2010

The Effect of Risk Factors on Coronary Heart Disease: An Age-Relevant Multivariate Meta Analysis

Yan Li

THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

THE EFFECT OF RISK FACTORS ON CORONARY HEART DISEASE: AN AGE-RELEVANT MULTIVARIATE META ANALYSIS

By

YAN LI

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Fall Semester, 2010

The members of the Committee approve the Dissertation of Yan Li defended on August 18, 2010.

Dan McGee
Professor Co-Directing Dissertation

Yiyuan She
Professor Co-Directing Dissertation

Ike Eberstein
University Representative

Xufeng Niu
Committee Member

The Office of Graduate Studies has verified and approved the above named committee members.

I dedicated this to my parents, my husband and my kids for their love and support.

ACKNOWLEDGEMENTS

I am very grateful to my major advisors Dr. Dan McGee and Dr. Yiyuan She. I would like to thank them for their guidance, encouragement, patience, consistent support, and the time they have spent on directing my dissertation. They led me into this research step by step, and enhanced my interest in continuing the Cardiovascular Outcomes research in my future career.

I would like to acknowledge my gratitude to my committee member Dr. Xufeng Niu who has been so helpful, willing and generous with his time. I have been benefitting so much from his suggestions and great help on my study and research in the past five years. I wish to thank my outside committee member Dr. Ike Eberstein for his support and encouragement.

I also wish to express my appreciation to the faculty and staff of the department for their support during my study. I especially benefitted from Dr. Fred Huffer's wonderful teaching and great help. My appreciation also extends to the great staff, Pamela McGhee, Chauncey Richburg, Jennifer Rivera, and James Stricherz for their very efficient and supportive assistance.

I am indebted to my parents, Zhiming and Rongxiang, for their selfless love and continuous support. I also would like to express special appreciation to my husband Zhenyu for his love, full support and encouragement. I also would like to thank my kids Michael and Karissa for their love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The importance of major risk factors, such as hypertension, total cholesterol, body mass index, diabetes, smoking, for predicting incidence and mortality of Coronary Heart Disease (CHD) is well known. In light of the fact that age is also a major risk factor for CHD death, a natural question is whether the risk effects on CHD change with age. This thesis focuses on examining the interaction between age and risk factors using data from multiple studies containing differing age ranges.

The aim of my research is to use statistical methods to determine whether we can combine these diverse results to obtain an overall summary, using which one can find how the risk effects on CHD death change with age. One intuitive approach is to use classical meta analysis based on generalized linear models. More specifically, one can fit a logistic model with CHD death as response and age, a risk factor and their interaction as covariates for each of the studies, and conduct meta analysis on every set of three coefficients in the multivariate setting to obtain 'synthesized' coefficients.

Another aspect of the thesis is a new method, meta analysis with respect to curves that goes beyond linear models. The basic idea is that one can choose the same spline with the same knots on covariates, say age and systolic blood pressure (SBP), for all the studies to ensure common basis functions. The knot-based tensor product basis coefficients obtained from penalized logistic regression can be used for multivariate meta analysis. Using the common basis functions and the 'synthesized' knot-based basis coefficients from meta analysis, a two-dimensional smooth surface on the age-SBP domain is estimated. By cutting through the smooth surface along two axes, the resulting slices show how the risk effect on CHD death change at an arbitrary age as well as how the age effect on CHD death change at an arbitrary SBP value. The application to multiple studies will be presented.

CHAPTER 1

INTRODUCTION

Extensive statistical data show that cardiovascular disease continues to be the largest killer in the United States [1]. Millions of people have developed Coronary Heart Disease (CHD) involving angina and myocardial infarction, of which nearly half a million people die every year. The study of the prevention and treatment of CHD is therefore of great importance in medical research and public health.

For years, many clinicians and practitioners have made efforts to explore what risk factors are predictive of CHD events or mortality in men or women in order to help people avoid the complications that result from the development of CHD. In the past, a great number of papers have focused on a specified follow-up study to analyze the effects of such risk factors on CHD as hypertension, hypercholesterolemia, diabetes, obesity, cigarette smoking, alcohol intake, physical inactivity, stress, etc, in order to investigate their associations with CHD events and mortality, develop CHD risk equations for use in predicting the development of CHD in individuals free of heart diseases, or even provide guidelines on CHD prevention in clinical practice [2, 3, 4, 5].

An issue we face is that aging is becoming a world-wide phenomenon. In the United States, people 65 and above are the fastest growing population group and most CHD deaths occur in this group. In light of the importance of age on increasing risk of CHD, some researchers have begun to focus on the study of risk factor effects on CHD incidence and mortality in the elderly. Some of them attempted to examine the relationship of total cholesterol level to CHD in old men and women in a follow-up study, some showing a positive association but others showing no significant association. Since this topic remained controversial, E. Anum and T. Adera [6] conducted a meta analysis to analyze the association between total cholesterol level and CHD events and mortality in persons aged 65 and above

and quantify the magnitude of the association with a summary relative risk, using data from 33 published follow-up studies. Meanwhile, R. Abbott et al. [7] used the Honolulu Heart Program to explore the relation between age and common risk factors for CHD; hypertension, cholesterol, diabetes, smoking, alcohol, and physical activity. Proportional hazard regression models including the risk factors and their interactions with age were conducted to assess the risk factor effect and the change of the effect with age on the incidence of CHD. Past studies, however, were limited to either only one follow-up study or a single risk factor, from which it is hard to unveil the essence of the risk factor effects individually.

In this thesis, I will analyze data from more than twenty original independent follow-up studies by means of logistic model and generalized additive models and integrate the findings to investigate the interaction between age and risk factors on CHD death by virtue of meta-analysis. The diverse nature of these various studies may help us clarify the associations of the risk factors with CHD mortality as age increases and further provide more comprehensive evidence about suggestions for guiding both young adults at lower risk and elderly people at higher risk on the prevention of CHD death.

CHAPTER 2

STATISTICAL BACKGROUND

2.1 Introduction

The exploration of the risk factors on CHD has remained popular as a topic in medical research for years. With an increasing trend of the number of older people, studies moved on to the risk factors of CHD in an elderly population.

The risk factors for CHD [8] generally include hypertension, cholesterol, diabetes, body mass index (BMI), cigarette smoking, alcohol intake, physical inactivity, etc. Among the risk factors, cigarette smoking, alcohol intake and physical activity are closely related to a person's living habits. On the other hand, many research results show that a person has higher chance to develop hypertension and diabetes as age increases while serum cholesterol levels and BMI often decline in the elderly. Thus, the effect of these factors on CHD may change with age. The effect of cholesterol, particularly, remains controversial. Some studies showed that cholesterol is an independent predictor of CHD among men older than 65 years [9, 10], while others suggested that there is no association in persons older than 70 years [11, 12]. In the oldest group the relationship was reported to become negative [13].

In 2004, E. Anum and T. Adera [6] conducted a meta analysis to investigate the association between total cholesterol level and CHD events and mortality in the elderly with a summary relative risk derived using data from 33 published follow-up studies. According to the results, in men followed from middle-age and above, total cholesterol has a significantly positive association with CHD incidence, CHD mortality as well as all-cause mortality. In men aged 65 years and above at entry, the positive associations between total cholesterol and CHD incidence as well as CHD mortality are significant while total cholesterol did not show any association with CHD mortality in women followed from 65 years and above. By contrast, for men aged 80 years and above at entry, the association with CHD mortality is

not significant and total cholesterol showed an inverse association with all-cause mortality. In addition, the relationship between HDL-C and CHD mortality in men did not show any significant association. Although this meta analysis can help to gain insights into medical research in CHD, it only used the results from published literature that seldom contains sufficient information to conduct more complex meta analysis. Further, the limitations of this study, such as excluding unpublished studies, and using a serum total cholesterol level ≥ 240 mg/dL vs. < 200 mg/dL, may introduce some bias on the estimation of true summary relative risk.

It is commonly known that the incidence of CHD increases with age, thus these factors always interact with age to some extent so that the comparison of the effects of risk factors on CHD among different age groups could be taken into account. In 2001, S. Franklin et al. [14] examined the relation of blood pressure including diastolic (DBP), systolic (SBP) and pulse pressure (PP) to CHD risk among three age groups using data combining the Framingham Heart Study cohort with the Framingham Offspring Study cohort. The Cox proportional hazard model was used to fit models with SBP, DBP or PP separately for three age groups (< 50, 50 to 59 and ≥ 60 years). Comparing the hazard ratios associated with a 10 mmHg increase in blood pressure indicated that DBP was the strongest predictor in the group < 50 years of age while SBP became stronger than DBP in the group 50 to 59 years of age. With increasing age, from 60 years of age on, PP became stronger compared to SBP and even DBP had inversely negative association with CHD incidence. In other words, as age increases, a gradual transition from DBP to SBP and then to PP as predictors of CHD events was found.

In 2002, R. Abbott et al. [7] used the Honolulu Heart Program to explore the relation between age and common risk factors for CHD: hypertension, cholesterol, diabetes, smoking, alcohol, and physical activity. Proportional hazard regression models including the risk factors and their interactions with age were used to provide assessments of a risk factor effect and the change of the effect with age on the incidence of CHD after adjusting other risk factors. The results showed that the positive association between hypertension severity and CHD became weaker with advancing age, while total cholesterol was not significantly related to CHD in the oldest men aged over 75. The effect of diabetes on the risk of CHD remained consistently 2-fold across all 10-year age ranges, and in contrast, the positive association between body mass index and CHD inversely changed to be significantly negative

from younger men groups (45 to 64) to the oldest men group (75 to 93). Furthermore, the protective effect of alcohol on CHD risk in men < 74 years old disappeared in the oldest group. In addition, active life-style and good life habit, such as doing exercise regularly especially in middle-age men and stoping smoking in earlier age, may decrease the risk of CHD.

While the previous studies provide considerable insight into the relationship between CHD and risk factors with increasing age, most of them only focused on a single study or one specified risk factor. Even though a few other studies considered multiple studies, only summary data from published research papers were integrated and analyzed, which may produce publication bias. No meta analyses have been performed based on individual patient data (IPD).

2.2 Logistic Regression Model

The logistic model is a useful way of describing the relationship between one or more risk factors and a dichotomous outcome such as disease events or mortality. The specific form of the logistic model is:

$$p = \Pr(Y = 1|X_1, ..., X_p) = \frac{e^{(\beta_0+\beta_1 X_1+...+\beta_p X_p)}}{1 + e^{(\beta_0+\beta_1 X_1+...+\beta_p X_p)}} \tag{2.1}$$

where Y is a binary outcome variable ($= 1$ for the outcome event, $= 0$ otherwise), $X_1, ..., X_p$ are independent variables, often called *covariates*, such as risk factors, and $\beta_0, \beta_1, ..., \beta_p$ are unknown parameters to be estimated using maximum likelihood, which yields values for unknown parameters that maximize the probability of obtaining the observed set of data [15].

For the logistic model, the likelihood function is used to express the probability of observed data as a function of the unknown parameters. This function is:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} {p_i}^{y_i} {p_i}^{1-y_i} \tag{2.2}$$

where $p_i = \frac{e^{(\beta_0+\beta_1 X_{i1}+...+\beta_p X_{ip})}}{1+e^{(\beta_0+\beta_1 X_{i1}+...+\beta_p X_{ip})}}$, and y_i denotes the outcome of event for i^{th} subject.

The log-likelihood is:

$$l(\beta) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \log(p_i) + (1 - y_i) \log(p_i)\}. \tag{2.3}$$

The p+1 likelihood equations are obtained by differentiating the log likelihood function with respect to the p+1 coefficients as following:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} p_i$$

and

$$\sum_{i=1}^{n} x_{ij} \{y_i - p_i\} = 0 \text{ for } j = 1, 2, ..., p.$$

It is common to use a numerical algorithm, such as Newton-Raphson algorithm [16], to obtain the estimated coefficients, $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$. The estimators of variances and covariances of these estimated coefficients can be obtained by the inverse of the information matrix, $\text{var}(\hat{\beta}) = I(\hat{\beta})^{-1}$, where $I(\hat{\beta})$ is called the information matrix, the matrix of a second partial derivatives of the log likelihood function. A single logistic regression coefficient is tested by Wald test statistic following asymptotically standard normal distribution, $W_j = \hat{\beta}_j / \text{se}(\hat{\beta}_j)$.

The logistic model can help to provide a clinically meaningful interpretation. The interpretation involves two aspects: determining the relationship between the outcome and the covariate according to the sign and quantity of estimated coefficients, $\hat{\beta}_i$ and predicting risk of that outcome using the estimated logistic model. Further, the logistic model is a powerful analytic tool because of the relationship between odds ratio (OR) and the estimated coefficient, $OR = e^{\hat{\beta}}$. For example, if y denotes the incidence of CHD and if x denotes whether the person has diabetes, then $OR = 2$ estimates that CHD is twice as likely to occur among people with diabetes than those without diabetes.

2.3 Meta Analysis

Meta analysis is a statistical method that integrates the findings from related but independent studies on the same topic for the purpose of obtaining an overall summary of the results [17, 18]. Meta analysis dated to 1976 [19], and has gained increasing popularity in medical research for recent years. In clinical trials, various clinical studies with similar treatment protocols are combined to reveal efficacy of a treatment, namely the effect size, such as log odds-ratio mentioned in Section 2.2. By means of meta analysis, one can amalgamate the effect sizes from different studies to provide findings that cannot be answered by a single study alone. Moreover, meta analysis can provide more statistical power to detect a treatment effect than an analysis based on a single study.

Clinical studies generally come from the diverse design and methods to some extent, such as controlled experiments or less well controlled, randomized or unrandomized design, and have different sample sizes and patient populations. Therefore, when combining the results from the studies in meta-analysis, weights are assigned for each study to remove the different level of sampling error. In 1986, R. DerSimonian and N. Laird [20] proposed a noniterative random-effects approach to estimate the overall effects by the observed effects from individual studies. In 1995, C. Berkey extended the random-effects model of DerSimonian and Laird to include relevant covariates which may explain heterogeneity [21] by performing meta regression. Besides single outcomes, meta analysis were also extended to two or more outcomes [22, 23, 24, 25, 26, 27, 28] when taking the correlation of multiple outcomes into account.

2.3.1 Univariate Meta Analysis

Fixed-effects Meta Analysis

A fixed-effects model assumes that all studies are drawn from a population with the same underlying parameter so that there is no between-study variation. The measure of effect size refers to the outcome measure of a study, commonly a summary statistic like the log odds-ratio, the log hazard-ratio, the risk difference, or the relative-risk.

Let Y_i be the observed measure of effect size in study i, $i = 1, 2, ..., k$, θ the unknown parameter of interest and $s_i^2 = \text{var}(Y_i)$ the variance of the measure of effect size in the i^{th} study. We assume $E(Y_i) = \theta$ and s_i^2 is known. According to the Central Limit Theorem, for moderately large study sizes, each Y_i should be asymptotically normally distributed and approximately unbiased [29]. Thus

$$Y_i|\theta, s_i^2 \sim N(\theta, s_i^2) \text{ for } i = 1, 2, ..., k.$$

When s_i^2 is assumed known, the log-likelihood for θ is proportional to $\sum_{i=1}^{k} (y_i - \theta)^2/s_i^2$, and the maximum likelihood estimator (MLE) for θ is:

$$\hat{\theta}_{MLE} = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i} \sim N(\theta, \frac{1}{\sum_{i=1}^{k} W_i}) \tag{2.4}$$

with $W_i = 1/s_i^2$.

Random-effects Meta Analysis

Often, the data do not support a fixed-effect model and a random-effects model is more suitable [30]. A random-effects model assumes that there is the between-study variation, i.e. the true effect sizes across studies are not identical.

Suppose that each observed effect size Y_i comes from a distribution with a study-specific mean, θ_i and variance, s_i^2.

$$Y_i|\theta_i, s_i^2 \sim N(\theta_i, s_i^2) \text{ for } i = 1, 2, ..., k \tag{2.5}$$

Under the assumption, each study-specific mean, θ_i, comes from a Normal distribution with mean, θ and variance, τ^2, i.e., $\theta_i \sim N(\theta, \tau^2)$. When τ^2=0, the random-effects model reduces to the fixed-effects model.

The Q statistic is a common way to evaluate whether the true effect sizes are identical across studies [31]. When the null hypothesis $H_0 : \tau^2 = 0$ (i.e. $H_0 : \theta_1 = ... = \theta_k$) is true,

$$\hat{\theta} = \frac{\sum_{i=1}^k W_i Y_i}{\sum_{i=1}^k W_i} \tag{2.6}$$

with $W_i = 1/s_i^2$ and $Q = \sum_{i=1}^k W_i(Y_i - \hat{\theta})^2$ is distributed as chi-squared with k-1 degrees of freedom. The null hypothesis will be rejected if the Q statistic exceeds the $100(1-\alpha)$th percentile of a chi-squared distribution. It is instructive to detect heterogeneity among studies in the stage of initial analysis and a random-effect meta-analysis is suitable if the hypothesis of homogeneity is rejected. However, it is generally agreed that the power of this test can be low [32] and therefore a non-significant result should be cautiously treated.

The MLE of θ is given by

$$\hat{\theta}_{MLE} = \frac{\sum_{i=1}^k W_i Y_i}{\sum_{i=1}^k W_i} \sim N\left(\theta, \frac{1}{\sum_{i=1}^k W_i}\right) \tag{2.7}$$

which has the same form as a fixed-effect model but with different weights, $W_i = 1/(s_i^2 + \tau^2)$.

For the random-effects model, several ways can be used to estimate the between study variance τ^2. DerSimonian and Laird [20] proposed a simple non-iterative procedure, the Method of Moments(MM), to estimate τ^2 based on the Q statistic. Under the assumptions of the random effects model, the expectation of Q is

$$E(Q) = (k-1) + \left(\sum_{i=1}^k W_i - \sum_{i=1}^k W_i^2 / \sum_{i=1}^k W_i\right)\tau^2 \tag{2.8}$$

and substituting the observed value of Q for $E(Q)$ provides the moments estimator

$$\hat{\tau}^2_{MM} = \max\left(0, \frac{Q-(k-1)}{\sum_{i=1}^{k} W_i - \frac{\sum_{i=1}^{k} W_i^2}{\sum_{i=1}^{k} W_i}}\right) \tag{2.9}$$

where truncation of $\hat{\tau}^2_{MM}$ at zero is used to ensure that the variance estimate is non-negative.

Maximum likelihood and restricted maximum likelihood are not as simple as the non-iterative MM estimator since they need iterative procedures to derive an estimator of τ^2. Assuming the within-study variance for the ith study, s_i^2, is known, the log likelihood is given by

$$\log L(\theta, \tau^2) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{k}\log(s_i^2+\tau^2) - \frac{1}{2}\sum_{i=1}^{k}\frac{(Y_i-\theta)^2}{s_i^2+\tau^2}$$

Since the k studies are assumed independent, the likelihood can be obtained and maximized to provide maximum likelihood estimates of unknown parameters, θ and τ^2. Starting with an initial estimate of $\hat{\tau}^2$, the maximum likelihood estimators can be obtained as follows:

$$\hat{\tau}^2_{ML} = \frac{\sum_{i=1}^{k} \hat{W}_i^{\,2}[(Y_i-\hat{\theta})^2 - s_i^2]}{\sum_{i=1}^{k} \hat{W}_i^{\,2}} \tag{2.10}$$

with $\hat{\theta} = \sum_{i=1}^{k} \hat{W}_i Y_i / \sum_{i=1}^{k} \hat{W}_i$ and $\hat{W}_i = 1/(s_i^2 + \hat{\tau}^2_{ML})$. The iteration stops at a specified convergence criterion.

The restricted maximum likelihood (REML) estimator can be found by maximizing the following log likelihood:

$$\log L_{REML}(\theta, \tau^2) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{k}\log(s_i^2+\tau^2) - \frac{1}{2}\log\left(\sum_{i=1}^{k}\frac{1}{s_i^2+\tau^2}\right) - \frac{1}{2}\sum_{i=1}^{k}\frac{(Y_i-\hat{\theta}^2)}{s_i^2+\tau^2}$$

Compared to ML estimation, the REML log likelihood function has a correction term in order to obtain an unbiased estimator for τ^2 [33], which is given by:

$$\hat{\tau}^2_{REML} = \frac{\sum_{i=1}^{k} \hat{W}_i^{\,2}[(Y_i-\hat{\theta})^2 + \sum_{i=1}^{k}\hat{W}_i - s_i^2]}{\sum_{i=1}^{k} \hat{W}_i^{\,2}} \tag{2.11}$$

with $\hat{\theta} = \sum_{i=1}^{k} \hat{W}_i Y_i / \sum_{i=1}^{k} \hat{W}_i$ and $\hat{W}_i = 1/(s_i^2 + \hat{\tau}^2_{REML})$. $\hat{\tau}^2_{RE}$ also starts with an initial estimate and the iteration stops at a specified convergence criterion.

The MM, ML and REML estimators are relatively acceptable when the between study variance is small [33]. Another estimation approach is the Bayesian method, which reflects

the uncertainty in the estimates [34]. Recall the random-effects meta-analysis model is given by

$$Y_i|\theta_i = \theta_i + e_i \text{ with } e_i \sim N(0, s_i^2) \text{ and } \theta_i \sim N(\theta, \tau^2)$$

To estimate the hyperparameters of θ and τ^2, prior distributions on the unknown parameters are specified in advance and the estimates are obtained by integrating out the unknown parameters over the joint distribution of all parameters. Let $\theta \sim N(0, a^2)$ and $\tau^{-2} \sim \text{Gamma}(c, d)$, then the joint posterior distribution for $V = (\theta, \theta_1, ..., \theta_n, \tau^2)$ is as following:

$$p(V|\mathbf{Y}, \mathbf{s}^2) \propto \prod_i p(\theta_i|y_i, s_i^2)p(\theta_i|\theta, \tau^2)p(\theta)p(\tau^2)$$

By integrating the above joint posterior distribution for θ and τ^2, then the estimators will be

$$\hat{\theta} = \int_{\theta} \int_{\theta_i, \tau^2} p(V) d\theta_i d\tau^2 d\theta$$

$$\hat{\tau}^2 = \int_{\tau^2} \int_{\theta_i, \theta} p(V) d\theta_i d\theta d\tau^2$$

Generally, the `BUGS` package is utilized to conduct the numerical computation of the integral, and 'vague' non-informative prior distributions are used for the unknown parameters θ and τ^2 [24].

Meta Regression

When one or more characteristics of the studies involved are associated with the effect size, meta-regression is preferable to simple meta-analysis since the inclusion of covariates explain some of the between-study variation in the observed effect size. The random-effects meta regression model is given by

$$Y_i = \boldsymbol{X_i}\boldsymbol{\beta} + \delta_i + e_i \tag{2.12}$$

with $\delta_i \sim N(0, \tau^2)$ and $e_i \sim N(0, s_i^2)$. Here, $\boldsymbol{X_i}$ is a vector of covariates from the ith study and $\boldsymbol{\beta}$ is the vector of regression coefficients. δ_i and e_i are the random effect and the sampling error, which account for the between-study variation and within-study variation respectively. Thus, assuming δ_i and e_i are independent, each study effect size, Y_i follows a normal distribution, i.e. $Y_i|\boldsymbol{\beta}, \tau^2 \sim N(\boldsymbol{X_i}\boldsymbol{\beta}, s_i^2 + \tau^2)$. When the between-study variance, τ^2, is zero, the model reduces to a fixed-effect meta regression model.

