{j=1}^{J} m_j(X_{ij\ell}) + Z_{ik\ell}^{\mathrm{T}} \beta_{zk\ell}\},$$

where $m_j(\cdot)$ for $j = 1, ..., J$ are modeled to satisfy some crucial constrains coming from the biology and nutrition, which are listed as follows.

- Monotonicity: The dietary components are chosen so that larger values are meant to denote a lower chance of disease, so that the functions $m_1(\cdot), ..., m_J(\cdot)$ are all required to be monotone nondecreasing. It makes sense to reorder the components so that increasing the score does not increase risk.

- Positivity: The functions $m_1(\cdot), ..., m_J(\cdot)$ are all required to be nonnegative, to mimic the HEI-2005 and other dietary pattern scores.

- Upper Bounds: The nutritionists believe that above a certain level, no extra benefit to health occurs by exceeding that level. Call the level for dietary component j as c_{Pj}, where the P stands for protective. This means that for

$x > c_{Pj}$, $m_j(x) = m_j(c_{Pj})$.

- Lower Bounds: The nutritionists also believe that below a certain level, no extra harm to health occurs by being below that level. Call the level for dietary component j as c_{Rj}, where the R stands for risk. This means that for $x < c_{Rj}$, $m_j(x) = m_j(c_{Rj})$.
- Bad Diets: To mimic the HEI-2005 score, there is the constraint that $m_j(c_{Rj}) = 0$.

In Section 5.2, we describe an expression of $m_j(x)$ with I-spline basis functions to satisfy these constraints, also the identifiability of the regression model; and then present methodology to obtain estimates and their estimated variance. Applications of these methodologies and discussions are illustrated in Section 5.3.

5.2 I-spline Basis Function and Regression Model

5.2.1 I-spline Basis Function

I-splines is defined as the integration of M-splines, which is always nonnegative and the full field integration is 1. Therefore, I-splines is always nonnegative and monotonic. Assume the p^{th} order I-splines is define in an interval with e interior knots, then totally there are $2p + e$ knots defined as

$$
\begin{aligned}
& t_1 \leq t_2 ... \leq t_{2p+e}, \\
& t_1 = t_2 = ... = t_p, t_{p+e+1} = t_{p+e+2} = ... = t_{2p+e}, \\
& t_m < t_{m+p} \qquad \text{for all } m,
\end{aligned}
$$

where t's are values of knots.

In the work of Ramsay (1988), expressions of I-spline basis functions for $p = 2$

are provided for $x \in [t_m, t_m + 1)$ with $t_m < t_{m+1}$, and they are piecewise quadratic functions as

$$\begin{aligned} I_m(x|p=2,t) &= (x - t_m)^2/\{(t_{m+1} - t_m)(t_{m+2} - t_m)\}, \\ I_{m-1}(x|p=2,t) &= 1 - (t_{m+1} - x)^2/\{(t_{m+1} - t_m)(t_{m+1} - t_{m-1})\}. \end{aligned}$$

Based on that, I derive the basis functions for the 3^{th} order I-splines for $x \in [t_m, t_m + 1)$ as follows, which are piecewise cubic

$$\begin{aligned} I_m(x|p=3,t) &= (x - t_m)^3/\{(t_{m+1} - t_m)(t_{m+2} - t_m)(t_{m+3} - t_m)\}, \\ I_{m-2}(x|p=3,t) &= 1 - (t_{m+1} - x)^3/\{(t_{m+1} - t_{m-2})(t_{m+1} - t_{m-1})(t_{m+1} - t_m)\}, \\ I_{m-1}(x|p=3,t) &= (t_m - t_{m-1})^2/\{(t_{m+1} - t_{m-1})(t_{m+2} - t_{m-1})\} + a_1/b_1 + a_2/b_2. \end{aligned}$$

where

$$\begin{aligned} a_1 &= 3/2(t_{m+1} + t_{m-1})(x^2 - t_m^2) - 3t_{m-1}t_{m+1}(x - t_m) - (x^3 - t_m^3), \\ a_2 &= 3/2(t_{m+2} + t_m)(x^2 - t_m^2) - 3t_m t_{m+2}(x - t_m) - (x^3 - t_m^3), \\ b_1 &= (t_{m+1} - t_m)(t_{m+1} - t_{m-1})(t_{m+2} - t_{m-1}), \\ b_2 &= (t_{m+1} - t_m)(t_{m+2} - t_m)(t_{m+2} - t_{m-1}). \end{aligned}$$

5.2.2 Regression Model

Assume we have the same number of interior points for all nutrition components, named as e. And the interior knots on the interval $[c_{Rj}, c_{Pj}]$ are equally spaced. Consequently, the distance of the interior points for j^{th} component is $d_j = (c_{Pj} - c_{Rj})/(e + 1)$. With these specifications, the special expression of I-splines functions are given in Appendix D.1 detailedly.

Furthermore, an easy way to satisfy the requirements of 'Upper Bounds' and

'Lower Bounds' for function $m_j(x)$ is to transfer the HEI density values x by the following expression

$$x \to \begin{cases} c_{Rj} & \text{for } x < c_{Rj}; \\ x & \text{for } x \in [c_{Rj}, c_{Pj}]; \\ c_{Pj} & \text{for } x > c_{Pj}, \end{cases}$$

before inputting it into $m_j(x)$.

Next, we apply the I-splines into our logistic regression model with p^{th} order and e interior points, where p is the order of I-splines, e is the number of interior knots and we assume p and e are the same for all nutrition components. Let the I-splines basis functions defined on the j^{th} component interval $[c_{Rj}, c_{Pj}]$ be $I_{jq}(\cdot)$ for $q = 1, ..., Q$, where $Q = p + e$. Therefore, after transforming all j^{th} component values into the interval $[c_{Rj}, c_{Pj}]$, the $m_j(x)$ is defined as $m_j(x) = \sum_{q=1}^{Q} I_{jq}(x)\exp(\gamma_{jq})$, where we have $\sum_{q=1}^{Q} I_{jq}(c_{Rj})\exp(\gamma_{jq}) = 0$ and $\sum_{q=1}^{Q} I_{jq}(c_{Pj})\exp(\gamma_{jq}) = \sum_{q=1}^{Q}\exp(\gamma_{jq})$ based on definitions of I-splines.

In consequence, we construct a logistic regression model as

$$\text{pr}(Y_{ik\ell} = 1 | X_{i1\ell}, ..., X_{iJ\ell}, Z_{ik\ell}) = H[\alpha_{k\ell} + \beta_{k\ell}\textstyle\sum_{j=1}^{J}\{\sum_{q=1}^{Q} I_{jq}(X_{ij\ell})\exp(\gamma_{jq})\} + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell}],$$

with $\beta = -1$ for identifiability.

Obviously, this regression model satisfies the requirements of 'Bad Diets' since $m_j(c_{Rj}) = \sum_{w=1}^{W} I_{jw}(c_{Rj})\exp(\gamma_{jq}) = 0$, 'Monotonicity' and 'Positivity'.

5.2.3 Iteration Algorithm and Variance Estimation

For the regression model in Section 5.2.2, $m_j(x)$ could be considered as a linear function of ω_{jq} with positive constrains of ω_{jq}, where $\omega_{jq} = \exp(\gamma_{jq})$. Therefore, the domain of ω_{jq} is $[0, +\infty)$ not R. The scoring method does not converge here because it always tends to make some ω_{jq} to the negative domain, which leads to some γ_{jq} go to negative infinity.

Rewrite the previous model as

$$\text{pr}(Y_{ik\ell} = 1|X_{i1\ell}, ..., X_{iJ\ell}, Z_{ik\ell}) = H[\alpha_{k\ell} + \beta_{k\ell}{\textstyle\sum_{j=1}^{J}}\{{\textstyle\sum_{q=1}^{Q}} I_{jq}(X_{ij\ell})\omega_{jq}\} + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell}] \quad (5.1)$$

with $\beta_{11} = -1$ and constrain $\omega_{jq} \geq 0$. Upon that, we introduce the following iteration algorithm which has good performances in convergence. First, define all unknown parameters in previous regression model as Θ with $\Theta = (\alpha_{11}, \alpha_{21}, ..., \beta_{21}, ..., \theta_{11}, ..., \omega_{11}, ..., \omega_{JQ})^{\mathrm{T}}$, $\omega_j = (\omega_{j1}, ..., \omega_{jQ})^{\mathrm{T}}$ and $\omega = (\omega_1^{\mathrm{T}}, ..., \omega_J^{\mathrm{T}})^{\mathrm{T}}$. Let $\Theta_{-\omega_j}$ denote all parameters except ω_j, and $\Theta_{-\omega}$ denote all parameters except all ω_j for $j = 1, ..., J$ as $\Theta_{-\omega} = (\alpha_{11}, \alpha_{21}, ..., \beta_{21}, ..., \theta_{11}, ..., \theta_{KL})^{\mathrm{T}}$. Here we give the steps of our iterative algorithm to obtain the estimated values for Θ.

1. Obtain the initial values of the Θ.

2. Define current estimate for ω as $\tilde{\omega}$, update $\Theta_{-\omega}$ for fixed $\tilde{\omega}$.

3. Define current estimate for $\Theta_{-\omega_j}$ as $\tilde{\Theta}_{-\omega_j}$, update ω_j for fixed $\tilde{\Theta}_{-\omega_j}$ with constrain $\omega_{jq} \geq 0$.

4. Repeat Step 3 for $j = 1, ..., J$.

5. Repeat Steps 2$\sim$4 until it converges.

Matlab program 'fmincon' is used in our real data analysis of the above algorithm.

Additionally, we could use the inverse Fisher matrix or the sandwich method to estimate the variance of $\widehat{\Theta}$. Depends on our analysis results, the estimated variance for the estimates are very close so that in this paper, we only use the previous one.

5.3 Results and Discussion

We apply the algorithm introduced in Section 5.2.3 to the real HEI-2005 data. In this data set, there are 219612 males with 3348 individuals have colorectal cancer and 4187 have lung cancer; and for 169480 females, there are 6647 individuals with breast cancer, 1846 with colorectal cancer and 2933 with lunch cancer, respectively. The concerned 12 nutrition components were presented in Table 1. Furthermore, this data set also included 38 smoking related covariates for both male and female samples.

The results from I-spline analysis are presented in Tables 5.1∼5.2, and Figures D.1∼D.2 in Appendix D.2. From these two tables, we can clearly see, for some nutrition components, their corresponding p-values will be changed as the intake amount changes. Such as 'Milk', comparing with the first quintile, the p-values for quintiles 2∼5 are (0.030, 0.003, 0.016, 0.000), which indicates this component has statistical significance on intake amount and suggests people should be better intake more milk to the recommended amount so that the best results can be achieved. Another two examples for such significant changes when comparing with quintile 1 is 'Oil' and 'Sodium'. Especially for 'Sodium', its p-values has significant decreases as the intake amount increases, and finally reaches the statistically significant influence level. Comparing quintiles 2-5 to quintile 1, we can still observe trends from some components, even without obtaining statistical significance. For example, the p-values of 'DOL' and 'Total Grains' are obviously decreasing with the increasing of intake amount, although there are no statistical significances. However, when compared with quintile 1, the significance in disease related to quintile 5 intake amount is clearly larger than the quintile 2 intake amount. Also, there are some components which are not so sensitive to the intake amount, such as 'Whole Gain', ‘Meats and Bean’, etc.

In order to compare between the results from segmented I-spline with positive constraint and the results from the following model as in Reedy’s paper, based on the different subpopulations, we present the likelihood ratio tests’ results for both two models of each disease. The results are summarized in Table 5.3. Obviously, for the lung cancer in female, and the colorectal cancer in both male and female, I-spline model with positive constraint has significant difference to the total score model. While for the rest two, lung cancer in male and breast cancer in female, there is no statistical significance between them. A possible reason for this is because the positive constraints have been applied.

Furthermore, to eliminate the multi-collinearity in the data set, (‘Total Fruit’, ‘Whole Fruit’, ‘Total Grains’, ‘Whole Grains’, ‘Total Vegetables’, ‘DOL’) in the nutrition components are changed to (‘Whole Fruit’, ‘Total Fruit- Whole Fruit’, ‘W-hole Grains’ , ‘Total Grains - Whole Grains’, ‘DOL’, ‘Total Vegetables - DOL’) And then Tables D.1∼D.3 and Figures D.3∼D.4 summarize and present the corresponding analysis results for them. From which one can easily see, the trends for these modified components are very similar to the original nutrition components.

		Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Total Fruit	Estimate	0.107	0.050	0.055	0.055	0.055
	s.e (vs quintile 1)		0.084	0.074	0.084	0.075
	p-value (vs quintile 1)		0.557	0.453	0.511	0.463
Whole Fruit	Estimate	0.000	0.007	0.035	0.041	0.041
	s.e (vs quintile 1)		0.082	0.075	0.086	0.070
	p-value (vs quintile 1)		0.933	0.644	0.631	0.556
Total Grains	Estimate	0.028	0.002	0.012	0.056	0.116
	s.e (vs quintile 1)		0.134	0.131	0.138	0.138
	p-value (vs quintile 1)		0.990	0.927	0.682	0.403
Whole Grains	Estimate	0.213	0.014	0.018	0.018	0.018
	s.e (vs quintile 1)		0.047	0.048	0.073	0.084
	p-value (vs quintile 1)		0.762	0.715	0.811	0.834
Total Vegetables	Estimate	0.467	0.005	0.024	0.028	0.028
	s.e (vs quintile 1)		0.105	0.102	0.111	0.112
	p-value (vs quintile 1)		0.964	0.817	0.798	0.799
DOL	Estimate	0.000	0.015	0.071	0.086	0.089
	s.e (vs quintile 1)		0.051	0.051	0.067	0.058
	p-value (vs quintile 1)		0.774	0.162	0.198	0.124

Table 5.1: I-spline analysis results for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$.

		Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Milk	Estimate	0.144	0.104	0.130	0.143	0.270
	s.e (vs quintile 1)		0.048	0.044	0.059	0.056
	p-value (vs quintile 1)		0.030	0.003	0.016	0.000
Meats and Beans	Estimate	0.443	0.005	0.031	0.059	0.074
	s.e (vs quintile 1)		0.165	0.155	0.160	0.157
	p-value (vs quintile 1)		0.975	0.841	0.714	0.636
Oil	Estimate	0.068	0.157	0.208	0.212	0.215
	s.e (vs quintile 1)		0.064	0.058	0.066	0.062
	p-value (vs quintile 1)		0.015	0.000	0.001	0.001
Saturated Fat	Estimate	0.047	0.038	0.082	0.093	0.241
	s.e (vs quintile 1)		0.049	0.049	0.074	0.169
	p-value (vs quintile 1)		0.440	0.096	0.211	0.153
Sodium	Estimate	0.000	0.003	0.031	0.112	0.190
	s.e (vs quintile 1)		0.079	0.066	0.070	0.073
	p-value (vs quintile 1)		0.975	0.643	0.110	0.010
SoFAAS	Estimate	0.198	0.034	0.042	0.042	0.059
	s.e (vs quintile 1)		0.065	0.057	0.069	0.075
	p-value (vs quintile 1)		0.605	0.465	0.542	0.436

Table 5.2: I-spline analysis results for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$.

