Florida State University Libraries

Electronic Theses, Treatises and Dissertations

The Graduate School

2010

Time-Varying Coefficient Models with ARMA-GARCH Structures for Longitudinal Data Analysis

Haiyan Zhao

THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

TIME-VARYING COEFFICIENT MODELS WITH ARMA-GARCH
STRUCTURES FOR LONGITUDINAL DATA ANALYSIS

By

HAIYAN ZHAO

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 Haiyan Zhao defended on September 28, 2010.

Xufeng Niu
Professor Co-Directing Dissertation

Fred Huffer
Professor Co-Directing Dissertation

Craig Nolder
University Representative

Dan McGee
Committee Member

Approved:

Dan McGee, Chair, Department of Statistics

Joseph Travis, Dean, College of Arts and Sciences

The Graduate School has verified and approved the above named committee members.

This thesis is dedicated to my family.

ACKNOWLEDGEMENTS

I would first like to acknowledge my gratitude to my major advisors Dr. Xu-Feng Niu and Dr. Fred Huffer for their support, guidance, and patience in my dissertation research. They directed me through the hard times during my research and provided precious suggestions for both my study and career. They are always willing to answer any type of questions even when I have to go down to details.

I would like to thank my committee members Dr. Daniel McGee and Dr. Craig Nolder for their generous support. I also want to give special thanks to Dr. McGee for providing the data set.

The statistics department is like a warm family. I spent a great time here and learned abundant things even beyond statistics. I also want to give thanks to Pamela, Chauncey, James, Jocelyne, and Evangelous who are always available to help me and answer my questions.

Finally, I want to give a heart full of thanks to my family for their support and love.

TABLE OF CONTENTS

List of Tables	vi
List of Figures	viii
Abstract	ix
1. INTRODUCTION	1
2. MODELS AND PROPERTIES AND ESTIMATION	5
2.1 Varying-coefficient Models	5
2.2 Time Series Models	9
2.3 Proposed model	17
3. SIMULATION STUDY	24
3.1 Data Generation	25
3.2 Kullback-Leibler Divergence	28
3.3 Simulation Results	29
3.4 Summary and Discussion	39
4. APPLICATION	42
4.1 The Framingham Heart Study	43
4.2 The Pooled Logistic Regression Method	45
4.3 Time-varying Coefficient Model with ARMA-GARCH Structure	50
4.4 Summary and Discussion	65
5. CONCLUSIONS AND FUTURE WORK	78
REFERENCES	81
BIOGRAPHICAL SKETCH	84

LIST OF TABLES

3.1	Simulation results for Case 1: $\phi = 0.8, \alpha_0 = 1, \alpha_1 = 0.3, \gamma_1 = 0.6$	31
3.2	Simulation results for Case 2: $\phi = 0.8, \alpha_0 = 1, \alpha_1 = 0.2, \gamma_1 = 0.2$	32
3.3	Estimates of AR-GARCH parameters for Case 1: $n = 200$	37
3.4	Estimates of AR-GARCH parameters for Case 1: $n = 500$	38
3.5	Estimates of AR-GARCH parameters for Case 1: $n = 1000$	38
3.6	Estimates of AR-GARCH parameters for Case 2: $n = 200$	39
3.7	Estimates of AR-GARCH parameters for Case 2: $n = 500$	39
3.8	Estimates of AR-GARCH parameters for Case 2: $n = 1000$	40
4.1	Variable Description	44
4.2	Disease rate for both gender at the end of FHS	44
4.3	Logistic Regression based on measurements at exam 4	45
4.4	Pooled Logistic Regression from exam 3 to exam 19 for both male and female	48
4.5	Pooled Logistic Regression from exam 3 to exam 19 for male	49
4.6	Pooled Logistic Regression from exam 3 to exam 19 for female	49
4.7	Model selection for the intercept β_{0t}	52
4.8	Model selection for the age effect	53
4.9	Model selection for the CSM effect	54
4.10	Model selection for the SBP effect	54
4.11	Model selection for the BMI effect	55
4.12	Model selection for the intercept β_{0t}	57
4.13	Model selection for the age effect	57

4.14	Model selection for the SBP effect	58
4.15	Model selection for the BMI effect	58
4.16	Model comparison for male	60
4.17	Estimates of time-series parameters in the proposed model for male by MLE	61
4.18	Estimates of time-series parameters in the proposed model for male by LA .	62
4.19	Model selection for female	63
4.20	Estimates of time-series parameters in the proposed model for female by MLE	65
4.21	Estimates of time-series parameters in the proposed model for female by LA	66

LIST OF FIGURES

3.1	Plot of the true β with $n = 1000$ and $T = 20, 50$, and 100	27
3.2	Time-varying parameters β : true (solid), MLE (dotted), LA (dashed)	34
3.3	Time-varying parameters β : true (solid), MLE (dotted), LA (dashed)	35
3.4	Time-varying parameters β : true (solid), MLE (dotted), LA (dashed)	36
4.1	Boxplots of potential risk factors	46
4.2	Pooling of repeated observations	47
4.3	Time series structure of intercept for male	68
4.4	Time series structure of the coefficients of age for male	69
4.5	Time series structure of the coefficients of CSM for male	70
4.6	Time series structure of the coefficients of SBP for male	71
4.7	Time series structure of the coefficients of BMI for male	72
4.8	Time series structure of intercept for female	73
4.9	Time series structure of the coefficients of age for female	74
4.10	Time series structure of the coefficients of SBP for female	75
4.11	Time series structure of the coefficients of BMI for female	75
4.12	Estimates of time-varying coefficients based on the proposed model for male	76
4.13	Estimates of time-varying coefficients based on the proposed model for female	77

ABSTRACT

The motivation of my research comes from the analysis of the Framingham Heart Study (FHS) data. The FHS is a long term prospective study of cardiovascular disease in the community of Framingham, Massachusetts. The study began in 1948 and 5,209 subjects were initially enrolled. Examinations were given biennially to the study participants and their status associated with the occurrence of disease was recorded. In this dissertation, the event we are interested in is the incidence of the coronary heart disease (CHD). Covariates considered include sex, age, cigarettes per day (CSM), serum cholesterol (SCL), systolic blood pressure (SBP) and body mass index (BMI, weight in kilograms/height in meters squared).

Statistical literature review indicates that effects of the covariates on Cardiovascular disease or death caused by all possible diseases in the Framingham study change over time. For example, the effect of SCL on Cardiovascular disease decreases linearly over time. In this study, I would like to examine the time-varying effects of the risk factors on CHD incidence. Time-varying coefficient models with ARMA-GARCH structure are developed in this research. The maximum likelihood and the marginal likelihood methods are used to estimate the parameters in the proposed models. Since high-dimensional integrals are involved in the calculations of the marginal likelihood, the Laplace approximation is employed in this study. Simulation studies are conducted to evaluate the performance of these two estimation methods based on our proposed models. The Kullback-Leibler (KL) divergence and the root mean square error are employed in the simulation studies to compare the results obtained from different methods. Simulation results show that the marginal likelihood approach gives more accurate parameter estimates, but is more computationally intensive.

Following the simulation study, our proposed models are applied to the Framingham

Heart Study to investigate the time-varying effects of covariates with respect to CHD incidence. To specify the time-series structures of the effects of risk factors, the Bayesian Information Criterion (BIC) is used for model selection. Our study shows that the relationship between CHD and risk factors changes over time. For males, there is an obviously decreasing linear trend for age effect, which implies that the age effect on CHD is less significant for elder patients than younger patients. The effect of CSM stays almost the same in the first 30 years and decreases thereafter. There are slightly decreasing linear trends for both effects of SBP and BMI. Furthermore, the coefficients of SBP are mostly positive over time, i.e., patients with higher SBP are more likely developing CHD as expected. For females, there is also an obviously decreasing linear trend for age effect, while the effects of SBP and BMI on CHD are mostly positive and do not change too much over time.

CHAPTER 1

INTRODUCTION

Longitudinal studies involve repeated observations of the same subjects during follow-up periods. The most commonly used methods of analyzing follow-up data include the Cox proportional hazard model [1] and the multiple logistic regression model [2, 3]. In these two types of models, coefficients of all covariates are usually assumed to be constants over time. Generalized varying-coefficient models (Hastie and Tibshirani 1993 [4]) relax this requirement by allowing regression coefficients to depend on certain covariates. Specifically, the varying-coefficient model defined in [4] has the form:

$$g(\mu) = \beta_0(R_0) + \beta_1(R_1)X_1 + \dots + \beta_p(R_p)X_p, \quad (1.1)$$

where $E(Y) = \mu$, $g(\cdot)$ is the link function, and $\{X_1, X_2, \dots, X_p\}$ are predictors for the response variable Y . In this model, coefficients $\{\beta_0, \beta_1, \dots, \beta_p\}$ are allowed to change smoothly with the values of some covariates $\{R_0, R_1, \dots, R_p\}$.

When the coefficients $\{\beta_0, \beta_1, \dots, \beta_p\}$ in model (1.1) are functions of time, the model is called a time-varying coefficient model. Time series models may be used to characterize the patterns of these functions. Time series analysis [5, 6, 7] is a powerful technique to model a sequence of dependent measurements. For stationary processes, widely used time series models include autoregressive (AR) models, moving-average (MA) models, and mixed autoregressive moving-average (ARMA) models. For nonstationary processes, there are two basic patterns: trend and seasonality. The former pattern represents a general systematic linear or nonlinear component that changes over time and does not repeat within the time range. The latter repeats itself in systematic intervals over time. These two components are frequently observed in real-life data. A decomposition method [6] may be used to remove trend and seasonal components in this type of time series.

Another type of important time series model is the volatility model. Engle [8] proposed the autoregressive conditional heteroscedasticity (ARCH) model, within which the variance of the current error term is a function of the variances of the previous time periods' error terms. Bollerslev [9] described an extension of the ARCH model known as the generalized ARCH (GARCH) model, which captures the ARMA structure for the error variance. These models are non-linear time series models and are used to model time-varying volatility clusters.

The dynamic linear model (DLM) proposed by West and Harrison [10] is an example of joining time-varying coefficient models and time series models. It consists of two equations:

$$\begin{aligned} Y_t &= X_t\beta_t + \mu_t, & \mu_t &\sim N(0, V_t), \\ \beta_t &= G_t\beta_{t-1} + \omega_t, & \omega_t &\sim N(0, W_t). \end{aligned} \tag{1.2}$$

The first equation in this model is the observation equation that specifies the relationship between response variable Y_t and covariate X_t . Since the coefficients β_t change with time, it is a time-varying coefficient model. The second equation is the evolution equation, where the time evolution of the state vector is considered. It is a Markov process for the regression parameters, that is, β_t depends on the previous values only through β_{t-1} . The sequence $\{\mu_t\}$ is called the observational error and ω_t is called the evolution error. Sequences $\{\beta_t\}$ and $\{\mu_t\}$ are assumed to be mutually independent. If the values of the sequences $\{X_t\}$ and $\{G_t\}$ are constants for all t , then model (1.2) is called a time series DLM. Furthermore, if the variance of the evolution error ω_t is constant in t , the evolution equation in DLM becomes an AR(1) model [5]. To capture the volatility of the covariate effects in the DLM model, we may consider an ARCH model or GARCH model for the evolution equation.