Here, $\hat{\boldsymbol{\beta}}$ and τ^2 are the parameters of interest. If τ^2 and s_i^2 were known, the coefficients $\hat{\boldsymbol{\beta}}$ could be estimated by a generalized least square (GLS) approach with $W_i = 1/(\tau^2 + s_i^2)$ as weights. However, τ^2 is an unknown parameter to be estimated. Morris [35] proposed an iterative approach for estimating τ^2 and regression coefficients $\hat{\boldsymbol{\beta}}$. Starting with an initial value for τ^2, say zero, the procedure iterates between estimating $\hat{\boldsymbol{\beta}}$ and $\hat{\tau}^2$ and finally the iteration stops at a predetermined converge criterion. The estimators for $\hat{\boldsymbol{\beta}}$ and $\hat{\tau}^2$ are as follows:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{X^T W X^{-1} X^T W Y} \tag{2.13}$$

with $\boldsymbol{W} = \text{diag}((s_1^2 + \tau^2)^{-1}, ..., (s_k^2 + \tau^2)^{-1})$ and

$$\hat{\tau}^2 = \frac{\sum_{i=1}^k W_i[(k/(k-p)(Y_i - \boldsymbol{X_i}\hat{\boldsymbol{\beta}})^2 - s_i^2]}{\sum_{i=1}^k W_i} \tag{2.14}$$

where p is the number of covariates in the model.

Example of BCG vaccine against Tuberculosis

An example of a set of trials on the efficacy of Bacillus Calmette-Guerin(BCG) vaccine against tuberculosis will be applied for illustrating the models above [28]. This set of trials consist of 13 studies, each having a vaccinated group and a non-vaccinated control group. Two covariates considered are geographic latitude of the study place and year of publication. Table 2.1 shows the data with logOR and var(logOR) for each trial.

The Q statistic for this example is Q=163.2 with df=12 and a p-value is 0.0000, which indicates a significant heterogeneity and a random-effect model is therefore appropriate. `Metafor` package in R is used to carry out meta analysis in this example. Table 2.2 shows the results for both the fixed-effects model and the random-effects model using ML, REML and MM. We can see the estimated overall logOR from the random-effects meta-analysis is lower than that from the fix-effects meta-analysis while the confidence interval is wider, which results from the fact that we consider between-study variation in the random-effects meta analysis. For the random-effects model, the estimated residual between-trial variances $\hat{\tau}^2$s appear a little different using different estimation methods while the estimates of the logOR and their 95% CIs have less difference among three methods. The REML and MM

Table 2.1: Data from the example of efficacy of BCG vaccine in the prevention of tuberculosis

Trial	Vaccinated		Not vaccinated		logOR	var(logOR)	Latitude	Year
	Disease	No disease	Disease	No disease				
1	4	119	11	128	-0.939	0.357	44	48
2	6	300	29	274	-1.666	0.208	55	49
3	3	228	11	209	-1.386	0.433	42	60
4	62	13536	248	12619	-1.456	0.020	52	77
5	33	5036	47	5761	-0.219	0.052	13	73
6	180	1361	372	1079	-0.958	0.010	44	53
7	8	2537	10	619	-1.634	0.227	19	73
8	505	87886	499	87892	0.012	0.004	13	80
9	29	7470	45	7232	-0.472	0.057	27	68
10	17	1699	65	1600	-1.401	0.075	42	61
11	186	50448	141	27197	-0.341	0.013	18	74
12	5	2493	3	2338	0.447	0.534	33	69
13	27	16886	29	17825	-0.017	0.072	33	76

Table 2.2: Results of the Meta Analysis

Model	$\hat{\theta}(95\%CI)$	$\hat{\tau}^2$
Fixed-effects	-0.4361(-0.5190,-0.3533)	-
Random-effects (ML)	-0.7420(-1.0907,-0.3932)	0.3025
Random-effects (REML)	-0.7452(-1.1098,-0.3806)	0.3378
Random-effects (MM)	-0.7474(-1.1242,-0.3706)	0.3663

estimations show closer results compared to ML. Moreover, the negative estimates of the logOR with 95% CI excluding 0 show an indication of the beneficial effect of BCG vaccine.

In meta analysis, a forest plot is a common graphical display to illustrate the relative strength of treatment effects in multiple studies on a same question. Figure 2.1 shows the forest plot of the BCG example. A horizontal line is used to represent the confidence interval of an effect estimate logOR, and the effect estimate is marked with a solid black square, the size of which represents the weight in the meta analysis, i.e. the inverse of variance corresponding to the study. The overall measure of effect logOR is plotted as a diamond, the lateral points of which indicate confidence intervals for this estimate. In addition, the

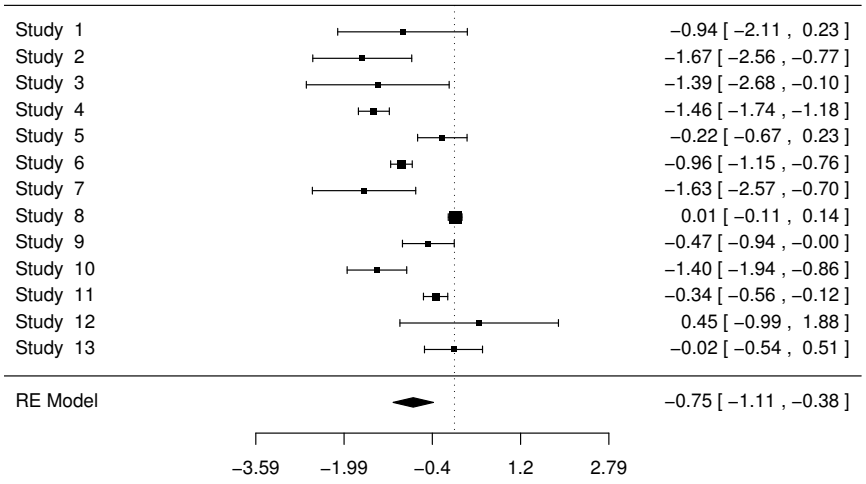

Figure 2.1: Forest Plot for BCG Example

vertical dotted line represents no effect. If the confidence intervals for individual studies overlap with this line, it demonstrates that at the given level of confidence their effect sizes do not differ from no effect for the individual study. From the forest plot, we can see that among 8 out of 13 studies the 95% CIs lies to the left of the line, which demonstrates the beneficial effect of BCG vaccine in these studies. The same conclusion is also drawn from the overall effect size and its confidence interval shown on the right bottom corner.

We will then consider the incorporation of covariates, Latitude and Year into the meta analysis. Table 2.3 shows the estimated regression coefficients ($\hat{\beta}$) and residual between-trial variance ($\hat{\tau}^2$) using ML estimation for three models. In the random-effect meta-regression on Latitude, $\hat{\tau}^2$ is 0.004, which is much smaller than 0.3025 in the random-effects meta-analysis, which indicates that Latitude explains 98.82% of the between-trial variation. In contrast, in the meta-regression on Year, we can see Year only explains 38.04% of the between-trial variance owing to the large $\hat{\tau}^2$ value of 0.2093. Comparing the $\hat{\tau}^2$ of 0.0021 in the model considering both covariates (Latitude and Year) to that of 0.004 in the univariate model, we can see that they are relatively close. Thus, latitude is a better covariate than Year to

Table 2.3: Results of Random-effects Meta-regression

Meta-regression	$\hat{\tau}^2$	$\hat{\beta}_{Latitude}(s.e.)$	$\hat{\beta}_{Year}(s.e.)$
Latitude	0.004	-0.0327(0.0034)	-
Year	0.2093	-	0.0305(0.015)
Latitude + Year	0.0021	-0.0335(0.0043)	-0.0013(0.0062)

explain the differences in BCG vaccination effect between the trials as it has explained most of between-trial variation.

2.3.2 Multivariate Meta Analysis with *Multiple Outcomes*

In clinical trials, when it is sometimes difficult to represent the treatment effect using only a single outcome, we need two or more outcomes. To perform a meta analysis of *multiple outcomes*, the within-study correlation across the different outcomes should be considered. In general, a random-effects *multiple-outcomes* model, called a multivariate random-effects model, is given by

$$\boldsymbol{Y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{\delta_i} + \boldsymbol{e_i} \tag{2.15}$$

where each $\boldsymbol{Y_i}$ is a $p \times 1$ vector with p outcomes for the i^{th} study $(i = 1, ..., k)$, $\boldsymbol{X_i}$ is the regression matrix containing the observed covariates for the i^{th} study; $\boldsymbol{\beta}$ is the vector of regression coefficients to be estimated; $\boldsymbol{\delta_i}$ is a vector of p random effects for the i^{th} study, which represents the heterogeneity unexplained by covariates and is assumed to follow a multivariate Normal distribution MVN(0, $\boldsymbol{D}$); and $\boldsymbol{e_i}$ is a vector of the sampling errors, which accounts for the within-study variation and is approximately distributed as MVN(0,$\boldsymbol{S_i}$). Assuming $\boldsymbol{\delta_i}$ and $\boldsymbol{e_i}$ are independent, $\boldsymbol{Y_i}$ follows a multivariate normal distribution:

$$\boldsymbol{Y_i} \sim N(\boldsymbol{X_i}\boldsymbol{\beta}, \boldsymbol{D} + \boldsymbol{S_i}) \tag{2.16}$$

In the model (2.16), all the within-study covariance matrices $\boldsymbol{S_i}$'s are assumed known and not necessarily diagonal while $\boldsymbol{D}$ and $\boldsymbol{\beta}$ are the two parameters of interest. Our goal is to estimate $\boldsymbol{\beta}$ as well as $\boldsymbol{D}$. It is easy to find the log-likelihood function given by the following

$$L = \text{constant} - \frac{1}{2}\sum_{i=1}^{k} \log|\boldsymbol{D} + \boldsymbol{S_i}| - \frac{1}{2}\sum_{i=1}^{k} \boldsymbol{e}_i^T(\boldsymbol{D} + \boldsymbol{S_i})^{-1}\boldsymbol{e}_i \tag{2.17}$$

where $\boldsymbol{e_i} = \boldsymbol{Y_i} - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$. Given $\boldsymbol{D}$, $\boldsymbol{\beta}$ is estimated by solving min $\sum_{i=1}^{k} \boldsymbol{e}_i^T(\boldsymbol{D}+\boldsymbol{S_i})^{-1}\boldsymbol{e}_i$, which is a weighted least square problem. However, $\boldsymbol{D}$ is always unknown and needs to be estimated. Currently, four approaches, Generalized Least Square (GLS), Maximum Likelihood (ML), Restricted Maximum Likelihood (REML) and Method of Moments (MM) are available to estimate $\boldsymbol{\beta}$ as well as $\boldsymbol{D}$.

Multivariate Fixed-effects Model

When all the $\boldsymbol{\delta_i}$ and the elements of the matrix $\boldsymbol{D}$ are zero, the above model reduces to a multivariate fixed-effects model. To clarify the multivariate random-effects model, first consider the fixed-effects model given by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

where $\boldsymbol{Y}$ is the $kp \times 1$ vector containing k vectors of $\boldsymbol{Y_i}$; $\boldsymbol{X}$ is the $kp \times p$ matrix containing the observed covariates for the k studies; $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients to be estimated; and $\boldsymbol{e}$ is the $kp \times 1$ vector of the sampling errors for the k studies. By fitting the model by GLS regression with $\boldsymbol{S_i}^{-1}$ as weights, the fixed-effects estimate for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T S^{-1} X})^{-1}\boldsymbol{X^T S^{-1} Y} \tag{2.18}$$

and

$$\text{cov}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X^T S^{-1} X})^{-1} \tag{2.19}$$

where $\boldsymbol{S}$ is a block diagonal matrix with k $p \times p$ $\mathbf{S}_i$ matrices as diagonal entries.

For the multivariate meta analysis, we also use a Q statistic to assess the homogeneity of the vector of the parameter estimates across studies. The Q statistic is shown below:

$$Q = \sum_{i=1}^{k} (\boldsymbol{Y_i} - \boldsymbol{X_i}\hat{\boldsymbol{\beta}})^T \boldsymbol{S_i}^{-1} (\boldsymbol{Y_i} - \boldsymbol{X_i}\hat{\boldsymbol{\beta}})$$

Under the null hypothesis, Q has an asymptotic χ^2 distribution with $p(k-1)$ degrees of freedom. If we reject the null hypothesis, the between-study variation should be considered and a multivariate random-effects model is therefore needed.

GLS Multivariate Random-effects Approach

Berkey *et al.* [23] proposed an iterative procedure to simultaneously estimate $\boldsymbol{\beta}$ and $\boldsymbol{D}$ with known $\boldsymbol{Y_i}$, $\boldsymbol{X_i}$ and $\boldsymbol{S_i}$. Firstly fit a fixed-effects model by GLS regression, and the estimates of $\boldsymbol{\beta}$ and its covariance matrix cov($\boldsymbol{\beta}$) can be accordingly obtained (see (2.18) and (2.19)). Then compute $kp \times 1$ vector of residuals ($\boldsymbol{Y}$-$\boldsymbol{X}\hat{\boldsymbol{\beta}}$) and rearrange them into a $k \times p$ matrix ($\boldsymbol{Y}$-$\boldsymbol{X}\hat{\boldsymbol{\beta}}$). The initial estimate of the covariance matrix D is

$$\hat{D} = (k-p)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - k^{-1}\sum_{i=1}^{k} S_i \tag{2.20}$$

Further, fit a new GLS model with $\hat{\boldsymbol{D}}+\boldsymbol{S_i}$ as weights and the estimate of $\boldsymbol{\beta}$ and its covariance matrix cov($\boldsymbol{\beta}$) can be obtained:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T}(\boldsymbol{D}+\boldsymbol{S})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X^T}(\boldsymbol{D}+\boldsymbol{S})^{-1}\boldsymbol{Y} \tag{2.21}$$

and

$$\text{cov}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X^T}(\boldsymbol{D}+\boldsymbol{S})^{-1}\boldsymbol{X})^{-1} \tag{2.22}$$

Then calculate new residuals ($\boldsymbol{Y}$-$\boldsymbol{X}\hat{\boldsymbol{\beta}}$) and a new $\hat{\boldsymbol{D}}$. In each iteration, update $\hat{\beta}$ from GLS models and $\hat{\boldsymbol{D}}$ from residuals until convergence for a specified stopping criterion. In the end, the estimates of unknown parameters, $\hat{\boldsymbol{D}}$, $\hat{\boldsymbol{\beta}}$ and cov($\hat{\boldsymbol{\beta}}$) are obtained. It is worth mentioning that in Equation (2.20), Berkey *et al.* did not force the $\hat{D}$ in the iteration to be positive semi-definite. However, in our analyses that will be presented in Chapter 3, this sometimes leads a negative definite $\hat{D}$.

ML and REML Multivariate Random-effects Approaches

The maximum likelihood and restricted Maximum Likelihood approaches are typically alternatives to multivariate meta-analysis. Since the k studies are assumed independent, the likelihood can be obtained and maximized to provide maximum likelihood estimates of all parameters. The log-likelihood function is given by Equation (2.17). By maximizing L_{ML}, the between-study covariance matrix $\hat{D}_{ML}$ and the p overall treatment effects, $\hat{\boldsymbol{\beta}}$=($\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$), are obtained (see (2.21) and (2.22)).

Restricted maximum likelihood (REML) provides a less biased estimator for the between-study covariance matrix and thus it is more popular than maximum likelihood. The log-

likelihood of the residuals, L_{REML}, is written as

$$L_{REML} = Const. - \frac{1}{2}\sum_{i=1}^{k} \log|\boldsymbol{D}+\boldsymbol{S_i}| - \frac{1}{2}\log|\sum_{i=1}^{k}(\boldsymbol{D}+\boldsymbol{S_i})^{-1}| - \frac{1}{2}\sum_{i=1}^{k} \boldsymbol{e_i^T}(\boldsymbol{D}+\boldsymbol{S_i})^{-1}\boldsymbol{e_i} \quad (2.23)$$

Compared to Equation (2.17), L_{REML} adds a correction term, $\frac{1}{2}\log|\sum_{i=1}^{k}(\boldsymbol{D}+\boldsymbol{S_i})^{-1}|$, so as to obtain a less biased between-study covariance matrix. By maximizing L_{REML}, the between-study covariance matrix $\hat{D}_{REML}$ and the p overall treatment effects $\hat{\boldsymbol{\beta}}$ as well as the covariance matrix are obtained. For both ML and REML estimations, the $(p \times 1)$ vector of overall treatment effects have the same form as GLS method (see (2.21) and (2.22)), but the between-study covariance matrices, $\hat{D}$, $\hat{D}_{ML}$ and $\hat{D}_{REML}$ have their own forms for the three estimation methods. For all these methods, forcing the negative eigenvalues at zero is needed in each iteration to ensure positive semi-definite between-study covariance matrices.

MM Random-effects Multivariate Approach

More recently, D. Jackson *et al.* [27] proposed a new approach by applying the method of moments in performing random-effects multivariate meta analysis. Compared to the three methods above, one of its appealing properties is that it avoids numerical maximization or iteration, thereby yielding the estimates for unknown parameters just by solving linear equations and matrix operations, which greatly improves the computational efficiency. The other advantage is that it does not require the assumption of normality to estimate the between-study covariance matrix. For simplicity, firstly consider a bivariate random effects meta-analysis model without covariates

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{Y_1} \\ \mu_{Y_2} \end{pmatrix}, \begin{pmatrix} s^2_{Y_{1i}} + \tau^2_{Y_1} & \rho_i s_{Y_{1i}} s_{Y_{2i}} + \kappa\tau_{Y_1}\tau_{Y_2} \\ \rho_i s_{Y_{1i}} s_{Y_{2i}} + \kappa\tau_{Y_1}\tau_{Y_2} & s^2_{Y_{2i}} + \tau^2_{Y_2}. \end{pmatrix}\right) \quad (2.24)$$

where μ_{Y_1} and μ_{Y_2} are the two overall effects; $s_{Y_{1i}}$, $s_{Y_{2i}}$ and ρ_i describe the within-study variation $\boldsymbol{S_i}$ and τ_{Y_1}, τ_{Y_2} and κ describes the between-study variation $\boldsymbol{D}$. Thus, $\boldsymbol{S_i}$ and $\boldsymbol{D}$ can be written as

$$\boldsymbol{S_i} = \begin{pmatrix} s^2_{Y_{1i}} & \rho_i s_{Y_{1i}} s_{Y_{2i}} \\ \rho_i s_{Y_{1i}} s_{Y_{2i}} & s^2_{Y_{2i}} \end{pmatrix}$$

and

$$\boldsymbol{D} = \begin{pmatrix} \tau^2_{Y_1} & \kappa\tau_{Y_1}\tau_{Y_2} \\ \kappa\tau_{Y_1}\tau_{Y_2} & \tau^2_{Y_2} \end{pmatrix}$$

Assume that the within-study covariance matrix for the i^{th} study $\boldsymbol{S_i}$ is known. The goal is to estimate the unknown parameters, μ_{Y_1}, μ_{Y_2}, τ_{Y_1}, τ_{Y_2} and κ. In univariate meta analysis, the method of moments based on the Q statistic was used to estimate the between-study variance (see (2.9)). Similarly, for a multivariate setting, the Q statistic can be written as a Q matrix:

$$Q = \begin{pmatrix} \sum W_{Y_{1i}}(Y_{1i} - \bar{Y}_{11})^2 & \sum (W_{Y_{1i}} W_{Y_{2i}})^{1/2}(Y_{1i} - \bar{Y}_{12})(Y_{2i} - \bar{Y}_{22}) \\ \sum (W_{Y_{1i}} W_{Y_{2i}})^{1/2}(Y_{1i} - \bar{Y}_{12})(Y_{2i} - \bar{Y}_{22}) & \sum W_{Y_{2i}}(Y_{2i} - \bar{Y}_{21})^2 \end{pmatrix}$$

where $W_{Y_{1i}}$=$1/s^2_{Y_{1i}}$, $W_{Y_{2i}}$=$1/s^2_{Y_{2i}}$; $\bar{Y}_{11}$ and $\bar{Y}_{12}$ denote the weighted averages of the Y_{1i} with $W_{Y_{1i}}$ and $(W_{Y_{1i}} W_{Y_{2i}})^{1/2}$ are weights. $\bar{Y}_{21}$ and $\bar{Y}_{22}$ denote weighted averages of the Y_{2i} with $W_{Y_{2i}}$ and $(W_{Y_{1i}} W_{Y_{2i}})^{1/2}$ as weights, respectively.

To estimate the between-study covariance matrix $\boldsymbol{D}$, we need to find the expectation of the Q matrix, $E(Q)$, and then solve the linear equations by equating $E(Q)$ to Q. The $E(Q)$ is given by

$$E(Q) = \begin{pmatrix} e_{11} & a + b\kappa\tau_{Y_1}\tau_{Y_2} \\ a + b\kappa\tau_{Y_1}\tau_{Y_2} & e_{22} \end{pmatrix}$$

where

$$e_{11} = (n_{Y_1} - 1) + \left(\sum W_{Y_{1i}} - \frac{\sum W^2_{Y_{1i}}}{\sum W_{Y_{1i}}} \right) \tau^2_{Y_1}$$

$$e_{22} = (n_{Y_2} - 1) + \left(\sum W_{Y_{2i}} - \frac{\sum W^2_{Y_{2i}}}{\sum W_{Y_{2i}}} \right) \tau^2_{Y_2}$$

$$a = \sum \rho_i - \frac{\sum \rho_i (W_{Y_{1i}} W_{Y_{2i}})^{1/2}}{\sum (W_{Y_{1i}} W_{Y_{2i}})^{1/2}}$$

and

$$b = \sum (W_{Y_{1i}} W_{Y_{2i}})^{1/2} - \frac{\sum (W_{Y_{1i}} W_{Y_{2i}})}{\sum (W_{Y_{1i}} W_{Y_{2i}})^{1/2}}$$

where the expectations of the diagonal entries of the Q matrix (e_{11} and e_{22}) are both linear functions of just one of the between-study variances, $\tau^2_{Y_1}$ and $\tau^2_{Y_2}$, which are the same as those in Equation (2.8) in the univariate case. The expectations of the off-diagonal entries of the Q matrix (e_{12} and e_{21}) are both linear functions of the between-study covariance $\kappa\tau_{Y_1}\tau_{Y_2}$. Hence the between-study covariance matrix $\hat{\boldsymbol{D}}_{DL}$ can be estimated by solving the linear equations $E(Q)[i,j] = Q[i,j]$, $\forall i,j$. For example, by solving the linear equations $E(Q)[1,1] = Q[1,1]$ and $E(Q)[2,2] = Q[2,2]$, $\tau^2_{Y_1}$ and $\tau^2_{Y_2}$ can be estimated; and similarly, by solving the linear

equation $E(Q)[2,1] = Q[2,1]$, $\kappa\tau_{Y_1}\tau_{Y_2}$ be estimated. To ensure $\hat{\boldsymbol{D}}_{DL}$ is estimated as positive semi-definite, the negative eigenvalues of $\hat{\boldsymbol{D}}_{DL}$ are forced to zero. After the between-study covariance matrix $\hat{\boldsymbol{D}}_{DL}$ is obtained, it is straightforward to estimate the two overall effects, μ_{Y_1} and μ_{Y_2}, by

$$\hat{\mu}_Y = \begin{pmatrix} \hat{\mu}_{Y_1} \\ \hat{\mu}_{Y_2} \end{pmatrix} = \left(\sum_{i=1}^{k} (\hat{\boldsymbol{D}}_{DL} + \boldsymbol{S_i})^{-1} \right)^{-1} \left(\sum_{i=1}^{k} (\hat{\boldsymbol{D}}_{DL} + \boldsymbol{S_i})^{-1} \boldsymbol{Y_i} \right)$$

and

$$\text{cov}(\hat{\mu}_Y) = \left(\sum_{i=1}^{k} (\hat{\boldsymbol{D}}_{DL} + \boldsymbol{S_i})^{-1} \right)^{-1}$$

where $\boldsymbol{Y_i}$ denotes $(Y_{1i}, Y_{2i})^T$. The equations for $\hat{\mu}_Y$ and $\text{cov}(\hat{\mu}_Y)$ are similar to (2.21) and (2.22) in GLS, ML and REML approaches. A simulation study suggested that this procedure not only gives similar results to iterative maximum likelihood approach but the computation speed is much faster than existing methods.