		Unweighted Model			Weighted Model			Model Comparison	
Gender	Disease	$\widehat{\beta}_{k\ell}$	s.e	p-value	$\widehat{\beta}_{k\ell}$	s.e	p-value	LR	p-value
Male	Colorectal Cancer	-0.018	0.002	0.000	-3.050	0.261	0.000	56.114	0.000
Male	Lung Cancer	-0.012	0.002	0.000	-1.470	0.230	0.000	-4.246	1.000
Female	Breast Cancer	0.001	0.002	0.722	0.322	0.227	0.156	1.896	0.169
Female	Colorectal Cancer	-0.013	0.003	0.000	-2.516	0.396	0.000	21.184	0.000
Female	Lung Cancer	-0.007	0.002	0.002	-1.273	0.302	0.000	7.942	0.005

Table 5.3: I-spine analysis results of $\widehat{\beta}_{k\ell}$ for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$. LR–Likelihood ratio test statistic.

6. CONCLUSIONS

In this dissertation, first I proposed a method named as Direct Integral Method for Ratios (DIMER) to construct confidence intervals for the ratio of two location parameters. The method, based on analytical results and further approximations to account for nuisance parameters, is computationally efficient. Compared to other methods in the literature, our simulations indicated that DIMER more nearly achieves nominal coverage levels while at the same time resulting in shorter confidence interval lengths. The most important reason why our DIMER method is better than the other compared methods is that there are severely heavy tail in the distribution of the ratio, our DIMER method avoid this by direct probability computation, while other methods are badly hindered at this part, especially for those methods which based on the assumption that use the normal distribution to approximate the Cauchy likely distribution.

Second, relative risk analysis is performed for the real HEI-2005 data set. The results from the DIMER and the modified nonparametric bootstrap method are well matched with each other. For the rest of methods, results of the inverse Fisher score method and the sandwich method are lower than the ones from the direct integral method and the modified nonparametric bootstrap method, while the nonparametric bootstrap method's results are higher than them.

For the aspect of computation time, due to the extremely large data amount, the time consumption ratio between the modified nonparametric bootstrap method and the direct integral method is around 7000, and this is a special case in which the data set only considers single disease and two subpopulations. In the future, if these formulas were applied to more complicated cases such as several different multiple diseases in multiple subpopulations, the time consumption ratio will consequently become even higher. In conclusion, regardless of reliability, accuracy and computation efficiency, DIMER is obviously the best among all the method presented in this paper to compute the relative risks and their confidence intervals.

Furthermore, variable selection methods are used for identify which nutrition component is more important for diseases across genders. And models comparison results are also provided. In addition, a model with I-spline basis function is built to satisfy some constraints from nutriology and biology. The results from I-spline analysis show the effect changing of the nutrition components on diseases.

REFERENCES

Akbaraly, T. N., Ferrie, J. E., Berr, C., Brunner, E. J., Head, J., Marmot, M. G., Singh-Manoux, A., Ritchie, K., Shipley, M. J. and Kivimaki. M. (2011). Alternative Healthy Eating Index and mortality over 18 y of follow-up: results from the Whitehall II cohort. *American Journal of Clinical Nutrition*, 94, 247-253.

Benton, D. and Krishnamoorthy, K. (2002). Performance of the parametric bootstrap method in small sample interval estimates. *Advances and Applications in Statistics*, 2, 269-285.

Beyene, J. and Moineddin, R. (2005). Methods for confidence interval estimation of a ratio parameter with application to location quotients. *BMC Medical Research Methodology*, 5, 1-7.

Brody, J. P., Williams, B. A., Wold, B. J. and Quake, S. R. (2002). Significance and statistical errors in the analysis of DNA microarray data. *The Proceedings of the National Academy of Sciences of the United States of America*, 99, 20, 12975-12978.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, Second Edition. Chapman and Hall CRC Press.

Cedilnik, A., Košmelj, K. and Blejec, A. (2004). The distribution of the ratio of jointly normal variables. *Metodološki zvezki*, 1, 99-108.

Chiuve, S. E., Fung, T. T., Rimm, E. B. Hu, F. B. McCullough, M. L., Wang, M., Stampfer, M. J. and Willett, W. C. (2012). Alternative dietary indices both strongly predict risk of chronic disease. *Journal of Nutrition*, online version at doi: 10.3945/jn.111.157222

Deaton, L. and Kamerud, D. (1978). The random variable X/Y, X, Y normal. *The American Mathematical Monthly*, 85, 206-207.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *The Canadian Journal of Statistics*, 9, 139-172.

Efron, B. and Tibshirani, R.J. (1994). *An Introduction to the Bootstrap*, First Edition. Chapman and Hall CRC Press.

Efron. B., Hastie, T., Johnstone, I. and Tibshirani, R.J. (2004). Leat Angle Regression. *Annals of Statistics*, 32, 407-499.

Fieller, E. C. (1932). The distribution of the index in a bivariate normal distribution. *Biometrika*, 24, 428-440.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society*, B 16, 175-185.

Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007) Pathwise Coordinate Optimization. *Annals of Statistics*, 1, 302332.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33, Issue 1.

Fungwe, T., Guenther, P. M., Juan, W. Y., Hiza, H, and Lino, M. (2009). The quality of children's diets in 2003-04 as measured by the Healthy Eating Index-2005. *Nutrition Insight*, 43, USDA Center for Nutrition Policy and Promotion.

Geary, R. C. (1930). The frequency distribution of the quotient of two normal variates.*Journal of the Royal Statistical Society*, 93, 442-446.

George, S. M., Neuhouser, M. L., Mayne, S. T., Irwin, M. L., Albanes, D., Gail, M. H., Alfano, C. M., Bernstein, L., McTiernan, A., Reedy, J., Smith, A. W., Ulrich, C. M. and Ballard-Barbash, R. (2010). Postdiagnosis diet quality is inversely related to a biomarker of inflammation among breast cancer survivors. *Cancer Epidemiology, Biomarkers & Prevention*, 19, 2220-2228.

Guenther, P. M., Reedy, J. and Krebs-Smith, S. M. (2008a). Development of the Healthy Eating Index-2005. *Journal of the American Dietetic Association*, 108, 1896-1901.

Guenther, P. M., Reedy, J., Krebs-Smith, S. M. and Reeve, B. B. (2008b). Evaluation of the Healthy Eating Index-2005. *Journal of the American Dietetic Association*, 108, 1854-1864.

Hayya, J., Armstrong, D. and Gressis, N. (1975). A note on the ratio of two normally distributed variables. *Management Science*, 21, 1338-1341.

Hinkley, D. V. (1969). On the ratio of two correlated normal random variables. *Biometrika*, 56, 635-639.

Kipnis, V., Midthune, D., Buckman, D. W., Dodd, K. W., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., Tooze, J. A., Carroll, R. J. and Freedman, L. S. (2009). Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics*, 65, 1003-1010.

Kipnis, V., Freedman, L. S., Carroll, R. J. and Midthune, D. (2010). A measurement error model for episodically consumed foods and energy. Preprint.

Kott, P. S., Guenther, P. M., Wagstaff, D. A., Juan W. Y. and Kranz, S. (2009). Fitting a linear model to survey data when the long-term average daily intake of a dietary component is an explanatory variable. *Survey Research Methods*, Vol 3, No 3, 157-165.

Marsaglia, G. (1965). Ratios of normal variables and ratios of sums of uniform variable. *Journal of the American Statistical Association*, 60, 193-204.

Pham-Gia, T., Turkkan, N. and Marchand, E. (2006). Density of the ratio of two normal random variables and applications. *Communications in Statistics: Theory and Methods*, 35, 1569-1591.

Qiao, C. G., Wood, G. R., Lai, C. D. and Luo, D. W. (2006). Comparison of two common estimators of the ratio of the means of independent normal variables in agricultural research. *Journal of Applied Mathematics and Decision Sciences*, Article ID 78375, 1-14.

Ramsay, J.O. (1988). Monotone Regression Splines in Action. *Statistical Science*, Vol 3, No 4, 425-441.

Reedy, J., Mitrou, P. N., Krebs-Smith, S. M., Wirfält, E., Flood, A., Kipnis, V., Leitzmann, M., Mouw, T., Hollenbeck, A., Schatzkin, A. and Subar, A. F. (2008). Index-based dietary patterns and risk of colorectal cancer: the NIH-AARP Diet and Health Study. *American Journal of Epidemiology*, 168, 38-48.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110-114.

Sherman, M., Maity, A. and Wang, S. (2011). Inferences for the ratio: Fieller's interval, log ratio, and large sample based confidence intervals. *AStA Advances in Statistical Analysis*, 95, 313-323.

Sinha, S., Mallick, B. K., Kipnis, V. and Carroll, R. J. (2010). Semiparametric Bayesian analysis of nutritional epidemiology data in the presence of measurement error. *Biometrics*, 66, 444-454.

Tooze, J. A., Grunwald, G. K., and Jones, R. H. (2002). Analysis of repeated measures data clumping at zero. *Statistical Methods in Medical Research* ,11, 341-355.

Tooze, J. A., Midthune, D., Dodd, K. W., Freedman, L. S., Krebs-Smith, S. M., Subar, A. F., Guenther, P. M., Carroll, R. J. and Kipnis, V. (2006). A new statistical method for estimating the distribution of usual intake of episodically consumed foods. *Journal of the American Dietetic Association*, 106, 1575-1587.

Wang, H. and Leng, C. (2007) Unified LASSO Estimation via Least Squares Approximation. *Journal of the American Statistical Association*, 102, 1039-1048.

Wang, Y. Q., Wang, S. J. and Carroll, R. J. The Direct Integral Method for Confidence Intervals for the Ratio of Two Location Parameters. (submitted)

Welch, B. L. (1947). The generalization of "student's" problem when several different population variances are involved. *Biometrika*, 34, 28-35.

Zhang, S., Midthune, D., Guenther, P. M., Krebs-Smith, S. M., Kipnis, V., Dodd, K. W., Buckman, D. W., Tooze, J. A., Freedman, L. S. and Carroll, R. J. (2011). A new multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment. *Annals of Applied Statistics*, 5, 1456-1487.

Zhang, S. Midthune, D., Pérez, A, Buckman, D. W., Kipnis, V., Freedman, L. S., Dodd, K. W., Krebs-Smith, S. M. and Carroll, R. J. (2011). Fitting a bivariate measurement error model for episodically consumed dietary components. *International Journal of Biostatistics*, Volume 7, Issue 1, Article 1, DOI:

10.2202/1557-4679.1267. Available at: http://www.bepress.com/ijb/vol7/iss1/1

Zhang, S., Midthune, D., Guenther, P. M., Krebs-Smith, S. M., Kipnis, V., Dodd, K. W., Buckman, D. W., Tooze, J. A., Freedman, L. S. and Carroll, R. J. (2011). Supplement to "A new multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment".

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101, 1418-1429.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

A.1 Fieller's Method

Consider a ratio $r = \mu_1/\mu_2$, where μ_1 and μ_2 are means from two correlated normal distributions $T_1 \sim \text{Normal}(\mu_1, v_1^2)$ and $T_2 \sim \text{Normal}(\mu_2, v_2^2)$. Let ρ denote the correlation coefficient between these two distributions. The estimated variance and covariance $\widehat{v}_1^2$, $\widehat{v}_2^2$ and $\widehat{\rho v_1 v_2}$ are jointly estimated with the same number of degrees of freedom d, and are independent of T_1 and T_2.

Introduce a latent variable $W = T_1 - rT_2$. Since $W/\sqrt{\widehat{v}_2^2 - 2r\widehat{\rho\widehat{v}_1}\widehat{v}_2 + r^2\widehat{v}_2^2}$ follows a t distribution with d degrees of freedom, a confidence interval with coverage probability $1-\alpha$ is calculated by using $-t_{d,\alpha/2} \leq W/\sqrt{\widehat{v}_2^2 - 2r\widehat{\rho\widehat{v}_1}\widehat{v}_2 + r^2\widehat{v}_2^2} \leq t_{d,\alpha/2}$, where $t_{d,\alpha/2}$ denotes the $(1-\alpha/2)100\%$ quantile for the t distribution with d degree freedom.

Rewrite the inequality and solve it as

$$r^2\{T_2^2 - t_{d,\alpha/2}^2\widehat{v}_2^2\} - 2r(T_1T_2 - t_{d,\alpha/2}^2\widehat{\rho\widehat{v}_1}\widehat{v}_2) + \{T_1^2 - t_{d,\alpha/2}^2\widehat{v}_1^2\} \leq 0.$$

Let $a = T_2^2 - t_{d,\alpha/2}^2\widehat{v}_2^2, b = (T_1T_2 - t_{d,\alpha/2}^2\widehat{\rho\widehat{v}_1}\widehat{v}_2)$ and $c = T_1^2 - t_{d,\alpha/2}^2\widehat{v}_1^2$. Following the inequality $ar^2 + br + c \leq 0$, two real roots $d_1 = (-b - \sqrt{b^2 - 4ac})/(2a)$ and $d_2 = (-b + \sqrt{b^2 - 4ac})/(2a)$ are obtained if $b^2 - 4ac \geq 0$.

A confidence interval of r which has coverage probability $1-\alpha$ is constructed as follows:

$$\text{Confidence Interval} = \begin{cases} (d_1, d_2) & \text{if } a \geq 0, \\ (-\infty, d_1) \bigcup (d_2, \infty) & \text{if } a < 0 \text{ and } t_{d,\alpha/2} < t_{com}, \\ (-\infty, \infty) & \text{if } t_{d,\alpha/2} \geq t_{com}, \end{cases}$$

where $t_{com} = (T_1^2 \widehat{v}_2^2 - 2T_1T_2\widehat{v}_{12} + T_2^2\widehat{v}_1^2)/(\widehat{v}_1^2\widehat{v}_2^2 - \widehat{v}_{12}^2)$ and it is certain that $a < 0$ when $t_{d,\alpha/2} > t_{com}$ as Fieller (1954) showed that $T_2^2/\widehat{v}_2^2 \leq t_{com}^2$.