The motivation of this study comes from the analysis of the Framingham Heart Study (FHS) [3, 11, 12]. The FHS is a long term prospective study of cardiovascular disease in the community of Framingham, Massachusetts. The study began in 1948 and 5,209 subjects were initially enrolled in the study. Examinations were given biennially to the study participants and their status associated with occurrences of disease was recorded. In this study, the event we are interested in is the incidence of coronary heart disease (CHD). The covariates include sex, age, cigarettes per day (CSM), serum cholesterol (SCL), systolic blood pressure (SBP) and body mass index (BMI, weight in kilograms/height in meters squared). There are a total of 20 exams, which means the follow-up time period is around 40 years.

Several papers provide evidence that the effects of covariates on Cardiovascular disease or death caused by all possible diseases in the Framingham study may change over time. For example, Wu and Ware [3] studied the occurrence or non-occurrence of the cardiovascular disease (CVD) as the dependent variable during each interval of the follow-up time. The authors estimated the risk of CVD as a function of the baseline SBP and repeated measurements of SCL among Framingham participants. For the general full model, Wu and Ware [3] used a multiple logistic regression model and allowed the risk in each interval to depend on all measurements made prior to the interval. Then the authors used a hierarchical model for the coefficients of SCL and assumed that these coefficients change linearly with the same slope for each interval. Reduced models based on the baseline SCL, most recent SCL, or average SCL were also fitted. The results showed that the model which used all previous SCL measurements and estimated their coefficients along a linear trend with the same slope gave a better fit than others. The estimated slope for coefficients of SCL was negative, which meant that the baseline measurements of SCL are more predictive than the most recent measurements. In addition, the results in this paper indicated that SBP was a statistically significant risk factor on the dependent variable CVD. In a recent paper, He [11] described some evidence that the effects of covariates with respect to all-cause mortality changed over time for the Framingham Heart Study. The author fitted a Bayesian dynamic survival model in which the coefficients of SBP, age, and sex were allowed to change over time. Results from He [11] showed that the coefficient function for age roughly increased with a linear trend over time. There was an overall linearly decreasing trend for the coefficients of SBP. Moreover, the coefficients of sex also changed over time but with a non-linear trend.

In this dissertation, we would like to examine the time-varying effects of the risk factors on the incidence of CHD in the Framingham Heart Study. Time-varying coefficient models with ARMA-GARCH structure are developed in this study to achieve this purpose. Different estimation methods will be discussed based on this model.

The time-varying coefficient models defined in (1.1) are a simple and useful extension of the classic linear models. There are many applications of this type of models in economics and finance [13, 14, 15] as well as in longitudinal data studies. The capital asset pricing model (CAPM) [16] is an example of this kind of model in finance. CAPM is used in finance to determine the required rate of return of an asset. The model says that the expected return of an asset or a portfolio equals the rate on a risk-free security plus the beta coefficient

multiplied by the expected excess return of the market portfolio. Here the beta coefficient is a measure of the volatility of a security or a portfolio relative to the rest of the financial market. Some recent studies [17, 18, 19] show that the beta-coefficients might change over time. Another example of the time-varying coefficient model is the relationship between the prototype electricity demand and covariates such as income, the real price of electricity, and temperature. Chang and Martinez-Chombo [20] found that the coefficients of covariates in this relationship may change over time.

In Chapter 2, I will review some classical time-varying coefficient models and time series models and discuss their properties and estimation procedures. Time-varying coefficient models with ARMA-GARCH structure are also proposed in this chapter. For the estimation of parameters in our proposed models, two different methods, maximum likelihood estimation and the marginal likelihood approach with Laplace approximation, are evaluated. In Chapter 3, simulation studies are conducted to investigate the performance of the two estimation methods for our proposed models. The Kullback-Leibler divergence and the root mean square error are employed in the simulation to compare the results from different methods. Our proposed models will be applied to the Framingham Heart Study to examine the time-varying effects of risk factors on CHD incidence in Chapter 4. Finally, in Chapter 5, I will discuss certain topics that will be considered in future research.

CHAPTER 2

MODELS AND PROPERTIES AND ESTIMATION

In the statistical and epidemiological literature, there are many models that can be used to address the relationship between events and risk factors. The two most commonly used models are the multiple logistic regression and the Cox proportional hazards models. In recent research, other modified or new models are proposed to estimate the effects of risk factors, such as time-varying coefficient models and time series models. In this chapter, we will review these models and discuss their related properties and methods of parameter estimation.

2.1 Varying-coefficient Models

In recent years, much progress has been made to increase the flexibility of generalized linear model. Varying-coefficient models proposed by Hastie and Tibshirani [4] are a very useful extension of the classic linear models. In this model, the coefficients are allowed to change smoothly with the value of other variables. Due to this advantage, varying-coefficient models are widely applied in many areas, such as longitudinal data studies [21], functional-coefficient nonlinear time series models [22], and finance and econometrics [23].

An expression of varying-coefficient models is given by model (1.1). For simplicity, a Gaussian model is considered, i.e. $g(\mu) = \mu$ and Y is normally distributed with mean μ . Then the varying-coefficient model becomes

$$Y = \beta_0(R_0) + \beta_1(R_1)X_1 + \cdots + \beta_p(R_p)X_p + \varepsilon, \quad (2.1)$$

where $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma^2$.

Varying-coefficient models are very general models and can be specified by choosing different kinds of coefficients $\beta_j(R_j)$ and covariates X_j . For example, if the coefficients

$\{\beta_0, \beta_1, \dots, \beta_p\}$ are functions of time, the model is called a time-varying coefficient model and has the form:

$$g(\mu_t) = \beta_0(t) + \beta_1(t)X_1(t) + \dots + \beta_p(t)X_p(t). \quad (2.2)$$

If the coefficients $\{\beta_0, \beta_1, \dots, \beta_p\}$ are fixed, i.e. $\beta_j(R_j) = \beta_j$ for $j = 0, \dots, p$, the model in (1.1) becomes a generalized linear model. The details for these two useful models will be discussed in details in the following two subsections.

2.1.1 Time-varying Coefficient Models

Time-varying coefficient models examine the effects of covariates on response variable over time. In model (2.2), the coefficients $\boldsymbol{\beta}(t) = (\beta_0(t), \dots, \beta_p(t))^T$ are arbitrary smooth functions of time t . In practice, there are two common procedures to estimate the coefficient functions $\boldsymbol{\beta}(t)$, nonparametric estimation and parametric estimation.

Hastie and Tibshirani [4] proposed a nonparametric estimation using smoothing spline method to estimate the coefficients $\beta_0(R_0), \dots, \beta_p(R_p)$ in the general varying-coefficient model (1.1). For the time-varying coefficient models in (2.2), the smoothing spline estimator of $\beta_0(t), \dots, \beta_p(t)$ is to minimize the penalized least squares,

$$J(\beta_1, \dots, \beta_p) = \sum_{i=1}^n \left\{ y_i - \sum_{j=1}^p x_{ij}\beta_j(t) \right\}^2 + \sum_{j=1}^p \lambda_j \int \beta_j''(t)^2 dr_j, \quad (2.3)$$

where we assume that the functions $\beta_0(t), \dots, \beta_p(t)$ are twice continuously differentiable and their second derivatives $\beta_0''(t), \dots, \beta_p''(t)$ are bounded and square integrable. In the above formula, $\boldsymbol{\lambda} = (\lambda_0, \dots, \lambda_k)$ are positive-valued smoothing parameters and they are considered to be fixed while estimating the coefficients. In the penalized least squares equation (2.3), the first term measures the goodness of fit and the second term penalizes the roughness of β_j by the second derivative β_j'' . It can be shown that the solution of the equation (2.3) is a natural cubic spline with knots at unique value of time t [24].

In time-varying coefficient models (2.2), one can specify structures of time-varying coefficients over time. For example, consider a Markov process for the coefficients $\boldsymbol{\beta}(t)$. Then it is a dynamic linear model (DLM) in (1.2). Also we may consider a simple linear coefficient function $\beta_j(t) = a_j + b_j t, j = 1, \dots, p$. There are many types of parametric functions which could be used to describe the structure of coefficients over time. After specifying the

parametric equation for coefficients, we can fit a parametric time-varying coefficients models using different estimation methods, such as the maximum likelihood method.

The standard Cox proportional hazards model can be extended to a varying-coefficient model [25], which is called time-varying coefficient Cox models. In the analysis of censored failure time observations, the standard Cox proportional hazards model [1] assumes that the regression coefficients are time-invariant during the follow-up time. Suppose that $\{T_1, T_2, \dots, T_n\}$ are survival times, $\{C_1, C_2, \dots, C_n\}$ are censoring times, and $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$ are covariate vectors. Let $Y_i = \min\{T_i, C_i\}$ and $\delta_i = 1_{\{T_i \leq C_i\}}$, $i = 1, \dots, n$. For the Cox model, the hazard function for the i^{th} observation is given by

$$h(t|\mathbf{X}_i) = h_0(t) \exp(\mathbf{X}_i' \boldsymbol{\beta}), \quad (2.4)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)'$ is the vector of regression coefficients. This model is a semi-parametric model because the baseline hazard $h_0(t)$ can take an arbitrary unspecified form. Now, consider two different observations i and j with covariates $\mathbf{X}_i$ and $\mathbf{X}_j$, respectively. The hazard ratio for these observations,

$$\frac{h(t|\mathbf{X}_i)}{h(t|\mathbf{X}_j)} = \frac{h_0(t) \exp(\mathbf{X}_i' \boldsymbol{\beta})}{h_0(t) \exp(\mathbf{X}_j' \boldsymbol{\beta})} = \frac{\exp(\mathbf{X}_i' \boldsymbol{\beta})}{\exp(\mathbf{X}_j' \boldsymbol{\beta})},$$

is constant over time t , and this is why the model is called proportional hazards model. However, the effects of covariates may vary over time in many real situations and these time-varying effects are of great interest. To make the Cox model more flexible, we can use the idea of varying-coefficient models and allow the coefficients β_j depending on time. The model becomes

$$\lambda(t|\mathbf{X}_i) = \lambda_0(t) \exp(\mathbf{X}_i' \boldsymbol{\beta}(t)). \quad (2.5)$$

This model is a non-proportional hazards model since the coefficients change over time.

2.1.2 Generalized Linear models

Let $\mathbf{Y}$ denote the response variable with observations $(y_1, \dots, y_n)$. Let $x_{ij}, j = 1, \dots, n$, denote the value of predictor j for subject i . Then a generalized linear model (GLM) has the form [26]:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, \quad (2.6)$$

where $i = 1, \dots, n$ and $\mu_i = E(Y_i)$. $g(\cdot)$ is a link function, which provides the relationship between the linear predictors and the mean of the distribution function. Link functions are usually assumed to be monotonic and differentiable.

In model (2.6), we assume that $\mathbf{Y} = (y_1, y_2, \dots, y_n)$ are independent observations from a distribution of natural exponential family. The density function of this family has the form:

$$f(y_i; \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\},$$

where a, b , and c are some functions to determine the specific distribution; θ_i is the natural parameter; and ϕ is called the dispersion parameter.