All the methods mentioned above need to employ spectral decomposition and truncation to ensure a positive semi-definite ($p.s.d$) between-study covariance matrix. The disadvantages may come from the concerns of optimality, convergence and $p.s.d$ constraint. For example, truncating the between-study covariance matrix on the boundary of its parameter space in the REML estimation results in upward bias in estimated between-study covariance matrix and -1 or 1 estimated between-study correlation [25]. They can be observed especially when the number of studies is small and the within-study variation is relatively large.

Example of Treatment of Periodontal Disease

An example from Berkey's paper [23] will be used to illustrate the methods in the multivariate meta analysis. It consists of five published trials comparing a surgical(S) procedure with a non-surgical(NS) procedure for moderate periodontal disease. The segments of each patient's mouth were divided into several sections that were randomly allocated to different treatment procedures. The two outcomes of central interest were the mm changes in probing depth and attachment level before and after the treatment, denoted by PD and AL. The goal of the treatment is to decrease PD and increase AL around the teeth. Table 2.4 presents the results from the five trials, including publication year, sample size, improvement in PD and

Table 2.4: Data from the example of the treatment of periodontic disease

Trial	Publication year	Number of patients	Improvement in Probing depth S - NS	Improvement in Attachment level S - NS	S_i PD AL
1	1983	14	+0.47	−0.32	$\begin{pmatrix} 0.0075 & 0.0030 \\ 0.0030 & 0.0077 \end{pmatrix}$
2	1982	15	+0.20	−0.60	$\begin{pmatrix} 0.0057 & 0.0009 \\ 0.0009 & 0.0008 \end{pmatrix}$
3	1979	78	+0.40	−0.12	$\begin{pmatrix} 0.0021 & 0.0007 \\ 0.0007 & 0.0014 \end{pmatrix}$
4	1987	89	+0.26	−0.31	$\begin{pmatrix} 0.0029 & 0.0009 \\ 0.0009 & 0.0015 \end{pmatrix}$
5	1988	16	+0.56	−0.39	$\begin{pmatrix} 0.0148 & 0.0072 \\ 0.0072 & 0.0304 \end{pmatrix}$

AL (positive values indicate that surgery provides the better patient outcome) and their covariance matrix.

The Q statistic of 128.2 indicates the need for a multivariate random-effects meta analysis. Since the covariance matrix of the two outcomes, PD and AL, is known, a bivariate meta analysis can be performed for the two outcomes. Table 2.5 shows the results for both univariate and multivariate models. We can see that the overall estimates and the 95%CIs in the fixed-effects model are smaller than the corresponding random-effects models. It seems that when we consider the between-study variation in the random-effects model, the difference between fixed-effects and random-effects models have a greater impact than the difference between univariate and multivariate meta analysis. There are three approaches available to perform multivariate meta analysis, of which GLS multivariate meta analysis is carried out by self-written R script, and REML and MM by the mvmeta function in STATA. Comparing the three approaches in terms of the overall estimates of PD and AL ($\hat{\beta}_{PD}$ and $\hat{\beta}_{AL}$) and their 95% CIs and the between-study covariance matrix $\hat{D}$, we can see that the overall estimates of $\hat{\beta}_{PD}$ and $\hat{\beta}_{AL}$ and 95% CI from GLS multivariate meta analysis are very close to the results from REML and MM in STATA. The between-study covariance matrix from GLS, however, is a little different from REML and MM, especially the off-diagonal entries, which comes from the difference among the estimates of the between-study

Table 2.5: Results from univariate and multivariate meta analysis

Methods	$\hat{\beta}_{PD}(95\%CI)$	$\hat{\beta}_{AL}(95\%CI)$	$\hat{D}$
Univariate			
GLS (Fixed)	0.3472 (0.2906, 0.4038)	-0.3926 (-0.4297, -0.3555)	-
GLS (Random)	0.3629 (0.2347, 0.4911)	-0.3455 (-0.4895, -0.2016)	0.0156 0.0215
Multivariate			
GLS (Fixed)	0.3072 (0.2512, 0.3632)	-0.3944 (-0.4309, -0.3578)	-
GLS (Random)	0.3605 (0.2320, 0.4889)	-0.3418 (-0.4858, -0.1979)	$\begin{pmatrix} 0.0157 & 0.0074 \\ 0.0074 & 0.0216 \end{pmatrix}$
REML (Random)	$0.3534(0.2333, 0.4735)$	$-0.3392(-0.5142, -0.1642)$	$\begin{pmatrix} 0.0117 & 0.0119 \\ 0.0119 & 0.0327 \end{pmatrix}$
MM (Random)	$0.3478(0.2385, 0.4572)$	$-0.3405(-0.5623, -0.1187)$	$\begin{pmatrix} 0.0121 & 0.0181 \\ 0.0181 & 0.0573 \end{pmatrix}$

covariance matrices in these methods.

CHAPTER 3

APPLICATION

3.1 Introduction

The importance of major risk factors, such as hypertension, total cholesterol, body mass index, diabetes, smoking, for predicting incidence and mortality of Coronary Heart Disease (CHD) is well known. In light of the fact that age is also a major risk factor for CHD and the population is aging, a natural question is whether the effects of risk factors on CHD changes with age, and in particular, whether the risk effects on CHD have a different impact in the elderly group as compared to the younger age group. Our aim is to examine the interaction between age and risk factors by examining the above question using data from multiple studies containing differing age ranges.

In past studies, the most common way is to estimate the CHD odds ratios (ORs) or hazard ratios (HRs) associated with the risk factors by means of a logistic regression or proportional hazard model in a single study. The main disadvantage is that the results are limited to a specified study and its specific age range. Thus, separate analyses on distinct studies may give rise to different conclusions to the same question. Additionally most studies have insufficient power to discern interaction. Meta-analysis has recently gained increasing popularity in a wide range of applications in medical research. Many studies on meta-analysis synthesize the results of related but independent studies originating from published literatures so as to obtain a combined estimate and a confidence interval for the effect. This method works well for a single outcome measure (e.g. an odds ratio or hazard ratio in clinical trials). However, the published literature seldom contains sufficient information to conduct more complex meta-analysis such as the interaction between age and risk-factors.

3.2 Data Description

The data in my thesis come from the Diverse Populations Collaboration (DPC) that examines variation in the results of epidemiological investigations in population samples, cohorts studies and clinical trials from many countries and cultures [36]. Since 1996, DPC has grown to consist of 27 studies from US, Europe and Asia, which include ARIC, BIP, Charleston Heart, CHS, Cordis, Evans County, Framingham Cohort, Framingham Offspring, Glostrup, Guangzhou, GOH, HDFP, Honolulu, Iceland, Israel, LRC, LRC-CPPT, MRFIT, NHANES I, NHANES II, NHIS, Norway, Puerto Rico, Renfrew-Paisley, Scottish Collaborative, Tecumseh, and Youguslavia. The full names of the 27 studies are listed in Appendix A. These studies contain person-level data that have natural subgroups stratified by sex (male and female), race (whites, blacks, and Hispanics), area of residence (urban and rural), treatment group (placebo and treatment) and other characteristics of the study samples (for example, random sample or hyperlipidaemic patients). According to these strata, we have 78 cohorts available for use in our analysis.

Table 3.1 presents the descriptive information for the 27 studies. The studies include 395,682 individuals with 60,374 deaths. The underlying cause of death was determined according to the International Classification of Diseases ICD-9 in most of the studies (codes 410-414 and 429.2 as CHD), in a few studies ICD-8 was used (codes 410-414 as CHD). Additionally, cause of death was assigned by a panel of physicians in a small number of studies. Table 3.2 and 3.3 show the number of participants and baseline characteristics of covariates (systolic blood pressure (SBP), diastolic blood pressure (DBP), serum total cholesterol (CHOL), body mass index (BMI), diabetes status (DIAB) and cigarette smoking (Smoking)) in women and men. The studies were included in the analysis if they have complete data on these covariates. Data for men and women were analyzed separately. Figure 3.1 presents the boxplot for age among 27 studies. From the range of age, we can see that most of studies contain patients from middle-adulthood to 65 years old and some studies have older patients aged over 75 years. Due to the broad age ranges of the studies, they can also be stratified by age groups ([25,40), [40,50), [50,60), [60,70), [70,80) and 80+) besides sex, race, site, treatment and other characteristic of the study mentioned above. Therefore, we have as many as 352 cohorts available for use of our analysis.

Table 3.1: Studies included in the analysis: Diverse Populations Collaboration

Study	Number of Patients	Total Deaths	CHD Deaths	Years Followed	Age	Strata
1.ARIC	15732	648	145	7	45-64	Race,Sex
2.BIP	14651	2715	981	9	36-74	Sex
3.Charleston Heart	2179	1245	391	31	35-97	Race,Sex
4.CHS	5795	821	250	6	65-90	Race,Sex
5.Cordis	5082	277	96	12	25-80	Sex
6.Evans County	2704	1657	524	32	25-79	Race,Sex
7.Framingham Cohort	4541	2775	716	38	34-69	Sex
8.Framingham Offspring	4461	446	95	23	25-62	Sex
9.Glostrup	10198	1101	275	17	29-80	Sex
10.GOH	5637	1606	281	28	29-59	Sex
11.Guangzhou	7060	-	-	-	25-71	Sex,Site
12.HDFP	10940	1424	494	8	30-69	Treatment, Race,Sex
13.Honolulu	8006	2466	419	23	45-68	
14.Iceland	18911	4753	1509	28	33-81	Sex
15.Israel	10059	3473	1097	24	40-75	
16.LRC	8623	996	345	15	30-97	Sex,Other
17.LRC-CPPT	3806	301	144	14	34-60	Treatment
18.MRFIT	12866	1033	428	12	35-58	Treatment
19.NHANES I	13721	4573	1419	22	25-75	Race,Sex
20.NHANES II	9087	2119	623	17	30-75	Race,Sex
21.NHIS	130063	11774	3175	9	25-90	Race,Sex
22.Norway	48646	3576	1009	19	35-49	Sex
23.Puerto Rico	9815	1737	376	16	35-79	Site
24.Renfrew-Paisley	15411	4447	1571	20	45-64	Sex
25.Scottish Collaborative	7003	2080	758	24	25-75	Sex
26.Tecumseh	4226	991	361	20	25-92	Sex
27.Yugoslavia	6459	1340	226	18	34-83	Site
Total	395682	60374	17708			

Table 3.2: Number of Participants and Baseline Characteristics in Women

Study	Number of Patients	Age	SBP, Mean $(SD)(mmHg)$	DBP, Mean $(SD)(mmHg)$	CHOL, Mean $(SD)(mg/dL)$	BMI, Mean $(SD)(kg/m^2)$	DIAB (%)	Smoking (%)
1.ARIC	8677	45-64	120(20)	72(11)	218.2(43.5)	27.9(6.1)	10.2	24.9
2.BIP	2772	43-74	140(21)	83(10)	239.5(42.6)	27.2(4.3)	24.8	7.3
3.Charleston Heart	1194	35-97	146(31)	86(13)	238.9(50.7)	25.5(5.5)	3.7	37.6
4.CHS	3329	65-90	137(22)	70(11)	221.0(38.3)	26.8(4.7)	14.6	12.7
5.Cordis	1422	25-71	121(19)	76(10)	198.3(42.8)	26.0(5.1)	2.2	24.1
6.Evans County	1419	25-79	155(35)	94(17)	222.8(47.7)	27.0(6.2)	4.7	15.2
7.Framingham Cohort	2532	34-68	134(25)	83(12)	241.2(46.9)	25.6(4.5)	1.9	38.2
8.Framingham Offspring	2276	25-62	119(16)	76(10)	195.1(38.7)	24.4(4.6)	1.1	44.2
9.Glostrup	5100	29-80	125(21)	77(11)	236.7(48.9)	24.3(4.3)	2.1	46.8
10.GOH	2893	29-59	124(18)	80(10)	-	25.4(4.7)	-	20.8
11.Guangzhou	3720	25-71	112(16)	71(10)	170.4(34.8)	20.4(2.7)	-	4.3
12.HDFP	5030	30-69	162(24)	102(9)	238.8(48.9)	29.1(6.8)	8.2	34.0
14.Iceland	9773	33-81	138(21)	84(10)	255.2(47.5)	25.2(4.3)	1.6	37.2
16.LRC	3986	30-97	124(21)	78(11)	224.6(49.7)	25.1(4.8)	1.9	33.0
19.NHANES I	8146	25-75	133(25)	82(13)	222.0(49.8)	25.7(5.7)	4.1	28.8
20.NHANES II	4823	30-75	132(24)	81(12)	231.2(51.3)	26.3(5.8)	6.3	28.3
21.NHIS	75791	25-90	-	-	-	24.9(5.3)	-	25.4
22.Norway	24015	35-49	131(18)	81(11)	242.4(48.0)	24.7(4.1)	0.5	36.9
24.Renfrew-Paisley	8353	45-64	150(25)	85(14)	248.1(42.4)	25.8(4.5)	1.2	46.7
25.Scottish Collaborative	1002	27-61	137(20)	83(11)	224.6(40.3)	24.7(3.8)	0.6	58.8
26.Tecumseh	2179	25-92	137(25)	82(14)	218.9(47.4)	25.5(5.3)	3.7	35.2
Total	217250							

Table 3.3: Number of Participants and Baseline Characteristics in Men

Study	Number of Patients	Age	SBP, Mean $(SD)(mmHg)$	DBP, Mean $(SD)(mmHg)$	CHOL, Mean $(SD)(mg/dL)$	BMI, Mean $(SD)(kg/m^2)$	DIAB (%)	Smoking (%)
1.ARIC	7055	45-64	123(18)	76(11)	211.0(39.9)	27.5(4.2)	9.7	27.8
2.BIP	11879	36-74	133(19)	81(10)	220.4(37,7)	26.6(3.3)	17.8	11.6
3.Charleston Heart	985	35-90	143(26)	86(12)	231.4(45.3)	25.1(3.9)	3.5	67.7
4.CHS	2466	65-90	136(21)	72(12)	197.9(35.8)	26.3(3.4)	19.1	11.2
5.Cordis	3660	25-80	126(17)	79(10)	208.9(44.1)	25.9(3.7)	3.9	38.0
6.Evans County	1285	25-76	149(30)	93(16)	207.2(40.4)	25.2(4.1)	3.1	56.9
7.Framingham Cohort	2009	34-69	132(20)	84(12)	234.7(41.2)	26.2(3.4)	2.2	69.2
8.Framingham Offspring	2185	25-62	127(16)	83(10)	205.2(39.1)	27.0(3.6)	2.7	45.3
9.Glostrup	5098	29-80	129(19)	81(11)	236.9(46.4)	25.4(3.5)	2.5	56.8
10.GOH	2744	29-59	128(16)	84(9)	-	24.9(3.7)	-	51.0
11.Guangzhou	3340	25-66	115(15)	74(10)	170.7(33.5)	20.2(2.2)	-	75.4
12.HDFP	5910	30-69	156(21)	101(9)	232.2(44.7)	27.8(4.6)	6.1	42.8
13.Honolulu	8006	45-68	134(21)	82(12)	218.3(38.3)	23.8(3.2)	9.5	43.8
14.Iceland	9138	33-79	140(19)	88(10)	245.9(41.1)	25.8(3.4)	2.1	28.6
15.Israel	10059	40-75	135(21)	84(11)	209.4(40.1)	25.7(3.3)	4.8	50.5
16.LRC	4637	30-91	127(17)	82(11)	222.8(45.3)	26.8(3.5)	1.6	37.1
17.LRC-CPPT	3806	34-60	124(14)	82(9)	295.8(31.4)	26.5(2.6)	-	37.5
18.MRFIT	12866	35-58	138(15)	93(10)	257.3(36.8)	27.7(3.5)	2.7	59.1
19.NHANES I	5575	25-75	137(22)	86(12)	220.8(46.3)	25.7(4.2)	3.9	38.1
20.NHANES II	4264	30-75	134(20)	84(12)	220.7(44.9)	25.9(4.0)	5.4	37.0
21.NHIS	54272	25-90	-	-	-	25.8(4.0)	-	30.8
22.Norway	24631	35-49	136(16)	85(11)	248.4(50.0)	25.2(3.0)	0.8	47.8
23.Puerto Rico	9815	35-79	132(23)	82(12)	202.0(41.3)	25.1(4.1)	5.4	43.7
24.Renfrew-Paisley	7058	45-64	149(23)	86(13)	226.4(37.2)	25.9(3.4)	1.3	56.6
25.Scottish Collaborative	6001	25-75	134(18)	84(10)	226.9(40.1)	25.1(3.1)	0.5	55.1
26.Tecumseh	2047	25-90	138(19)	84(13)	219.9(42.6)	25.7(3.8)	2.8	61.2
27.Yugoslavia	6459	34-83	133(19)	81(10)	186.3(45.1)	23.3(3.3)	0.7	69.9
Total	178432							

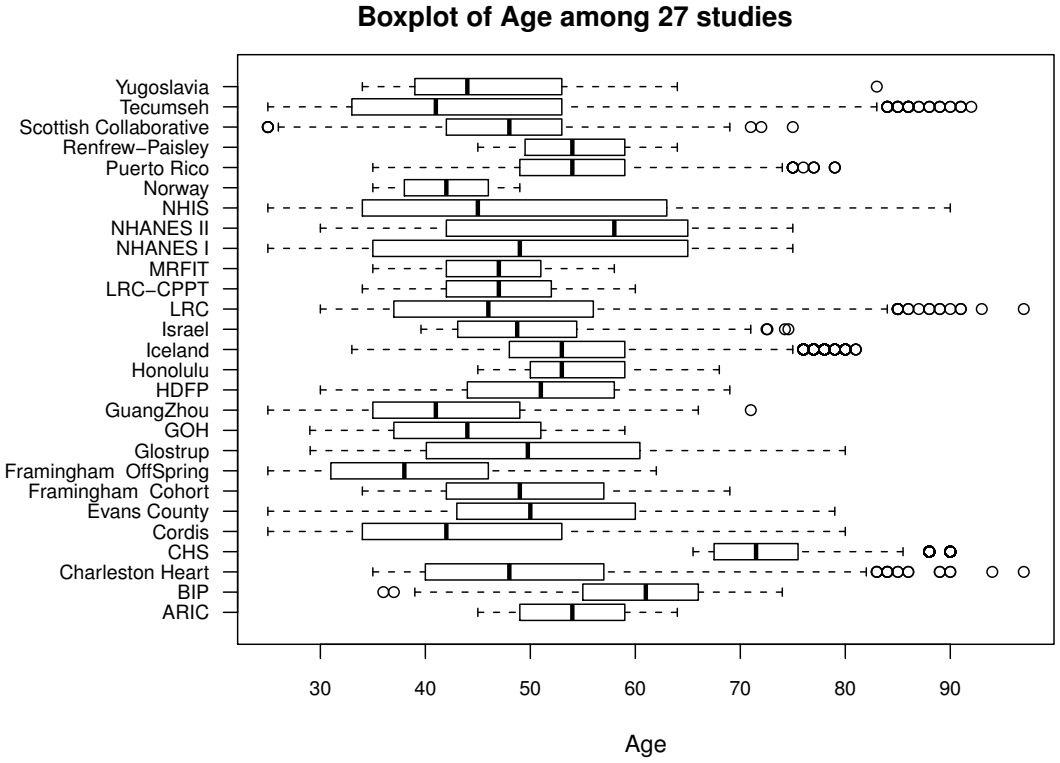

Figure 3.1: Boxplot of age among 27 studies

3.3 Statistical Methods

We first consider the case of univariate meta analysis based on a single covariate. In order to examine the change of the association of CHD death with one covariate as age increases among the studies, the patients were divided into several groups according to studies (1-27), sex (male and female) and age groups ([25,40), [40,50), [50,60), [60,70), [70,80) and 80+). Thus, we have 213 cohorts available for use of our analysis, 119 for male and 94 for female. Considering that there are so few observations for CHD death, those cohorts with less than 10 of CHD death are excluded and the remaining 148 cohorts (92 for male and 56 for female) will be analyzed. A series of simple logistic models with CHD death as a response on a covariate, for example, SBP, were carried out for the 148 cohorts. An estimated coefficient can be interpreted as logOR of CHD death in terms of one-unit change in the covariate, and

therefore we perform a series of univariate meta analyses on the coefficients to obtain the overall logORs for 12 subgroups determined by sex and age groups, 6 for male and 6 for female. By fitting weighted regression on overall logORs vs. age groups, the change of the overall logORs with age can be examined. At the same time, we can compare the difference between female group and male group.

In univariate meta analysis, we examine the impact of one covariate on CHD death in different age groups and two sex groups among studies. The other approach is to consider age as a continuous covariate and add the main effect of age and the interaction between age and other covariates simultaneously into a logistic model. In this way, we can both regard the coefficients as multiple outcomes to perform multivariate meta analysis and incorporate the correlation among the coefficients into consideration.

Recall the multiple logistic model introduced in Section 2.2. In general, the model is as follows:

$$Pr(Y = 1|\boldsymbol{X}) = \frac{e^{\boldsymbol{X'}\hat{\boldsymbol{\beta}}}}{1 + e^{\boldsymbol{X'}\hat{\boldsymbol{\beta}}}}$$

where $\hat{\boldsymbol{\beta}}$'s are estimates of the true vector of parameters, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)'$. Note that the baseline $\hat{\beta}_0$ is not included to perform meta analysis. Since each estimated coefficient can be interpreted as logOR of CHD death in terms of one-unit change in the corresponding covariate, the coefficients $\hat{\boldsymbol{\beta}}$ can be used as *multiple outcomes* to perform the multivariate meta analysis. Assuming that the estimated coefficients and their covariance matrices for all the k studies are known, the p estimates $\hat{\boldsymbol{\beta}}_i$ for the i^{th} study are such that

$$\hat{\boldsymbol{\beta}}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{S}_i)$$

In our investigation, we do not incorporate covariates into meta analysis. When no covariates are considered, each of the $\boldsymbol{X_i}$'s is a $p \times p$ identity matrix. Therefore, the weighted least-square estimate for the average vector of p parameters is given by:

$$\bar{\boldsymbol{\beta}} = \left(\sum_{i=1}^{k} \boldsymbol{S}_i^{-1}\right)^{-1} \left(\sum_{i=1}^{k} \boldsymbol{S}_i^{-1}\hat{\boldsymbol{\beta}}_i\right) \tag{3.1}$$

The covariance matrix for this average vector is

$$\text{cov}(\bar{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{k} \boldsymbol{S}_i^{-1}\right)^{-1} \tag{3.2}$$

Thus, Equations (3.1) and (3.2) are parameters estimates for the fixed-effects multivariate model introduced in Chapter 2. As univariate meta analysis, we employ a large sample Q test to assess the homogeneity of the vectors of parameter estimates across studies. The Q statistic is shown below:

$$Q = \sum_{i=1}^{k} (\hat{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}})' \mathbf{S}_i^{-1} (\hat{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}})$$

Under the null hypothesis, Q has an asymptotic χ^2 distribution with $p(k-1)$ degrees of freedom. If we reject the null hypothesis, the between-study variation should be considered and a multivariate random-effects model is therefore needed.