There are several limitations of Fieller's algorithm. First, $b^2 - 4ac \geq 0$ is required; otherwise the inequality function has two complex roots. Second, when a decreases to 0, the interval range increases rapidly and can become infinite. Finally, if a is negative, the confidence interval is deterministic to have infinite length.

A.2 Proofs of Lemmas 1-2

Proof of Lemma 1:

Lemma 1 follows because

$$\begin{aligned}
\mathrm{pr}(\widehat{r} \le x) &= \mathrm{pr}(T_1 \le xT_2, T_2 > 0) + \mathrm{pr}(T_1 \ge xT_2, T_2 < 0) \\
&= \int_0^\infty \int_{-\infty}^{xt_2} (v_1 v_2)^{-1} f_1\{(t_1-\mu_1)/v_1\} f_2\{(t_2-\mu_2)/v_2\} dt_1 dt_2 \\
&\quad + \int_{-\infty}^0 \int_{xt_2}^{\infty} (v_1 v_2)^{-1} f_1\{(t_1-\mu_1)/v_1\} f_2\{(t_2-\mu_2)/v_2\} dt_1 dt_2 \\
&= \int_0^\infty F_1\{(xt_2-\mu_1)/v_1\} v_2^{-1} f_2\{(t_2-\mu_2)/v_2\} dt_2 \\
&\quad + \int_{-\infty}^0 [1 - F_1\{(xt_2-\mu_1)/v_1\}] v_2^{-1} f_2\{(t_2-\mu_2)/v_2\} dt_2 \\
&\overset{z=(t_2-\mu_2)/v_2}{=} \int_{-\mu_2/v_2}^\infty F_1[\{x(\mu_2+v_2 z)-\mu_1\}/v_1] f_2(z) dz \\
&\qquad + \int_{-\infty}^{-\mu_2/v_2} (1 - F_1[\{x(\mu_2+v_2 z)-\mu_1\}/v_1]) f_2(z) dz.
\end{aligned}$$

For simplicity, define $g(z|x,\mu_1,\mu_2,v_1,v_2)$ as in Section 2.2.2, then we have the cumulative distribution function of $\widehat{r} = T_1/T_2$ as

$$\mathrm{pr}(\widehat{r} \le x) = \int_{-\infty}^\infty g(z|x,\mu_1,\mu_2,v_1,v_2) \mathrm{exp}(-z^2) dz,$$

Proof of Lemma 2:

Similarly, by letting $V = \mathrm{cov}(T_1, T_2)$ Lemma 2 follows from the fact that

$$\begin{aligned}
&\mathrm{pr}(\widehat{r} \le x) \\
&= \int_0^\infty \int_{-\infty}^{xt_2} (2\pi|v_1^2 v_2^2 - v_{12}^2|^{1/2})^{-1} \mathrm{exp}\{-(t_1-\mu_1, t_2-\mu_2) V^{-1} (t_1-\mu_1, t_2-\mu_2)^{\mathrm{T}}/2\} dt_1 dt_2 \\
&+ \int_{-\infty}^0 \int_{xt_2}^\infty (2\pi|v_1^2 v_2^2 - v_{12}^2|^{1/2})^{-1} \mathrm{exp}\{-(t_1-\mu_1, t_2-\mu_2) V^{-1} (t_1-\mu_1, t_2-\mu_2)^{\mathrm{T}}/2\} dt_1 dt_2 \\
&\overset{z=(t_2-\mu_2)/v_2}{=} (2\pi)^{-1/2} \int_{-\mu_2/v_2}^\infty \Phi[\{x(\mu_2+v_2 z) - (\mu_1 + z v_{12}/v_2)\} v_2/\sqrt{v_1^2 v_2^2 - v_{12}^2}] \, \mathrm{exp}(-z^2/2) dz \\
&+ (2\pi)^{-1/2} \int_{-\infty}^{-\mu_2/v_2} (1 - \Phi[\{x(\mu_2+v_2 z) - (\mu_1 + z v_{12}/v_2)\} v_2/\sqrt{v_1^2 v_2^2 - v_{12}^2}]) \, \mathrm{exp}(-z^2/2) dz.
\end{aligned}$$

Similarly, define $g(z|x, \mu_1, \mu_2, v_1^2, v_2^2, v_{12})$ as in Section 2.2.3. Therefore, the cumulative distribution function for $\widehat{r}$ is

$$\mathrm{pr}(\widehat{r} \leq x) = \int_{-\infty}^{\infty} g(z|x, \mu_1, \mu_2, v_1^2, v_2^2, v_{12})\mathrm{exp}(-z^2)dz,$$

Proof of Algorithm in Section 2.2.4:

Since Z_1 and Z_2 are independent and both have t distributions with degree freedom of d, which are defined as $Z_1 = \{(T_1 - rT_2) - (\mu_1 - r\mu_2)\}/\sqrt{\widehat{v}_1^2 - 2\eta\widehat{v}_{12} + \eta^2\widehat{v}_2^2}$ and $Z_2 = (T_2 - \mu_2)/\widehat{v}_2$ in Section 2.2.4, respectively. Therefore, the jointly density distribution of Z_1 and Z_2 is

$$f(Z_1, Z_2) = f_{t,d}(Z_1)f_{t,d}(Z_2),$$

where $f_{t,d}$ is the standard student t density with degree of freedom d.

Based on that, by the density transform method, the jointly distribution of T_1, T_2, that is

$$\begin{aligned} f(T_1, T_2) = {} & \widehat{v}_2^{-1}(\widehat{v}_1^2 - 2\eta\widehat{v}_{12} + \eta^2\widehat{v}_2^2)^{-1/2} \\ & f_{t,d}[\{(t_1 - dt_2) - (\mu_1 - \eta\mu_2)\}/\sqrt{\widehat{v}_1^2 - 2\eta\widehat{v}_{12} + \eta^2\widehat{v}_2^2}]f_{t,d}\{(t_2 - \mu_2)/\widehat{v}_2\}. \end{aligned}$$

Similarly, let $z = (t_2 - \mu_2)/\widehat{v}_2$, and define $g(z|x, \mu_1, \mu_2, \widehat{v}_1^2, \widehat{v}_2^2, \widehat{v}_{12}, \eta)$ as in Section 2.2.4, we get the cumulative distribution function of $\widehat{r} = T_1/T_2$ as

$$\mathrm{pr}(\widehat{r} \leq x) \approx \int_{-\infty}^{\infty} g(z|x, \mu_1, \mu_2, \widehat{v}_1^2, \widehat{v}_2^2, \widehat{v}_{12}, \eta)\mathrm{exp}(-z^2)dz.$$

A.3 More Details of Simulation Results in Chapter 2

	Mean of Coverage			Mean of Length			Median of Length			IQR of Length			90% Quantile of Length		
Method	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI
$n_1 = n_2 = 18, \text{mean}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.01, 0.01, 1.10, 1.00), \text{median}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.01, 0.01, 1.00, 1.00), \text{cv}(\widehat{\omega}) = 0.260, \text{cv}(\widehat{\lambda}) = 0.265.$															
IF	84.05	89.40	94.60	1.63	1.95	2.56	1.09	1.30	1.71	0.91	1.09	1.43	2.83	3.38	4.44
HM	88.50	92.90	96.70	1.74	2.08	2.73	1.15	1.37	1.80	0.70	0.83	1.09	2.31	2.75	3.61
NB	92.15	94.50	97.75	20.66	24.62	32.35	1.67	1.98	2.61	4.47	5.33	7.00	31.39	37.40	49.15
PB	92.00	94.20	97.35	38.84	46.28	60.83	1.49	1.78	2.34	2.21	2.64	3.46	22.75	27.10	35.62
FI	89.85	95.05	99.35	∞	∞	∞	1.39	1.80	3.08	1.08	1.62	5.27	4.28	8.25	∞
DIMER	91.45	95.90	99.50	2.69	4.92	63.53	1.43	1.88	3.35	1.09	1.63	4.98	3.74	6.12	37.32
$b^2 - 4ac < 0$	0.00	0.05	0.45												
$a < 0$	2.90	5.65	14.47												
$n_1 = n_2 = 25, \text{mean}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.00, 1.05, 1.00), \text{median}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.00, 1.00, 1.00), \text{cv}(\widehat{\omega}) = 0.211, \text{cv}(\widehat{\lambda}) = 0.210.$															
IF	86.15	92.15	96.50	1.35	1.60	2.11	0.95	1.13	1.49	0.66	0.79	1.04	2.17	2.59	3.41
HM	90.15	94.75	98.20	1.12	1.33	1.75	0.97	1.16	1.52	0.47	0.56	0.73	1.65	1.97	2.59
NB	92.55	95.30	98.45	10.25	12.21	16.05	1.17	1.40	1.83	1.04	1.24	1.63	7.55	8.99	11.82
PB	92.55	95.45	98.40	7.23	8.61	11.31	1.13	1.35	1.77	0.78	0.93	1.23	3.84	4.57	6.01
FI	90.15	95.90	99.60	∞	∞	∞	1.10	1.38	2.12	0.63	0.87	1.75	2.17	3.02	6.92
DIMER	91.15	96.30	99.70	1.78	2.75	10.96	1.12	1.42	2.23	0.65	0.90	1.91	2.20	3.06	6.74
$b^2 - 4ac < 0$	0.00	0.00	0.05												
$a < 0$	0.50	1.00	4.45												
$n_1 = n_2 = 50, \text{mean}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.00, 1.02, 1.00), \text{median}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.00, 1.00, 1.00), \text{cv}(\widehat{\omega}) = 0.144, \text{cv}(\widehat{\lambda}) = 0.148.$															
IF	90.00	93.10	97.25	0.94	1.12	1.47	0.67	0.79	1.04	0.36	0.43	0.56	1.38	1.64	2.15
HM	90.65	95.50	98.65	0.71	0.84	1.11	0.67	0.79	1.04	0.23	0.27	0.36	0.95	1.13	1.48
NB	92.15	95.40	98.55	0.84	1.00	1.32	0.70	0.84	1.10	0.29	0.34	0.45	1.10	1.31	1.72
PB	92.15	96.10	98.65	0.80	0.95	1.25	0.71	0.84	1.11	0.28	0.33	0.43	1.09	1.29	1.70
FI	91.20	95.75	99.00	0.76	0.93	1.35	0.70	0.86	1.19	0.25	0.32	0.50	1.04	1.29	1.90
DIMER	91.50	95.80	99.10	0.77	0.94	1.36	0.71	0.87	1.21	0.26	0.33	0.51	1.05	1.31	1.94
$b^2 - 4ac < 0$	0.00	0.00	0.00												
$a < 0$	0.00	0.00	0.00												

Table A.1: Confidence intervals for β_{21} in a simulation study with 2000 replications and true parameter values $(\beta_{10}, \beta_{20}, \beta_{21}, \omega) = (0.00, 0.00, 1.00, 1.00)$ for the linear regression model $Y_{1i} = \beta_{10} - X_{1i}\omega + \varepsilon_{1i}; Y_{2j} = \beta_{20} + \beta_{21}X_{2j}\omega + \varepsilon_{2j}$. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval and DIMER–Direct Integral Method for Ratios.

	Mean of Coverage			Mean of Length			Median of Length			IQR of Length			90% Quantile of Length		
Method	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI
$n_1 = n_2 = 18, \text{mean}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.01, 0.01, 2.85, 0.75), \text{median}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.01, 0.01, 1.00, 0.75), \text{cv}(\widehat{\omega}) = 0.346, \text{cv}(\widehat{\lambda}) = 0.353.$															
IF	83.60	88.60	94.15	4.51	5.38	7.07	1.46	1.74	2.29	1.35	1.61	2.12	4.59	5.47	7.19
HM	86.45	91.65	95.55	2974.62	3544.48	4658.24	1.54	1.83	2.41	1.22	1.45	1.90	4.06	4.84	6.36
NB	93.35	95.10	97.75	74.05	88.24	115.97	4.54	5.41	7.11	22.45	26.75	35.16	105.09	125.22	164.57
PB	93.05	94.55	97.35	1634.42	1947.53	2559.48	3.55	4.23	5.56	22.09	26.32	34.59	94.25	112.31	147.60
FI	91.44	96.04	99.35	∞	∞	∞	2.13	2.97	7.75	3.01	7.41	∞	∞	∞	∞
DIMER	92.80	96.55	99.55	7.57	15.87	56.16	2.15	3.05	8.28	2.44	4.55	39.34	10.03	25.88	105.14
$b^2 - 4ac < 0$	0.65	1.60	7.05												
$a < 0$	11.17	16.92	34.64												
$n_1 = n_2 = 25, \text{mean}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.00, 1.17, 0.75), \text{median}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.00, 1.00, 0.75), \text{cv}(\widehat{\omega}) = 0.281, \text{cv}(\widehat{\lambda}) = 0.280.$															
IF	85.65	91.40	96.30	2.02	2.40	3.16	1.27	1.51	1.99	0.99	1.19	1.56	3.20	3.81	5.01
HM	89.45	94.15	97.75	5.06	6.03	7.92	1.30	1.55	2.03	0.80	0.96	1.26	2.69	3.20	4.21
NB	93.40	95.55	98.25	42.36	50.47	66.33	2.07	2.47	3.24	8.10	9.65	12.68	45.18	53.84	70.75
PB	93.10	95.50	98.30	53.39	63.62	83.61	1.91	2.28	2.99	5.49	6.54	8.60	35.70	42.54	55.90
FI	91.05	96.49	99.65	∞	∞	∞	1.59	2.11	3.90	1.42	2.40	12.88	5.72	15.03	∞
DIMER	92.40	96.95	99.75	4.53	9.82	35.54	1.62	2.16	4.15	1.36	2.20	8.46	4.64	7.96	61.76
$b^2 - 4ac < 0$	0.05	0.15	1.10												
$a < 0$	4.20	6.96	19.26												
$n_1 = n_2 = 50, \text{mean}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.00, 1.04, 0.75), \text{median}(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega}) = (0.00, 0.01, 1.00, 0.75), \text{cv}(\widehat{\omega}) = 0.192, \text{cv}(\widehat{\lambda}) = 0.197.$															
IF	89.20	93.00	97.25	1.19	1.41	1.86	0.89	1.07	1.40	0.57	0.67	0.89	2.00	2.38	3.13
HM	90.80	95.30	98.45	0.99	1.17	1.54	0.89	1.06	1.39	0.39	0.46	0.61	1.43	1.70	2.23
NB	93.00	95.60	98.35	2.57	3.06	4.02	1.01	1.20	1.58	0.61	0.72	0.95	2.33	2.77	3.64
PB	92.75	96.10	98.65	3.79	4.52	5.93	1.00	1.19	1.57	0.59	0.70	0.92	2.19	2.61	3.43
FI	91.30	95.80	99.10	∞	∞	∞	0.97	1.21	1.77	0.49	0.65	1.17	1.73	2.28	4.13
DIMER	91.55	96.05	99.10	1.16	1.52	3.70	0.98	1.22	1.81	0.50	0.66	1.22	1.75	2.31	4.23
$b^2 - 4ac < 0$	0.00	0.00	0.00												
$a < 0$	0.05	0.25	1.30												

Table A.2: Confidence intervals for β_{21} in a simulation study with 2000 replications and true parameter values $(\beta_{10}, \beta_{20}, \beta_{21}, \omega) = (0.00, 0.00, 1.00, 0.75)$ for the linear regression model $Y_{1i} = \beta_{10} - X_{1i}\omega + \varepsilon_{1i}; Y_{2j} = \beta_{20} + \beta_{21}X_{2j}\omega + \varepsilon_{2j}$. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval and DIMER–Direct Integral Method for Ratios.