The maximum likelihood estimation (MLE) can be employed to estimate the parameters in generalized linear models. Let

$$l_i = \log f(y_i; \theta_i, \phi) = \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi).$$

Since the observations are independent, the log-likelihood function for a GLM can be expressed by

$$l(\beta_1, \dots, \beta_p) = \sum_i l_i = \sum_i \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + \sum_i c(y_i, \phi). \quad (2.7)$$

Newton-Raphson algorithm may be used to maximize the log-likelihood function with respect to the regression parameters.

The logistic regression model is a special case of the generalized linear models, in which the response variable y is a binary response variable which usually represents the occurrence (coded as 1) or non-occurrence (coded as 0) of some outcome event. Let $p_i(\mathbf{x}_i) = p(y_i = 1 | \mathbf{x}_i)$, where $\mathbf{x}_i$ is a vector of the explanatory variables for the i^{th} observation y_i . Using the logit link function, the logistic regression model has the following form:

$$\text{logit}(p_i(\mathbf{x}_i)) = \log \left(\frac{p_i(\mathbf{x}_i)}{1 - p_i(\mathbf{x}_i)} \right) = \mathbf{x}_i' \boldsymbol{\beta}, \quad (2.8)$$

or

$$p_i(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i' \boldsymbol{\beta})}. \quad (2.9)$$

If there are p predictors, we set $\mathbf{x}_i^T = (x_{i1}, \dots, x_{ip})$ and $\boldsymbol{\beta}^T = (\beta_1, \dots, \beta_p)$. The interpretation of the logistic regression model is that the parameter β_j refers to the effect of x_{ij} on the log odds of event $y_i = 1$ controlling for the other covariates, i.e. $\exp(\beta_j)$ can be interpreted as the odds ratio of having an event associated with a one-unit increase in x_{ij} during the study period.

2.2 Time Series Models

Time series analysis is one of the most important and widely used statistical techniques. Time series applications cover economic forecasting, study of biological data, control systems, signal processing and communications, and vibrations engineering etc. In this subsection, we will discuss some time series models including the ARIMA, ARCH, and GARCH models.

2.2.1 ARIMA models

Stationarity of stochastic processes plays an important role in time series analysis. A time series process $\{X_t\}$ is called strictly stationary if the joint distribution of $(X_{t_1}, \dots, X_{t_k})$ is the same with that of $(X_{t_1+t}, \dots, X_{t_k+t})$ for all k and t . In other words, strict stationarity requires that the joint distribution of $(X_{t_1}, \dots, X_{t_k})$ is invariant under time shift. This condition is very strong and hard to be satisfied empirically. There is a weaker version of stationarity. A time series process $\{X_t\}$ is said to be weakly stationary if it satisfies the following conditions:

- (i) $E(X_t) \equiv \mu$;
- (ii) $\text{Var}(X_t) \equiv \sigma^2$;
- (iii) $\gamma(k) = \text{Cov}(X_t, X_{t+k})$ for all integers k and t .

Most statistical forecasting methods are based on the assumption that the time series in study are weakly stationary or can be stationarized through the use of mathematical transformations. In general, when we state that the time series process $\{X_t\}$ is stationary we mean that it's weakly stationary. A white noise is a simple type of stochastic process and is defined as an uncorrelated series with zero mean and constant variance, which implies that white noise is an stationary process.

Assume that $\{X_t\}$, $t = \dots, -1, 0, 1, \dots$, is a time series process. An ARIMA(p,d,q) model is given by

$$\left(1 - \sum_{i=1}^p \phi_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{i=1}^q \theta_i L^i\right) \varepsilon_t, \quad (2.10)$$

where L is the lag operator, that is $L^k X_t = X_{t-k}$; coefficients ϕ_i 's are the parameters of the autoregressive part of the model; θ_i 's are the parameters in the moving average part; sequence $\{\varepsilon_t\}$ is usually assumed to be a white noise stochastic process with mean zero and variance σ^2 ; p is the order of the autoregressive term; d is a positive integer that controls the order of differencing; q is the order of the moving average part. Autoregressive

models (ARIMA(p,0,0) or AR(p)), moving-average models (ARIMA(0,0,q) or MA(q)), and autoregressive moving-average models (ARIMA(p,0,q) or ARMA(p,q)) are all special cases of ARIMA models.

Let $\phi(L) = (1 - \sum_{i=1}^p \phi_i L^i)$ and $\theta(L) = (1 + \sum_{i=1}^q \theta_i L^i)$. Then the ARIMA(p,d,q) model can be expressed in the form:

$$\phi(L)(1 - L)^d X_t = \theta(L)\varepsilon_t.$$

ARIMA models are a very general class of models for time series data. If $d \neq 0$, the process $\{X_t\}$ is not stationary. Stationarity can usually be accomplished by differencing the series, that is, we perform a difference transformation upon the original series. The differenced series is a new series consisting of the increments between consecutive items in the original series, which is less likely to exhibit a trend pattern than the original series. Correlograms should then be constructed for the differenced series. If the simple autocorrelations for the differenced series still exhibit nonstationarity, it may be subjected to a second differencing. The differencing process should be continued until one is satisfied that the resulting series is sufficiently free of trend (or stationary) before proceeding to the specification of an ARIMA model.

In the model (2.10), we work on $W_t = (1 - L)^d X_t$ after differencing the series. Process $\{W_t\}$ follows an ARMA(p, q) model. If the roots of the characteristic equation, i.e.

$$1 - \phi_1 x - \phi_2 x^2 - \dots - \phi_p x^p = 0$$

lie outside the unit circle, the ARMA(p, q) process $\{W_t\}$ is stationary and causal, which means that the inverse of $\phi(L)$ exists. Then the model (2.10) can be expressed in the form:

$$W_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad t = \dots, -1, 0, 1, \dots$$

Since an ARIMA(p,d,q) process can be reduced to ARMA(p,q) model by differencing the series $\{X_t\}$, we consider how to estimate the parameters in an ARMA(p,q) model. Suppose that $\{X_t\}$ is a zero-mean stationary ARMA(p,q) process with model expressed as

$$X_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}, \quad (2.11)$$

where $\varepsilon_t \sim WN(0, \sigma^2)$. The parameters in the model are $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)'$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)'$ and the white noise variance σ^2 . The algorithm usually used to estimate the parameter

is the maximum likelihood (ML) method. Two approaches may be used to maximize the likelihood function. The first one is maximizing the unconditional likelihood function [6]. Another approach is maximizing the conditional likelihood function [7].

First let's consider the unconditional approach. Assume that $\{X_t\}$ is a zero-mean Gaussian process. Let $\Gamma_n = E(X_n X_n')$ denote the covariance matrix which is assumed to be non-singular. Let $\mathbf{X}_n = (X_1, \dots, X_n)'$. The unconditional likelihood of $\mathbf{X}_n$ is

$$L(\Gamma_n) = (2\pi)^{-n/2} (\det(\Gamma_n))^{-1/2} \exp\left(-\frac{1}{2} \mathbf{X}_n' \Gamma_n^{-1} \mathbf{X}_n\right). \quad (2.12)$$

Let $\hat{X}_{n+1}$ denote the one-step predictors and $v_n = E(X_{n+1} - \hat{X}_{n+1})^2$ denote the mean squared error of predictions. Define $r_n = v_n/\sigma^2$. Then the unconditional likelihood (2.12) can be expressed in the form:

$$L(\phi, \theta, \sigma^2) = (2\pi\sigma^2)^{-n/2} (r_0 \cdots r_{n-1})^{-1/2} \exp\left[-\frac{1}{2}\sigma^{-2} \sum_{j=1}^n (X_j - \hat{X}_j)^2 / r_{j-1}\right]. \quad (2.13)$$

The estimated parameters $(\hat{\phi}, \hat{\theta}, \hat{\sigma}^2)$ will maximize this unconditional likelihood function. Differentiating the log-likelihood function of $\log L(\phi, \theta, \sigma^2)$ partially with respect to σ^2 , we get

$$\hat{\sigma}^2 = n^{-1} S(\hat{\phi}, \hat{\theta}),$$

where

$$S(\hat{\phi}, \hat{\theta}) = \sum_{j=1}^n (X_j - \hat{X}_j)^2 / r_{j-1},$$

and $\hat{\phi}, \hat{\theta}$ are the values of ϕ, θ which minimize

$$l(\phi, \theta) = \ln(n^{-1} S(\phi, \theta)) + n^{-1} \sum_{j=1}^n \ln r_{j-1}.$$

For the unconditional likelihood function above, it is very difficult to estimate the one-step predictor $\hat{X}_{n+1}$. Therefore in practice, the conditional likelihood function is conveniently used in the ML estimation. Suppose that we know the initial values of X_t and ε_t before the series beginning. Let $\mathbf{X}_* = (X_0, X_{-1}, \dots, X_{-p+1})'$ denote the initial values of X_t and $\boldsymbol{\varepsilon}_* = (\varepsilon_0, \varepsilon_1, \dots, \varepsilon_{-q+1})'$ the initial values of ε_t . Notice that the ARMA(p,q) process in (2.11) can be written as

$$\varepsilon_t = X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}. \quad (2.14)$$

Using the initial values and observations, $\{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n\}$ could be calculated in turn from equation (2.14).

Assume that $\{\varepsilon_t\}$ is normally distributed. Then joint probability density of $\{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n\}$ is

$$p(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n) = (2\pi)^{-n/2} \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{t=1}^n \varepsilon_t^2 \right\}.$$

Then the log-likelihood, conditional on the initial values $(\mathbf{X}_*, \boldsymbol{\varepsilon}_*)$, would be

$$l_*(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = -\frac{n}{2} \ln(2\pi) - n \ln(\sigma) - \frac{S_*(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma^2}, \quad (2.15)$$

where $S_*(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^n \varepsilon_t^2(\boldsymbol{\phi}, \boldsymbol{\theta} | \mathbf{X}_*, \boldsymbol{\varepsilon}_*, \mathbf{X}_n)$ is called conditional sum of squares function.

For the equation above, without regard to the value of σ^2 , maximizing the conditional log-likelihood is equivalent to minimizing the conditional sum of squares function $S_*(\boldsymbol{\phi}, \boldsymbol{\theta})$. The parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$ can be obtained by the conditional least squares estimates. After $\boldsymbol{\phi}, \boldsymbol{\theta}$ have been estimated, σ^2 can be estimated by maximizing the log-likelihood l_* with respect to σ^2 . Then we have

$$\begin{aligned} -\frac{n}{2\hat{\sigma}^2} + \frac{S_*(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})}{2\hat{\sigma}^4} &= 0 \\ \hat{\sigma}^2 &= \frac{S_*(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})}{n}. \end{aligned}$$

2.2.2 ARCH and GARCH models

Recent developments in financial econometrics suggest the use of volatility models. Usually, volatility refers to the standard deviation of value changes of a financial instrument with a specific time horizon. It is often used to quantify the risk of the instrument over that time period. For financial data, the volatility tends to occur in clusters, which means that large changes tend to be followed by large changes; small changes tend to be followed by small changes. To capture the time-varying volatility clustering, Engle [8] proposed the autoregressive conditionally heteroscedastic (ARCH) model and Bollerslev [9] proposed the generalized autoregressive conditionally heteroscedastic (GARCH) model. An ARCH(m) model is assumed to have the form:

$$\begin{aligned} \varepsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \dots + \alpha_m \varepsilon_{t-m}^2, \end{aligned} \quad (2.16)$$

where the sequence $\{z_t\}$ is independent and identically distributed random variables with mean zero and variance one; $\alpha_0 > 0$ and $\alpha_i \geq 0$ for $i > 0$, which ensure that σ_t^2 is positive. The coefficients α_i 's must satisfy some other conditions to ensure that the unconditional variance of ε_t is finite.