To derive the random-effects model, we assume

$$E(\hat{\boldsymbol{\beta}}_i | \boldsymbol{\theta}_i) = \boldsymbol{\theta}_i, \mathrm{cov}(\hat{\boldsymbol{\beta}}_i | \boldsymbol{\theta}_i) = \boldsymbol{S}_i$$

$$E(\boldsymbol{\theta}_i) = \boldsymbol{\beta}, \mathrm{cov}(\boldsymbol{\theta}_i) = \boldsymbol{D}$$

where $\boldsymbol{\theta}_i$ are i.i.d. random vectors and D is the $p \times p$ between-study covariance matrix, not necessarily diagonal. From these assumptions, we obtain

$$\hat{\boldsymbol{\beta}}_i \sim N(\boldsymbol{\beta}, \boldsymbol{D} + \boldsymbol{S}_i)$$

The initial estimate of $\boldsymbol{D}$ was given by

$$\hat{\boldsymbol{D}} = \frac{1}{k-1} \sum_{i=1}^{k} \left(\hat{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}}\right) \left(\hat{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}}\right)' - \frac{1}{n} \sum_{i=1}^{k} \boldsymbol{S}_i \tag{3.3}$$

Equation (3.3) is akin to Equation (2.20) in Section 2.3. Once we derive this initial estimate of $\boldsymbol{D}$, we will use it to obtain a new estimate of the average coefficient vector and its covariance by replacing $\boldsymbol{D}$ with our estimate $\hat{\boldsymbol{D}}$ in Equations (3.4) and (3.5) as follows:

$$\hat{\boldsymbol{\beta}}^* = \left(\sum_{i=1}^{k} (\boldsymbol{D} + \boldsymbol{S}_i)^{-1} \right)^{-1} \left(\sum_{i=1}^{k} (\boldsymbol{D} + \boldsymbol{S}_i)^{-1} \hat{\boldsymbol{\beta}}_{\boldsymbol{i}} \right) \tag{3.4}$$

and

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}^*) = \left(\sum_{i=1}^{k} (\boldsymbol{D} + \boldsymbol{S})_i^{-1} \right)^{-1} \tag{3.5}$$

In the second iteration, $\bar{\boldsymbol{\beta}}$ in Equation (3.3) is replaced by $\hat{\boldsymbol{\beta}}^*$. Then, iterate between estimating $\hat{\boldsymbol{\beta}}^*$ and estimating $\hat{\boldsymbol{D}}$ until convergence for a specified stopping criterion. When there exists negative variances in the covariance matrix $\hat{\boldsymbol{D}}$, we carry out a spectral decomposition with respect to $\hat{\boldsymbol{D}}$ and truncate the negative eigen values at zero to ensure a *p.s.d* $\hat{\boldsymbol{D}}$. Accordingly, the truncated between-study covariance matrix $\hat{D}_+$ is given by

$$\hat{D}_+ = \sum_{i=1}^{p} \max(0, \lambda_i) e_i e_i^T$$

where λ_i and e_i are the ith eigenvalue and the corresponding eigenvector of $\hat{D}$. The procedure for fixed-effects and random-effects multivariate models described above has been implemented in R by script written by the author.

In this investigation, we start by fitting a simple logistic model of CHD death as response variable and age, blood pressure and their interaction as predictors for each of the 27 studies. However, of 27 studies, the two studies, Guangzhou and NHIS were excluded from our analysis since there are missing values for the CHD death variable in the Guangzhou study and for blood pressure in the NHIS study. In order to remove the potential confounding effect of sex, the subjects were stratified by sex, 25 cohorts for males and 19 cohorts for females. Accordingly, gender-specific fitted logistic models were obtained individually for the 44 cohorts. After fitting the logistic models for all the strata, estimation of the coefficients for the main effects of age, blood pressure and their interaction were done for each cohort.

The Wald test or likelihood ratio test can be used to test the significance of the main effects and the interaction term. In order to evaluate the significance of the age-sbp interaction simultaneously in the studies, the Fisher's inverse chi-square method is used to combine the p-values [37]. Suppose that p-values of n tests are $p_1, ..., p_n$, when the null hypothesis, 'there are no significant age-sbp interaction in the studies' is true, p_i $(i = 1, ..., n)$ is uniformly distributed in the interval (0,1), then -2log(p_i) follows a chi-square distribution with 2 degrees of freedom. Since the k studies are independent of each other, $-2\sum_{i=1}^{k}$ log(p_i) has a chi-square distribution with $2k$ degrees of freedom. The rejection criteria is then $-2\sum_{i=1}^{k}$ log(p_i) $\geq C$, where the critical value C is obtained from the upper tail of the chi-squares distribution with $2k$ degrees of freedom.

After testing the significance of the age-sbp interaction, multivariate fixed-effects and random-effects models were both applied in order to compare the results of the estimates of

overall coefficients and their covariance matrix. In addition, homogeneity of the between-study variation was tested.

3.4 Results

3.4.1 Univariate Meta Analysis

The results of random-effects meta analysis for logORs of CHD death with SBP in different groups stratified by sex and age groups are shown in Table 3.4. We can see that the overall logORs show a clear decreasing trend with age for both females and males. Based on the 95%CIs, the positive association of CHD death with SBP becomes insignificant for age groups $[70, 80)$ and 80+ in female, and the same result can be found for age groups 80+ in male. Furthermore, the decreasing trends can be seen in Figure 3.2, which shows the weighted regression of overall logORs of CHD associated with 1 mmHg increase in SBP vs age group. Even though the two lines have different slopes of 0.0059 and 0.0046 for females and males, they are close to each other. The analysis was also done for total cholesterol level (CHOL). Figure 3.3 presents the weighted regression of overall logORs of CHD with 1 mmol/L increase in CHOL vs age group. It can be seen that logORs of CHD with CHOL declines faster in females than in males by comparing the slopes of weighted regression line, and females also have a higher starting point compared to males.

3.4.2 Multivariate Meta Analysis

Testing the Significance of Interaction

In the multiple logistic models of CHD as response and age, SBP, and their interaction as covariates for all the 44 cohorts, the coefficients and the covariance matrix were estimated. Based on the Wald-test for the interaction of age and sbp, only 9 out of 19 cohorts have p-values that are greater than 0.05 in females, i.e., 0.29(ARIC), 0.161(BIP), 0.089(Cordis), 0.189(Glostrup), 0.436(HDFP), 0.34 (LRC), 0.241(Norway), 0.35 (Renfrew-Paisley) and 0.192(Scottish Collaborative) while in males 12 out of 25 cohorts the interaction of age and sbp appears insignificant with large p-values, i.e., 0.167(ARIC), 0.142(BIP), 0.238(CHS), 0.375(Cordis), 0.673(Framingham Cohort), 0.694(Framingham Offspring), 0.244(Glostrup), 0.525(LRC-CPPT), 0.586(MRFIT), 0.606(Norway), 0.169 (Renfrew-Paisley) and 0.06(Scottish Collaborative). The high values of the Fisher's statistic, 168.72 and 180.96, for women

Table 3.4: Results of Random-effects Meta Analysis for SBP

Gender	Age Group	Studies	logOR	s.e.(logOR)	95%CI	p-value(Q test)
female	25-40	4	0.0310	0.0052	(0.0208,0.0413)	0.3702
female	40-50	11	0.0245	0.0028	(0.0191,0.0299)	0.0221
female	50-60	14	0.0192	0.0021	(0.0151,0.0234)	0.0088
female	60-70	14	0.0113	0.0022	(0.0071,0.0155)	0.0005
female	70-80	10	0.0079	0.0025	(-0.0030,0.1280)	0.1115
female	80+	3	0.0081	0.0061	(-0.0040,0.0201)	0.5492
male	25-40	14	0.0224	0.0043	(0.0141,0.0308)	0.1232
male	40-50	23	0.0224	0.0016	(0.0193,0.0255)	0.0084
male	50-60	23	0.0163	0.0015	(0.0134,0.0193)	<0.0001
male	60-70	20	0.0113	0.0015	(0.0084,0.0141)	0.0005
male	70-80	10	0.0094	0.0020	(0.0055,0.0132)	0.2217
male	80+	2	0.0055	0.0062	(-0.0067,0.0177)	0.6621

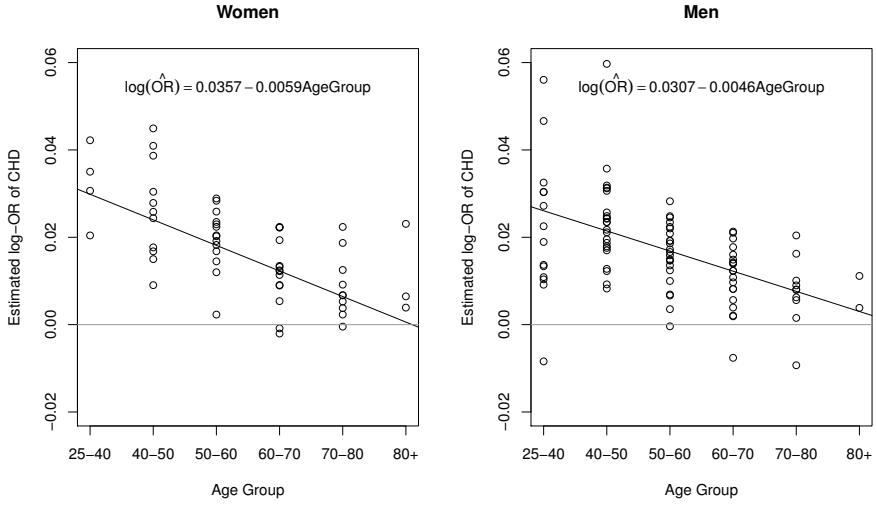

Figure 3.2: WLS Regression of Overall logOR of CHD with SBP vs Agegroup

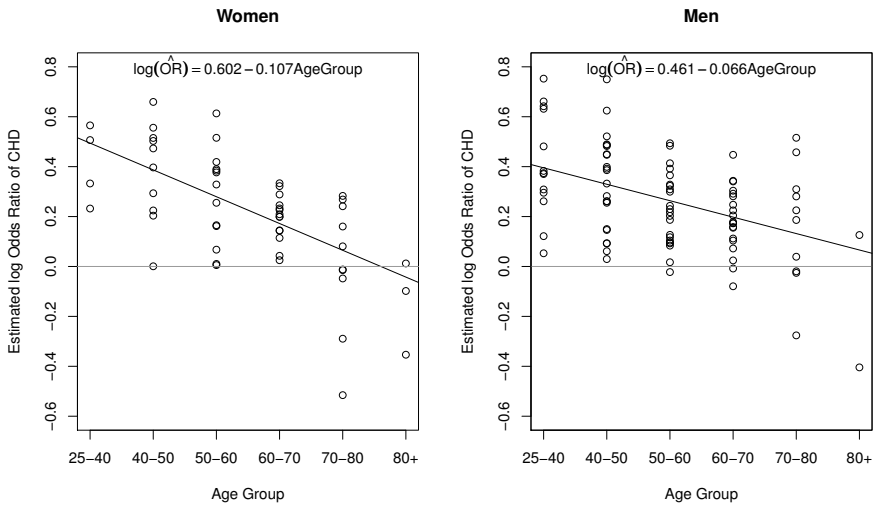

Figure 3.3: WLS Regression of Overall logOR of CHD with CHOL(mmol/L) vs Agegroup

Table 3.5: Results of Multivariate Fixed-effects Model for Age and SBP

Sex	Terms	Overall Coefs	Std Error	95% Lower CI	95% Upper CI
Female	age	0.176	0.0099	0.167	0.205
	sbp	0.055	0.0038	0.051	0.066
	agesbp	-0.00068	0.000066	-0.00079	-0.00053
Male	age	0.139	0.0069	0.127	0.153
	sbp	0.046	0.0027	0.041	0.052
	agesbp	-0.00055	0.000047	-0.00064	-0.00046

and men, indicate that a strong interaction exists between age and sbp in the studies.

Multivariate Fixed-effects Models

Table 3.5 shows the results from the multivariate fixed-effects model. We can see that the positive overall main effects of age and sbp and the negative interaction are all significant due to the 95% confidence intervals excluding 0. Moreover, the negative interaction between age and sbp indicates that the impact of high SBP on CHD death declines as age increases when the between-study variation is not considered. However, the tests of homogeneity for between-study variation are both significant for the two gender groups with high values of

Table 3.6: Results of Multivariate Random-effects Model for Age and SBP

Sex	Terms	Overall Coefs	Std Error	95% Lower CI	95% Upper CI
Female	age	0.202	0.0327	0.137	0.266
	sbp	0.064	0.0103	0.044	0.084
	agesbp	-0.0008	0.000194	-0.0012	-0.0004
Male	age	0.145	0.0111	0.124	0.167
	sbp	0.047	0.0054	0.037	0.058
	agesbp	-0.00058	0.000076	-0.00073	-0.00043

Table 3.7: Results of $\hat{D}$ for GLS Multivariate Random-effects Model for SBP

Terms	Between-study Covariance $\hat{D}$
Female + SBP	$\begin{pmatrix} 0.01559 & 0.00475 & -8.98\text{e-}5 \\ 0.00475 & 0.001446 & -2.74\text{e-}5 \\ -8.98\text{e-}5 & -2.74\text{e-}5 & 5.27\text{e-}7 \end{pmatrix}$
Male + SBP	$\begin{pmatrix} 0.00138 & 0.000706 & -7.569\text{e-}6 \\ 0.000706 & 0.00044 & -5.14\text{e-}6 \\ -7.57\text{e-}6 & -5.14\text{e-}6 & 6.28\text{e-}8 \end{pmatrix}$

chi-square statistic, respectively, 207.37 and 322.52. Therefore, a random-effects model is appropriate.

Multivariate Random-effects Models

Tables 3.6 and 3.7 show the results from the multivariate random-effects models. Compared to the results of the fixed-effects model in Table 3.5, we can see that the random-effects models give wider confidence intervals for both women and men. Moreover, the difference between fixed-effects and random-effects models have a greater impact in women than in men, which indicates that the between-study variance in women is relatively larger. In addition, the overall estimates for age, sbp and the interaction are larger in women than in men. Figures 3.4 and 3.5 show the logORs of CHD associated with specified increase in SBP for different ages in women and men. It is obvious that for a fixed increase in SBP, say 20mmHg, the logORs of CHD associated with SBP decrease as age increases for both

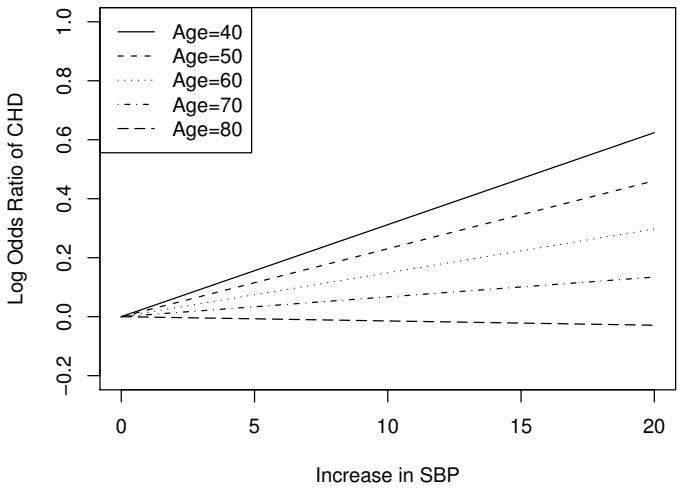

Figure 3.4: Log Odds Ratio of CHD death associated with specified increase in SBP for different ages in women

women and men. In other words, high SBP has less impact on CHD death for older people than for younger people. Figure 3.6 also shows the declining trend of the logORs of CHD death with 10mmHg increase in SBP with age. Moreover, while women have higher logORs of CHD associated with SBP than men shown in Figures 3.4 and 3.5, it seems that there is no association between CHD death and SBP in men and women around age 75 and 80 respectively due to a zero slope, and the relationship switches direction by becoming negative after these ages. Based on the 95% confidence intervals shown in Figure 3.6, it is indicated that CHD death is no longer significantly related to the elevation of SBP in women over 70 years old and in men over around 75 years old. This is actually consistent to that in univariate meta analysis. Therefore, based on the above results, we can conclude that aging plays an important role in influencing the effect of blood pressure on CHD death.

Compare GLS to MM and REML methods

We also compared three methods of estimation, GLS to MM and REML. Table 3.9 shows the results from multivariate random-effects models using the three methods. Comparing the results, we can see that the overall effects are quite close to each other while the

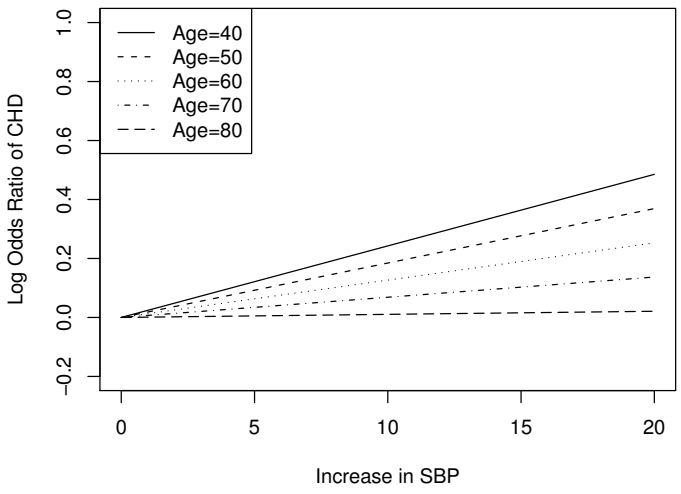

Figure 3.5: Log Odds Ratio of CHD death associated with specified increase in SBP for different ages in men

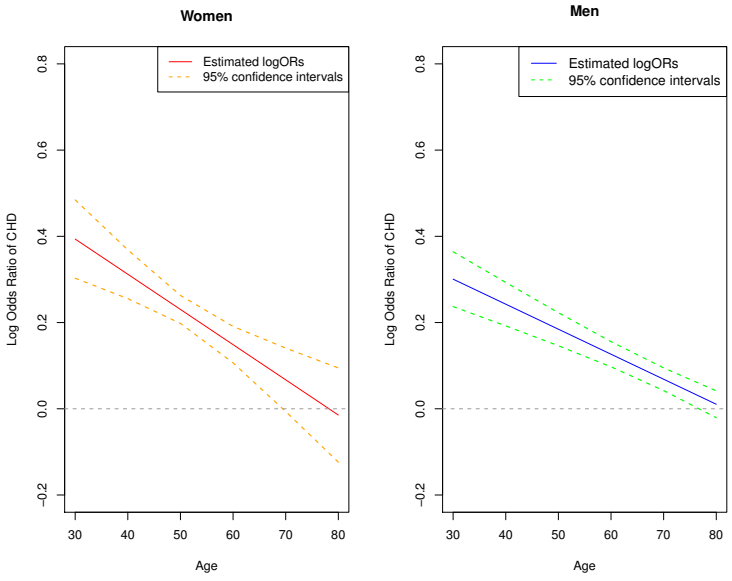

Figure 3.6: LogORs of CHD death with 10mmHg increase in SBP vs. age (Random-effects Model)

Table 3.8: Comparison of Results from Multivariate Random-effects Models using MM, REML and GLS methods

Method	Sex	Terms	Overall Coefs	Std Error	95% Lower CI	95% Upper CI
MM	Female	age	0.192	0.0159	0.161	0.224
		sbp	0.061	0.0069	0.048	0.075
		agesbp	-0.0008	0.00010	-0.00096	-0.00056
REML	Female	age	0.191	0.0121	0.167	0.215
		sbp	0.060	0.0055	0.049	0.071
		agesbp	-0.0007	0.00008	-0.00091	-0.00059
GLS	Female	age	0.202	0.0327	0.137	0.266
		sbp	0.064	0.0103	0.044	0.084
		agesbp	-0.0008	0.000194	-0.0012	-0.0004
MM	Male	age	0.145	0.0164	0.125	0.165
		sbp	0.048	0.0041	0.040	0.056
		agesbp	-0.0006	0.000065	-0.00071	-0.00045
REML	Male	age	0.146	0.0164	0.127	0.165
		sbp	0.048	0.0041	0.041	0.055
		agesbp	-0.0006	0.000057	-0.00070	-0.00047
GLS	Male	age	0.145	0.0111	0.124	0.167
		sbp	0.047	0.0054	0.037	0.058
		agesbp	-0.00058	0.000076	-0.00073	-0.00043

standard errors of overall effects for MM and REML methods are more consistent than for the GLS method. Moreover, compared to MM and REML, the GLS method gives higher standard errors in Female+SBP and Male+DBP, and lower standard errors in Male+SBP and Female+DBP. The difference between the standard errors of overall effects among three methods may be due to the different estimates of between-study covariance matrices $\hat{D}$. After the negative eigen values of $\hat{D}$'s are forced to zero, the resulting estimate becomes $\hat{D}_+$, which might give rise to larger differences for the standard errors of overall effects. Further simulation study is needed in the future to compare the three methods.

CHAPTER 4

META ANALYSIS OF CURVES

4.1 Introduction

In the previous chapters, we focused on logistic models with linear terms to investigate how the risk effect on CHD death changes with age. Yet there is no reason to believe that the risk effect on CHD death is perfectly linear in age. We thus come up with the natural question of how we can go beyond linear models. One may use age grouping as a solution, but it is not always the best choice. When age is divided into multiple nonoverlapping age groups, no smoothness is imposed and the risk effects may demonstrate abrupt change. It is more reasonable to have gradually changed risk effect. Fitting the models regarding separately different age groups does not use all the information provided by the data. For instance, within each age group, say [40,50), we only use a subset of the data to build a model and the analysis is likely to be sensitive to the choice of bins. In fact, the choices of the bin size and endpoints of age groups often turn out to be an ad-hoc job in our experience.

We propose to use a smooth model, such as splines, to get an effect curve varying with age. In this way, the true effect curve is possibly nonlinear and the natural smoothness borrowing the neighborhood information may also improve the accuracy of estimates. This gives us a generalized additive model (GAM), which is an extension of a generalized linear model (GLM). Chapter 4 will center around GAM and then extend to multivariate meta analysis with respect to GAMs.

4.2 Generalized Additive Models

Before introducing GAM, recall the general form of binomial GLM given by

$$g(\mu) = \log(\frac{\mu}{1-\mu}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p, \tag{4.1}$$

where μ is the mean of the binary response equal to Pr(Y=1), and $g(\cdot)$ is called the link function to relate μ to covariates, $x_1, ..., x_p$. Although such a model solves the problems of bounded outcome and heterogeneous variances by transforming the expected value of response with a link function, it is still a linear model. The construction of the higher order terms typically need our *prior* knowledge of the curve shape. In some situations, the logistic regression model encounters difficulty in fitting data that does not follow a simple parametric curve shape. Therefore, GAM is more flexible since it not only 'borrows' the strength of GLM but allows more flexibility by fitting non-parametric models.

Compared to the model in Equation (4.1), an additive logistic regression model has a more general form given by

$$g(\mu) = \log(\frac{\mu}{1-\mu}) = \sum_{i=1}^{p} f_i(x_i), \tag{4.2}$$

where f_i's are 'arbitrary' smooth functions of the covariates x_i's. Note that $g(\cdot)$ in Equation (4.2) is not restricted to *logit* link function, but it can also be other classical link functions associated with GLM, such as the identity link and the log link corresponding to additive linear models and log-additive models respectively. The non-parametric form for the functions f_i's avoids the need of making parametric assumptions about the form of f_i and therefore makes the model more flexible. The more attractive properties of GAM include that we can easily handle the mixture of linear and nonlinear terms together and at the same time the nonlinear interaction among two or more covariates can be incorporated into the model. In addition, the additivity property of GAM allows us to interpret the model in the same way as before.