	Mean of Coverage			Mean of Length			Median of Length			IQR of Length			90% Quantile of Length		
Method	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI
$n = 10, (\beta, \mu) = (1.00, 1.00), \text{cv}(\widehat{\beta}) = 0.385, \text{mean}(\widehat{\beta}, \widehat{\mu}) = (0.99, 1.06), \text{median}(\widehat{\beta}, \widehat{\mu}) = (1.01, 0.98), \text{cv}(\widehat{\beta}\widehat{\mu}) = 0.343, \rho(\widehat{\beta}, \widehat{\beta}\widehat{\mu}) = -0.008.$															
IF	89.70	93.05	97.05	77.28	92.09	121.02	1.51	1.80	2.37	1.36	1.63	2.14	5.13	6.11	8.03
HM	78.55	84.90	90.70	51.90	61.84	81.27	1.45	1.73	2.27	1.32	1.58	2.07	4.63	5.52	7.25
NB	93.00	94.55	96.70	236.37	281.65	370.15	6.96	8.29	10.89	35.93	42.82	56.27	132.71	158.14	207.83
PB	90.95	93.15	95.90	193.31	230.35	302.73	3.80	4.53	5.95	27.61	32.90	43.23	129.34	154.12	202.55
FI	91.39	95.74	99.30	∞	∞	∞	2.35	3.76	57.35	7.67	∞	∞	∞	∞	∞
DI	92.70	95.95	99.15	11.29	23.58	96.24	2.25	3.36	12.71	3.45	8.22	55.51	15.18	36.78	124.01
$b^2 - 4ac < 0$	2.40	5.00	21.15												
$a < 0$	18.44	27.37	48.57												
$n = 25, (\beta, \mu) = (1.00, 1.00), \text{cv}(\widehat{\beta}) = 0.214, \text{mean}(\widehat{\beta}, \widehat{\mu}) = (0.99, 1.08), \text{median} = (\widehat{\beta}, \widehat{\mu}) = (0.99, 1.01), \text{cv}(\widehat{\beta}\widehat{\mu}) = 0.205, \rho(\widehat{\beta}, \widehat{\beta}\widehat{\mu}) = 0.005.$															
IF	91.70	94.95	97.90	1.72	2.05	2.70	0.95	1.13	1.49	0.47	0.56	0.74	1.68	2.00	2.63
HM	87.70	93.10	97.40	1.53	1.82	2.40	0.94	1.12	1.47	0.48	0.57	0.75	1.60	1.91	2.50
NB	91.80	94.85	97.85	9.57	11.40	14.98	1.09	1.30	1.71	1.21	1.44	1.89	7.37	8.78	11.53
PB	91.75	95.00	98.20	8.63	10.28	13.51	1.09	1.30	1.71	0.87	1.04	1.37	4.34	5.17	6.80
FI	89.90	95.05	99.20	∞	∞	∞	1.08	1.38	2.19	0.70	0.99	2.28	2.28	3.29	10.03
DIMER	90.35	94.95	99.10	1.92	2.71	17.66	1.08	1.37	2.15	0.70	0.98	2.13	2.24	3.17	7.57
$b^2 - 4ac < 0$	0.00	0.00	0.10												
$a < 0$	0.60	1.20	5.41												
$n = 50, (\beta, \mu) = (1.00, 1.00), \text{cv}(\widehat{\beta}) = 0.141, \text{mean}(\widehat{\beta}, \widehat{\mu}) = (1.00, 1.02), \text{median}(\widehat{\beta}, \widehat{\mu}) = (0.99, 1.00), \text{cv}(\widehat{\beta}\widehat{\mu}) = 0.144, \rho(\widehat{\beta}, \widehat{\beta}\widehat{\mu}) = -0.011.$															
IF	91.90	95.60	98.25	0.70	0.84	1.10	0.66	0.79	1.04	0.21	0.25	0.32	0.93	1.11	1.46
HM	90.45	95.10	98.15	0.70	0.83	1.09	0.66	0.79	1.03	0.23	0.27	0.35	0.95	1.13	1.49
NB	92.10	95.30	98.10	0.91	1.08	1.42	0.69	0.82	1.08	0.30	0.36	0.47	1.14	1.35	1.78
PB	93.00	95.55	98.30	0.78	0.93	1.22	0.70	0.83	1.10	0.28	0.33	0.44	1.09	1.30	1.71
FI	90.60	95.05	99.50	0.76	0.94	1.38	0.70	0.86	1.21	0.27	0.34	0.52	1.07	1.33	1.98
DIMER	90.65	95.00	99.40	0.76	0.94	1.36	0.70	0.85	1.20	0.27	0.34	0.52	1.07	1.33	1.97
$b^2 - 4ac < 0$	0.00	0.00	0.00												
$a < 0$	0.00	0.00	0.00												

Table A.3: Confidence intervals for μ in a simulation study with 2000 replications for linear regression model $Y_i = \beta(X_i - \mu) + \epsilon_i$ with Setting I: $(\beta, \mu) = (1.00, 1.00)$. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval and DIMER–Direct Integral Method for Ratios.

	Mean of Coverage			Mean of Length			Median of Length			IQR of Length			90% Quantile of Length		
Method	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI
n = 10, (β,μ)=(2.00,1.00), mean$(\widehat{\beta},\widehat{\mu})$=(2.01,1.03), median$(\widehat{\beta},\widehat{\mu})$=(2.00,0.99), cv$(\widehat{\omega})$=0.190, cv$(\widehat{\lambda})$=0.174, $\rho(\widehat{\omega},\widehat{\lambda})$=0.021.															
IF	91.05	93.60	97.65	0.90	1.07	1.41	0.77	0.92	1.21	0.38	0.45	0.59	1.35	1.61	2.12
HM	81.75	88.60	94.75	0.80	0.96	1.26	0.71	0.84	1.11	0.40	0.48	0.63	1.26	1.50	1.97
NB	91.00	94.15	97.35	8.18	9.75	12.82	0.94	1.12	1.48	1.48	1.77	2.32	8.81	10.50	13.80
PB	88.30	93.15	97.00	5.10	6.08	7.99	0.79	0.94	1.24	0.62	0.74	0.97	2.39	2.85	3.74
FI	89.64	94.98	98.84	∞	∞	∞	0.89	1.15	1.94	0.63	0.91	2.40	2.05	3.08	19.86
DIMER	90.15	94.95	98.80	1.54	2.36	7.16	0.89	1.14	1.86	0.62	0.89	2.12	1.98	2.87	8.15
LR	91.95	95.95	99.35												
$b^2-4ac<0$	0.05	0.40	1.05												
$a<0$	0.90	1.86	7.93												
n = 25, (β,μ)=(2.00,1.00), mean$(\widehat{\beta},\widehat{\mu})$=(2.01,1.01), median$(\widehat{\beta},\widehat{\mu})$=(2.01,1.00), cv$(\widehat{\omega})$=0.107, cv$(\widehat{\lambda})$=0.105, $\rho(\widehat{\omega},\widehat{\lambda})$=-0.013.															
IF	90.20	94.75	97.85	0.50	0.59	0.78	0.47	0.56	0.74	0.13	0.16	0.20	0.63	0.76	0.99
HM	87.80	92.35	97.45	0.48	0.57	0.75	0.46	0.54	0.72	0.15	0.18	0.24	0.63	0.75	0.99
NB	88.20	93.35	97.75	0.51	0.60	0.79	0.47	0.56	0.74	0.18	0.21	0.28	0.70	0.83	1.09
PB	89.65	93.95	97.75	0.51	0.61	0.80	0.48	0.57	0.74	0.18	0.21	0.28	0.70	0.83	1.09
FI	89.75	94.30	98.90	0.52	0.64	0.91	0.49	0.60	0.84	0.18	0.22	0.32	0.72	0.89	1.28
DIMER	89.80	94.30	98.80	0.53	0.64	0.90	0.49	0.60	0.83	0.18	0.22	0.32	0.72	0.88	1.26
LR	90.70	95.80	99.20												
$b^2-4ac<0$	0.00	0.00	0.00												
$a<0$	0.00	0.00	0.00												
n = 50, (β,μ)=(2.00,1.00), mean$(\widehat{\beta},\widehat{\mu})$=(2.00,1.01), median$(\widehat{\beta},\widehat{\mu})$=(2.00,1.00), cv$(\widehat{\omega})$=0.072, cv$(\widehat{\lambda})$=0.072, $\rho(\widehat{\omega},\widehat{\lambda})$=-0.021.															
IF	89.90	94.35	98.85	0.34	0.40	0.53	0.33	0.40	0.52	0.06	0.08	0.10	0.41	0.48	0.64
HM	88.90	93.95	98.65	0.33	0.40	0.52	0.33	0.39	0.51	0.08	0.09	0.12	0.41	0.49	0.64
NB	88.35	93.40	98.25	0.33	0.40	0.52	0.33	0.39	0.51	0.08	0.10	0.13	0.41	0.49	0.65
PB	89.75	94.40	98.90	0.34	0.41	0.53	0.33	0.40	0.52	0.08	0.10	0.13	0.43	0.51	0.67
FI	89.50	94.60	98.75	0.35	0.42	0.56	0.34	0.41	0.55	0.08	0.10	0.14	0.43	0.52	0.71
DIMER	89.55	94.55	98.65	0.35	0.42	0.56	0.34	0.41	0.55	0.08	0.10	0.14	0.43	0.52	0.70
LR	89.90	95.15	99.30												
$b^2-4ac<0$	0.00	0.00	0.00												
$a<0$	0.00	0.00	0.00												

Table A.4: Confidence intervals for μ in a simulation study with 2000 replications for linear regression model $Y_i = \beta(X_i - \mu) + \epsilon_i$ with $(\beta, \mu) = (2.00, 1.00)$, where ϵ_i follows a skew normal distribution with mean 0, variance 1 and skewness 0.78. Values for $b^2 - 4ac < 0$ indicate percents in simulation that Fieller's interval is invalid and values for $a < 0$ represent percents of infinite lengths obtained by Fieller's interval. IF–Inverse Fisher Score method, HM–Hayya's method, NB–Nonparametric Bootstrap, PB–Parametric Bootstrap, FI–Fieller's Interval, DIMER–Direct Integral Method for Ratios and LR–Likelihood ratio test.

A.4 Bootstrapping Details

Here are the general steps we used for the nonparametric bootstrap and parametric bootstrap.

- Procedure for the nonparametric bootstrap:
  - For given data set (Y_1, Y_2, X_1, X_2), obtain the estimates $(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega})$.
  - Generate $B = 400$ bootstrap data sets with replacement for two subgroups separately.
  - For the b^{th} generated data set $(Y_{1,b}, Y_{2,b}, X_{1,b}, X_{2,b})$, obtain $(\widehat{\beta}_{10,b}$, $\widehat{\beta}_{20,b}$, $\widehat{\beta}_{21,b}$, $\widehat{\omega}_b)$. Repeat this process for all resampled data sets.
  - Compute the standard error of $\widehat{\beta}_{21,b}$ as $\text{se}_{\beta_{21},\text{nonpara,boot}}$.
  - Construct the $(1-\alpha)100\%$ confidence interval:
    $(\widehat{\beta}_{21} - z_{\alpha/2}\text{se}_{\beta_{21},\text{nonpara,boot}}, \widehat{\beta} + z_{\alpha/2}\text{se}_{\beta_{21},\text{nonpara,boot}})$.
- Procedure for the parametric bootstrap:
  - For given data set (Y_1, Y_2, X_1, X_2), obtain the estimates $(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega})$.
  - Fix $(\widehat{\beta}_{10}, \widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\omega})$ and (X_1, X_2), we generate $B = 400$ data sets of $(Y_{1,b}, Y_{2,b})$ using a parametric model.
  - For the b^{th} generated data set $(Y_{1,b}, Y_{2,b})$ with (X_1, X_2), obtain $(\widehat{\beta}_{10,b}$, $\widehat{\beta}_{20,b}$, $\widehat{\beta}_{21,b}$, $\widehat{\omega}_b)$. Repeat this process for all resampled data sets.
  - Compute the estimated standard error of $\widehat{\beta}_{21,b}$ as $\text{se}_{\beta_{21},\text{para,boot}}$.
  - Construct the $(1-\alpha)100\%$ confidence interval:
    $(\widehat{\beta}_{21} - z_{\alpha/2}\text{se}_{\beta_{21},\text{para,boot}}, \widehat{\beta} + z_{\alpha/2}\text{se}_{\beta_{21},\text{para,boot}})$.

APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

B.1 Proof of Lemma 3

We have

$$N^{1/2}\{\widehat{S}_{\alpha,k\ell}(\widehat{\Lambda}_{k\ell}) - \widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})\} = \{\partial\widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}^{\mathrm{T}}\}N^{1/2}(\widehat{\Lambda}_{k\ell} - \Lambda_{k\ell}) + o_p(1),$$

and asymptotically, $N^{1/2}(\widehat{\Lambda}_{k\ell} - \Lambda_{k\ell})$ follows a normal distribution as

$$N^{1/2}(\widehat{\Lambda}_{k\ell} - \Lambda_{k\ell}) = \mathrm{Normal}(0, V_{\Lambda_{k\ell}}),$$

where the estimate of $V_{\Lambda_{k\ell}}$ can be directly obtained from $A^{-1}(\widehat{\Theta})\widehat{V}_{\Psi}(\widehat{\Theta})A^{-\mathrm{T}}(\widehat{\Theta})$ in Section 3.2.3 which acquired by using the sandwich method.