From the structure of the ARCH model, it is seen that large past ε_{t-i} tend to be followed by large ε_t , which is similar to the volatility clustering in real financial data. Also the model shows that ε_t is dependent and can be described by a simple quadratic function of its lagged values. Because the model can be written as

$$\begin{aligned}\varepsilon_t^2 &= \sigma_t^2 z_t^2 \\ \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_m \varepsilon_{t-m}^2 &= \sigma_t^2.\end{aligned}$$

Subtracting these two equations, we get

$$\varepsilon_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_m \varepsilon_{t-m}^2 + \gamma_t$$

where $\gamma_t = \sigma_t^2(z_t^2 - 1)$.

Now we consider some properties of the ARCH model. First, we consider the conditional distribution of ε_t given the information available up to time $(t-1)$. Let F_{t-1} denote this information and assume that $\{z_t\}$ are i.i.d normally distributed with mean zero and variance one. Then we have

$$\varepsilon_t | F_{t-1} \sim N(0, \sigma_t^2),$$

where σ_t^2 is determined by the available information F_{t-1} . In other words, given the available information F_{t-1} , the next observation ε_t has a normal distribution with conditional mean zero and conditional variance σ_t^2 .

Next we consider unconditional properties of ε_t . The unconditional mean of ε_t is zero since

$$E(\varepsilon_t) = E[E(\varepsilon_t | F_{t-1})] = E[\sigma_t E(z_t | F_{t-1})] = E[\sigma_t E(z_t)] = 0.$$

The unconditional variance of ε_t can be obtained as

$$\begin{aligned}\text{Var}(\varepsilon_t) = E(\varepsilon_t^2) &= E[E(\varepsilon_t^2 | F_{t-1})] \\ &= E[E(\sigma_t^2 Z_t^2 | F_{t-1})] \\ &= E(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_m \varepsilon_{t-m}^2) \\ &= \alpha_0 + \alpha_1 E(\varepsilon_{t-1}^2) + \cdots + \alpha_m E(\varepsilon_{t-m}^2)\end{aligned}$$

Since ε_t is a stationary process with $E(\varepsilon_t) = 0$, we have $\text{Var}(\varepsilon_t) = E(\varepsilon_{t-1}^2) = \dots = E(\varepsilon_{t-m}^2)$. Therefore, we get $\text{Var}(\varepsilon_t) = \frac{\alpha_0}{1-\alpha_1-\dots-\alpha_m}$ under the condition $\sum_{i=1}^m \alpha_i < 1$. For the unconditional covariance of ε_i and ε_j , $i \neq j$ and assuming $i < j$, the following equation is derived

$$\begin{aligned} \text{Cov}(\varepsilon_i, \varepsilon_j) &= E(\varepsilon_i \varepsilon_j) = E[E(\sigma_i z_i \sigma_j z_j | F_{j-1})] \\ &= E(\sigma_i \sigma_j E(z_i z_j | F_{j-1})) \\ &= E(\sigma_i \sigma_j E(z_i z_j)) = 0 \end{aligned}$$

Therefore, $\{\varepsilon_t\}$ is a white noise process with mean zero and constant variance. For the ARCH(m) model, there is another very useful property. The excess kurtosis of ε_t is positive and the tail distribution of ε_t is heavier than that of a normal distribution. Heavy tails are a common aspect of financial data, and hence ARCH models are so popular in this field.

Although the ARCH(m) model in (2.16) is simple and easy to handle, it often requires a high order m to adequately describe the volatility process, and that means many parameters need to be considered. Bollerslev [9] proposed an extension of ARCH known as the generalized ARCH (GARCH) model. A GARCH(m, s) model is defined by the equations

$$\begin{aligned} \varepsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^m \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^s \beta_j \sigma_{t-j}^2, \end{aligned} \quad (2.17)$$

where again sequence $\{z_t\}$ are independent and identically distributed random variables with mean zero and variance one; $\alpha_0 > 0$, $\alpha_i \geq 0$ and $\beta_j \geq 0$. The condition $\sum_{i=1}^{\max(m,s)} (\alpha_i + \beta_i) < 1$ ensures that the unconditional variance of ε_t is finite, where $\alpha_i = 0$ for $i > m$ and $\beta_j = 0$ for $j > s$. If $s = 0$, the model reduces to the ARCH(m) model.

For a GARCH(1,1) model, Nelson [27] showed that (2.17) has a unique stationary solution if and only if $E[\log(\beta_1 + \alpha_1 \varepsilon_0^2)] < 0$. In the case of a general GARCH(m, s) model, Bougerol and Picard [28, 29] gave necessary and sufficient conditions for the existence of a unique stationary solution of model (2.17). For the discussion later, we always assume that a GARCH(m, s) process is stationary.

Let $\alpha(L) = \sum_{i=1}^m \alpha_i L^i$ and $\beta(L) = \sum_{j=1}^s \beta_j L^j$. Then model (2.17) can be written in

terms of the lag operator L

$$\begin{aligned}\varepsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= \alpha_0 + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2.\end{aligned}\tag{2.18}$$

If the roots of the characteristic equation, i.e.

$$1 - \beta_1 x - \beta_2 x^2 - \dots - \beta_s x^s = 0$$

lie outside the unit circle and the process $\{\varepsilon_t\}$ is stationary, then we have

$$\begin{aligned}\sigma_t^2 &= \frac{\alpha_0}{1 - \beta(1)} + \frac{\alpha(L)}{1 - \beta(L)} \varepsilon_t^2 \\ &= \alpha_0^* + \sum_{i=1}^{\infty} \gamma_i \varepsilon_{t-i}^2,\end{aligned}$$

where $\alpha_0^* = \frac{\alpha_0}{1 - \beta(1)}$. This expression tells us that a GARCH(m, s) process is an ARCH process with infinite order.

The quasi maximum likelihood estimator (QMLE) is the most common procedure for estimating parameters in model (2.17) [30]. The difficulty for estimating GARCH parameters is that there is no explicit expression of the probability density of ε_t since the distribution of σ_t is hard to know. Therefore, one can consider the conditional distribution of ε_t given $\{\varepsilon_{t-1}, \dots, \varepsilon_{-m+1}\}$ and $\{\sigma_0^2, \dots, \sigma_{-s+1}^2\}$.

Let $\boldsymbol{\theta} = (\alpha_0, \dots, \alpha_m, \beta_1, \dots, \beta_s)^T$ denote the parameters which need to be estimated. Assume that the distribution of $\{z_t\}$ are i.i.d. normally distributed. Then the conditional distribution of ε_t given $\{\varepsilon_{t-1}, \dots, \varepsilon_{-m+1}\}$ and $\{\sigma_0^2, \dots, \sigma_{-s+1}^2\}$ is also Gaussian distributed with mean zero and variance $\tilde{\sigma}_t^2(\boldsymbol{\theta})$ given by

$$\tilde{\sigma}_t^2(\boldsymbol{\theta}) = \begin{cases} \sigma_t^2, & t \leq 0 \\ \alpha_0 + \sum_{i=1}^m \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^s \beta_j \tilde{\sigma}_{t-j}^2(\boldsymbol{\theta}), & t > 0. \end{cases}$$

The conditional Gaussian log-likelihood function has the form:

$$\begin{aligned}l_n(\boldsymbol{\theta}) &= \log f_{\boldsymbol{\theta}}(\varepsilon_1, \dots, \varepsilon_n | \varepsilon_0, \dots, \varepsilon_{-m+1}, \sigma_0^2, \dots, \sigma_{-s+1}^2) \\ &= \sum_{t=1}^n \log f_{\boldsymbol{\theta}}(\varepsilon_t | \varepsilon_{t-1}, \dots, \varepsilon_{-m+1}, \sigma_0^2, \dots, \sigma_{-s+1}^2) \\ &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^n \left(\frac{\varepsilon_t^2}{\tilde{\sigma}_t^2(\boldsymbol{\theta})} + \log \tilde{\sigma}_t^2(\boldsymbol{\theta}) \right).\end{aligned}\tag{2.19}$$

Since $\tilde{\sigma}_t^2(\boldsymbol{\theta})$ are defined recursively, we need to choose a certain initialization for $\{\varepsilon_{t-1}, \dots, \varepsilon_{-m+1}\}$ and $\{\sigma_0^2, \dots, \sigma_{-s+1}^2\}$. Berkes et al. [31] showed that the choice of initial values does not matter for the asymptotic properties of the QMLE. Then the quasi maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta} l_n(\boldsymbol{\theta}), \quad (2.20)$$

where $\Theta \subset (0, \infty) \times [0, \infty)^{m+s}$

2.2.3 ARMA-GARCH model

Just as its name implies, the ARMA-GARCH model is a combination of ARMA and GARCH models: i.e. in an ARMA process, the error term follows a GARCH process. Specifically, The ARMA-GARCH model has the form:

$$\begin{aligned} X_t &= \sum_{i=1}^p \phi_i X_{t-i} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j} \\ \varepsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^m \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^s \beta_j \sigma_{t-j}^2, \end{aligned} \quad (2.21)$$

where the sequence $\{z_t\}$ is assumed independent and normally distributed with mean zero and variance one. To ensure the stationarity and invertibility of the ARMA process and the positive conditional variance of the GARCH process, we need the following constraints on the parameters:

- (i) all roots of $1 - \sum_{i=1}^p \phi_i x^i = 0$ are outside the unit circle;
- (ii) all roots of $1 + \sum_{j=1}^q \theta_j x^j = 0$ are outside the unit circle;
- (iii) $\alpha_0 > 0$, $\alpha_i \geq 0$ and $\beta_j \geq 0$, for $i = 1, \dots, m$ and $j = 1, \dots, s$;
- (iv) $\sum_{i=1}^p \alpha_i + \sum_{j=1}^q \beta_j < 1$.

The last condition ensures that the unconditional variance of ε_t is finite.