4.2.1 Modeling with basis functions

For simplicity of representation, consider a simple model containing one smooth function of one covariate as follows:

$$y_i = f(x_i) + \epsilon_i$$

where y_i is a response, x_i a covariate, f a smooth function and ϵ_i i.i.d. $N(0, \sigma^2)$ random variables. Normally, a set of bases are chosen to define the space of functions of f. Choosing bases is equivalent to choosing some basis functions in order to further estimate f. In other words, once we have bases, basis functions are treated known. Assume that f can be

represented as:

$$f(x) = \sum_{j=1}^{m} \beta_j s_j(x), \tag{4.3}$$

where $s_j(x)$ are the basis functions and β_j are the basis coefficients [38]. Thus, $f(x)$ is a linear combination of s_j, and β_j are unknown parameters to be estimated. In other words, estimating f amounts to finding m unknown parameters β_j's.

Suppose that (x_i, y_i) are pairs of observations. The β_j's can be obtained by minimizing:

$$\begin{aligned} \sum_{i=1}^{n} (y_i - f(x_i))^2 &= \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \beta_j s_j(x_i) \right)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \boldsymbol{s}(x_i)^T \boldsymbol{\beta} \right)^2, \end{aligned}$$

where $\boldsymbol{\beta}$ is the vector of coefficients β_j, and $\boldsymbol{s}(x_i)$ is the vector containing each basis function evaluated at x_i. It is easy to see that this is like a standard linear model fitting problem. Define the model matrix $\boldsymbol{X}$ as follows:

$$\boldsymbol{X} = \begin{bmatrix} s_1(x_1) & s_2(x_1) & \cdot\cdot & s_m(x_1) \\ s_1(x_2) & s_2(x_2) & \cdot\cdot & s_m(x_2) \\ s_1(x_3) & s_2(x_3) & \cdot\cdot & s_m(x_3) \\ \cdot & \cdot & \cdot\cdot & \cdot \\ \cdot & \cdot & \cdot\cdot & \cdot \\ s_1(x_n) & s_2(x_n) & \cdot\cdot & s_m(x_n) \end{bmatrix} = \begin{bmatrix} \boldsymbol{s}(x_1)^T \\ \boldsymbol{s}(x_2)^T \\ \boldsymbol{s}(x_3)^T \\ \cdot \\ \cdot \\ \boldsymbol{s}(x_n)^T \end{bmatrix} \tag{4.4}$$

Thus, the fitting problem is equivalent to minimizing $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ and the least squares estimates of $\boldsymbol{\beta}$ will be:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

Note that given bases, the model matrix $\boldsymbol{X}$ has its specific form. Suppose that a polynomial basis with order 3 was chosen to represent $f(x)$. The bases for the space of polynomials of order 3 and below are $s_1(x) = 1$, $s_2(x) = x$, $s_3(x) = x^2$, and $s_4(x) = x^3$, so that $f(x)$ can be written as:

$$f(x) = \sum_{j=1}^{4} \beta_j s_j(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3.$$

Then we minimize:

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \left\| \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_1 & {x_1}^2 & {x_1}^3 \\ 1 & x_2 & {x_2}^2 & {x_2}^3 \\ 1 & x_3 & {x_3}^2 & {x_3}^3 \\ . & . & . & . \\ . & . & . & . \\ 1 & x_n & {x_n}^2 & {x_n}^3 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} \right\|^2$$

Although the polynomial bases tend to be easy to use, it has poor approximation properties in theory and the polynomial bases possibly become collinear as the dimension of the basis increases. By contrast, the spline bases have good theoretical and practical properties [39]. For univariate smooth function, a cubic spline basis is popular. Let $\{x_j^* : j = 1, \ldots, q\}$ be a set of points chosen in the range of x, where x_1^* and x_q^* are two boundary points. $f(x)$ using a cubic spline can be represented as a curve made up sections of cubic polynomial joined together at chosen knots, so that $f(x)$ is continuous and also has continuous first and second derivatives at the interior knots, i.e., for $j = 2, \ldots, q-1$:

$$f(x_{j-}^*) = f(x_{j+}^*),$$

$$f'(x_{j-}^*) = f'(x_{j+}^*),$$

$$f''(x_{j-}^*) = f''(x_{j+}^*),$$

which are three constraints imposed at each interior knot. Given $q-2$ interior knots, a cubic spline basis can be written as [40]:

$$s_1(x) = 1, s_2(x) = x, s_3(x) = x^2, s_4(x) = x^3,$$

$$s_{4+l} = (x - x_{l+1}^*)_+^3 \text{ for } l = 1, \ldots, q-2.$$

There are $q+2$ basis functions corresponding to $q+2$ dimensional linear space of functions due to the fact: ($q-1$ regions) $\times$ (4 parameters per region) $-$ ($q-2$ knots) $\times$ (3 constraints per knot) $= q+2$. When polynomials are used to fit data, it may lead to strange behaviors near the boundary knots. Thus, additional constraints, namely *'natural'* spline constraints are often necessary to make the function linear beyond the boundary knots. In other words, the spline has zero second derivative outside the interval $[x_2^*, x_{q-1}^*]$, which applies $f''(x) = 0$ for $x < x_2^*$ and $x > x_{q-1}^*$. Taking the two constraints in account, the number of the resulting parameters turns out to be q.

It is well worth being aware that there are many alternative ways of writing down a basis for cubic splines, not limited to the above form. In addition, besides cubic splines, there are other splines for use, such as cubic regression splines, cyclic cubic regression spline, P-splines, etc. Since the choice of splines is not of the main interest in our research, the detailed information of them will not be discussed here. Moreover, if we want to do meta analysis for a set of studies, we have to choose the same spline with the same knots. Once the same knots are chosen, all the studies have the same basis functions that allow for use in further meta analysis.

4.2.2 Controlling the degree of smoothing

When the least square method is used to fit a GAM model, one obvious possibility is that the resulting curve would not be smooth at all, that is, the model might be too complex. Fortunately, adaptive shrinkage has been fully taken into consideration in GAM so that the drawback can be overcome by imposing a penalty during model fitting. By adding a wiggliness term, the fitting problem thus amounts to finding $\boldsymbol{\beta}$ by minimizing:

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \int [f''(x)]^2 dx, \tag{4.5}$$

where λ is a tunable smoothing parameter. The first term measures the closeness of the function $f(x)$ to the data, and the second term penalizes curvature in the function using the integral of second derivative squared. The smoothing parameter λ is selected to control the tradeoff between closely matching the data and having a smooth model. As λ is getting larger, the data will be more smoothed; very large values for λ leads to a straight line.

The wiggliness term can always written as a quadratic from in $\boldsymbol{\beta}$ with known coefficient matrix $\boldsymbol{\Omega}$ as weighting matrix as follows:

$$\int [f''(x)]^2 dx = \boldsymbol{\beta}^T \boldsymbol{\Omega} \boldsymbol{\beta},$$

where $\boldsymbol{\Omega}$ is a penalty matrix with a general form as follows:

$$\boldsymbol{\Omega} = \begin{bmatrix} \int s_1''(x)^2 \mathrm{d}x & \int s_1''(x) s_2''(x) \mathrm{d}x & \cdot\cdot & \int s_1''(x) s_m''(x) \mathrm{d}x \\ \int s_2''(x) s_1''(x) \mathrm{d}x & \int s_2''(x)^2 \mathrm{d}x & \cdot\cdot & \int s_2''(x) s_m''(x) \mathrm{d}x \\ \cdot & \cdot & \cdot\cdot & \cdot \\ \cdot & \cdot & \cdot\cdot & \cdot \\ \int s_m''(x) s_1''(x) \mathrm{d}x & \int s_m''(x) s_2''(x) \mathrm{d}x & \cdot\cdot & \int s_m''(x)^2 \mathrm{d}x \end{bmatrix}$$

In theory, each entry of $\mathbf{\Omega}$ can be calculated because s'' does depend on the specification of the splines and knots we choose.

S. Wood [41] provided a knot-based spline that not only facilitates us to obtain basis functions when a set of knots are placed, but has a simple form for penalty matrix. Suppose that $q-2$ interior knots $\{x_j^* : j = 2, \ldots, q-1\}$ are chosen within the range of a covariate. The basis is as follows: $s_1(x) = 1, s_2(x) = x, s_{i+2}(x) = R(x, x_j^*)$ for $i = 1, \ldots, q-2$ and$j = 2, \ldots, q-1$ where

$$\begin{aligned} R(x, x_j^*) = \quad & [(x_j^* - 1/2)^2 - 1/12][(x - 1/2)^2 - 1/12]/4 \\ & -[(|x - x_j^*| - 1/2)^4 - 1/2((|x - x_j^*| - 1/2)^2) + 7/240]/24 \end{aligned} \tag{4.6}$$

The first basis function is the intercept term, the second is the linear term, and all other higher order terms are given by the $R(x, x_j^*)$ equation. For each knot, we have an extra predictor based on the equation. Note that the basis here is suitable on $x \in [0, 1]$. When we want to use this basis, we need to scale x to ensure it belongs to the interval $[0, 1]$. As a result, the model matrix $\boldsymbol{X}$ is:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & R(x_1, x_2^*) & R(x_1, x_3^*) & \cdot\cdot & R(x_1, x_{q-2}^*) \\ 1 & x_2 & R(x_2, x_2^*) & R(x_2, x_3^*) & \cdot\cdot & R(x_2, x_{q-2}^*) \\ \cdot & \cdot & \cdot & \cdot & \cdot\cdot & \cdot \\ 1 & x_n & R(x_n, x_2^*) & R(x_n, x_3^*) & \cdot\cdot & R(x_n, x_{q-2}^*) \end{bmatrix} \tag{4.7}$$

Given the basis and knots location, the penalty matrix $\mathbf{\Omega}$ is known of the following form:

$$\mathbf{\Omega} = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdot\cdot & 0 \\ 0 & 0 & 0 & 0 & \cdot\cdot & 0 \\ 0 & 0 & R(x_2^*, x_2^*) & R(x_2^*, x_3^*) & \cdot\cdot & R(x_2^*, x_{q-2}^*) \\ 0 & 0 & R(x_3^*, x_2^*) & R(x_3^*, x_3^*) & \cdot\cdot & R(x_3^*, x_{q-2}^*) \\ 0 & 0 & \cdot & \cdot & \cdot\cdot & \cdot \\ 0 & 0 & R(x_{q-2}^*, x_2^*) & R(x_{q-2}^*, x_3^*) & \cdot\cdot & R(x_{q-2}^*, x_{q-2}^*) \end{bmatrix} \tag{4.8}$$

There are a lot of zeros in the penalty matrix $\mathbf{\Omega}$. The reason is because the second derivatives of the intercept and linear terms are zero. It indicates that no penalties are imposed on the intercept and linear terms. Therefore, the fitting problem becomes minimizing:

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^T \mathbf{\Omega} \boldsymbol{\beta}. \tag{4.9}$$

The penalized least squares estimator of $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \mathbf{\Omega})^{-1} \boldsymbol{X}^T \boldsymbol{y}, \tag{4.10}$$

and the influence matrix $\boldsymbol{A}$ is:

$$\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}^T. \tag{4.11}$$

Another problem in fitting is how large the λ value should be to control the wiggliness of a model. A selection criteria of λ commonly used is to minimize prediction error by generalized cross validation (GCV) [42] that is derived from ordinary cross validation (OCV). The GCV score often has the following form:

$$V_g = \frac{n\sum_{i=1}^{n}(\hat{f}_i - y_i)^2}{[n - \mathrm{tr}(\boldsymbol{A})]^2},$$

where $\hat{f}_i$ is the estimate for the i^{th} observation, and $\boldsymbol{A}$ is the influence matrix defined in Equation (4.11). The term $\mathrm{tr}(\boldsymbol{A})$ is the estimated degrees of freedom of the model. Note then that the lower the GCV score, the better the model. In addition to GCV, there is another alternative criterion for choosing λ, which attempts to minimize the expected mean square error by Mallow's C_p, BIC (Bayesian information criterion) or AIC (Akaike's information criterion) [41].

4.2.3 Fitting by matrix decompositions

In practice, orthogonal matrix methods are often used in solving the least squares problem in order to improve computational efficiency. Generally, the calculation of least square estimates begins with first forming the QR decomposition of the model matrix, $\boldsymbol{X}$. Specifically, the model matrix $\boldsymbol{X}$ can always be decomposed

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R},$$

where $\boldsymbol{Q}$ is the first p columns of an $n \times n$ orthogonal matrix, and $\boldsymbol{R}$ is a $p \times p$ upper triangular matrix. Orthogonal matrices just rotate vectors without changing their length. Also recall the property of orthogonality, $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$. Then, both the response vector $\boldsymbol{y}$ and the columns of $\boldsymbol{X}$ can be rotated by pre-multiplication with $\boldsymbol{Q}^T$. Thus, applying $\boldsymbol{Q}^T$ to $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ yields

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \|\boldsymbol{Q}^T\boldsymbol{y} - \boldsymbol{Q}^T\boldsymbol{X}\boldsymbol{\beta}\|^2 = \left\|\boldsymbol{Q}^T\boldsymbol{y} - \boldsymbol{R}\boldsymbol{\beta}\right\|^2.$$

Therefore, minimizing $\left\|\boldsymbol{Q}^T\boldsymbol{y} - \boldsymbol{R}\boldsymbol{\beta}\right\|^2 + \lambda\boldsymbol{\beta}^T\boldsymbol{\Omega}\boldsymbol{\beta}$ yields

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{R}^T\boldsymbol{R} + \lambda\boldsymbol{\Omega})^{-1}\boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{y},$$

where $\hat{\boldsymbol{\beta}}$ is the least square estimate of $\boldsymbol{\beta}$. When $\lambda = 0$, $\hat{\boldsymbol{\beta}} = \boldsymbol{R}^{-1}\boldsymbol{Q}^T\boldsymbol{y}$. Using the orthogonal decomposition method, the computational efficiency can be greatly improved.

4.2.4 Extensions to generalized linear models

The previous sections only dealt with penalized linear models where the response is a continuous variable rather than penalized generalized linear models (GLM), but the generalization is straightforward. For binomial GLM, we can define a penalized likelihood for the model:

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\lambda\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{\Omega}\boldsymbol{\beta}, \tag{4.12}$$

where λ is a nonnegative smoothing parameter and $\boldsymbol{\Omega}$ is a penalty matrix. When a spline is given, $\boldsymbol{\Omega}$ is treated as known. Meanwhile, $l(\boldsymbol{\beta})$ is the unpenalized binomial log-likelihood. For estimating $\boldsymbol{\beta}$, we denote $\boldsymbol{p}$ a $n \times 1$ vector of probability estimates that the responses are equal to 1. Further define the $\boldsymbol{\beta}$-dependent weights as $w_i = p_i(1 - p_i)$ for $i = 1, \ldots, n$. It is not hard to show that the first derivative (the gradient) of $l(\boldsymbol{\beta})$ as a function of $\boldsymbol{\beta}$ is

$$\nabla l(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{p}),$$

and the Hessian matrix of $l(\boldsymbol{\beta})$ is

$$H(l(\boldsymbol{\beta})) = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\boldsymbol{T}}} = -\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X},$$

where $\boldsymbol{W}$ is the $n \times n$ diagonal matrix with the diagonal elements of w_i for $i = 1, \ldots, n$, and $\boldsymbol{X}$ is the model matrix determined by the chosen spline bases and the data (for the spline given in Equation (4.6), $\boldsymbol{X}$ is defined as the form in Equation (4.7)). When a penalty is imposed, the first and second derivatives of $l_p(\boldsymbol{\beta})$ turn out to be the following form:

$$\nabla l_p(\boldsymbol{\beta}) = \frac{\partial l_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{p}) - \lambda\boldsymbol{\Omega}\boldsymbol{\beta},$$

and the second derivative of $l_p(\boldsymbol{\beta})$ is

$$H(l_p(\boldsymbol{\beta})) = \frac{\partial^2 l_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\boldsymbol{T}}} = -(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X} + \lambda\boldsymbol{\Omega}).$$

To fit penalized logistic regression models, we use the Newton-Raphson algorithm:

$$\begin{aligned} \boldsymbol{\beta}^{new} &= \boldsymbol{\beta}^{old} - \left(\frac{\partial^2 l_p}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\boldsymbol{T}}}\right)^{-1}\frac{\partial l_p}{\partial \boldsymbol{\beta}} \\ &= \boldsymbol{\beta}^{old} + (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X} + \lambda\boldsymbol{\Omega})^{-1}(\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{p}) - \lambda\boldsymbol{\Omega}\boldsymbol{\beta}^{old}). \end{aligned}$$

The coefficient estimation is controlled by the nonnegative penalty regularization parameter λ, thus the value of λ needs to be determined by the data. The criteria, such as GCV score, Mallow's C_p, BIC or AIC [43] can be effective approaches to specify the penalty parameter. In this work, we minimize AIC:

$$AIC = -2 \times l(\hat{\boldsymbol{\beta}}) + 2 \times df(\lambda), \tag{4.13}$$

where $df(\lambda)$ is the effective degrees of freedom approximated by

$$df(\lambda) = tr[(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X} + \lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X}]. \tag{4.14}$$

Using the values of $\boldsymbol{W}$ and $\hat{\boldsymbol{\beta}}$ from the final Newton-Raphson step, we can estimate the $df(\lambda)$ and AIC, thus the optimal λ corresponding to the minimum AIC can be found. In addition, the covariance matrix of $\hat{\boldsymbol{\beta}}$ is also estimated from the final iteration:

$$\begin{aligned} \text{cov}(\hat{\boldsymbol{\beta}}) = & \quad (\frac{\partial^2 l_p}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\boldsymbol{T}}})^{-1} I(\boldsymbol{\beta}) (\frac{\partial^2 l_p}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\boldsymbol{T}}})^{-1} \\ = & \quad (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X} + \lambda\boldsymbol{\Omega})^{-1} (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X}) (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X} + \lambda\boldsymbol{\Omega})^{-1}, \end{aligned}$$

where $I(\boldsymbol{\beta})$ denotes the information in $\boldsymbol{y}$. The covariance estimate is often called a sandwich estimate.

Like in the ordinary GLM, the penalized GLM may also have dispersion effect. In general, dispersion is not an issue in ordinary linear regression with normally distributed y, whereas for binomial and Poisson distributions, dispersion deserves to be taken into account because the variance of y is a function of the mean. Dispersion is especially common for a Poisson distribution due to the variances much larger than the means [44], but Poisson distributions have identical mean and variance. The dispersion parameter, ϕ, can be estimated by the Pearson-like scale estimator, Pearson chi-squared statistic divided by residual degree of freedom. In general, the estimate of ϕ is defined as

$$\hat{\phi} = \frac{\sum_i V(\hat{\mu}_i)^{-1}(y_i - \hat{\mu}_i)^2}{n - df(\lambda)}, \tag{4.15}$$

where $V(\hat{\mu}_i)$ is the variance of $\hat{\mu}_i$. When y has a binomial distribution, $V(\hat{\mu}_i)$ equals to $p_i(1 - p_i)$. The magnitude of $\hat{\phi}$ reflects the degree of dispersion effect. $\hat{\phi} > 1$ reflects overdispersion; $\hat{\phi} = 1$ reflects no dispersion; and $\hat{\phi} < 1$ indicates underdispersion. Overdispersion typically occurs for a Poisson distribution while underdispersion might occur

for a Binomial distribution. Therefore, by taking dispersion effect into consideration, the covariance matrix of $\boldsymbol{\beta}$ turns out to be [45]:

$$\text{cov}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X^T W X} + \lambda\boldsymbol{\Omega})^{-1}(\boldsymbol{X^T W X})(\boldsymbol{X^T W X} + \lambda\boldsymbol{\Omega})^{-1}\hat{\phi}. \tag{4.16}$$

According to the above description, we suggest the following recipe for univariate penalized logistic regression:

1. Choose a spline basis and q knots within the range of covariate;
2. Vary λ within a range, and for each value of λ use the Newton-Raphson algorithm to estimate $\hat{\boldsymbol{\beta}}(\lambda)$;
3. Compute AIC and estimate $\hat{\phi}$;
4. Find the optimal λ to minimize the AIC, and the optimal $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ are obtained, and $\text{cov}(\hat{\boldsymbol{\beta}})$ can be estimated as well.

4.2.5 Smooth of multiple covariates

We discussed univariate GAM, which can be represented using a basis with associated penalty regularizing wiggliness of the model. The methods can also be used to construct smooth functions of any number of covariates. Suppose here that two covariates, x_1 and x_2, are available for a binary response y, and that a logistic additive model has the following form:

$$\text{logit}(\mu) = f(x_1, x_2), \tag{4.17}$$

where $f(x_1, x_2)$ is a two-dimensional function of two variables, which can be represented as a smooth surface on the two-dimensional domain of x_1 and x_2. Without x_2, $f(x_1, x_2)$ is actually a one-dimensional function. For a fixed value of x_2, we can do basis expansion on x_2 using basis functions a_i:

$$f(\boldsymbol{x_1}, x_2) = \sum_{i=1}^{I} \alpha_i(x_2) a_i(\boldsymbol{x_1})$$

where the basis coefficients α_i are not constant but a coefficient function of x_2. Thus, we can do further basis expansion on x_2 using basis functions b_j:

$$\alpha_i(x_2) = \sum_{j=1}^{J} \beta_{ij} b_j(x_2).$$

Note that here the basis functions a_i and b_j are different. Combining the above two steps yields the final two-dimensional model $f(\boldsymbol{x_1}, x_2)$:

$$f(\boldsymbol{x_1}, \boldsymbol{x_2}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \beta_{ij} b_j(\boldsymbol{x_2}) a_i(\boldsymbol{x_1}).$$

where I and J are the number of chosen knots in the intervals of x_1 and x_2, respectively. In the above model, β_{ij} are the two-dimensional basis coefficients to be estimated and $b_j(\boldsymbol{x_2})a_i(\boldsymbol{x_1})$ are the two-dimensional basis functions. In this way, a smooth surface on the two-dimensional domain of x_1 and x_2 can be obtained.

There is another much easier way to construct two-dimensional basis functions. Suppose that we use the standard notation of $\otimes$ to denote Kronecker product and $\odot$ as element-wise multiplication of matrices, respectively. We can reexpress the surface in matrix notation as $f(x_1, x_2) = \boldsymbol{X}\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is the $IJ \times 1$ vector of unknown basis coefficients and $\boldsymbol{X}$ is the model matrix given by

$$\boldsymbol{X} = (\boldsymbol{X}_1 \otimes \mathbf{1}'_J) \odot (\mathbf{1}'_I \otimes \boldsymbol{X}_2), \tag{4.18}$$

and is of dimension $n \times IJ$. In Equation (4.18), $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are the model matrices constructed from the basis functions a_i and b_j.

It is then easy to show that the ith row of $\boldsymbol{X}$ is simply:

$$\boldsymbol{X}_i = \boldsymbol{X}_{1i} \otimes \boldsymbol{X}_{2i}.$$

These are indeed **tensor product** bases. In other words, once we specify a spline for the two covariates and choose a set of knots on the grid of x_1 and x_2, the two-dimensional basis functions are known. Given the bases, it is straightforward to create model matrices, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}$. It allows us to reduce a complex two-dimensional smoothing problem to an ordinary parameter estimation.