In consequence, the asymptotic limit distribution of $N^{1/2}\{\widehat{S}_{\alpha,k\ell}(\widehat{\Lambda}_{k\ell}) - \widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})\}$ is

$$N^{1/2}\{\widehat{S}_{\alpha,k\ell}(\widehat{\Lambda}_{k\ell}) - \widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})\} = \mathrm{Normal}\left(0, \{\partial\widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}^{\mathrm{T}}\}V_{\Lambda_{k\ell}}\{\partial\widehat{S}_{\alpha,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}\}\right).$$

And

$$N^{1/2}(\widehat{\mathcal{V}}_{k\ell} - \mathcal{V}_{k\ell}) = N^{1/2}[\{\widehat{S}_{0.90,k\ell}(\widehat{\Lambda}_{k\ell}) - \widehat{S}_{0.90,k\ell}(\Lambda_{k\ell})\} - \{\widehat{S}_{0.10,k\ell}(\widehat{\Lambda}_{k\ell}) - \widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})\}]$$

, therefore asymptotically

$$N^{1/2}(\widehat{\mathcal{V}}_{k\ell} - \mathcal{V}_{k\ell}) \sim \mathrm{Normal}(0, D_{k\ell}),$$

where

$$
\begin{aligned}
D_{k\ell} = \{\partial \widehat{S}_{0.90,k\ell}(\Lambda_{k\ell})/\partial \Lambda_{k\ell}^{\mathrm{T}}\}\mathrm{var}(\widehat{\Lambda}_{k\ell})\{\partial \widehat{S}_{0.90,k\ell}(\Lambda_{k\ell})/\partial \Lambda_{k\ell}\} \\
+\{\partial \widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})/\partial \Lambda_{k\ell}^{\mathrm{T}}\}\mathrm{var}(\widehat{\Lambda}_{k\ell})\{\partial \widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})/\partial \Lambda_{k\ell}\} \\
-2\{\partial \widehat{S}_{0.90,k\ell}(\Lambda_{k\ell})/\partial \Lambda_{k\ell}^{\mathrm{T}}\}\mathrm{var}(\widehat{\Lambda}_{k\ell})\{\partial \widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})/\partial \Lambda_{k\ell}\}.
\end{aligned}
$$

B.1.1 Proof of Lemma 4

We have defined $\widehat{\mathcal{V}}_{k\ell}^{*} = B^{-1}\sum_{b=1}^{B}\widehat{\mathcal{V}}_{k\ell}^{b}$ and $\widehat{D}_{k\ell}^{*} = (B-1)^{-1}\sum_{b=1}^{B}\left(\widehat{\mathcal{V}}_{k\ell}^{b} - \widehat{\mathcal{V}}_{k\ell}^{*}\right)^{2}$. Obviously as B $\to \infty$, by central limit theorem, we could write

$$\mathrm{B}^{1/2}(\widehat{D}_{k\ell}^{*})^{-1/2}(\widehat{\mathcal{V}}_{k\ell}^{*} - \widehat{\mathcal{V}}_{k\ell}) \backsim \text{Normal(0,1)}.$$

And based on previous section, we get ${n_\ell}^{1/2}(D_{k\ell})^{-1/2}(\widehat{\mathcal{V}}_{k\ell} - \mathcal{V}_{k\ell}) \backsim$ Normal(0,1). Write

$$\mathrm{B}^{1/2}(\widehat{D}_{k\ell}^{*})^{-1/2}(\widehat{\mathcal{V}}_{k\ell}^{*} - \mathcal{V}_{k\ell}) = \mathrm{B}^{1/2}(\widehat{D}_{k\ell}^{*})^{-1/2}(\widehat{\mathcal{V}}_{k\ell}^{*} - \widehat{\mathcal{V}}_{k\ell}) + \mathrm{B}^{1/2}(\widehat{D}_{k\ell}^{*})^{-1/2}(\widehat{\mathcal{V}}_{k\ell} - \mathcal{V}_{k\ell}).$$

Hence if we could prove that $\mathrm{B}^{1/2}(\widehat{D}_{k\ell}^{*})^{-1/2}(\widehat{\mathcal{V}}_{k\ell} - \mathcal{V}_{k\ell})$ weakly converge to 0, by Slutsky's theorem it is enough to write $\mathrm{B}^{1/2}(\widehat{D}_{k\ell}^{*})^{-1/2}(\widehat{\mathcal{V}}_{k\ell}^{*} - \mathcal{V}_{k\ell}) \backsim$ Normal(0,1).

It can be expanded as

$$\mathrm{B}^{1/2}(\widehat{D}_{k\ell}^{*})^{-1/2}(\widehat{\mathcal{V}}_{k\ell} - \mathcal{V}_{k\ell}) = (\mathrm{B}^{1/2}/n_{\ell}^{1/2})\{n_{\ell}^{1/2}(D_{k\ell})^{-1/2}(\widehat{\mathcal{V}}_{k\ell} - \mathcal{V}_{k\ell})\}\{D_{k\ell}^{1/2}/(\widehat{D}_{k\ell}^{*})^{1/2}\}.$$

By Chebyshev's inequality

$$\mathrm{P}\{|\widehat{D}_{k\ell}^{*} - \mathrm{E}(\widehat{D}_{k\ell}^{*})| < \epsilon\} > 1 - \mathrm{var}(\widehat{D}_{k\ell}^{*})/\epsilon^{2}.$$

As the data from bootstrap are identically and independently distributed, when $\mathrm{B} \to \infty$, we have $\mathrm{var}(\widehat{D}^*_{k\ell}) \to 0$ w.p.1 and $\mathrm{E}(\widehat{D}^*_{k\ell}) = \widehat{D}_{k\ell}$, hence $\widehat{D}^*_{k\ell}$ weakly converges to $D_{k\ell}$. Also we have ${n_\ell}^{1/2}{D_{k\ell}}^{-1/2}(\widehat{\mathcal{V}}_{k\ell} - V_{k\ell}) = O_p(1)$. Therefore, if n_ℓ is much larger than B as $n_\ell/\mathrm{B} \to \infty$, the asymptotic distribution for $\widehat{\mathcal{V}}^*_{k\ell}$ as follows

$$\mathrm{B}^{1/2}(\widehat{D}^*_{k\ell})^{-1/2}(\widehat{\mathcal{V}}^*_{k\ell} - \mathcal{V}_{k\ell}) \backsim \text{Normal}(0,\ 1).$$

B.1.2 Procedures

B.1.2.1 Procedures for Compute the Relative Risk by the Direct Integral Method

In the model (5.1), after obtaining $\widehat{\omega}_j$ for $j = 1, ..., J$, write $T_{i\ell} = \sum_{j=1}^{J} X_{ij\ell}\widehat{\omega}_j$. For the first disease in the first subpopulation (where we set $\beta_{11} = -1$ for identifiability), the relative risk and its confidence intervals are given by the following procedures

- Compute 10th and 90th percentile of T_{i1} as $T_{1,_{90th}}$ and $T_{1,_{10th}}$, respectively.

- Run logistic regression model as $\mathrm{pr}(Y_{i11} = 1|T_{i1}, Z_{i11}) = H(\alpha^*_{11} + \beta^*_{11}T_{i1} + Z^{\mathrm{T}}_{i11}\theta^*_{11})$, obtain estimate $\widehat{\beta}^*_{11}$ of β^*_{11} and its estimated standard deviation $\sigma_{\widehat{\beta}^*_{11}}$.

- Calculate $100(1-\alpha)\%$ CI for β^*_{11} as $(\widehat{\beta}^*_{11} - Z_{\alpha/2}\sigma_{\widehat{\beta}^*_{11}}, \widehat{\beta}^*_{11} + Z_{\alpha/2}\sigma_{\widehat{\beta}^*_{11}})$, which defined as $(a_{1,11,\alpha}, a_{2,11,\alpha})$ and where $Z_{\alpha/2}$ is the $(\alpha/2)^{th}$ quantile of the standard normal distribution.

- The relative risk estimate is $\exp\{\widehat{\beta}^*_{11}(T_{1,_{90th}} - T_{1,_{10th}})\}$.

- The confidence interval of the relative risk is $\big(\exp\{a_{1,11,\alpha}(T_{1,_{90th}} - T_{1,_{10th}})\}, \exp\{a_{2,11,\alpha}(T_{1,_{90th}} - T_{1,_{10th}})\}\big)$.

For any other $\beta_{k\ell}$ which $(k, \ell) \neq (1, 1)$, here we give the steps to obtain the estimate of the relative risk and its confidence intervals.

- Compute 10th and 90th percentile of $T_{i\ell}$ as $T_{\ell,_{90th}}$ and $T_{\ell,_{10th}}$, respectively.
- Calculate $100(1-\alpha)\%$ CI for $\beta_{k\ell}$ as $(a_{1,k\ell,\alpha}, a_{2,k\ell,\alpha})$ by the direct integral method.
- The relative risk for the k^{th} disease in the ℓ^{th} subpopulation is $\exp\{\widehat{\beta}_{k\ell}(T_{\ell,_{90th}} - T_{\ell,_{10th}})\}$.
- The confidence interval of the relative risk is $\big(\exp\{a_{1,k\ell,\alpha}(T_{\ell,_{90th}} - T_{\ell,_{10th}})\}, \exp\{a_{2,k\ell,\alpha}(T_{\ell,_{90th}} - T_{\ell,_{10th}})\}\big)$.

B.1.2.2 Bootstrapping Details

Procedure for the nonparametric bootstrap:

1. Generate $B = 2500$ bootstrap data sets with replacement for two subgroups, male and female, respectively.
2. For the b^{th} generated data set $(Y_{ik\ell,b}, X_{ij\ell,b}, Z_{ik\ell,b})$ for $\ell = 1,\ldots,L$, $k = 1,\ldots,K_\ell$ and $i = 1,\ldots,n_\ell$, run regression model $\mathrm{pr}(Y_{ik\ell,b} = 1|X_{ij\ell,b}, Z_{ik\ell,b}) = H(\alpha_{k\ell} + \beta_{k\ell}\sum_{j=1}^{J} X_{ij\ell,b}\omega_j + Z_{ik\ell,b}^{\mathrm{T}}\theta_{k\ell})$ with $\beta_{11} = -1$ to obtain $(\widehat{\alpha}_{11,b}, \widehat{\theta}_{11,b}, \widehat{\alpha}_{12,b}, \ldots, \widehat{\omega}_b)$
3. Define $T_{i\ell,b} = \sum_{j=1}^{J} X_{ij\ell,b}\widehat{\omega}_{j,b}$, compute 10th and 90th percentile of $T_{i\ell,b}$.
4. Run logistic regression model as $\mathrm{pr}(Y_{ik\ell,b} = 1|T_{i\ell,b}, Z_{ik\ell,b}) = H(\alpha^*_{k\ell,b} + \beta^*_{k\ell,b}T_{i\ell} + Z_{ik\ell}^{\mathrm{T}}\theta^*_{k\ell,b})$ to estimate $\widehat{\beta}^*_{k\ell,b}$.
5. Compute log(relative risk) as $\mathcal{V}_{k\ell,b} = \widehat{\beta}^*_{k\ell,b}(T_{i\ell,b_{90th}} - T_{i\ell,b_{10th}})$.
6. Repeat steps $(2 \sim 5)$ for all resampled data sets.
7. Construct the 95% confidence interval of the relative risk by method I $\big(\exp\{\mathcal{V}_{(k\ell,b)_{2.5th}}\}, \exp\{\mathcal{V}_{(k\ell,b)_{97.5th}}\}\big)$.

8. Construct the 95% confidence interval of the relative risk by method II

   $\big(\exp(\overline{\mathcal{V}_{k\ell,b}} - 1.96\mathrm{se}_{\mathcal{V}_{k\ell,b}}), \exp(\overline{\mathcal{V}_{k\ell,b}} + 1.96\mathrm{se}_{\mathcal{V}_{k\ell,b}})\big)$,

   where $\overline{\mathcal{V}_{k\ell,b}}$ is mean of $\mathcal{V}_{k\ell,b}$ and $\mathrm{se}_{\mathcal{V}_{k\ell,b}}$ is its standard error.

B.2 Plots of Confidence Intervals Lengths

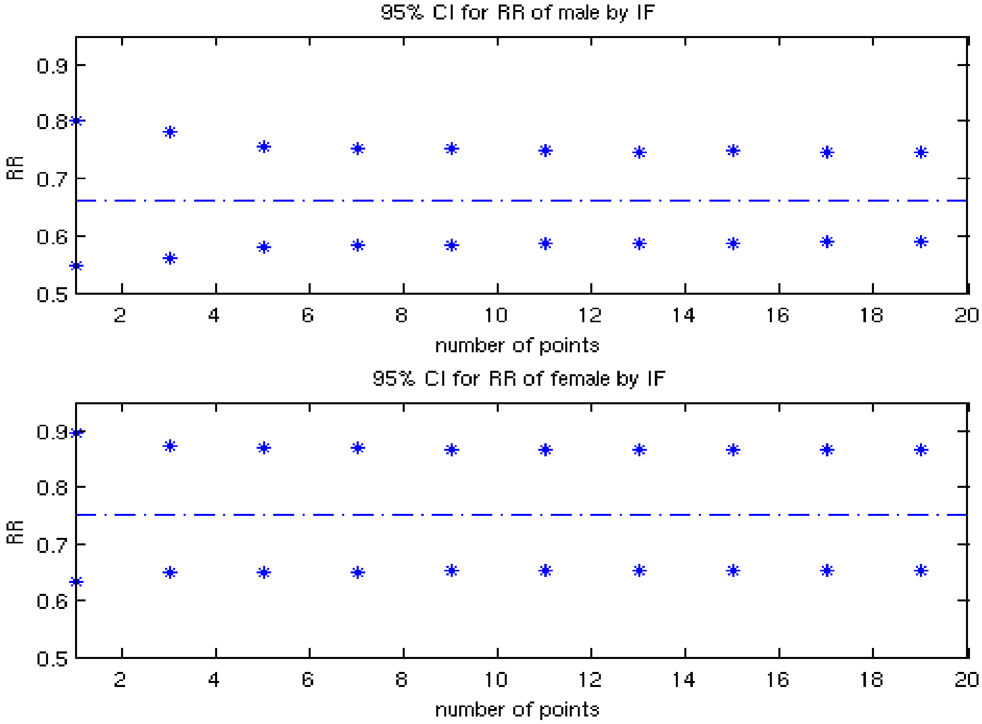

Figure B.1: Relationship between the 95% confidence interval by using the inverse Fisher matrix and the number of points which involved in computing the derivative $\partial\widehat{S}_{0.90,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}^{\mathrm{T}}$ and $\partial\widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}^{\mathrm{T}}$.