Let $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)'$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)'$, $\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_m)'$, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_s)'$. The parameters which need to be estimated are $\boldsymbol{\varphi} = \{\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$. After reviewing the estimation of the ARIMA and GARCH models in section 2.2.1 and 2.2.2, the ideas can be extended to the ARMA-GARCH model. The conditional likelihood function will be used so that some initial values must be given. Let $\mathbf{X}_* = (X_0, X_{-1}, \dots, X_{-p+1})'$ denote the initial values of X_t . Let $\boldsymbol{\varepsilon}_* = (\varepsilon_0, \varepsilon_{-1}, \dots, \varepsilon_{1-\max(m,q)})'$ denote the initial values of ε_t . The observations

are denoted by $\mathbf{X}_n = \{X_1, X_2, \dots, X_n\}$. Then $\{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n\}$ can be calculated by the following equation:

$$\tilde{\varepsilon}_t = X_t - \sum_{i=1}^p \phi_i X_{t-i} - \sum_{j=1}^q \theta_j \tilde{\varepsilon}_{t-j}.$$

If the initial values of σ_t are also given, the procedure of estimation for the ARMA-GARCH model is the same as that for the GARCH model. Let $\boldsymbol{\sigma}_* = (\sigma_0, \sigma_{-1}, \dots, \sigma_{-s+1})'$ denote the initial values of σ_t . The conditional log-likelihood function is

$$\begin{aligned} l_n(\boldsymbol{\varphi}) &= \log f_{\boldsymbol{\varphi}}(X_1, \dots, X_n | \mathbf{X}_*, \boldsymbol{\varepsilon}_*, \boldsymbol{\sigma}_*) \\ &= \log f_{\boldsymbol{\varphi}}(\tilde{\varepsilon}_1, \dots, \tilde{\varepsilon}_n | \mathbf{X}_*, \boldsymbol{\varepsilon}_*, \boldsymbol{\sigma}_*, \mathbf{X}_n) \\ &= \sum_{t=1}^n \log f_{\boldsymbol{\varphi}}(\tilde{\varepsilon}_t | \mathbf{X}_*, \boldsymbol{\varepsilon}_*, \boldsymbol{\sigma}_*, \mathbf{X}_n) \\ &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^n \left(\frac{\tilde{\varepsilon}_t^2}{\tilde{\sigma}_t^2(\boldsymbol{\varphi})} + \log \tilde{\sigma}_t^2(\boldsymbol{\varphi}) \right), \end{aligned} \quad (2.22)$$

where $\tilde{\sigma}_t^2(\boldsymbol{\varphi}) = \alpha_0 + \sum_{i=1}^m \alpha_i \tilde{\varepsilon}_{t-i}^2 + \sum_{j=1}^s \beta_j \tilde{\sigma}_{t-j}^2$. Then the QMLE of $\boldsymbol{\varphi}$ is

$$\hat{\boldsymbol{\varphi}} = \arg \max_{\boldsymbol{\varphi}} l_n(\boldsymbol{\varphi}). \quad (2.23)$$

2.3 Proposed model

The literature provides evidence suggesting that the effects of covariates in a medical data analysis, such as gender, age, SCL, and SBP, with respect to CHD disease, might change with time. Therefore, it is reasonable to use time-varying coefficient models in analyzing the effect of covariates. Since the coefficients of covariates change with time, there might exist serial correlations among the coefficients and time series models may be applied to characterize the patterns of the coefficients. Based on these considerations, the following model is proposed. Assume that the observed data is given by $(y_{it}, \mathbf{x}_{it})$, $i = 1, \dots, n$, $t = 1, \dots, T$, where y_{it} denotes the response variable for the i^{th} subject at time t and $\mathbf{x}_{it}$ denotes a vector of covariates for the i^{th} subject at time t . Our proposed model has the form:

$$g(\mu_{it}) = \beta_{0t} + \beta_{1t}x_{1it} + \dots + \beta_{Kt}x_{Kit}, \quad (2.24)$$

where

$$\begin{aligned}
\beta_{kt} &= \sum_{i=1}^p \phi_{ki} \beta_{k,t-i} + \varepsilon_{kt} + \sum_{j=1}^q \theta_{kj} \varepsilon_{k,t-j} \quad k = 0, 1, \dots, K \\
\varepsilon_{kt} &= \sigma_{kt} z_{kt} \\
\sigma_{kt}^2 &= \alpha_{k0} + \sum_{i=1}^m \alpha_{ki} \varepsilon_{k,t-i}^2 + \sum_{j=1}^s \gamma_{kj} \sigma_{k,t-j}^2,
\end{aligned} \tag{2.25}$$

where $\{z_{kt}, t = 1, \dots, T, k = 0, \dots, K\}$ are independent and identically-distributed random variables with mean zero and variance one. Usually, we assume z_{kt} has a standard normal distribution, that is, $z_{kt} \sim N(0, 1)$.

There are two components in the proposed model, generalized linear model in (2.24) and an ARMA-GARCH model in (2.25). For the generalized linear model, the coefficients are allowed to change with time. Therefore, it is also a time-varying coefficient model. For the coefficient function of each covariate, we apply an ARMA-GARCH model to capture the time series patterns. The constraints on the parameters in the ARMA-GARCH part are the same as we discussed in the previous section.

For simplicity, we may just use AR-GARCH models for the coefficients β_{kt} . And for the error term ε_{kt} , the most widely used specification is the GARCH(1,1) model. The link function $g(\cdot)$ can specify various generalized linear models, such as logistic regression for binary response and a log-linear model when the response follows a Poisson distribution.

Our proposed model consists of the generalized linear model with time-varying coefficients that follow ARMA-GARCH models. There are many parameters that need to be estimated. First, let's define some notation. Let $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \dots, \beta_{kT})', k = 0, 1, \dots, K$. For each $\boldsymbol{\beta}_k$, the parameters need to be estimated are $\boldsymbol{\phi}_k = (\phi_{k1}, \dots, \phi_{kp})', \boldsymbol{\theta}_k = (\theta_{k1}, \dots, \theta_{kq})', \boldsymbol{\alpha}_k = (\alpha_{k0}, \dots, \alpha_{km})'$, and $\boldsymbol{\gamma}_k = (\gamma_{k1}, \dots, \gamma_{ks})'$. Let $\boldsymbol{\varphi}_k = (\boldsymbol{\phi}'_k, \boldsymbol{\theta}'_k, \boldsymbol{\alpha}'_k, \boldsymbol{\gamma}'_k)'$, $\boldsymbol{\varphi} = (\boldsymbol{\varphi}'_0, \boldsymbol{\varphi}'_1, \dots, \boldsymbol{\varphi}'_K)'$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}'_0, \dots, \boldsymbol{\beta}'_K)'$. In our model, $\boldsymbol{\beta}$ is unobserved and generated by ARMA-GARCH models. The joint probability density function of $\mathbf{Y} = \{y_{it}\}$ and $\boldsymbol{\beta}$ is

$$f(\mathbf{Y}, \boldsymbol{\beta}) = f(\mathbf{Y}|\boldsymbol{\beta})f(\boldsymbol{\beta})$$

We assume that given $\boldsymbol{\beta}$, the observations in $\mathbf{Y}$ are conditionally independent and the distribution of y_{it} has exponential form. Then we have

$$f(\mathbf{Y}|\boldsymbol{\beta}) = \prod_{t=1}^T \prod_{i=1}^n \exp \left\{ \frac{y_{it}\theta_{it} - b(\theta_{it})}{a(\phi)} + c(y_{it}, \phi) \right\}.$$

The log-likelihood function is

$$l(\boldsymbol{\beta}|\mathbf{Y}) = \sum_{i=1}^n \sum_{t=1}^T \left\{ \frac{y_{it}\theta_{it} - b(\theta_{it})}{a(\phi)} + c(y_{it}, \phi) \right\},$$

where θ_{it} is the natural parameter which is a function of $\boldsymbol{\beta}$, and ϕ is called the dispersion parameter. We also assume that the coefficients of the different covariates are independent. Then we have

$$f(\boldsymbol{\beta}) = \prod_{k=1}^K f(\boldsymbol{\beta}_k)$$

The conditional log-likelihood for ARMA-GARCH is discussed in section 2.2.3. Therefore, we have the following joint log-likelihood function for our proposed model,

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\varphi}) &= l(\boldsymbol{\beta}|Y) + \sum_{k=1}^K l(\boldsymbol{\varphi}_k|\boldsymbol{\beta}_k) \\ &= \sum_{i=1}^n \sum_{t=1}^T \left\{ \frac{y_{it}\theta_{it} - b(\theta_{it})}{a(\phi)} + c(y_{it}, \phi) \right\} - \sum_{k=1}^K \sum_{t=1}^T \left\{ \frac{1}{2} \log(\sigma_{kt}^2) + \frac{\varepsilon_{kt}^2}{2\sigma_{kt}^2} \right\} + \text{constant}. \end{aligned} \quad (2.26)$$

There are a total of $(T+p+q+m+s)(K+1)$ parameters in our model. After we have the joint log-likelihood function, we can estimate the parameters using the standard maximum likelihood estimation (MLE) by treating the $\boldsymbol{\beta}$ as fixed parameters. Another method that we will discuss is based on the marginal distribution of $\mathbf{Y}$, which is obtained by integrating out $\boldsymbol{\beta}$ from the joint density function $f(\mathbf{Y}, \boldsymbol{\beta})$. The Laplace approximation is used to estimate the integral. Details of these two methods will be discussed in the following two subsections.

2.3.1 Maximum Likelihood Estimation

For most applications of longitudinal data, the number of subjects is very large, usually more than 1000. However, the follow-up time does not last very long due to a variety of reasons, such as expense of experiment or disease status during a clinical trial. In the Framingham Heart Study (FSH), the follow-up time is about 40 years that is much longer than follow-up periods in other studies. Therefore time series models can be considered for the time-varying coefficients β_{kt} 's. In our proposed model, let $\boldsymbol{\delta} = (\boldsymbol{\beta}, \boldsymbol{\varphi})$. The parameter estimates can be obtained by maximizing the joint log-likelihood function in equation (2.26),

$$\hat{\boldsymbol{\delta}}_{MLE} = \arg \max_{\boldsymbol{\delta}} l(\boldsymbol{\delta}|Y, X). \quad (2.27)$$

In the optimization process, it is usually more efficient to use the gradient of the log-likelihood function with respect to all parameters in the model. However, the equations for the gradient are quite complicated since there are many parameters in our proposed model. In the following paragraphs, a simple model will be shown as an illustration. Let's assume that y_{it} has a Bernoulli distribution with success probability p_{it} and there is only one covariate in our model. In addition, we assume that there is no intercept in the model. For the time-varying coefficient of covariate, an AR(1)-GARCH(1,1) model is used. Then the model becomes

$$\log \frac{p_{it}}{1 - p_{it}} = \beta_t x_{it} \quad \left(p_{it} = \frac{e^{\beta_t x_{it}}}{1 + e^{\beta_t x_{it}}} \right) \quad (2.28)$$

where

$$\begin{aligned} \beta_t &= \phi \beta_{t-1} + \varepsilon_t \\ \varepsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 \sigma_{t-1}^2, \end{aligned} \quad (2.29)$$

where $z_t \sim N(0, 1)$. For initial values, we set $\beta_0 = 0$, $\varepsilon_0 = 0$, and $\sigma_0^2 = 0$ so that $\beta_1 = \varepsilon_1$ and $\sigma_1^2 = \alpha_0$. Based on the joint log-likelihood function of our proposed model (2.26), the joint log-likelihood function for this simplified model is

$$\begin{aligned} &l(\{\beta_t\}, \phi, \alpha_0, \alpha_1, \gamma_1) \\ &= \sum_{i=1}^n \sum_{t=1}^T \left\{ \beta_t x_{it} y_{it} - \log(1 + e^{\beta_t x_{it}}) \right\} - \sum_{t=1}^T \left\{ \frac{1}{2} \log \sigma_t^2 + \frac{\varepsilon_t^2}{2\sigma_t^2} \right\} + \text{constant} \\ &= \sum_{t=1}^T \left\{ \beta_t \left(\sum_{i=1}^n x_{it} y_{it} \right) \right\} - \sum_{i=1}^n \sum_{t=1}^T \left\{ \log(1 + e^{\beta_t x_{it}}) \right\} \\ &\quad - \frac{1}{2} \sum_{t=1}^T \left\{ \log v_t + \frac{\varepsilon_t^2}{v_t} \right\} + \text{constant}, \end{aligned}$$

where $v_t = \sigma_t^2$. Then the first partial derivatives of this log-likelihood function with respect to the β 's can be computed by

$$\begin{aligned} \frac{\partial l}{\partial \beta_s} &= \sum_{i=1}^n x_{is} y_{is} - \sum_{i=1}^n \frac{x_{is} e^{\beta_s x_{is}}}{1 + e^{\beta_s x_{is}}} \\ &\quad - \frac{1}{2} \sum_{t=1}^T \left\{ \frac{\partial v_t / \partial \beta_s}{v_t} + \frac{2\varepsilon_t \frac{\partial \varepsilon_t}{\partial \beta_s}}{v_t} + \varepsilon_t^2 \left(\frac{-\partial v_t / \partial \beta_s}{v_t^2} \right) \right\}. \end{aligned} \quad (2.30)$$