Having derived the two-dimensional basis functions for representing the function $f(x_1, x_2)$, another crucial question is how to regularize the wiggliness of the two-dimensional function because we want it to be smooth in both x_1 and x_2. A natural way of measuring wiggliness of $f(x_1, x_2)$ is to start from one-dimensional problem and then extend to the two covariates case. Suppose that the penalty term for each marginal smooth can be expressed as a quadratic form in the marginal parameters:

$$J(f_{x_1}) = \boldsymbol{\alpha}^T \boldsymbol{\Omega}_1 \boldsymbol{\alpha} \text{ and } J(f_{x_2}) = \boldsymbol{\beta}^T \boldsymbol{\Omega}_2 \boldsymbol{\beta},$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of basis coefficients of $f(x_1)$ and $f(x_2)$, and $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ are the penalty matricies of dimension $I \times I$ and $J \times J$. In order to measure wiggliness of $f(x_1, x_2)$, we define:

$$J(f) = \boldsymbol{\gamma}^T(\lambda_1\tilde{\boldsymbol{\Omega}}_1 + \lambda_2\tilde{\boldsymbol{\Omega}}_2)\boldsymbol{\gamma},$$

where λ_1 and λ_2 are smoothing parameters controlling the tradeoff between wiggliness in two directions on the domain of x_1 and x_2, and $\tilde{\boldsymbol{\Omega}}_1$ and $\tilde{\boldsymbol{\Omega}}_2$ are the penalty matrices transformed by the following way [41]:

$$\tilde{\boldsymbol{\Omega}}_1 = \boldsymbol{\Omega}_1 \otimes \boldsymbol{I}_J \text{ and } \tilde{\boldsymbol{\Omega}}_2 = \boldsymbol{I}_I \otimes \boldsymbol{\Omega}_2. \tag{4.19}$$

where $\boldsymbol{I}_I$ and $\boldsymbol{I}_J$ are the rank I and J identity matrices. Based on the definition of Kronecker product, $\tilde{\boldsymbol{\Omega}}_1$ and $\tilde{\boldsymbol{\Omega}}_2$ can be written as the following forms :

$$\tilde{\boldsymbol{\Omega}}_1 = \begin{bmatrix} \boldsymbol{\Omega}_1^{11}\boldsymbol{I}_J & \boldsymbol{\Omega}_1^{12}\boldsymbol{I}_J & \cdot\cdot & \boldsymbol{\Omega}_1^{1I}\boldsymbol{I}_J \\ \boldsymbol{\Omega}_1^{21}\boldsymbol{I}_J & \boldsymbol{\Omega}_1^{22}\boldsymbol{I}_J & \cdot\cdot & \boldsymbol{\Omega}_1^{2J}\boldsymbol{I}_J \\ \cdot & \cdot & \cdot\cdot & \cdot \\ \boldsymbol{\Omega}_1^{I1}\boldsymbol{I}_J & \boldsymbol{\Omega}_1^{I2}\boldsymbol{I}_J & \cdot\cdot & \boldsymbol{\Omega}_1^{II}\boldsymbol{I}_J \end{bmatrix}_{IJ\times IJ}$$

and

$$\tilde{\boldsymbol{\Omega}}_2 = \begin{bmatrix} \boldsymbol{\Omega}_2 & & & \\ & \boldsymbol{\Omega}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{\Omega}_2 \end{bmatrix}_{IJ\times IJ},$$

where $\boldsymbol{\Omega}_1^{ij}$ denotes the $(i,j)^{th}$ entry of $\boldsymbol{\Omega}_1$, and each block in $\tilde{\boldsymbol{\Omega}}_1$ is a diagonal matrix (proportional to the identity matrix). Clearly, $\tilde{\boldsymbol{\Omega}}_2$ is a block diagonal matrix with $\boldsymbol{\Omega}_2$ as diagonal blocks. Hence, once we specify a spline for the two covariates and choose a set of knots on the grid of x_1 and x_2, $\tilde{\boldsymbol{\Omega}}_1$ and $\tilde{\boldsymbol{\Omega}}_2$ are known.

To give a better view of how the different penalties are imposed on the coefficients, we rearrange the $IJ \times 1$ vector of $\boldsymbol{\gamma}$ as a $I \times J$ matrix:

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdot\cdot & \gamma_{1J} \\ \gamma_{21} & \gamma_{22} & \cdot\cdot & \gamma_{2J} \\ \cdot & \cdot & \cdot\cdot & \cdot \\ \gamma_{I1} & \gamma_{I2} & \cdot\cdot & \gamma_{IJ} \end{bmatrix}_{I\times J}$$

Then $J(f)$ becomes the following:

$$J(f) = \lambda_1 \sum_{j=1}^{J} \boldsymbol{\gamma}_{\cdot j}^T \boldsymbol{\Omega}_1 \boldsymbol{\gamma}_{\cdot j} + \lambda_2 \sum_{i=1}^{I} \boldsymbol{\gamma}_{i\cdot}^T \boldsymbol{\Omega}_2 \boldsymbol{\gamma}_{i\cdot}$$

where $\boldsymbol{\gamma}_{\cdot j}$ is the j^{th} column of $\boldsymbol{\gamma}$ containing I coefficients, and $\boldsymbol{\gamma}_{i\cdot}$ is the i^{th} row of $\boldsymbol{\gamma}$ containing J coefficients. It is not hard to see that λ_1 and λ_2 control the smoothness along x_1 and x_2 directions, respectively. In other words, with the different penalties imposed along two directions, the smoothness of $f(x_1, x_2)$ can accordingly be well controlled.

Given a wiggliness measure for the smooth function, we can define a penalized likelihood of the model in Equation (4.17):

$$l_p(\boldsymbol{\gamma}) = l(\boldsymbol{\gamma}) - \frac{1}{2}\boldsymbol{\gamma}^T(\lambda_1\tilde{\boldsymbol{\Omega}}_1 + \lambda_2\tilde{\boldsymbol{\Omega}}_2)\boldsymbol{\gamma}, \tag{4.20}$$

where λ_1 and λ_2 are smoothing parameters, regularizing the influence of the penalty: the larger the λ, the smoother the surface. Given values of λ_1 and λ_2, $\boldsymbol{\gamma}$ can be estimated by maximizing l_p using numerical methods. Much like what we saw in the univariate GLM case, the basis coefficients $\boldsymbol{\gamma}$ can be estimated by maximizing the penalized likelihood in Equation (4.20) using Newton-Raphson algorithm:

$$\begin{aligned}
\boldsymbol{\gamma}^{new} = \quad & \boldsymbol{\gamma}^{old} - (\frac{\partial^2 l_p}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\boldsymbol{T}}})^{-1}\frac{\partial l_p}{\partial\boldsymbol{\gamma}} \\
= \quad & \boldsymbol{\gamma}^{old} + (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X} + \lambda_1\tilde{\boldsymbol{\Omega}}_1 + \lambda_2\tilde{\boldsymbol{\Omega}}_2)^{-1}(\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{y}-\boldsymbol{p}) - (\lambda_1\tilde{\boldsymbol{\Omega}}_1 + \lambda_2\tilde{\boldsymbol{\Omega}}_2)\boldsymbol{\gamma}^{old}).
\end{aligned}$$

To determine λ_1 and λ_2, GCV score, Mallow's C_p, BIC or AIC can be used as criteria. Again, we chose AIC in our work:

$$AIC = -2 \times l(\hat{\boldsymbol{\gamma}}) + 2 \times df(\lambda_1, \lambda_2),$$

where $df(\lambda_1, \lambda_2)$ is the effective degrees of freedom approximated by

$$df(\lambda_1, \lambda_2) = tr[(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X} + \lambda_1\tilde{\boldsymbol{\Omega}}_1 + \lambda_2\tilde{\boldsymbol{\Omega}}_2)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X}],$$

where $\boldsymbol{W}$ is the diagonal matrix with the diagonal elements $p_i(1-p_i)$ for $i = 1, \ldots, n$. Using the values of $\boldsymbol{W}$ and $\hat{\boldsymbol{\gamma}}$ from the final Newton-Raphson step, we can estimate $df(\lambda_1, \lambda_2)$ and AIC, thus the optimal λ_1 and λ_2 correspond to the model with the minimum AIC value. In addition, the covariance matrix of $\boldsymbol{\gamma}$ can be estimated from the final iteration:

$$\begin{aligned}
\text{cov}(\hat{\boldsymbol{\gamma}}) = \quad & (\frac{\partial^2 l_p}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\boldsymbol{T}}})^{-1} I(\boldsymbol{\gamma}) (\frac{\partial^2 l_p}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\boldsymbol{T}}})^{-1}\hat{\phi} \\
= \quad & (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X} + \lambda_1\tilde{\boldsymbol{\Omega}}_1 + \lambda_2\tilde{\boldsymbol{\Omega}}_2)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X} + \lambda_1\tilde{\boldsymbol{\Omega}}_1 + \lambda_2\tilde{\boldsymbol{\Omega}}_2)^{-1}\hat{\phi},
\end{aligned} \tag{4.21}$$

where $I(\hat{\boldsymbol{\gamma}})$ denotes the information in $\boldsymbol{y}$ and $\hat{\phi}$ is the estimate of dispersion parameter.

In summary, the recipe we suggest for bivariate penalized logistic regression is as follows:

1. Choose a spline basis and place a series of knots on a rectangular grid on the plane of two covariates;

2. Generate a tensor product bases from two marginal bases in order to obtain a 'combined' model matrix;

3. Vary (λ_1, λ_2) over an exponential grid, and for each pair of (λ_1, λ_2), use the Newton-Raphson algorithm to estimate $\hat{\boldsymbol{\gamma}}$;

4. Compute AIC and estimate $\hat{\phi}$, and by grid search find the optimal λ_1 and λ_2 to minimize AIC;

5. The $\hat{\boldsymbol{\gamma}}$ and $\hat{\phi}$ corresponding to the optimal λ_1 and λ_2 are obtained, and $\text{cov}(\hat{\boldsymbol{\gamma}})$ can be estimated as well.

There are alternative ways to impose penalty proposed in literature, for example, using a single penalty of the form [41]:

$$\boldsymbol{\gamma}^T \lambda \boldsymbol{\Omega} \boldsymbol{\gamma} = \boldsymbol{\gamma}^T \lambda (\boldsymbol{\Omega}_1 \otimes \boldsymbol{\Omega}_2) \boldsymbol{\gamma},$$

It is essentially akin to the formation of tensor product spline bases directly using a Kronecker product of two marginal bases. Although the single penalty is computationally efficient and easy to implement, it often results in undersmoothing because $\boldsymbol{\Omega}$ is a Kronecker product of $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$, suggesting that the rank of $\boldsymbol{\Omega}$ being the product of ranks of $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$. The rank of $\boldsymbol{\Omega}$ is always too low and the resulting model might be too complex. For instance, we want to construct a smooth of 2 covariates as a tensor product of 2 spline bases, each of rank 7. Since the tensor product bases are obtained by a Kronecker product of two marginal bases, the resulting smooth would have $7 \times 7 = 49$ free parameters. For the penalty matrix $\boldsymbol{\Omega}$, its rank equals the product of the ranks of $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$, i.e., $5 \times 5 = 25$. The effective degrees of freedom for the smooth would then vary between 24 and 49, which indicates a too complex model. Nevertheless, when we use two penalties imposed on $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ respectively, the penalty matrix would have rank 45. The effective degrees of freedom would then vary between 4 and 49 that greatly decreases the complexity of the resulting model. In particular, for high-dimensional tensor product of spline bases, the multiple term penalties would give much more useful range of effective degrees of freedom compared to the single penalty.

4.3 Meta Analysis with respect to GAM

Although the GAM has widespread applications in nonparametric studies, its application in meta analysis remains a new topic that has not been studied in the literature. As mentioned previously, a challenging problem is how to perform meta-analysis with respect to curves.

4.3.1 One-dimensional smoothing on age

Proposed method

For one-dimensional smoothing on age, our proposed method is as follows:

1. Set spline bases for age within a reasonable range by choosing a spline and q knots on age;

2. Fit a GAM for each study in order to capture the possibly nonlinear trends of the risk effects; that is,

$$\text{CHD death} \sim s(age)$$

3. When a series of the GAMs are fitted, every set of the basis coefficients can be regarded as multiple outcomes to conduct multivariate meta analysis with *multiple outcomes*;

4. Finally, using the estimated basis coefficients from meta analysis, the effect curve varying with age can be obtained to show the overall trend.

Therefore, in addition to the fact that the linearity can be easily captured with flexible modeling of predictor effects, the underlying nonlinearity can be detected in order to provide more information on how the effects of risk factors on CHD death change with age.

The key point in the proposed method is to ensure common age basis functions for all studies. It can be achieved by choosing the same spline with the same knots on age simultaneously. When a GAM is fitted, m basis functions can be represented as the $n \times m$ model matrix $\boldsymbol{X}$ in Equation (4.4), which maps the parameters associated with such a smooth to the predicted values of the smooth at a set of n covariate values. Here, we define $\boldsymbol{S}$ as the basis functions. Using the same spline and the same set of knots, $\boldsymbol{S}$ should be the same for all studies but basis coefficients $\boldsymbol{\beta}$ are different. Once we have the comparable coefficients $\boldsymbol{\beta}$ independent of the data, we can regard these coefficients as multiple outcomes to conduct

multivariate meta analysis. Using the overall basis coefficients from meta analysis, call $\hat{\boldsymbol{\beta}}_{meta}$, and the common basis functions $\boldsymbol{S}$, the effect curves varying with age can be easily plotted.

Implementation difficulty with `mgcv`

It is well known that `mgcv` is a professional and popular spline package in R [41]. Yet there are difficulties in the implementation of meta analysis with respect to curves. Recall that in practice the first step is to form QR decomposition of the model matrix $\boldsymbol{X}$ in order to improve computational efficiency. In this way, the common basis function $\boldsymbol{S}$ is reparameterized to $\boldsymbol{SQ}$. Furthermore, there is an important fact that the first column in model matrix $\boldsymbol{S}$ to which QR is applied is 1's and thus it is absorbed into the intercept term in GAM. In other words, the `mgcv` package of R provides the data-dependent $\boldsymbol{S}_x$, in which the first column of $\boldsymbol{SQ}$ is discarded.

$$\boldsymbol{S}_x = \boldsymbol{SQ}[,-1] = \begin{bmatrix} s_2^*(x_1) & \cdot\cdot & s_m^*(x_1) \\ s_2^*(x_2) & \cdot\cdot & s_m^*(x_2) \\ s_2^*(x_3) & \cdot\cdot & s_m^*(x_3) \\ \cdot & \cdot\cdot & \cdot \\ \cdot & \cdot\cdot & \cdot \\ s_2^*(x_k) & \cdot\cdot & s_m^*(x_k) \end{bmatrix} = \begin{bmatrix} \boldsymbol{s}^*(x_1)^T \\ \boldsymbol{s}^*(x_2)^T \\ \boldsymbol{s}^*(x_3)^T \\ \cdot \\ \cdot \\ \boldsymbol{s}^*(x_k)^T \end{bmatrix}$$

Especially, when there are more than two covariates in GAM, the first columns of the basis functions correpsonding to those covariates are all absorbed into the global intercept term in GAM. Considering further meta analysis, we want to obtain the common basis functions $\boldsymbol{S}$ rather than $\boldsymbol{S}_x$ from `mgcv`. We thus need to find the transformation matrix that relates common basis functions $\boldsymbol{S}$ and data-dependent basis functions $\boldsymbol{S}_x$ from `mgcv`. Nevertheless, discarding the first column of $\boldsymbol{SQ}$ makes it impossible to achieve $\boldsymbol{S}$ by $\boldsymbol{S}_x$.

Moreover, what we still need are the coefficients for the knot-based nontransformed design. However, the fact that the first column of $\boldsymbol{SQ}$ is absorbed into the intercept term in GAM makes the first parameter in basis coefficients after reparameterization also discarded. Define the resulted basis coefficients after reparameterization output from GAM as $\boldsymbol{\beta}_x$. Note that similar to $\boldsymbol{S}_x$, $\boldsymbol{\beta}_x$ is the vector of data-dependent basis coefficients without the intercept from `mgcv`. Thus, discarding the first column of $\boldsymbol{SQ}$ also makes it impossible to obtain $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_x$.

• **How to get the effect curves varying with age**

Regarding the above difficulties on the practical implementation using `mgcv` package of R, we implement our own algorithms for further meta analysis. For one-dimensional smoothing on age, we propose to choose the same spline with an exact bases form and the same knots on age for all the studies, and then follow the recipe of univariate penalized logistic regression to obtain comparable knot-based basis coefficients. In this way, every set of coefficients can be regarded as multiple outcomes for multivariate meta analysis. Once we have the overall estimates and their covariance matrix of the basis coefficients from meta analysis, it is straightforward to plot the effect curve varying with age and the 95% confidence interval curves. The overall trend of the risk effect on CHD death varying with age can be demonstrated from the curves.

4.3.2 Two-dimensional smooth surface

In order to investigate interaction and nonlinear effects of age and other risk factors, a natural approach is to build a two-dimensional smooth model. We can choose a spline for two covariates and place a series of knots on the covariates for all the studies; then according to the recipe of bivariate penalized logistic regression, knot-based basis coefficients can be obtained. What is critical here is to construct a tensor product basis representing smooth functions of two covariates. Thus, having obtained the basis coefficients and the tensor product basis, it is straightforward to estimate a two-dimensional smooth surface on the domain of two covariates.

Similar to the spirit of meta analysis on one-dimensional smoothing of age, for two-dimensional smoothing surface, we can also choose the same splines and the same knots location for all the studies. By constructing tensor product bases, they also have common basis functions. Every set of tensor product bases coefficients from bivariate penalized logistic regression can be regarded as multiple outcomes for multivariate meta analysis. In the end, given the coefficients from meta analysis and the tensor product bases, an overall two-dimensional smooth surface can be easily obtained.

4.4 Data Analysis

The previous sections stated how the problem of estimating a GAM becomes the problem of estimating basis coefficients once a basis for the smooth functions has been chosen, together with our proposed method of performing meta analysis with respect to curve. The purpose of this section is to apply our proposed method into analysis, and to illustrate its feasibility. The section starts by giving examples of one-dimensional smoothing on age, including the interaction of age and three SBP levels. There follows data analysis on two-dimensional smooth surface on age and SBP to investigate their interaction in affecting CHD death. We close with a discussion on our proposed methods.

4.4.1 One-dimensional smoothing on Age

We first examine the age effect on CHD death among three SBP groups, high SBP ($\geq$ 160mmHg), medium SBP ($[140, 160)$mmHg), and low SBP ($<$ 140mmHg). In order to illustrate the algorithm of univariate penalized logistic regression, we first use an individual dataset, then move onto all studies to perform meta analysis.

For Study 3 (the Charleston Heart Study), the age variable ranges from 35 to 97. Considering further meta analysis, we only use a subset of data with age ranged from 35 to 75 to ensure most studies have similar age ranges. Thus, the male subjects aged 35 to 75 are used to fit penalized logistic regression in this section.

The foremost job is to choose a type of spline and set a series of knots within the age range [35,75]. We use the spline basis given in Equation (4.6) because it is easy to implementation, suffices in the real world, and is knot-based so as to facilitate us to obtain a common basis function when the same knots are placed for all the studies. We perform a scaling step on age, scaled age = (age-35)/(75-35), which ensures the scaled age ranged between 0 and 1. Since the basis dimension depends on the number of knots we choose, we choose 9 interior knots on age, $0.1 - 0.9$ by 0.1, and the resulting basis dimension is 11. Once the spline is chosen and the knots selection is done, the model matrix $\boldsymbol{X}$ and the penalty matrix $\boldsymbol{\Omega}$ are treated as known.

The optimal tuning parameter λ with AIC criterion in penalized logistic regression. Figures 4.1, 4.2 and 4.3 show the change of AIC values with λ among SBP groups. In the low SBP group, we can see that the AIC value tends to decrease sharply for $\lambda < 0.5$,

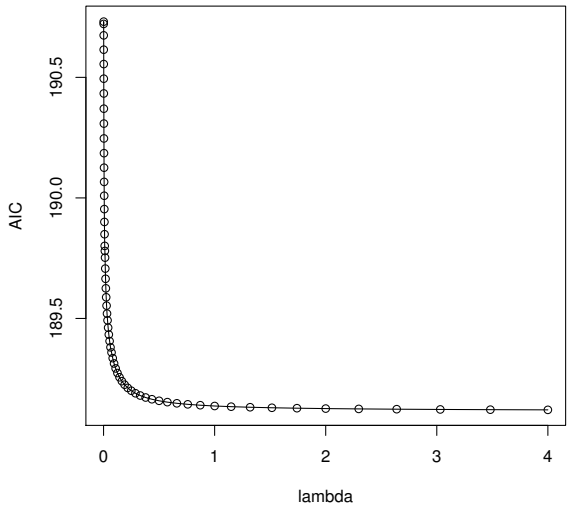

Figure 4.1: AIC vs. λ in low SBP group

but its change gets flat when λ is large. This indicates that the fitted curve in the low SBP group would be nearly linear because the optimal λ tends to be infinite. The high SBP group shows the same result as the low SBP group. Nevertheless, for the medium SBP group, the optimal λ is around $2^{-4.6}$. It can also be verified by the fitted curves as shown in Figure 4.4. Log-odds of CHD death linearly increases with age in the low and high SBP groups, but the medium SBP group show clearly nonlinear. Besides, the estimated dispersion parameters, $\hat{\phi}$, are 0.99, 1.02 and 1.01, very close to 1, for the three groups respectively, suggesting no dispersion effects on our data.

Another example is from Study 7, the Framingham Cohort Study. Figure 4.5 presents the fitted curves of log-odds of CHD death vs. age among the three groups in Study 7. It can be seen that low and medium SBP groups show nonlinear relation while in high SBP group log-odds of CHD death linearly increases with age. From the two examples above, we can see that different studies show distinct results, thus meta analysis would be a good approach to synthesize the results so as to obtain an overall effect curve.