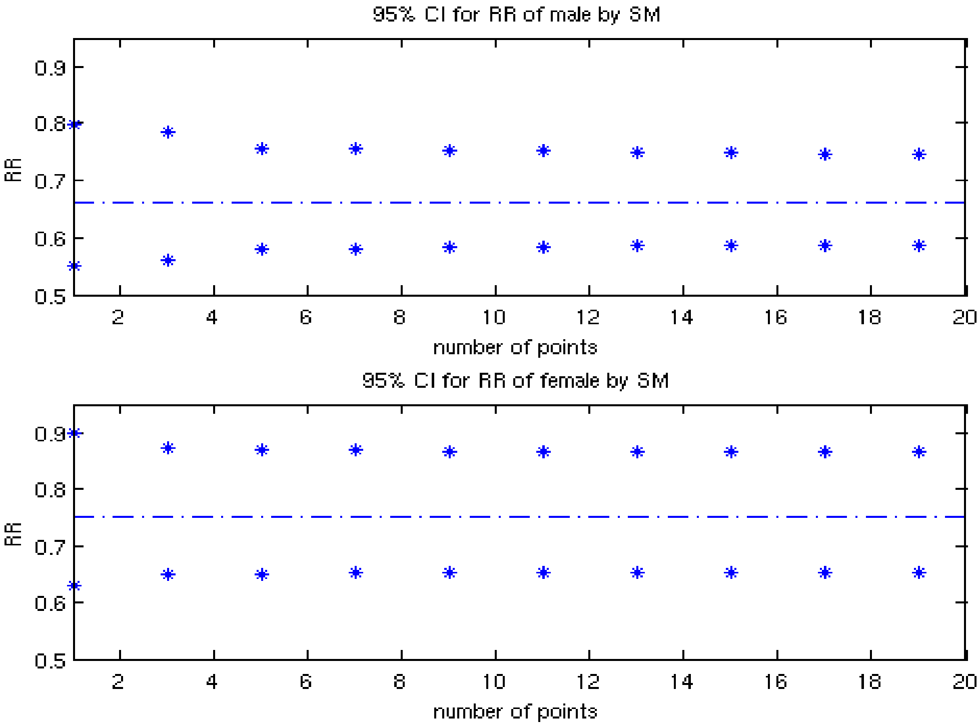

Figure B.2: Relationship between the 95% confidence interval by using the sandwich method and the number of points which involved in computing the derivative $\partial\widehat{S}_{0.90,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}^{\mathrm{T}}$ and $\partial\widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}^{\mathrm{T}}$.

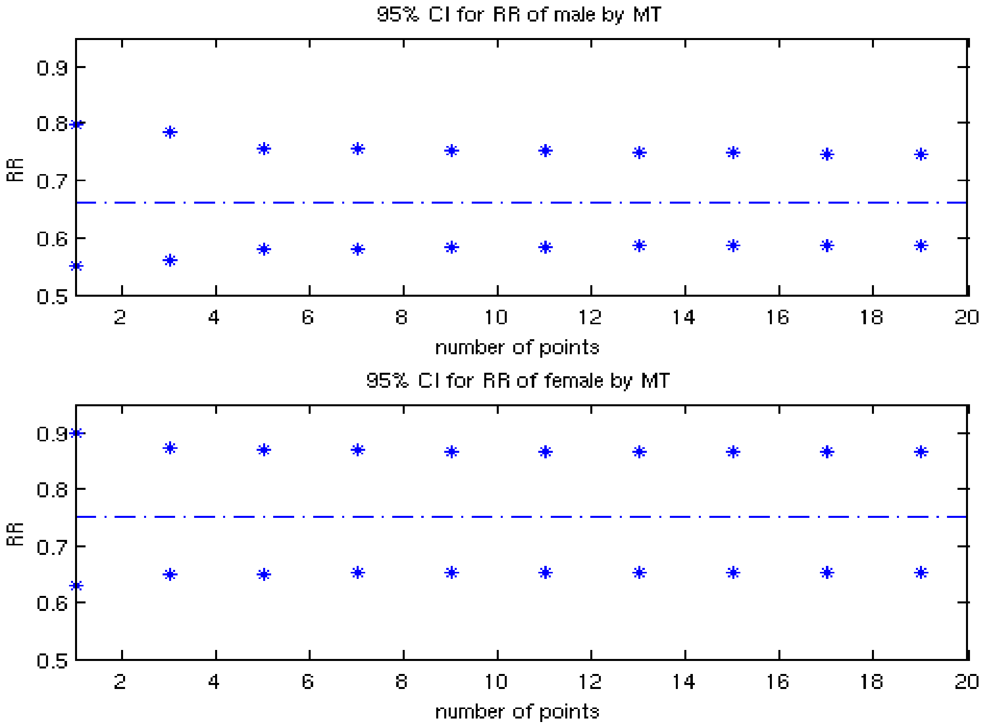

Figure B.3: Relationship between the 95% confidence interval by using the model transformation method and the number of points which involved in computing the derivative $\partial\widehat{S}_{0.90,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}^{\mathrm{T}}$ and $\partial\widehat{S}_{0.10,k\ell}(\Lambda_{k\ell})/\partial\Lambda_{k\ell}^{\mathrm{T}}$.

APPENDIX C

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

C.1 Proof of Lemma 6

We first prove ii) of Lemma 6. Let $\alpha_N = \alpha + \mathbf{v}\sqrt{N}$, $\theta_N = \theta + \mathbf{s}/\sqrt{N}$ and $\omega_N = \omega + \mathbf{u}/\sqrt{N}$, where $\mathbf{v} = (v_{11}, \ldots, v_{K_L L})^{\mathrm{T}}$, $\mathbf{s} = (s_{11}^{\mathrm{T}}, \ldots, s_{K_L L}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{u} = (u_1, \ldots, u_J)^{\mathrm{T}}$. Let $\Psi_N(\mathbf{v}, \mathbf{s}, \mathbf{u}) = 2\mathcal{L}(\alpha + \mathbf{v}\sqrt{N}, \theta + \mathbf{s}/\sqrt{N}, \omega + \mathbf{u}/\sqrt{N}) + \lambda_N \sum_{j=1}^{J} \widehat{t}_j |\omega_j + u_j/\sqrt{N}|$. Define $(\widehat{\mathbf{v}}_N^{\mathrm{T}}, \widehat{\mathbf{s}}_N^{\mathrm{T}}, \widehat{\mathbf{u}}_N^{\mathrm{T}})^{\mathrm{T}} = \arg\min \Psi_N(\mathbf{v}, \mathbf{s}, \mathbf{u})$, then $\widehat{\alpha}_N = \alpha + \widehat{\mathbf{v}}_N/\sqrt{N}$, $\widehat{\theta}_N = \theta + \widehat{\mathbf{s}}_N/\sqrt{N}$, $\widehat{\omega}_N = \omega + \widehat{\mathbf{u}}/_N\sqrt{N}$, and $\widehat{\mathbf{v}}_N = \sqrt{N}(\widehat{\alpha}_N - \alpha)$, $\widehat{\mathbf{s}}_N = \sqrt{N}(\widehat{\theta}_N - \theta)$, $\widehat{\mathbf{u}}_N = \sqrt{N}(\widehat{\omega}_N - \omega)$. Define $\varepsilon_{ik\ell} = Y_{ik\ell} - p_{ik\ell}$ and $\varepsilon = \{\varepsilon_{ik\ell} : 1 \le i \le n_\ell, 1 \le k \le K_\ell, 1 \le \ell \le L\}^{\mathrm{T}}$. Then by Taylor expansion

$$
\begin{aligned}
&\Psi_N(\mathbf{v}, \mathbf{s}, \mathbf{u}) - \Psi_N(\mathbf{0}) \\
&= 2\{\mathcal{L}(\alpha + \mathbf{v}\sqrt{N}, \theta + \\
&\qquad \mathbf{s}/\sqrt{N}, \omega + \mathbf{u}/\sqrt{N}) - \mathcal{L}(\alpha, \theta, \omega)\} + \lambda_N \textstyle\sum_{j=1}^{J} \widehat{t}_j (|\omega_j + u_j/\sqrt{N}| - |\omega_j|) \\
&= \textstyle\sum_{\ell=1}^{2} \sum_{k=1}^{K_\ell} \sum_{i=1}^{n_\ell} V_{ik\ell} (\mathbf{v}^{\mathrm{T}}, \mathbf{s}^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}}) Q_{ik\ell} Q_{ik\ell}^{\mathrm{T}} (\mathbf{v}^{\mathrm{T}}, \mathbf{s}^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}})^{\mathrm{T}}/N \\
&\quad -2 \textstyle\sum_{\ell=1}^{2} \sum_{k=1}^{K_\ell} \sum_{i=1}^{n_\ell} \varepsilon_{ik\ell} Q_{ik\ell}^{\mathrm{T}} (\mathbf{v}^{\mathrm{T}}, \mathbf{s}^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}})^{\mathrm{T}}/\sqrt{N} \\
&\quad +\lambda_N/\sqrt{N} \textstyle\sum_{j=1}^{J} \widehat{t}_j \sqrt{N}(|\omega_j + u_j/\sqrt{N}| - |\omega_j|) + O(N^{-1/2}) \\
&= (\mathbf{v}^{\mathrm{T}}, \mathbf{s}^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}})(\mathbf{Q}^{\mathrm{T}}\mathbf{V}\mathbf{Q}/N)(\mathbf{v}^{\mathrm{T}}, \mathbf{s}^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}})^{\mathrm{T}} - 2(\mathbf{v}^{\mathrm{T}}, \mathbf{s}^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}})(\mathbf{Q}^{\mathrm{T}}\varepsilon/\sqrt{N}) \\
&\quad +\lambda_N/\sqrt{N} \textstyle\sum_{j=1}^{J} \widehat{t}_j \sqrt{N}(|\omega_j + u_j/\sqrt{N}| - |\omega_j|) + O(N^{-1/2}).
\end{aligned}
$$

By the Central Limit Theorem $\mathbf{Q}^{\mathrm{T}}\varepsilon/\sqrt{N} \to_d \mathbf{W} = \mathrm{Normal}(\mathbf{0}, \mathbf{\Sigma})$. Following the same argument as the proofs of Theorem 2 in Zou (2006), if $\omega_j \neq 0$, then $\widehat{t}_j \to_p |\omega_j|^{-\gamma}$ and $\sqrt{N}(|\omega_j + u_j/\sqrt{N}| - |\omega_j|) \to u_j \mathrm{sgn}(\omega_j)$. By Slutsky' theorem, one has

$\lambda_N/\sqrt{N}\sum_{j=1}^{J}\widehat{t}_j\sqrt{N}(|\omega_j+u_j/\sqrt{N}|-|\omega_j|)\rightarrow_p 0$. If $\omega_j=0$, then $\sqrt{N}(|\omega_j+u_j/\sqrt{N}|-|\omega_j|)=|u_j|$ and $\lambda_N\widehat{t}_j/\sqrt{N}=(\lambda_N/\sqrt{N})N^{\gamma/2}(|\sqrt{N}\widehat{\omega}_{j,0}|)^{-\gamma}\rightarrow\infty$ with probability approaching 1, since $\sqrt{N}\widehat{\omega}_{j,0}=O_p(1)$ and $\lambda_N N^{(\gamma-1)/2}\rightarrow\infty$.

Therefore by Slutsky's theorem, one has $\Psi_N(\mathbf{v},\mathbf{s},\mathbf{u})-\Psi_N(\mathbf{0})\rightarrow_d\Psi(\mathbf{v},\mathbf{s},\mathbf{u})$ where $\Psi(\mathbf{v},\mathbf{s},\mathbf{u})=(\mathbf{v}^{\mathrm{T}},\mathbf{s}^{\mathrm{T}},\mathbf{u}_{\mathcal{A}}^{\mathrm{T}})\mathbf{\Sigma}_{11}(\mathbf{v}^{\mathrm{T}},\mathbf{s}^{\mathrm{T}},\mathbf{u}_{\mathcal{A}}^{\mathrm{T}})^{\mathrm{T}}-2(\mathbf{v}^{\mathrm{T}},\mathbf{s}^{\mathrm{T}},\mathbf{u}_{\mathcal{A}}^{\mathrm{T}})\mathbf{W}_{\mathcal{A}}$ if $u_j=0$ for all $j\notin\mathcal{A}$ and $\Psi(\mathbf{v},\mathbf{s},\mathbf{u})=\infty$ otherwise, and where $\mathbf{W}_{\mathcal{A}}=\mathrm{Normal}(\mathbf{0},\mathbf{\Sigma_{11}})$. The unique minimum of $\Psi(\mathbf{v},\mathbf{s},\mathbf{u})$ is $\{(\mathbf{\Sigma_{11}^{-1}}\mathbf{W}_{\mathcal{A}})^{\mathrm{T}},\mathbf{0}_{\mathcal{A}}^{\mathrm{T}}\}^{\mathrm{T}}$, and thus $\widehat{\mathbf{u}}_{\mathcal{A}}\rightarrow_d\mathbf{\Sigma}_{11}^{-1}\mathbf{W}_{\mathcal{A}}$ and $\widehat{\mathbf{u}}_{\mathcal{A}^C}\rightarrow_d\mathbf{0}$.

Next we prove i) of Theorem 6. For all $j\in\mathcal{A}$, by the weak law of convergence, $\widehat{\omega}_j\rightarrow\omega_j$ in probability and thus $P(j\in\mathcal{A}_n^*)\rightarrow 1$. Then it suffices to show that for all $j'\notin\mathcal{A}$, $P(j'\in\mathcal{A}_n^*)\rightarrow 0$. Let $\mathbf{Q}=\{(\mathbf{Q}_\alpha)_{N\times(\sum K_l)},(\mathbf{Q}_\theta)_{N\times(\sum d_{kl})},(\mathbf{Q}_\omega)_{N\times J}\}$, where $\mathbf{Q}_\omega=(Q_{\omega,1},\cdots,Q_{\omega,J})$. For $j'\in\mathcal{A}^C$ and $j'\in\mathcal{A}_n^*$, one has $2Q_{\omega,j'}^{\mathrm{T}}(Y-\widehat{p})=\lambda_N\widehat{t}_{j'}$, where $\widehat{p}=\{\widehat{p}_{ikl}:1\leq i\leq n_\ell,1\leq k\leq K_L,1\leq l\leq L\}^{\mathrm{T}}$ and

$$\widehat{p}_{ik\ell}=H\{\widehat{\alpha}_{k\ell}+\beta_{k\ell}{\textstyle\sum_{j=1}^{J}}X_{ij\ell}\widehat{\omega}_j+Z_{ik\ell}^{\mathrm{T}}\widehat{\theta}_{k\ell}\}.$$

By the above results in the proof of part ii), $2Q_{\omega,j'}^{\mathrm{T}}(Y-\widehat{p})/\sqrt{N}=O_p(1)$, and $\lambda_N\widehat{t}_{j'}/\sqrt{N}\rightarrow_p\infty$. Thus for all $j'\in\mathcal{A}^C$, $P(j'\in\mathcal{A}_N^*)\leq P(2Q_{\omega,j'}^{\mathrm{T}}(Y-\widehat{p})=\lambda_N\widehat{t}_{j'})\rightarrow 0$.