The first partial derivatives relative to the parameters specified in the AR(1)-GARCH(1,1) model, i.e. $\phi, \alpha_0, \alpha_1, \gamma_1$, will have the form:

$$\begin{aligned}\frac{\partial l}{\partial \phi} &= -\frac{1}{2} \sum_{t=1}^T \left\{ \frac{\partial v_t / \partial \phi}{v_t} + \frac{2\varepsilon_t \frac{\partial \varepsilon_t}{\partial \phi}}{v_t} + \varepsilon_t^2 \left(\frac{-\partial v_t / \partial \phi}{v_t^2} \right) \right\} \\ \frac{\partial l}{\partial \alpha_0} &= -\frac{1}{2} \sum_{t=1}^T \left\{ \frac{\partial v_t / \partial \alpha_0}{v_t} + \varepsilon_t^2 \left(\frac{-\partial v_t / \partial \alpha_0}{v_t^2} \right) \right\} \\ \frac{\partial l}{\partial \alpha_1} &= -\frac{1}{2} \sum_{t=1}^T \left\{ \frac{\partial v_t / \partial \alpha_1}{v_t} + \varepsilon_t^2 \left(\frac{-\partial v_t / \partial \alpha_1}{v_t^2} \right) \right\} \\ \frac{\partial l}{\partial \gamma_1} &= -\frac{1}{2} \sum_{t=1}^T \left\{ \frac{\partial v_t / \partial \gamma_1}{v_t} + \varepsilon_t^2 \left(\frac{-\partial v_t / \partial \gamma_1}{v_t^2} \right) \right\}.\end{aligned}$$

Based on the AR(1)-GARCH(1,1) model (2.29), we have $\varepsilon_t = \beta_t - \phi\beta_{t-1}$ and $\varepsilon_1 = \beta_1$ by the assumption. Thus the first partial derivatives of ε_t with respect to all parameters are

$$\begin{aligned}\frac{\partial \varepsilon_t}{\partial \beta_s} &= \delta_{s,t} - \phi \delta_{s,t-1}, \quad \text{where} \quad \delta_{s,t} = \begin{cases} 1 & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} \\ \frac{\partial \varepsilon_t}{\partial \phi} &= -\beta_{t-1}, \quad \frac{\partial \varepsilon_t}{\partial \alpha_0} = 0 \\ \frac{\partial \varepsilon_t}{\partial \alpha_1} &= 0, \quad \frac{\partial \varepsilon_t}{\partial \gamma_1} = 0.\end{aligned}$$

Similarly, according to the equation $v_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 v_{t-1}$ in the AR-GARCH model, we can calculate the first partial derivatives of v_t for all parameters by the following equations,

$$\begin{aligned}\frac{\partial v_t}{\partial \beta_s} &= 2\alpha_1 \varepsilon_{t-1} \frac{\partial \varepsilon_{t-1}}{\partial \beta_s} + \gamma_1 \frac{\partial v_{t-1}}{\partial \beta_s}, \\ \frac{\partial v_t}{\partial \phi} &= 2\alpha_1 \varepsilon_{t-1} \frac{\partial \varepsilon_{t-1}}{\partial \phi} + \gamma_1 \frac{\partial v_{t-1}}{\partial \phi}, \\ \frac{\partial v_t}{\partial \alpha_0} &= 1 + \gamma_1 \frac{\partial v_{t-1}}{\partial \alpha_0}, \\ \frac{\partial v_t}{\partial \alpha_1} &= \varepsilon_{t-1}^2 + \gamma_1 \frac{\partial v_{t-1}}{\partial \alpha_1}, \\ \frac{\partial v_t}{\partial \gamma_1} &= v_{t-1} + \gamma_1 \frac{\partial v_{t-1}}{\partial \gamma_1}.\end{aligned}$$

Finally, the partial derivatives of the log-likelihood function can be computed recursively using the above formulas.

2.3.2 Marginal likelihood approach

In our proposed model, time-varying coefficient model (2.24) involves random parameter β , which follows an ARMA-GARCH model with deterministic parameters φ . The marginal likelihood function of the ARMA-GARCH parameters can be obtained by integrating out the random parameter β from the joint log-likelihood function $l(\beta, \varphi)$ in section 2.3. Mathematically the marginal likelihood function for φ can be expressed by

$$L(\varphi) = \int \exp [l(\beta, \varphi)] d\beta. \quad (2.31)$$

This likelihood function involves a high-dimensional integral. It is a difficult computational problem to calculate this marginal likelihood function. The Laplace approximation [32] can be used to estimate this integral.

Let

$$\hat{\beta}(\varphi) = \arg \max_{\beta} l(\beta, \varphi)$$

and

$$H(\varphi) = \frac{\partial^2}{\partial \beta^2} l(\beta, \varphi) \big|_{\beta=\hat{\beta}(\varphi)}.$$

A Taylor expansion of the joint log-likelihood function around the point $\hat{\beta}(\varphi)$ yields

$$l(\beta, \varphi) \simeq l(\hat{\beta}(\varphi), \varphi) - \frac{1}{2}(\beta - \hat{\beta})^T H(\beta - \hat{\beta}) + \dots$$

Then we have

$$\begin{aligned} L(\varphi) &\simeq \int \exp\{l(\hat{\beta}(\varphi), \varphi)\} \exp\left\{-\frac{1}{2}(\beta - \hat{\beta})^T H(\beta - \hat{\beta})\right\} d\beta \\ &= \exp\{l(\hat{\beta}(\varphi), \varphi)\} \sqrt{\frac{(2\pi)^T}{\det H}} \\ &\propto \exp\{l(\hat{\beta}(\varphi), \varphi)\} \sqrt{\frac{1}{\det H}} \end{aligned}$$

Therefore, the Laplace approximation is defined to have the expression

$$L^*(\varphi) = |\det\{H(\varphi)\}|^{-1/2} \exp [l\{\hat{\beta}(\varphi), \varphi\}]. \quad (2.32)$$

Let

$$l^*(\varphi) = \log L^*(\varphi) = -\frac{1}{2} \log |\det\{H(\varphi)\}| + l\{\hat{\beta}(\varphi), \varphi\}.$$

The problem becomes to maximize the Laplace approximation $l^*(\varphi)$ with respect to φ ,

$$\hat{\varphi} = \arg \max_{\varphi} l^*(\varphi).$$

The expression of the Laplace approximation includes a Hessian matrix of $l(\beta, \varphi)$ with respect to β . For our simplified model (2.28) and (2.29), we can calculate this Hessian matrix by differentiating the first partial derivatives of $l(\beta, \delta)$ with respect to β given in (2.30). Let $H = (\frac{\partial^2 l}{\partial \beta_m \partial \beta_s})$.

$$\begin{aligned} \frac{\partial^2 l}{\partial \beta_m \partial \beta_s} = & -\delta_{m,s} \sum_{i=1}^n \left\{ \frac{x_{is}^2 e^{\beta_s x_{is}}}{1 + e^{\beta_s x_{is}}} - \left(\frac{x_{is} e^{\beta_s x_{is}}}{1 + e^{\beta_s x_{is}}} \right)^2 \right\} \\ & - \frac{1}{2} \sum_{t=1}^T \left\{ \frac{\partial^2 v_t}{\partial \beta_m \partial \beta_s} \frac{1}{v_t} - \frac{\partial v_t}{\partial \beta_s} \frac{\partial v_t}{\partial \beta_m} \frac{1}{v_t^2} + 2 \frac{\partial \varepsilon_t}{\partial \beta_m} \frac{\partial \varepsilon_t}{\partial \beta_s} \frac{1}{v_t} + 2 \frac{\varepsilon_t}{v_t} \frac{\partial^2 \varepsilon_t}{\partial \beta_m \partial \beta_s} \right. \\ & \left. - 2 \frac{\varepsilon_t}{v_t^2} \frac{\partial \varepsilon_t}{\partial \beta_s} \frac{\partial v_t}{\partial \beta_m} - 2 \frac{\varepsilon_t}{v_t^2} \frac{\partial \varepsilon_t}{\partial \beta_m} \frac{\partial v_t}{\partial \beta_s} - \frac{\varepsilon_t^2}{v_t^2} \frac{\partial^2 v_t}{\partial \beta_m \partial \beta_s} + 2 \frac{\varepsilon_t^2}{v_t^3} \frac{\partial v_t}{\partial \beta_s} \frac{\partial v_t}{\partial \beta_m} \right\}, \end{aligned}$$

$$\text{where } \delta_{m,s} = \begin{cases} 1 & \text{if } m = s \\ 0 & \text{if } m \neq s. \end{cases}$$

This Hessian matrix involves the first and second derivatives of v_t and ε_t with respect to β . In the previous section, we have

$$\frac{\partial v_t}{\partial \beta_s} = 2\alpha_1 \varepsilon_{t-1} \frac{\partial \varepsilon_{t-1}}{\partial \beta_s} + \gamma_1 \frac{\partial v_{t-1}}{\partial \beta_s},$$

so that

$$\frac{\partial^2 v_t}{\partial \beta_m \partial \beta_s} = 2\alpha_1 \frac{\partial \varepsilon_{t-1}}{\partial \beta_m} \frac{\partial \varepsilon_{t-1}}{\partial \beta_s} + 2\alpha_1 \varepsilon_{t-1} \frac{\partial^2 \varepsilon_{t-1}}{\partial \beta_m \partial \beta_s} + \gamma_1 \frac{\partial^2 v_{t-1}}{\partial \beta_m \partial \beta_s}.$$

Also,

$$\frac{\partial \varepsilon_t}{\partial \beta_s} = \delta_{s,t} - \phi \delta_{s,t-1} \implies \frac{\partial^2 \varepsilon_t}{\partial \beta_m \partial \beta_s} = 0.$$

Now, the Hessian matrix H can be computed recursively for any particular value of β .