When we repeat the procedure using the same spline with the same knots for other studies, we can obtain a set of knots-based coefficients for each study, which can be regarded as multiple outcomes to conduct multivariate meta analysis. Considering the different age

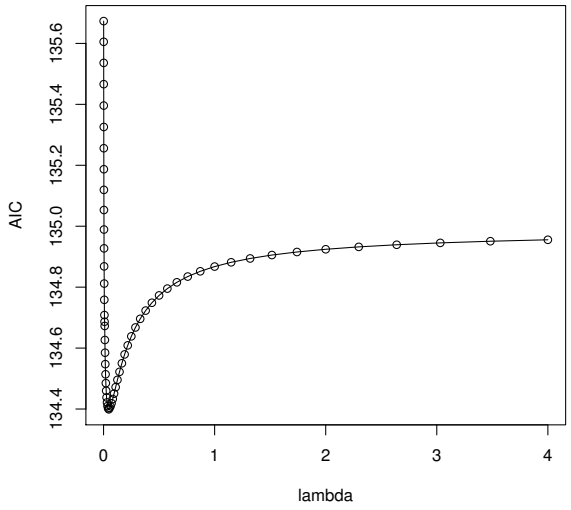

Figure 4.2: AIC vs. λ in medium SBP group

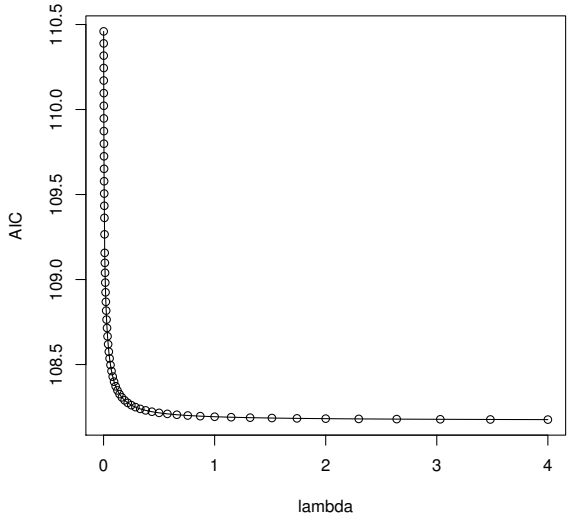

Figure 4.3: AIC vs. λ in high SBP group

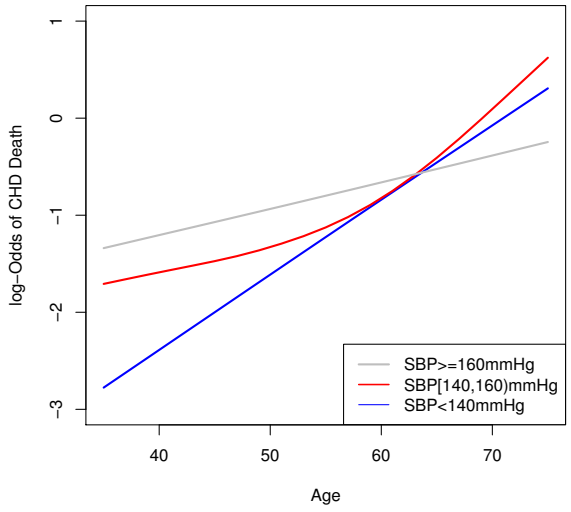

Figure 4.4: Log-odds of CHD death on age among three SBP groups in Study 3

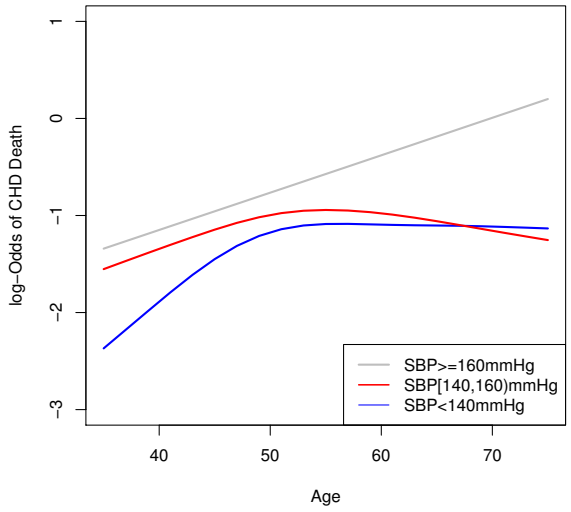

Figure 4.5: Log-odds of CHD death on age among three SBP groups in Study 7

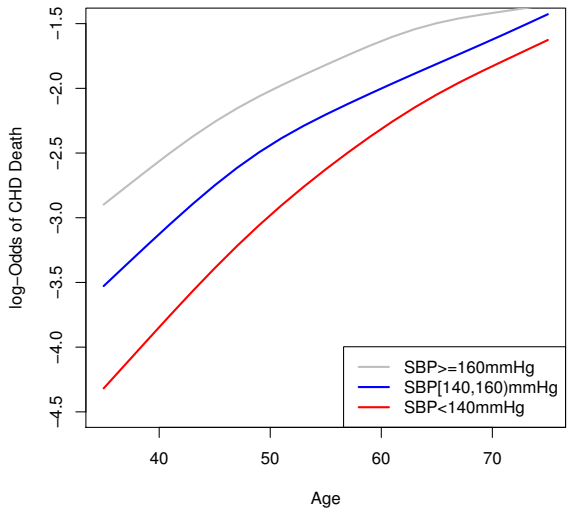

Figure 4.6: Log-odds of CHD death on age for three SBP groups from meta analysis

range of the 27 studies, we only select subjects at aged between 35 and 75 years old. Figure 4.6 shows the fitted curves of log-odds of CHD death vs. age for three SBP levels. We can see that log-odds of CHD death increases with age among three SBP groups, and also the higher the SBP levels, the larger the log-odds of CHD death associated with age. When we compare the slopes of the three curves, the age effect on CHD death in the high SBP group appears weaker than in other two groups.

4.4.2 Two-dimensional smooth surface

Another strategy to examine interaction of age and SBP is to introduce a smooth surface on the two-dimensional age-SBP domain. Similar to the one-dimensional smoothing analysis on age in Section 4.4.1, subjects aged 35 to 75 and with SBP between 100 and 200mmHg are selected for our analysis. We again choose the spline basis given in Equation (4.6) for both age and SBP and place a certain number of knots on age and SBP. Before knots location, a scaling step on covariates is required, that is, scaled age = (age-35)/(75-35), and scaled SBP = (SBP-100)/(200-100), which ensure the scaled covariates ranged between 0 and 1. The knots on age and SBP are both 0.1-0.9 by 0.2, which result in 49 unknown basis coefficients to be estimated.

In the algorithm of bivariate penalized logistic regression, given the marginal bases, we can generate a two-dimensional tensor product spline by the method presented in Equation (4.18). The resulting tensor product spline exists in the age-SBP plane, and the knots are placed on a grid of age and SBP, carving the plane into subrectangles. The unknown tensor product spline coefficients $\boldsymbol{\gamma}$ can be estimated by maximizing likelihood. To regularize coefficients sestimation, appropriate penalties can be put on $\boldsymbol{\gamma}$ to ensure smoothness in the surface. We here use two methods mentioned in Section 4.2.5 to penalize estimation of the surface: (1) single penalty; (2) double penalties on the two axes of age and SBP.

We first try the simpler penalty. It needs to generate a 'combined' penalty matrix $\boldsymbol{\Omega}$ of the following form:

$$\boldsymbol{\Omega} = \boldsymbol{\Omega}_1 \otimes \boldsymbol{\Omega}_2,$$

and the fitting problem amounts to maximizing the penalized likelihood:

$$l_p(\boldsymbol{\gamma}) = l(\boldsymbol{\gamma}) - \frac{1}{2}\lambda \boldsymbol{\gamma}^{\boldsymbol{T}} \boldsymbol{\Omega} \boldsymbol{\gamma}.$$

Given a value of λ, the Newton-Raphson algorithm yields the estimate of $\boldsymbol{\gamma}$. Similarly, by specifying a series of λ's, say 2^{-10} to 2^2 with an increment of $2^{0.1}$, we can use AIC criteria to find the optimal λ. For Study 3, the optimal λ is $2^{-3.3}$. When we look at the two-dimensional smooth surface in Figure 4.7, we see the surface is clearly nonlinear. The reason is that the simple penalty would often lead to severe undersmoothing. As we stated in Section 4.2.5, the rank of the penalty matrix is the product of the ranks of the marginal penalty matrices, which is always too low to be useful. Due to the low rank, the number of free parameters is still large, thus making the model too complex.

Alternatively, we can use two penalties on two smooths, which provide more flexibility in regularization. It also allows us to follow the recipe of bivariate penalized logistic regression shown in Section 4.2.5. An important task is also to get optimal λ_1 and λ_2 corresponding to smallest AIC value. Given a grid of λ_1 and λ_2, the two-dimensional grid search yields a minimum AIC and the corresponding λ_1 and λ_2 are optimal. For example, for Study 3, both λ_1 and λ_2 vary within 2^{-10} to 2^2 with an increment of $2^{0.5}$. The resulting optimal λ_1 and λ_2 are 4 and $2^{-8.5}$. Interestingly, in this particular study, the optimal value λ_1 is always the upper bound of the grid we specified for λ_1. This indicates that the relatively large penalty on the coefficients along age leads to essentially a linear trend. We think it is reasonable to adopt $[2^{-10}, 2^2]$ as a proper range for λ_1 because although the optimal λ_1 can be arbitrarily

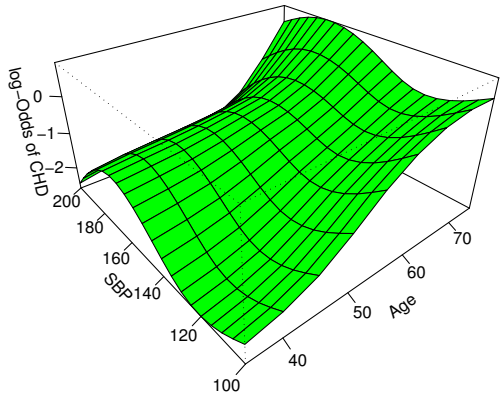

Figure 4.7: Two-dimensional smooth surface using simple penalty in Study 3

large, the decrease in AIC is mild. Meanwhile, from Figure 4.8 presenting two-dimensional smooth surface using two penalties in Study 3, we can see the method of double penalties imposed on two smooths yields a smoother surface and a less complex model as compared to the simple penalty.

Further, Figures 4.9, 4.10 and 4.11 show the optimal estimated coefficient surfaces for males in Studies 5, 7 and 16. Although the studies all display pronounced interaction between age and SBP as shown above, they suggest different results in the interaction due to distinct estimated surfaces. Meta analysis is a useful tool to integrate the results from different studies. Thus, we rely on the algorithm of bivariate penalized logistic regression and use the same spline with the same knots on covariates for all the studies to fit a series of penalized logistic models. Then, treat the tensor product bases coefficients as multiple outcomes to conduct multivariate meta analysis. Figure 4.12 presents two-dimensional smooth surface in males from meta analysis. The surface shows that log-odds of CHD death increases as age increases and the same thing for SBP. In some sense, the two-dimensional surface can visually reflect the interaction of age and SBP. Nevertheless, another way is to analyze their interaction by aid of projection plots. Specifically, one can cut through the surface to get effect curves at specific ages or SBPs. Figures 4.13 and 4.14 show projection plots

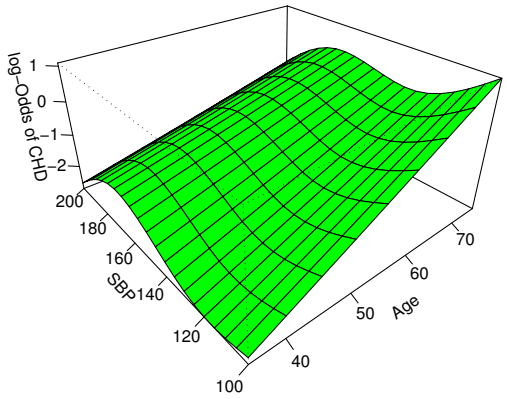

Figure 4.8: Two-dimensional smooth surface using two penalties in Study 3

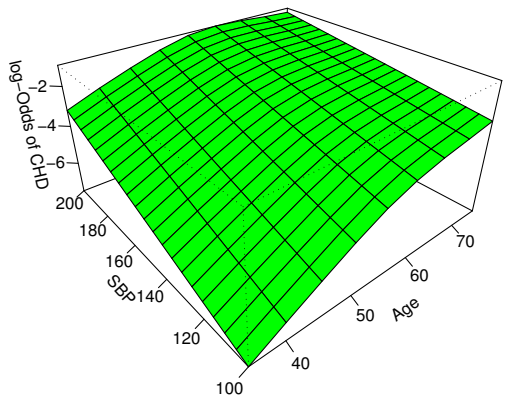

Figure 4.9: Two-dimensional smooth surface using two penalties in Study 5

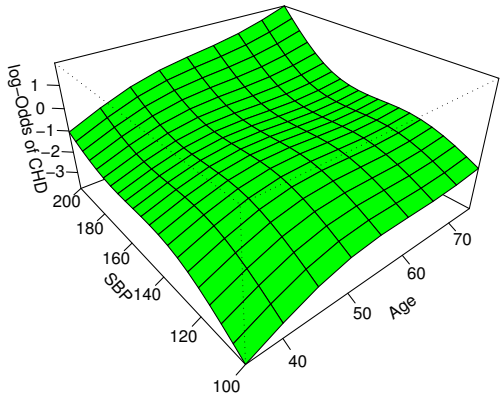

Figure 4.10: Two-dimensional smooth surface using two penalties in Study 7

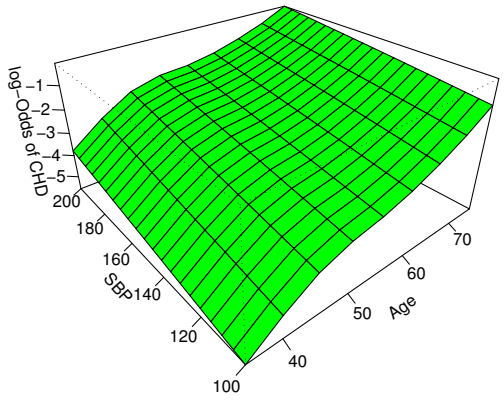

Figure 4.11: Two-dimensional smooth surface using two penalties in Study 16

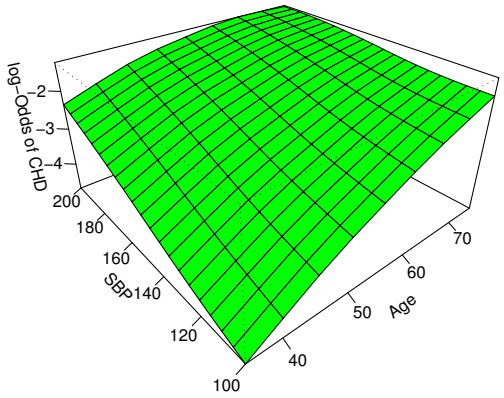

Figure 4.12: Two-dimensional smooth surface using two penalties from meta analysis in males

corresponding to the axes of age and SBP, respectively. From Figure 4.13, we can see that log-odds of CHD death increases with age for all the specific SBP values while for men having higher SBP value, log-odds of CHD death increase with age at a lower rate. When looking at the curves of SBP effect on CHD death at specific ages in males shown in Figure 4.14, we can see that the SBP effects at different ages display stronger linearity compared to the age effects in Figure 4.13. Compared to young people, the log-odds of CHD death increase much slowly with the elevation of SBP among old people over 65 years old, suggesting that SBP has less impact on CHD death in the elderly. According to the analysis so far, the interaction of age and SBP on CHD death in males group can be clearly seen. It is worth mentioning that the confidence intervals of the curves were not shown in Figures 4.13 and 4.14 because they heavily overlap each other that may be resulted from the relatively large covariance estimates.

However, we can still argue the differences of the effects in CHD death among different SBP groups are significant enough in most cases. Figure 4.15 presents how logORs of CHD death associated with SBP change with age including 95% confidence intervals. From the plots, we can see that logORs of CHD death associated with 20 mmHg increase in SBP

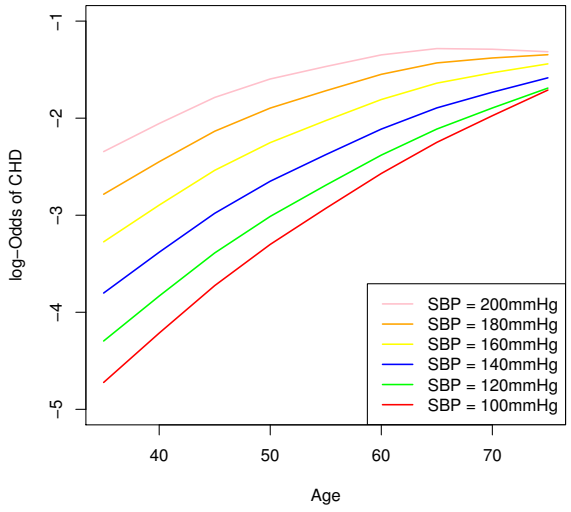

Figure 4.13: Curves of age effect on CHD death at specific SBP values in males

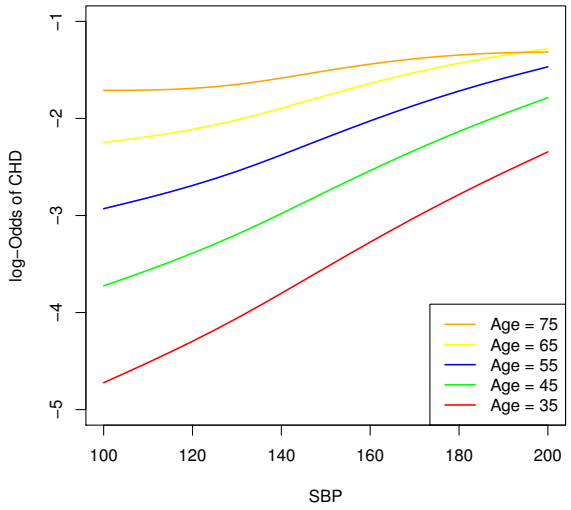

Figure 4.14: Curves of SBP effect on CHD death at specific ages in males

decline with age. Moreover, the 95% confidence intervals indicate that for males having SBP of 140 vs. 120mmHg, 160 vs. 140mmHg, and 180 vs. 160mmHg, the association between CHD death and SBP is significant at least before 70 years old. Besides, from Figure 4.15, the difference of logOdds in CHD death for males having SBP of 120 vs. 100mmHg is only significant among young men below 50 years old. It might be due to a sample size of the male subjects having SBP less than 120mmHg. Therefore, we can see that the results from our proposed method, for the most part, is consistent to the results based on classical meta analysis with respect to generalized linear models as shown in Figure 3.6.

We may also follow the same procedure as shown above in males to get individual two-dimensional estimated surfaces for all the females cohorts as well as a 'synthesized' two-dimensional surface for females from meta analysis. Figure 4.16 presents two-dimensional smooth surface in females from meta analysis, and Figures 4.17 and 4.18 show projection plots corresponding to the axes of age and SBP, respectively. From the curves of age effect on CHD death at specific SBP values in females shown in figures 4.17, we can see a similar trend of age effects on CHD death as shown in males. Yet, it seems that males have stronger linearity than females. Meanwhile, Figure 4.18 shows quite similar curves to Figure 4.14. Therefore, the interaction of age and SBP on CHD death exists in both males and females groups. Moreover, Figure 4.19 presents how logORs of CHD death associated with the 20mmHg increase in SBP change with age for females. We can see that among females the logORs of CHD death associated with a 20 mmHg increase in SBP also decline with age, and the association between CHD death and SBP for females is significant at least before 70 years old. It is also consistent to the result shown in Figure 3.6.

4.4.3 Discussion

We first applied our proposed meta analysis method of one-dimensional smooth curve on age to analyze the interaction of age and SBP on CHD death amongst the studies when subjects were divided into three groups, low SBP, medium SBP, and high SBP. Based on the curves from meta analysis, we can not only look at the age effects within each of the SBP groups, but compare the relative strength of the age effects from the slopes of the curves. We also achieved the estimation of a two-dimensional smooth surface on the age-SBP domain by aid of our proposed meta analysis method of two-dimensional smooth surfaces. To get a smooth surface, a two-dimensional tensor product spline needed to be generated by the Kronecker

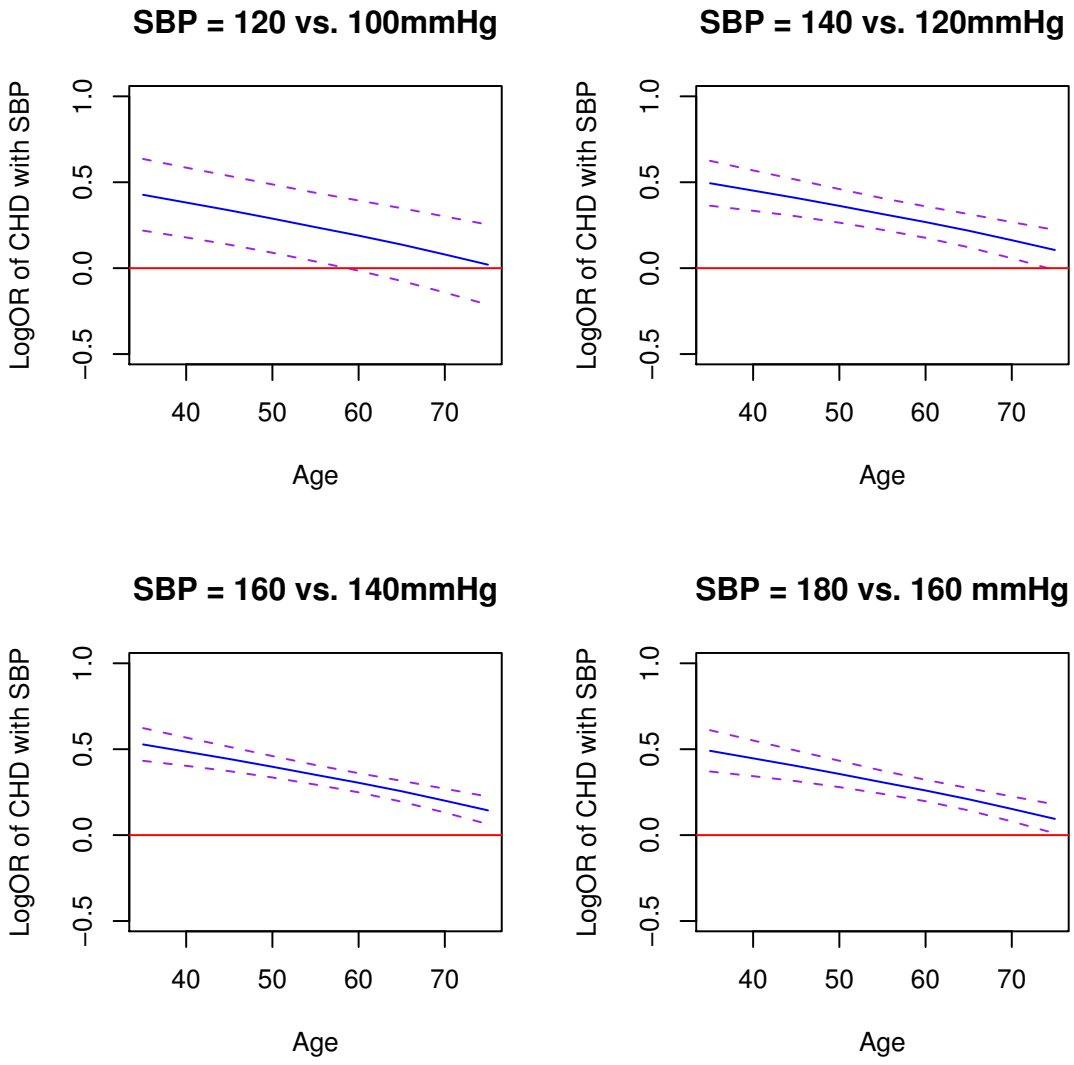

Figure 4.15: LogORs of CHD death with SBP vs. age for males

product of marginal bases, thereby all the studies had common basis functions. By virtue of fitting procedure and regularization estimation, the knot-based unknown tensor product spline coefficients can be obtained for use in further multivariate meta analysis.

The foremost job for fitting procedure is the choice of spline as well as knots placement. It should be aware that the spline is not limited to the one we used, and other types of knots-based splines can also be used, such as B-splines. Also, in light of computational cost, we chose 7 knots on both age and SBP, resulting in 49 tensor product bases coefficients.