C.2 Figures of Cross Validation Results for Colorectal Cancer Data

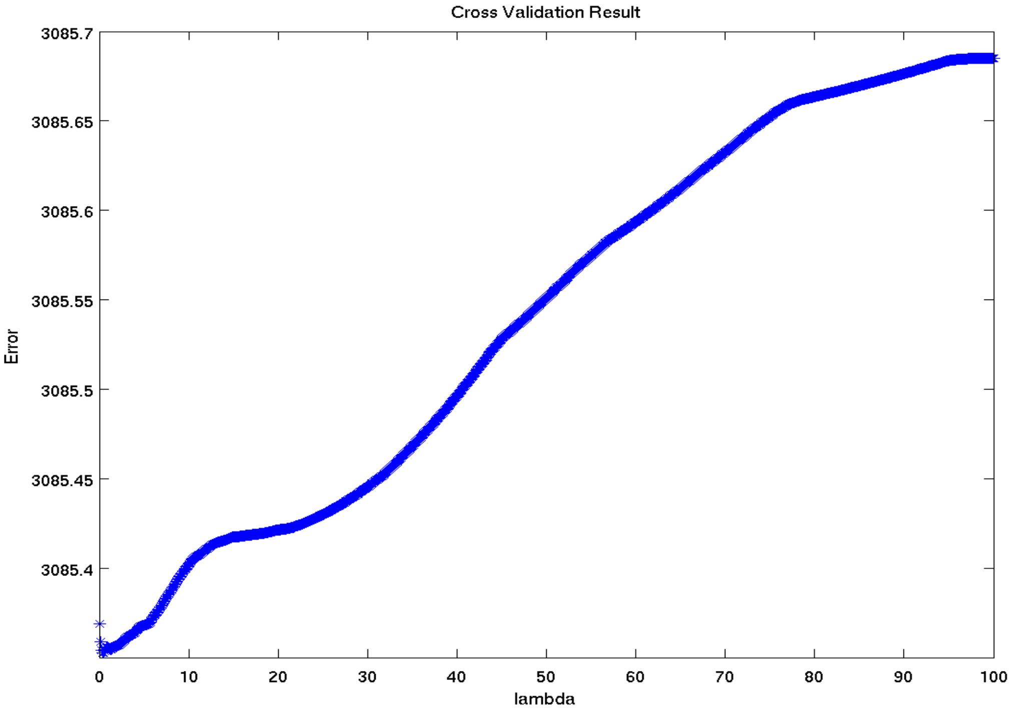

Figure C.1: Prediction error of cross-validation method in the adaptive lasso for the logistic regression of the colorectal cancer on HEI component scores of men (293615 with 2151 cases) and women together (198245 with 959 cases), where β_{11} = -1, $\widehat{\beta}_{12} = -0.7411$ and $\gamma = 1.5$.

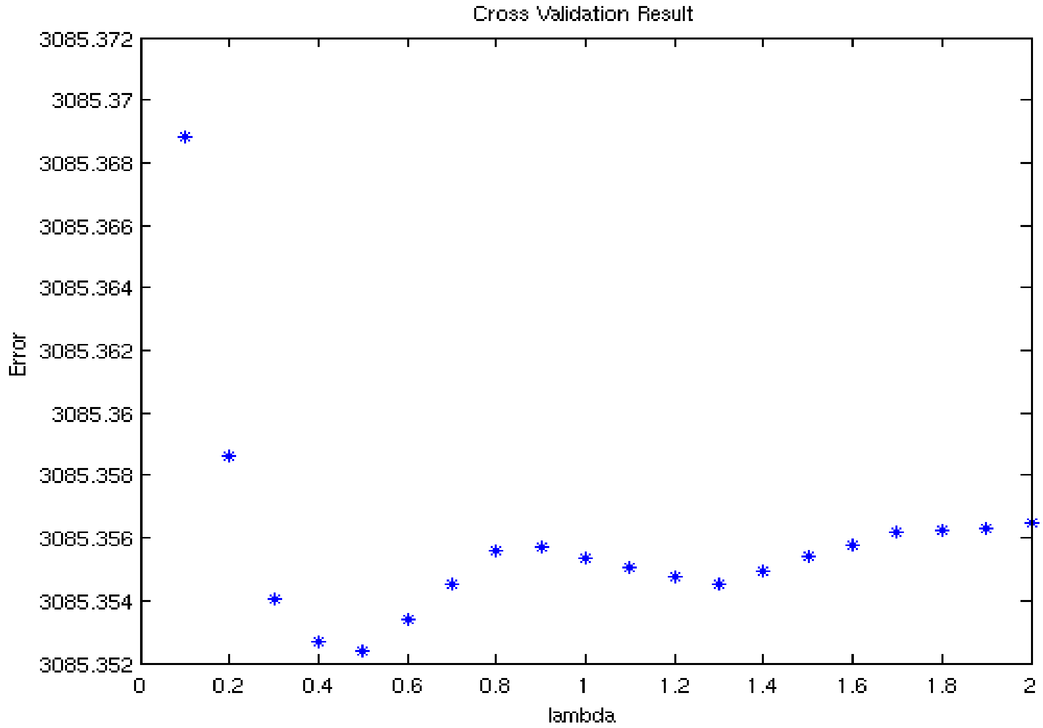

Figure C.2: Prediction error of cross-validation method in the adaptive lasso for the logistic regression of the colorectal cancer on HEI component scores of men (293615 with 2151 cases) and women together (198245 with 959 cases), where β_{11} = -1, $\widehat{\beta}_{12} = -0.7411$ and $\gamma = 1.5$.

APPENDIX D

SUPPLEMENTARY MATERIAL FOR CHAPTER 5

D.1 Expression of I-splines basis function with specifications

Based on the specifications in Section 5.2.2, expressions of $I_m(x|p,t)$ for $p = 2$ in $[t_{jm}, t_{jm+1})$ are as follows.

- when $m = 2$ $(t_m = c_{Rj})$,

$$\begin{aligned} I_s(x|p=2,t) &= 0 \qquad \text{for } s > m, \\ I_m(x|p=2,t) &= (x-t_m)^2/(2d_j^2), \\ I_{m-1}(x|p=2,t) &= 1-(t_{m+1}-x)^2/d_j^2. \end{aligned}$$

- when $3 \leq m \leq e+p-1$, or which written as $3 \leq m \leq e+1$,

$$\begin{aligned} I_s(x|p=2,t) &= 0 \qquad \text{for } s > m, \\ I_m(x|p=2,t) &= (x-t_m)^2/(2d_j^2), \\ I_{m-1}(x|p=2,t) &= 1-(t_{m+1}-x)^2/(2d_j^2), \\ I_s(x|p=2,t) &= 1 \qquad \text{for } s < m-1. \end{aligned}$$

- when $m = e + p = e + 2$,

$$\begin{aligned} I_s(x|p=2,t) &= 0 \qquad \text{for } s > m, \\ I_m(x|p=2,t) &= (x-t_m)^2/d_j^2, \\ I_{m-1}(x|p=2,t) &= 1-(t_{m+1}-x)^2/(2d_j^2), \\ I_s(x|p=2,t) &= 1 \qquad \text{for } s < m-1. \end{aligned}$$

For I-spline with $p = 3$ and x is in $[t_{jm}, t_{jm+1})$,

- when $m = 3$ $(t_m c_{Rj})$,

$$\begin{aligned} I_s(x|p=3,t) &= 0 \qquad \text{for } s > m, \\ I_m(x|p=3,t) &= (x-t_m)^3/(6d_j^3), \\ I_{m-2}(x|p=3,t) &= 1-(t_{m+1}-x)^3/d_j^3, \\ I_{m-1}(x|p=3,t) &= a_1/(2d_j^3) + a_2/(4d_j^3). \\ \text{where} & \\ a_1 &= 3/2(2t_m+d)(x^2-t_m^2) - 3t_m(t_m+d_j)(x-t_m) - (x^3-t_m^3), \\ a_2 &= 3(t_m+d_j)(x^2-t_m^2) - 3t_m(t_m+2d_j)(x-t_m) - (x^3-t_m^3). \end{aligned}$$

- when $m = 4$,

$$\begin{aligned}
I_s(x|p=3,t) &= 0 \qquad \text{for } s > m, \\
I_m(x|p=3,t) &= (x - t_m)^3/(6d_j^3), \\
I_{m-2}(x|p=3,t) &= 1 - (t_{m+1} - x)^3/(4d_j^3), \\
I_{m-1}(x|p=3,t) &= 1/6 + a_1/(6d_j^3) + a_2/(6d_j^3), \\
I_s(x|p=3,t) &= 1 \qquad \text{for } s < m-2.
\end{aligned}$$

where

$$\begin{aligned}
a_1 &= 3t_m(x^2 - t_m^2) - 3(t_m - d_j)(t_m + d_j)(x - t_m) - (x^3 - t_m^3), \\
a_2 &= 3(t_m + d_j)(x^2 - t_m^2) - 3t_m(t_m + 2d_j)(x - t_m) - (x^3 - t_m^3).
\end{aligned}$$

- when $5 \le m \le e + p - 2$, or expressed as $5 \le m \le e + 1$,

$$\begin{aligned}
I_s(x|p=3,t) &= 0 \qquad \text{for } s > m, \\
I_m(x|p=3,t) &= (x - t_m)^3/(6d_j^3), \\
I_{m-2}(x|p=3,t) &= 1 - (t_{m+1} - x)^3/(6d_j^3), \\
I_{m-1}(x|p=3,t) &= 1/6 + a_1/(6d_j^3) + a_2/(6d_j^3), \\
I_s(x|p=3,t) &= 1 \qquad \text{for } s < m-2.
\end{aligned}$$

where

$$\begin{aligned}
a_1 &= 3t_m(x^2 - t_m^2) - 3(t_m - d_j)(t_m + d_j)(x - t_m) - (x^3 - t_m^3), \\
a_2 &= 3(t_m + d_j)(x^2 - t_m^2) - 3t_m(t_m + 2d_j)(x - t_m) - (x^3 - t_m^3).
\end{aligned}$$

- when $m = e + p - 1 = e + 2$,

$$\begin{aligned}
I_s(x|p=3,t) &= 0 \qquad \text{for } s > m, \\
I_m(x|p=3,t) &= (x - t_m)^3/(4d_j^3), \\
I_{m-2}(x|p=3,t) &= 1 - (t_{m+1} - x)^3/(6d_j^3), \\
I_{m-1}(x|p=3,t) &= 1/6 + a_1/(6d_j^3) + a_2/(6d_j^3), \\
I_s(x|p=3,t) &= 1 \qquad \text{for } s < m - 2.
\end{aligned}$$

where

$$\begin{aligned}
a_1 &= 3t_m(x^2 - t_m^2) - 3(t_m - d_j)(t_m + d_j)(x - t_m) - (x^3 - t_m^3), \\
a_2 &= 3(t_m + d_j)(x^2 - t_m^2) - 3t_m(t_m + 2d_j)(x - t_m) - (x^3 - t_m^3).
\end{aligned}$$

- when $m = e + p = e + 3$,

$$\begin{aligned}
I_s(x|p=3,t) &= 0 \qquad \text{for } s > m, \\
I_m(x|p=3,t) &= (x - t_m)^3/d_j^3, \\
I_{m-2}(x|p=3,t) &= 1 - (t_{m+1} - x)^3/(6d_j^3), \\
I_{m-1}(x|p=3,t) &= 1/4 + a_1/(4d_j^3) + a_2/(2d_j^3), \\
I_s(x|p=3,t) &= 1 \qquad \text{for } s < m - 2.
\end{aligned}$$

where

$$\begin{aligned}
a_1 &= 3t_m(x^2 - t_m^2) - 3(t_m - d_j)(t_m + d_j)(x - t_m) - (x^3 - t_m^3), \\
a_2 &= 3/2(2t_m + d_j)(x^2 - t_m^2) - 3t_m(t_m + d_j)(x - t_m) - (x^3 - t_m^3).
\end{aligned}$$

After these specifications, the I-spline functions are now much easier to be applied into our model.

D.2 Figures for I-spline Analysis in HEI-2005

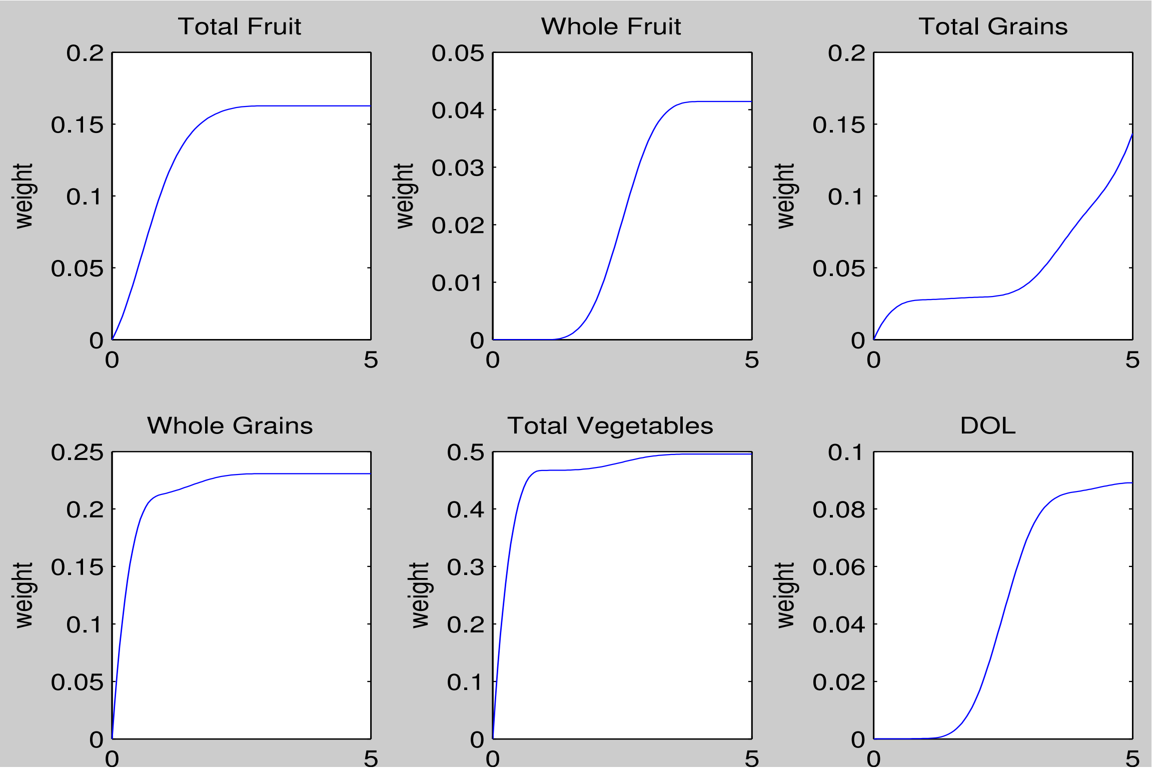

Figure D.1: I-spline analysis results for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$.