After we get the estimate of $\hat{\varphi}$, the estimation for β can be obtained by

$$\hat{\beta} = \arg \max_{\beta} l(\beta, \hat{\varphi}).$$

CHAPTER 3

SIMULATION STUDY

In Section 2.3, we proposed time-varying coefficient models with ARMA-GARCH structures. It can be used to evaluate the time-varying effects of covariates during follow-up periods. Two estimation methods, the maximum likelihood estimation method and the marginal likelihood approach with Laplace approximation, are discussed in Section 2.3.1 and 2.3.2, respectively. They are used to estimate the parameters in the proposed models. To evaluate the performance of these two estimation methods, we will conduct simulation study based on the simplified models (2.28) and (2.29) described in Section 2.3 in the previous chapter. First, in Section 3.1, we will discuss the data generation procedure based on the described models above. We consider two different sets of the true parameters in the AR-GARCH model (2.29). There are considered as two cases and we generate two data sets accordingly. Then in Section 3.2, we will introduce the Kullback-Leibler (KL) divergence which is used to compare the simulation results between different estimation methods. In Section 3.3, the MLE and the marginal likelihood approach are used to estimate the parameters $\{\{\beta_t\}, \phi, \alpha_0, \alpha_1, \gamma_1\}$ in models (2.28) and (2.29). The KL divergence is calculated to compare the estimation of time-varying coefficients $\{\beta_t\}$ from the two estimation methods. To evaluate the estimation of time-series parameters $\{\phi, \alpha_0, \alpha_1, \gamma_1\}$, the root mean square error (RMSE) is employed. Model (2.28) without the ARMA-GARCH part (2.29) is also considered and the estimation of parameters $\{\beta_t\}$ will be discussed. In the last Section 3.4, we will summarize and compare the results obtained from the simulation study.

The simulation study is based on longitudinal data. That is, individuals are measured repeatedly during follow-up periods. To illustrate the performance of the proposed model, we will consider the sample sizes with the number of subjects $n = 200, 500$, and 1000 . Different follow-up times $T = 20, 50$, and 100 will be used for each sample size n . The simulation is

repeated 100 times for each combination of the sample size of subject and the follow-up time and the corresponding KL divergence is calculated and shown for each replication. We will calculate the mean of the KL divergence values for each set of the 100 replications and all of the means will be summarized to compare the performance of the two estimation methods.

3.1 Data Generation

Our considered models (2.28) and (2.29) consist of the logistic regression model and the AR(1)-GARCH(1,1) model. In the logistic regression model, only one covariate is included and the coefficient of the covariate is allowed to change with time. Also intercept is not included in the model. To examine the time-varying coefficients, the AR(1)-GARCH(1,1) model is employed for the estimated coefficients at different time periods. The parameters in the models can be divided into two categories, time-varying coefficients $\beta = \{\beta_t; t = 1, \dots, T\}$ and time-series parameters $\varphi = \{\phi, \alpha_0, \alpha_1, \gamma_1\}$. We first generate the parameters $\{\beta_t; t = 1, \dots, T\}$ according to the following AR(1)-GARCH(1,1) model,

$$\begin{aligned}\beta_t &= \phi\beta_{t-1} + \varepsilon_t \\ \varepsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 \sigma_{t-1}^2,\end{aligned}$$

where z_t is normally distributed with mean zero and variance one. Based on the stationarity constraints of ARMA-GARCH model, we choose the true parameters $\phi = 0.8$ and $\alpha_0 = 1$. For the values of α_1 and γ_1 , we will consider the following two cases:

Case 1: $\alpha_1 = 0.3, \gamma_1 = 0.6$;

Case 2: $\alpha_1 = 0.2, \gamma_1 = 0.2$.

For these two cases, the conditional variance process $\{\sigma_t^2\}$ is stationary and all the unconditional moments of the GARCH(1,1) process are finite. In Case 1, the sum of α_1 and γ_1 is close to one. However, in Case 2, the sum of α_1 and γ_1 is much smaller than that in case 1, which implies that the process $\{\beta_t\}$ is more stable in case 2 than that in case 1. Differences between these two cases will be evaluated in the simulation.

For the initial values of simulation study, we set $\beta_1 = \varepsilon_1$ and $\sigma_1^2 = \alpha_0$, which are obtained by assuming $\beta_0 = 0, \varepsilon_0 = 0$, and $\sigma_0^2 = 0$. Then $\{\beta_t\}$ can be generated recursively based on the AR(1)-GARCH(1,1) model for each of the two cases mentioned above. In each step t ,

first we compute σ_t^2 based on the values in step $(t - 1)$, then ε_t can be generated and used to compute the time-varying coefficient β_t . Figure 3.1 shows the plot of $\{\beta_t\}$ for both Case 1 and Case 2 with $n = 1000$ and $T = 20, 50$, and 100 . Since the simulation is repeated 100 times for each sample size, the values in each plot of the figure are the means of the 100 generated β 's. From this Figure, we can see that the range of the β parameter in Case 1 is wider than that in Case 2 for $T = 20, 50$, and 100 . In addition, the variation of the coefficients over time in Case 1 is larger than that in Case 2 for all considered follow-up time T . These results are expected since GARCH parameters α_1 and γ_1 in Case 1 result in larger variance for β than that in Case 2. Therefore, we should get more stable series of parameters β generated based on the true values in Case 2.

The covariate, $\mathbf{X} = \{x_{it}; i = 1, \dots, n; t = 1, \dots, T\}$, in model (2.29) is simulated by the uniform distribution on the interval $[-1, 1]$. Notice that $\mathbf{X}$ in the simulation study is a matrix with rows denoting different individuals and columns denoting follow-up times. For example, x_{it} is the covariate value for the i^{th} subject at time t . After we have the covariate values $\{x_{it}\}$ and the covariate coefficients $\{\beta_t\}$, the success probability, $p = \{p_{it}\}$, can be calculated using the following formula,

$$p_{it} = \frac{e^{\beta_t x_{it}}}{1 + e^{\beta_t x_{it}}}.$$

Given $\{\beta_t\}$, the responses, $\mathbf{Y} = \{y_{it}\}$, are independent for different individuals and follow-up times and are assumed to follow a Bernoulli distribution with success probability $p = \{p_{it}\}$. Naturally, the response variable $\mathbf{Y}$ is generated by the success probability $\{p_{it}\}$. In this study, $\mathbf{Y}$ is also a matrix in which element y_{it} denotes “success” ($y_{it} = 1$) or “failure” ($y_{it} = 0$) for the i^{th} subject at time t .

In the following sections, two estimation methods, the maximum likelihood method and the marginal likelihood approach with Laplace approximation, are used to estimate the parameters in models (2.28) and (2.29) based on the simulated data. During the estimation procedure, the following constraints are imposed on parameters:

$$-1 < \phi < 1, \quad \alpha_0 > 0, \quad \alpha_1 \geq 0, \quad \gamma_1 \geq 0.$$

To evaluate the performance of the models, we also estimate parameters $\{\beta_t\}$ in the model (2.28) without considering the ARMA-GARCH part, which is a misspecified model with respect to our proposed model. Comparisons among the models are made by calculating the Kullback-Leibler divergence distance that will be introduced in the following section.

Figure 3.1: Plot of the true β with $n = 1000$ and $T = 20, 50$, and 100 .

3.2 Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence is introduced by Kullback and Leibler in 1951 [33] to measure the difference between two probability distributions. The KL divergence can be considered as a kind of distance between the two probability densities. However it is not a real distance measure since it is not symmetric. Specifically, let P represent the “true” distribution of observations, or a precise calculated theoretical distribution. Let Q represent a theory, model, description, or approximation of P . Here we just consider the discrete probability distributions, $P = \{P(1), \dots, P(n)\}$ and $Q = \{Q(1), \dots, Q(n)\}$. Then the definition of discrete KL divergence [33] is

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}, \quad (3.1)$$

where i ranges over all possible data sets.

One of the properties of the KL divergence is that it is non-negative and equal to zero if $P = Q$. The smaller the KL divergence, the more similar are the distributions of the two variables. Therefore, in the simulation study and model comparison, models with smaller values of the KL divergence are better approximations of the true distribution.

For our particular model in (2.29), let $Y = (y_{it})$, P_β denotes the true probability distribution, and $P_{\hat{\beta}}$ denotes the estimated probability distribution. We want to calculate $D_{KL}(P_\beta||P_{\hat{\beta}})$. From the definition of the KL divergence, we have

$$D_{KL}(P||Q) = \sum_Y P(Y) \log \frac{P(Y)}{Q(Y)} = E_P \log \frac{P(Y)}{Q(Y)} = E_P [\log P(Y) - \log Q(Y)].$$

On the other hand,

$$P_{\hat{\beta}}(Y) = \prod_{i=1}^n \prod_{t=1}^T \frac{e^{\hat{\beta}_t x_{it} y_{it}}}{1 + e^{\hat{\beta}_t x_{it}}},$$

$$\log P_{\hat{\beta}}(Y) = \sum_{i=1}^n \sum_{t=1}^T [\hat{\beta}_t x_{it} y_{it} - \log(1 + e^{\hat{\beta}_t x_{it}})].$$

Therefore, we get

$$\begin{aligned} D_{KL}(P_\beta||P_{\hat{\beta}}) &= E_\beta [\log P_\beta(Y) - \log P_{\hat{\beta}}(Y)] \\ &= \sum_{i=1}^n \sum_{t=1}^T [\beta_t x_{it} p_{it}(\beta) - \log(1 + e^{\beta_t x_{it}}) - \hat{\beta}_t x_{it} p_{it}(\beta) + \log(1 + e^{\hat{\beta}_t x_{it}})], \end{aligned} \quad (3.2)$$

where $p_{it}(\beta) = \frac{e^{\beta_t x_{it}}}{1 + e^{\beta_t x_{it}}}$.

The formula of the KL divergence in equation (3.2) will be used to measure the difference between the estimated probability distribution and the true probability distribution and to compare the parameter estimates obtained from different methods.

3.3 Simulation Results

Based on the simulated data in Section 3.1, parameters $\beta = \{\beta_t\}$ and $\varphi = \{\phi, \alpha_0, \alpha_1, \gamma_1\}$ in models (2.28) and (2.29) will be estimated by the maximum likelihood estimation and the marginal likelihood method discussed in Section 2.3.1 and 2.3.2. The MLE method estimates two parts of the parameters together. However, the marginal likelihood method first estimates the time series parameters φ , then parameters β can be estimated based on the estimates of φ . Both methods are based on the joint likelihood function. In Section 2.3, we derived the joint likelihood function $l(\{\beta_t\}, \phi, \alpha_0, \alpha_1, \gamma_1)$ and the gradient of the likelihood function for models (2.28) and (2.29). For the MLE method, the parameters can be estimated by solving the following constrained maximum likelihood problem using a quasi-Newton method. Let $\delta = (\{\beta_t\}, \phi, \alpha_0, \alpha_1, \gamma_1)$.

$$\hat{\delta}_{MLE} = \arg \max_{\delta \in \Delta} l(\delta|Y, X), \quad (3.3)$$

where $\Delta = \mathbb{R}^T \times (-1, 1) \times (0, \infty) \times [0, \infty) \times [0, \infty)$. The L-BFGS-B algorithm in the ‘‘R’’ package is used to solve the constrained nonlinear optimization problem. During the optimization procedure, we use zeros as the initial values for $\{\beta_t\}$ and $(0, 1, 0, 0)$ as the initial values for $(\phi, \alpha_0, \alpha_1, \gamma_1)$.

For the marginal likelihood method, we integrate the random variable β out of the joint likelihood function to get the marginal likelihood function for φ . The Laplace approximation is employed to estimate this integral, which is calculated using the following steps:

Step 1: Evaluate $\hat{\beta}(\varphi) = \arg \max_{\beta} l(\beta, \varphi)$.