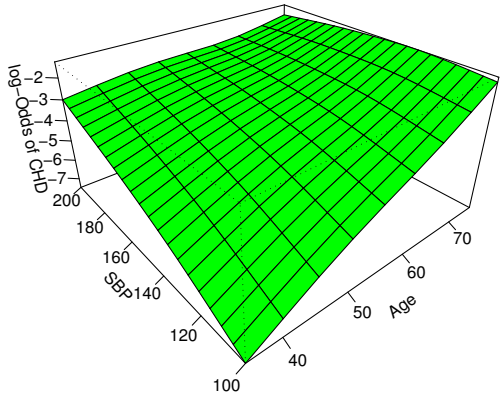

Figure 4.16: Two-dimensional smooth surface using two penalties from meta analysis in females

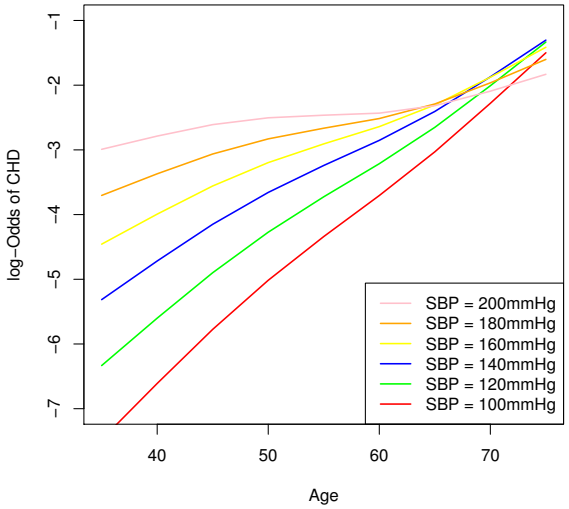

Figure 4.17: Curves of age effect on CHD death at specific SBP values in females

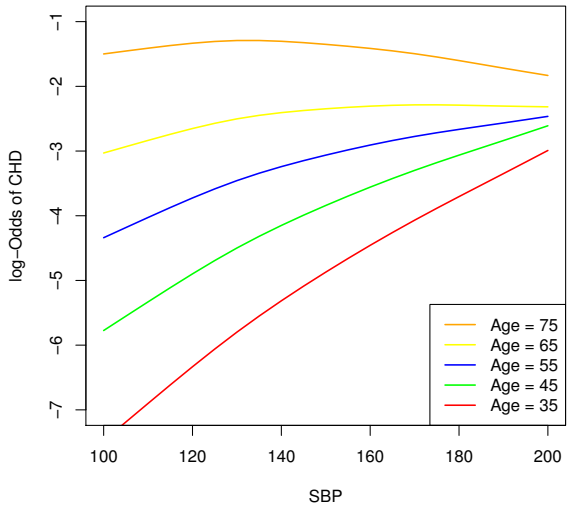

Figure 4.18: Curves of SBP effect on CHD death at specific ages in females

Under other situations, one can still carry out our procedure by placing a generous number of knots and setting knots location in a different manner, followed by regularizing the coefficients estimation with optimal tuning smoothing parameters to suppress the smooth term complexity. For estimation regularization, we do not recommend the simpler penalty due to its main drawback that the penalty matrix always has too low rank, leading to a too complex model as well as severe undersmoothing. By contrast, imposing two penalties on two smooths of age and SBP can increase the rank of penalty matrix, yielding a wider and more useful range of the effective degree of freedoms that makes the surface smoother. Meanwhile, in our work we achieved an automatic smoothing parameter selection procedure by AIC criterion to determine one or two optimal tuning parameters. One can alternatively select other reasonable optimization criteria, such as GCV score, Mallow's C_p, and BIC. Furthermore, after obtaining unknown tensor product spline coefficients through fitting procedure and regularization estimation, we can use them to perform multivariate meta analysis using our GLS algorithm. Using the common basis functions and the 'synthesized' knot-based basis coefficients from meta analysis, a two-dimensional smooth surface on the age-SBP domain was estimated. By cutting through the smooth surface along two axes, the resulting slices show how the risk effect on CHD death changes at an arbitrary age as well

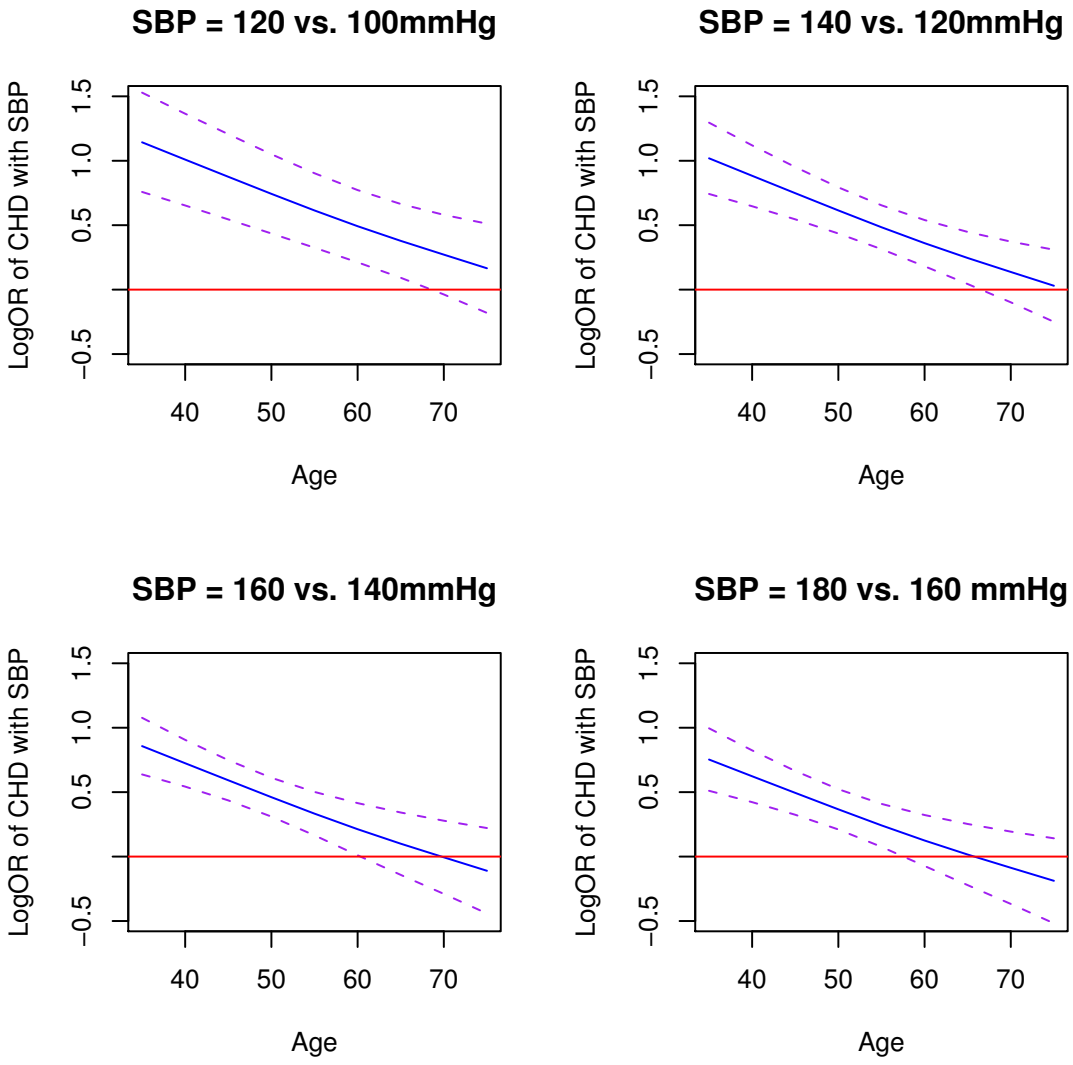

Figure 4.19: LogORs of CHD death with SBP vs. age for females

as how the age effect on CHD death changes at an arbitrary SBP value. Moreover, it was shown that how the logORs of CHD death for people having two different SBP values change with age. Therefore, our proposed methods provide more detailed and useful information about how the SBP effect on CHD death changes with age.

Our proposed methods still have some limitations. First, the issue that is worth mentioning is the choice of the range of the tuning smoothing parameter. We found that when the fitted curve is nearly linear, the optimal smooth parameter tends to be large, always

attaining the maximum value we specify for the smoothing parameter. This originates from the essential feature of smooth parameters, as stated in last section. In our examples, it has been also noticed that although the optimal smoothing parameter may be very large, AIC decreases mildly after some value of the smoothing parameter. Our solution was to specify a reasonably large range for the smooth parameter based on AIC values. Yet, the shape of the estimated function or surface is probably affected by how well we are able to tune the smoothing parameters. This demonstrates the importance of more analyses on the choices of optimization criteria and smoothing parameters. Second, as to the estimation regularization in our algorithm, the construction of penalty matrices in two-dimensional smooth on age and SBP is just achieved by simple transformation shown in Equation (4.19). To better tune the degree of smoothing for smooth terms, one may need to apply reparametrization to marginal smoothes [41]. Third, it is also noted that multivariate meta analysis was conducted by our GLS approach in our application. When negative variances in the covariance matrix exist at some iterations, the GLS algorithm involves a spectral decomposition with respect to covariance matrix as well as forcing the negative eigen values at zero to ensure a positive semi-definite covariance matrix. Considering that the results we obtained may be affected by the drawback of the performance of the GLS approach as mentioned earlier, a comparative simulation study on those algorithms of multivariate meta analysis, such as GLS, Maximum likelihood, and the Method of moments, is needed. In addition, although our proposed meta analysis method of one-dimensional smooth curve on age performs well for analyzing the age effects of CHD death among two or more groups, the results are likely to be sensitive to the choice of bins, such as bin size and endpoints of groups. Therefore, it is more suitable for naturally categorized groups, such as males/females, smokers/nonsmokers, and having diabetes/not having diabetes, etc.

Overall, our proposed methods not only can be easy to implementation in practice, but have many nice features. For instance, the marriage of penalized regression splines and multivariate meta analysis allows us to gain more clinically meaningful results about the nonlinear relationship between CHD dealth and SBP changing with age, supported by our multiple studies. The methods are certainly not limited to the age-SBP interaction, and can be used for other age-risk interaction, such as age-CHOL and age-BMI. The analysis on age-risk interaction after adjustment of other risk factors is also tractable by creating a model matrix including the bases related to all the covariates. Rather than two dimensions,

one can use such methods to extend to higher dimensions that smoothly account for other covariates, thereby allowing us to investigate more complex interactions among more than two covariates. Although our proposed methods of meta analysis of curves is grounded on penalized logistic regression, the methods can also be applied into penalized proportional hazard model. Another attractive feature of the proposed approach is that the smoothness can be well controlled by imposing penalties on smoothes even when a generous number of knots are placed. In a nutshell, the proposed methods provide a new method, meta analysis with respect to curves to get a 'synthesized' smooth curve or a 'synthesized' smooth surface that can be a great advantage over conventional meta analysis based on linear models.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

Meta analysis is a statistical method that is widely used for synthesis of results across related but independent studies for the purpose of examining the sources of variability between study results and to derive summary measures. In this thesis, we aimed to examine the interaction between age and systolic blood pressure (SBP) using data from multiple studies containing differing age ranges by means of meta analysis.

We first fit a logistic model with CHD death as response and age, SBP and their interaction as covariates for each of the studies, and conducted multivariate meta analysis on every set of three coefficients to obtain 'synthesized' coefficients. The analysis of the coefficients from meta analysis indicated a negative interaction of age and SBP, in other words, SBP has less impact on CHD death in older people than in younger people. Based on our data, it was found that the association of CHD death with SBP is significant for men younger than 75 years and for women younger than 70 years.

We also went beyond the linear models because there is no reason to believe that the risk effect on CHD death is perfectly linear in age. We presented a new approach by performing meta analysis with respect to curves. The basic idea for one-dimensional smoothing on age is that by choosing the same spline with the same knots on age for all the studies, they have common basis functions and the knot-based basis coefficients obtained from univariate penalized logistic regression can be used for further multivariate meta analysis. The age effect curve on CHD death can be easily plotted using common basis functions and the 'synthesized' knot-based basis coefficients from meta analysis. One-dimensional smoothing on age is well suitable for interaction of age and a dichotomous variable on CHD death. Meta analysis of two-dimensional tensor product smooth surface was also proposed to explore the interaction

of age and SBP. From the projection curves of the smooth surface, we can analyze the SBP effects along age or the age effects along SBP. Based on our data, a negative interaction of age and SBP on CHD death was found, which agrees with the result from logistic models. Also, the change of the log-odds of CHD death demonstrates noticeable nonlinearity, that is, the log-odds of CHD death increase faster in younger men than in older men, especially for men having higher SBP. Meanwhile, men have stronger linearity of the age effects on CHD death than women, whereas the SBP effects on CHD death look more linear in both men and women. Furthermore, for males having SBP above 120mmHg, the logORs in CHD death associated with SBP significantly decline with age at least before 70 years old, whereas among females, the logORs of CHD death associated with SBP also decline with age, and the association between CHD death and SBP for females is significant at least before 70 years old. On this point, they, for the most part, are consistent to the results from classical meta analysis based on logistic regression models.

5.2 Future work

In this thesis, we mainly carried out multivariate meta analyses using our eigenvalues-truncated GLS algorithm. In some sense, the maximum likelihood approach is probably an alternative way, but it may cause expected problems, such as long computation time and non-convergence. Additionally, the method of moments recently proposed by D. Jackson *et al.* [27] is also an improved approach in performing random-effects multivariate meta analysis. Although it can dramatically reduce computation cost and avoid having the normality assumption, the accuracy of theoretically unbiased estimates still deserves to be studied, especially when the within-study covariance is nearly singular. Thus, a systematic simulation study on these approaches needs to be done to compare the accuracy of estimates, convergence issue and computational efficiency. Moreover, we would like to carry out other algorithms, such as a hierarchical Bayesian modeling approach using Markov Chain Monte Carlo (MCMC) techniques to conduct multivariate meta analysis [28].

On the other hand, there are still some issues that require further investigation in the proposed method of meta analysis with respect to curves. First, we would like to introduce regularization to estimate the covariance of basis coefficients in order to improve the standard errors of the coefficients. Second, the construction of the penalty matrices in two-dimensional

smooth of two covariates is also needed to be improved by reparameterization rather than the simple transformation we used in Equation (4.19). Third, in addition to AIC, other criteria, such as BIC and GCV, can also be reasonable in regularization. It suggests an importance of comparison in different criteria to tune nonnegative smoothing parameters. Fourth, our work will extend to more complex additive models by incorporating both linear and nonlinear terms, as well as study the interaction of age and other risk factors, such as CHOL and BMI. Finally, we would like to implement meta analysis based on penalized proportional hazard model so as to apply our method in survival data.

APPENDIX A

The Diverse Populations Collaboration

The Diverse Populations Collaboration (DPC) contains 27 studies from several countries and cultures and the participation in the collaboration has continued to grow since 1996. The full names of 27 studies are as follows:

1. **ARIC**: The Atherosclerosis Risk in Communities Study
2. **BIP**: The Bezafibrate Infarction Prevention
3. **Charleston Heart**: The Charleston Heart Study
4. **CHS**: The Cardiovascular Health Study
5. **Cordis**: The Cardiovascular Occupational Risk Factor Determination in Israel Study
6. **Evens County**: The Evans County Study
7. **Framingham Cohort**: The Framingham Heart Study
8. **Framingham Offspring**: The Framingham Offspring Study
9. **Glostrup**: The Glostrup Population Study
10. **GOH**: The Glucose Intolerance, Obesity, and Hypertension Study
11. **Guangzhou**: The Guangzhou Chinese Cohort
12. **HDFP**: The Hypertension Detection and Follow-up Program
13. **Honolulu**: The Honolulu Heart Program
14. **Iceland**: The Iceland Reykjavik Study
15. **Israel**: The Israeli Ischemic Heart Disease Study
16. **LRC**: The Lipid Research Clinics Prevalence Study
17. **LRC-CPPT**: The Lipid Research Clinics Primary Prevention Trial
18. **MRFIT**: The Multiple Risk Factor Intervention Trial
19. **NHANES I**: The First National Health and Nutrition Examination Survey Epidemio-

logic Follow-up Study
20. **NHANES II**: The Second National Health and Nutrition Examination Survey Mortality Follow-up Study
21. **NHIS**: The National Health Interview Survey
22. **Norway**: The Norwegian Countries Study
23. **Puerto Rico**: The Puerto Rico Study
24. **Renfrew-Paisley**: The Renfrew-Paisley Survey
25. **Scottish Collaborative**: The Scottish Collaborative Study
26. **Tecumseh**: The Tecumseh Community Health study
27. **Yugoslavia**: The Yugoslavia Cardiovascular Health Study

Data for the 27 studies were either obtained from public data sources or from the main investigators. The studies from public sources are shown below.

1. **The National Heart, Lung, and Blood Institute**: ARIC, CHS, Framingham Cohort, Framingham Offspring, Honolulu, LRC, LRC-CPPT, HDFP, MRFIT, and Puerto Rico.
2. **The National Center for Health Statistics (USA)**: NHANES I and NHANES II
3. **The Inter-university Consortium for Political and Social Research (ICPSR)**: Tecumseh.

REFERENCES

[1] W. Rosamond, K. Flegal, and G. Friday et. al. Heart disease and stroke statistics - 2007 update. *Circulation*, 115:e69–e171, 2007. 1

[2] W. B. Kannel, D. McGee, and T. Gordon. A general cardiovascular risk profile: The framingham study. *The American Journal of Cardiology*, 38:46–51, July 1976. 1

[3] K. M. Anderson, P. W. F. Wilson, and W. B. Kannel. Cardiovascular disease risk profiles. *American Heart Journal*, 121:293–8, 1990. 1

[4] K. M. Anderson, P. Odell P. Wilson, and W. B. Kannel. An updated cronary risk profile. *Circulation*, 83(1):356–62, 1991. 1

[5] P. W. F. Wilson, R. B. D'Agostino, D. Levy, A.M. Belanger, H. Silbershatz, and W.B. Kaannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97:1837–47, 1998. 1

[6] E. Anum and T. Adera. Hypercholesterolemia and coronary heart disease in the elderly: A meta-analysis. *Annals of Epidemiology*, 14:705–21, 2004. 1, 2.1

[7] R. Abbott, J. Curb, B. Rodriguez, and K. Masaki et. al. Age-related changes in risk factor effects on the incidence of coronary heart disease. *Annals of Epidemiology*, 12:173–81, 2002. 1, 2.1

[8] W. B. Kannel and T. R. Dawber et. al. Risk factors in coronary heart disease: An evaluation of several serum lipids as predictors of coronary heart disease the framingham study. *Annals of Internal Medicine*, 61:888–99, 1964. 2.1

[9] M. Weijenberg, E. Feskens, and D. Kromhout. Total and high density lipoprotein cholesterol as risk factors for coronary heart disease in elderly men during 5 years of follow-up. *American Journal of Epidemiology*, 143(2):151–8, 1996. 2.1

[10] R. Benfante and D. Reed. Is elevated serum cholesterol level a risk factor for coronary heart disease in the elderly. *JAMA*, 263(3):393–6, 1990. 2.1

[11] H. Krumholz, T. Seeman, and S. Merrill et. al. Lack of association between cholesterol and coronary heart disease mortality and morbidy and all-cause mortality in persons older than 70 years. *JAMA*, 272(17):1335–40, 1994. 2.1

[12] M. Corti, J. Guralnik, and M. Salive et. al. HDL cholesterol predicts coronary heart disease mortality in older persons. *JAMA*, 274(7):539–44, 1995. 2.1

[13] A. Weverling-Rijnsburger, G. Blauw, and A. Lagaay et. al. Total cholesterol and risk of mortality in the oldest old. *Lancet*, 350:1119–23, 1995. 2.1

[14] S. Franklin, M. Larson, and S. Khan et. al. Does the relation of blood pressure to coronary heart disease risk change with aging. *Circulation*, 103:1245–9, 2001. 2.1

[15] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley Series In Probability and Statistics, second edition, 2000. 2.2

[16] A. J. Dobson. *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC, first edition, 2002. 2.2

[17] L. V. Hedges. A random effects model for effect sizes. *Psychological Bulletin*, 93:388–95, 1983. 2.3

[18] N. Laird and F. Mosteller. Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care*, 6:5–30, 1990. 2.3

[19] G. V. Glass. Primary, secondary and meta-analysis of research. *Educational Researcher*, 5:3–8, 1976. 2.3

[20] R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7:177–88, 1986. 2.3, 2.3.1

[21] C. Berkey, D. Hoaglin, F. Mosteller, and G. Colditz. A random-effects regression model for meta-analysis. *Statistics in Medicine*, 14:395–411, 1995. 2.3

[22] C. Berkey and J. Anderson. Multiple-outcome meta-analysis of clinical trials. *Statistics in Medicine*, 15:537–57, 1996. 2.3

[23] C. Berkey, D. Hoaglin, A. Antczak-Bouckoms, F. Mosteller, and G. Colditz. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine*, 17:2537–50, 1998. 2.3, 2.3.2, 2.3.2

[24] In-Sun Nam, K. Mengersen, and P. Garthwaite. Multivariate meta-analysis. *Statistics in Medicine*, 22:2309–33, 2003. 2.3, 2.3.1

[25] R. D. Riley, K. R. Abrams, A. J . Sutton, and et.al. Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology*, 7, 2007. 2.3, 2.3.2

[26] R. D. Riley, K. R. Abrams, P. C. Lambert, and et.al. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *BMC Medical Research Methodology*, 26:78–97, 2007. 2.3

[27] D. Jackson, I. R. White, and S. G. Thompson. Extending Dersimonian and Laird's methodology to perform multivariate random effects meta-analysis. *Statistics in Medicine*, 29:1282–97, 2009. 2.3, 2.3.2, 5.2

[28] H. Houwelingen, L. Arends, and T. Stijnen. Tutorial in biostatistics advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine*, 21:589–624, 2002. 2.3, 2.3.1, 5.2

[29] S. Normand. Tutorial in biostatistics meta-analysis: Formulating, evaluating, combining and reporting. *Statistics in Medicine*, 18:321–59, 1999. 2.3.1

[30] J. E. Schmid, G. G. Koch, and L. M. LaVange. An overview of statistical issues and methods of meta-analysis. *Journal of Biopharmaceutical Statistics*, 1:103–20, 1991. 2.3.1

[31] A. Sutton, K. Abrams, D. Jones, T. Sheldon, and F. Song. *Methods for meta-analysis in medical research*. Wiley, first edition, 2000. 2.3.1

[32] R. J. Hardy and S. G. Thompson. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, 17:841–56, 1998. 2.3.1

[33] K. Sidik and J. N. Jonkman. A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine*, 26:1964–81, 2007. 2.3.1, 2.3.1

[34] S. Senn and V. Barnett. *Meta-Analysis of Controlled Clinical Trials*. John Wiley & Sons, Ltd., England, first edition, 2002. 2.3.1

[35] C. N. Morris. Parametric empirical bayes inference: Theory and applicatiton. *Journal of American Statistical Association*, 78:47–55, 1983. 2.3.1

[36] D. L. McGee and the Diverse Populations Collaboration. Body mass index and mortality: A meta-analysis based on person-level data from twenty-six observational studies. *Ann Epidemiol*, 15:87–97, 2005. 3.2

[37] L. V. Hedges and I. Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, California, first edition, 1985. 3.3

[38] T. Hastie and R. Tibshirani. Generalized additive models. *Journal of Statistical Science*, 1:297–318, 1986. 4.2.1

[39] S. N. Wood and N. H. Augustin. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, 157:157–177, 2002. 4.2.1

[40] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition, 2009. 4.2.1

[41] S. N. Wood. *Generalized Additive Models*. Chapman & Hall/CRC, Taylor & Francis Group, first edition, 2006. 4.2.2, 4.2.2, 4.2.5, 4.2.5, 4.3.1, 4.4.3

[42] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC, first edition, 1990. 4.2.2

[43] S. N. Wood. Fast stable direct ditting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:495–518, 2008. 4.2.4

[44] Alan Agresti. *Categorical Data Analysis*. John Wiley & Sons, Inc., second edition, 2002. 4.2.4

[45] G. Marra and R. Radice. Penalised regression splines: theory and application to medical research. *Statistical Methods in Medical Research*, 19:107–25, 2010. 4.2.4

BIOGRAPHICAL SKETCH

Yan Li

Yan Li was born in Shanxi Province, China. In 2005, she enrolled into the Master program in the Department of Statistics of Florida State University. She obtained her Master degree in Biostatistics in the spring of 2007, and then continued her doctoral study in Biostatistics in the fall of 2007.

Yan Li's education background and experience have shaped her research interests covering multivariate meta analysis, generalized additive models, meta analysis with respect to curves, and epidemiology.