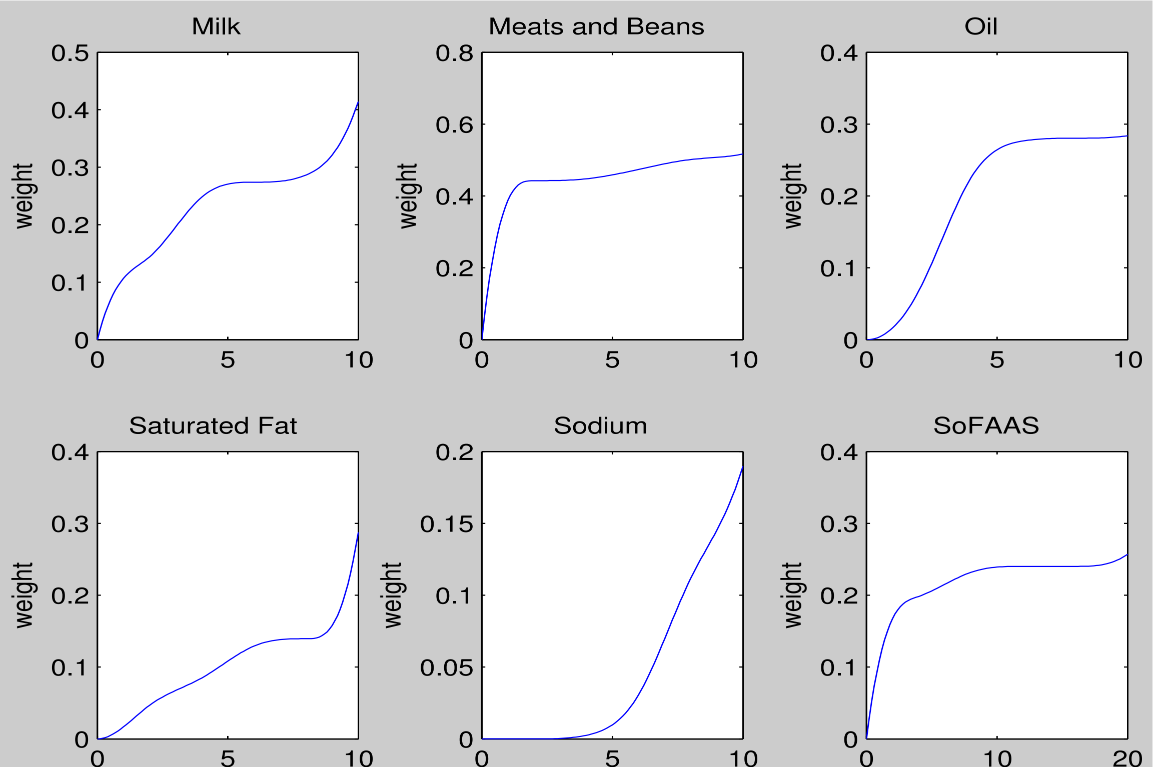

Figure D.2: I-spline analysis results for the logistic regression of the colorectal and lung cancers on HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$.

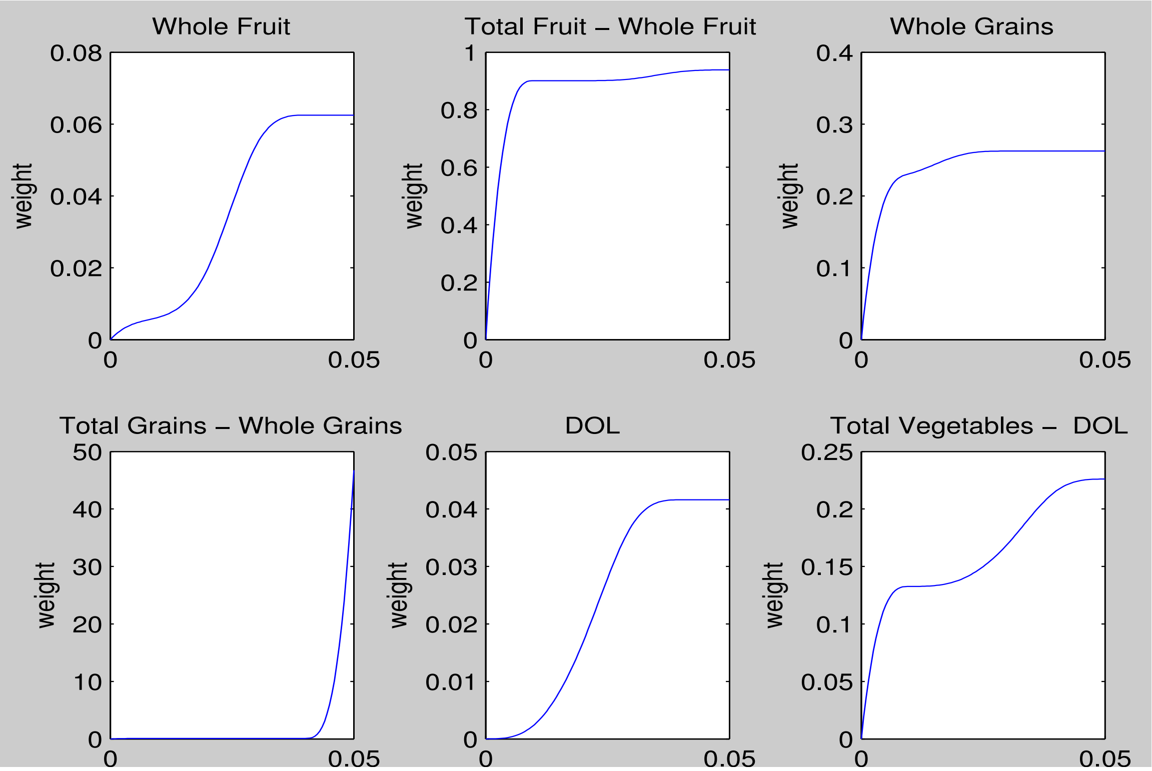

Figure D.3: I-spline analysis results for the logistic regression of the colorectal and lung cancers on modified HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$.

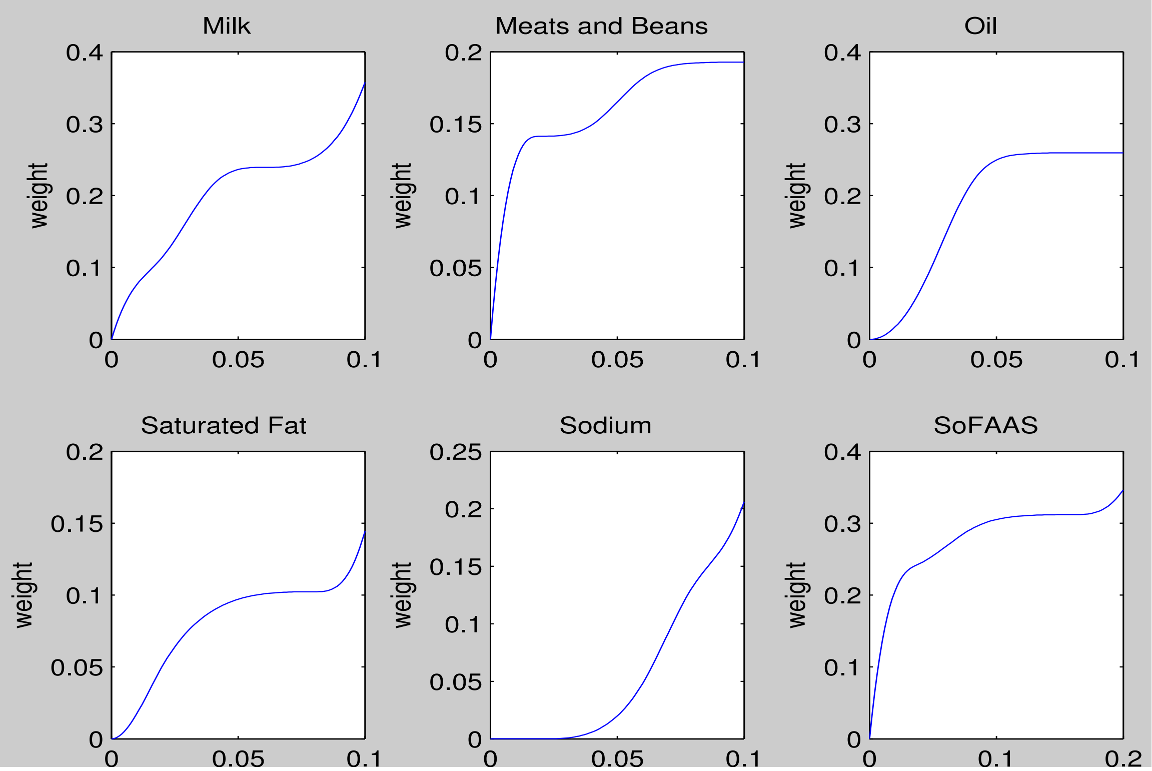

Figure D.4: I-spline analysis results for the logistic regression of the colorectal and lung cancers on modified HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$.

D.3 I-Spline Analysis Results for Modified HEI-2005 Components

		Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Whole Fruit	Estimate	0.006	0.014	0.048	0.056	0.056
	s.e (vs quintile 1)		0.079	0.071	0.084	0.064
	p-value (vs quintile 1)		0.864	0.498	0.503	0.380
Total Fruit - Whole Fruit	Estimate	0.901	0.000	0.006	0.031	0.038
	s.e (vs quintile 1)		0.113	0.096	0.103	0.242
	p-value (vs quintile 1)		1.000	0.948	0.760	0.877
Whole Grains	Estimate	0.231	0.025	0.031	0.031	0.031
	s.e (vs quintile 1)		0.050	0.057	0.093	0.187
	p-value (vs quintile 1)		0.614	0.583	0.736	0.867
Total Grains - Whole Grains	Estimate	0.098	0.000	0.000	0.000	46.629
	s.e (vs quintile 1)		0.050	0.061	0.347	57.045
	p-value (vs quintile 1)		1.000	1.000	1.000	0.414
DOL	Estimate	0.002	0.014	0.035	0.039	0.039
	s.e (vs quintile 1)		0.052	0.051	0.072	0.088
	p-value (vs quintile 1)		0.782	0.497	0.586	0.658
Total Vegetables - DOL	Estimate	0.133	0.005	0.037	0.083	0.093
	s.e (vs quintile 1)		0.054	0.062	0.165	2.118
	p-value (vs quintile 1)		0.922	0.556	0.614	0.965

Table D.1: I-spline analysis results for the logistic regression of the colorectal and lung cancers on modified HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$.

		Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Milk	Estimate	0.114	0.101	0.125	0.139	0.243
	s.e (vs quintile 1)		0.048	0.044	0.060	0.056
	p-value (vs quintile 1)		0.037	0.005	0.020	0.000
Meats and Beans	Estimate	0.141	0.008	0.040	0.051	0.052
	s.e (vs quintile 1)		0.166	0.156	0.161	0.158
	p-value (vs quintile 1)		0.962	0.797	0.752	0.744
Oil	Estimate	0.069	0.147	0.189	0.190	0.190
	s.e (vs quintile 1)		0.065	0.059	0.067	0.063
	p-value (vs quintile 1)		0.024	0.001	0.004	0.002
Saturated Fat	Estimate	0.050	0.039	0.051	0.052	0.094
	s.e (vs quintile 1)		0.049	0.049	0.074	0.162
	p-value (vs quintile 1)		0.431	0.302	0.483	0.562
Sodium	Estimate	0.000	0.006	0.048	0.133	0.206
	s.e (vs quintile 1)		0.080	0.067	0.070	0.074
	p-value (vs quintile 1)		0.942	0.469	0.059	0.005
SoFAAS	Estimate	0.244	0.047	0.066	0.068	0.102
	s.e (vs quintile 1)		0.065	0.058	0.069	0.075
	p-value (vs quintile 1)		0.468	0.253	0.324	0.174

Table D.2: I-spline analysis results for the logistic regression of the colorectal and lung cancers on modified HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1$.

		Unweighted Model			Weighted Model			Model Comparison	
Gender	Disease	$\widehat{\beta}_{k\ell}$	s.e	p-value	$\widehat{\beta}_{k\ell}$	s.e	p-value	LR	p-value
Male	Colorectal Cancer	-0.021	0.003	0.000	-48.156	4.158	0.000	63.746	0.000
Male	Lung Cancer	-0.015	0.002	0.000	-22.022	3.686	0.000	-2.762	1.000
Female	Breast Cancer	-0.001	0.002	0.737	-0.534	2.875	0.853	-0.078	1.000
Female	Colorectal Cancer	-0.015	0.004	0.000	-42.585	6.225	0.000	27.174	0.000
Female	Lung Cancer	-0.008	0.003	0.003	-18.845	4.785	0.000	6.736	0.009

Table D.3: I-spine analysis results of $\widehat{\beta}_{k\ell}$ for the logistic regression of the colorectal and lung cancers on modified HEI component scores of men (219612 with 3348 and 4187 cases, respectively) and the breast, colorectal and lung cancers on women (169480 with 6647, 1846 and 2933 cases, respectively) together, where $\beta_{11} = -1, \gamma = 1.5$. Unweighted Model–$\mathrm{pr}(Y_{ik\ell} = 1|X_{ij\ell}, Z_{ik\ell}) = H(\alpha_{k\ell} + \beta_{k\ell}\sum_{j=1}^{J} X_{ij\ell} + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell})$, Weighted Model–$\mathrm{pr}(Y_{ik\ell} = 1|X_{i1\ell}, ..., X_{iJ\ell}, Z_{ik\ell}) = H[\alpha_{k\ell}+\beta_{k\ell}\sum_{j=1}^{J}\{\sum_{q=1}^{Q} I_{jq}(X_{ij\ell})\widehat{\omega}_{jq}\}+Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell}]$ with constrain $\omega_{jq} \geq 0$. LR–Likelihood ratio test statistic.