Step 2: Evaluate $H(\varphi) = \frac{\partial^2}{\partial \beta^2} l(\beta, \varphi)|_{\beta=\hat{\beta}(\varphi)}$.

Step 3: Compute

$$L^*(\varphi) = |\det\{H(\varphi)\}|^{-1/2} \exp [l\{\hat{\beta}(\varphi), \varphi\}].$$

$\hat{\beta}(\varphi)$ can be computed by maximizing the joint log-likelihood function $l(\beta, \varphi)$ with respect to fixed parameter φ . This can be done by finding a standard nonlinear function optimizer.

Since there is no constraint for β , we use the BFGS algorithm in “R” to find the optimizer. The Hessian matrix H of the log-likelihood function $l(\beta, \varphi)$ with respect to β was derived in Section 2.3.2. Then $H(\varphi)$ can be computed for $\beta = \hat{\beta}(\varphi)$. Finally, we can numerically calculate the Laplace approximation $L^*(\varphi)$. The estimates of φ can be found using the quasi-Newton method by maximizing the Laplace approximation. During the optimization process of the quasi-Newton method, $\hat{\beta}(\varphi)$ will be re-evaluated at each iteration step. After we have the estimates of time series parameters, $\hat{\varphi}$, the estimates of β can be evaluated by

$$\hat{\beta} = \arg \max_{\beta} l(\beta, \hat{\varphi}).$$

As a comparison, we also fit the logistic regression model in (2.28) using the data generated from Section 3.1. This is a misspecified model since the data are generated based on models (2.28) and (2.29) together. The model is fitted at each time t , $t = 1, \dots, T$, using the measurements of the covariate and the dependent variable at time t . The parameter β_t can be estimated by the maximum likelihood estimation. Therefore, we can get the estimated time-varying coefficients, $\hat{\beta} = \{\hat{\beta}_1, \dots, \hat{\beta}_T\}$. Comparing the estimates of β based on this misspecified model and the model used to generate the data, we will assess if our proposed model gets any performance lift using the AR-GARCH model.

Using the formula in equation (3.2), the KL divergence will be calculated in each iteration based on different estimation methods and models. The simulation results for Case 1 and Case 2 are presented in the following Table 3.1 and Table 3.2, respectively. In the tables, three different means of the KL divergence are listed for each combination of number of subjects and follow-up time. GLM denotes the results based on the misspecified model (2.28). Two estimation methods, the maximum likelihood estimation and the marginal likelihood method with Laplace approximation, are used to jointly estimate the two parts of parameters, β and φ , in models (2.28) and (2.29). The means of the KL divergences from these two methods are denoted by MLE and LA, respectively.

Table 3.1 shows that the KL divergence values by the MLE method are usually smaller than these by GLM, and the KL divergence values by the marginal likelihood method are usually smaller than these by the MLE method for each sample size combination of the number of subjects and follow-up times. Paired difference t -test (one-sided test) is used to compare the results of the KL divergence for different methods. The (L-G) t -test is the statistical test between the KL divergence values by LA and GLM, which is used to determine

Table 3.1: Simulation results for Case 1: $\phi = 0.8, \alpha_0 = 1, \alpha_1 = 0.3, \gamma_1 = 0.6$.

subjects	follow-up time	KL Divergence			(L-G)	(L-M)
		GLM	MLE	LA	t-test	t-test
200	20	10.31	10.08	9.17	-6.04	-3.75
	50	27.39	24.92	23.97	-10.05	-2.23
	100	53.55	48.40	47.02	-13.37	-2.42
500	20	10.64	10.65	10.19	-3.72	-2.69
	50	26.04	25.25	24.38	-7.49	-2.53
	100	51.17	48.53	48.42	-8.72	-2.62
1000	20	9.71	9.76	9.64	-0.82	-1.37
	50	25.57	24.80	24.75	-1.92	-1.96
	100	49.56	48.74	48.38	-6.44	-3.30

if the mean of KL divergence by LA is statistically significantly smaller than that by GLM. The results indicate that we get improvements of the β estimates by our proposed model through introducing the AR-GARCH component under the 95% confidence level since most of the t -test values are smaller than -1.66 , which is the one-sided 5% critical value for the t -distribution with 99 degrees of freedom. Similarly, the (L-M) t -test is the test between the LA and MLE methods and is used to compare these two estimation methods. The values of the (L-M) t -test show that method LA is statistically significantly better than the MLE method. On the other hand, we notice that the magnitudes of the (L-G) t -test values become smaller as the number of subjects increases for each fixed follow-up time. That is, the difference between the means of the KL divergence by GLM and LA becomes less significant as the sample size n increases. Actually, this is supposed to be seen because we can obtain rather accurate estimates of parameters by GLM for large sample size. Similarly, for the same number of subjects, the absolute differences between the means of the KL divergence for GLM and the proposed model (LA) becomes larger as the follow-up time T increases, which implies that including the ARMA-GARCH part in the proposed model will improve the goodness-of-fit of the model more significantly when the follow-up time is getting longer. Finally, the (L-M) t -test shows that the marginal likelihood approach with Laplace approximation gives more accurate parameter estimates than those by the MLE method.

Results for Case 2 are presented in Table 3.2, which shows similar patterns as those for

Table 3.2: Simulation results for Case 2: $\phi = 0.8, \alpha_0 = 1, \alpha_1 = 0.2, \gamma_1 = 0.2$.

subjects	follow-up time	KL Divergence			(L-G)	(L-M)
		GLM	MLE	LA	t-test	t-test
200	20	10.29	10.63	9.33	-5.51	-4.52
	50	25.69	26.26	23.41	-9.96	-2.97
	100	50.21	56.49	44.30	-15.30	-4.05
500	20	10.22	10.04	9.88	-4.42	-2.28
	50	25.46	25.04	24.05	-8.60	-1.58
	100	50.02	47.55	47.52	-9.33	0.74
1000	20	10.44	10.24	10.11	-3.20	1.20
	50	26.65	26.38	25.82	-7.21	-1.07
	100	50.43	49.01	49.00	-7.58	-0.51

Case 1. The KL divergence values have the following order $\text{GLM} > \text{MLE} > \text{LA}$, except that the means of the KL divergence by MLE are slightly bigger than that by GLM for the sample size $n = 200$. The (L-G) t -test shows that the improvement of the proposed model comparing to GLM is highly significant based on the marginal likelihood estimation with Laplace approximation. It should be pointed out that the means of the KL divergence from MLE and LA methods for Case 2 are not statistically significantly different for large sample sizes of subjects ($n=1000$), which is different from the results for Case 1. In other words, both the MLE and LA methods give very similar parameter estimates when the number of subjects is large.

As we discussed above, for the estimation of time-varying parameters β in models (2.28) and (2.29), two different estimation methods, the maximum likelihood estimation and the marginal likelihood approach with Laplace approximation, were employed. To gauge the accuracy of the estimation, the KL divergence values were calculated and results were presented in Tables 3.1 and 3.2. In the tables, we did not list all the parameter estimates based on the tow methods for the combinations of the number of subjects and follow-up times since there are too many estimated values to be listed. To compare the estimation by MLE and LA methods, the estimates of time-varying parameters from these two methods are plotted in one figure with true values as the references. Figures 3.2, 3.3, and 3.4 are plots of the true time-varying parameters β (solid line) and estimated parameters $\hat{\beta}$

from the two estimation methods, the maximum likelihood estimation (dotted line) and the marginal likelihood approach with Laplace approximation (dashed line), for the sample sizes of subjects $n = 200, 500$, and 1000 , respectively. From these figures, we can see that the parameter estimates by both methods are very closer to the true values as the follow-up time T increases for all sample sizes of subjects, which is consistent with the testing results listed in Tables 3.1 and 3.2. Overall, the estimates of time-varying parameters by both the MLE and LA methods are close to true values for both cases, and the MLE and LA methods give very similar parameter estimates, especially for large number of subjects.

After we compare the estimates of β from the two estimation methods, we will focus on the estimation of the AR-GARCH parameters, $\varphi = \{\phi, \alpha_0, \alpha_1, \gamma_1\}$. To measure the accuracy of the estimation results, the root mean square error (RMSE) will be employed. The RMSE is defined as the square root of the mean square error (MSE). We know that the MSE is the average of the squared difference between the true and estimated parameters. For example, let $\theta = \{\theta_j, j = 1, 2, \dots, n\}$ denote the true parameter and $\hat{\theta}$ be an estimate of θ , then the MSE of $\hat{\theta}$ with respect to the estimated value θ is defined as,

$$\text{MSE} = \frac{1}{n} \sum_{j=1}^n (\hat{\theta}_j - \theta_j)^2.$$

Therefore, it follows that the RMSE has the form,

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n} \sum_{j=1}^n (\hat{\theta}_j - \theta_j)^2}.$$

The parameter estimates in the AR(1)-GARCH(1,1) model are recorded based on each method and the combinations of the sample size of subjects and the follow-up time for each replication in the simulation study. The RMSE values are calculated based on the parameter estimates. Tables 3.3, 3.4, and 3.5 list the parameter estimates and the corresponding RMSE (values in parenthesis) using the true values in Case 1 for the number of subjects $n = 200, 500$, and 1000 , respectively. From the tables, we can see that the marginal likelihood approach with Laplace approximation (LA) usually produces smaller RMSE than the maximum likelihood method (MLE) for parameters α_1 and γ_1 . In other words, the LA method provides more accurate estimates for parameters in the GARCH component. On the other hand, in many cases it seems that the MLE method gives slightly better estimation results for the parameter ϕ than the LA method. In addition, we can see that the parameter

Figure 3.2: Time-varying parameters β : true (solid), MLE (dotted), LA (dashed)

Figure 3.3: Time-varying parameters β : true (solid), MLE (dotted), LA (dashed)

Figure 3.4: Time-varying parameters β : true (solid), MLE (dotted), LA (dashed)

estimates from both methods are closer to the true values as the follow-up time T increases while the number of subjects does not change. For example, for parameter γ_1 and when $n = 1000$, the RMSE's of the estimates by LA will decrease as the follow-up time increases from 0.395 to 0.282, and then to 0.206. Furthermore, for the same follow-up time, the parameters estimates from both methods improve as the number of subjects n increases. Comparing the estimation results by both the methods for the four estimated parameters, the estimates of α_0 are not less accurate since the values of RMSE for α_0 are much larger than those of other parameters. Overall, both LA and MLE methods work very well to estimate the parameters in the AR-GARCH model.

Table 3.3: Estimates of AR-GARCH parameters for Case 1: $n = 200$.

parameters	true value	T=20		T=50		T=100	
		MLE	LA	MLE	LA	MLE	LA
ϕ	0.8	0.755 (0.218)*	0.719 (0.234)	0.768 (0.133)	0.744 (0.130)	0.801 (0.074)	0.773 (0.084)
α_0	1	1.176 (1.289)	1.341 (1.194)	1.314 (1.049)	1.289 (1.056)	1.149 (0.909)	1.371 (0.812)
α_1	0.3	0.501 (1.759)	0.172 (0.413)	0.487 (0.598)	0.202 (0.266)	0.441 (0.597)	0.255 (0.195)
γ_1	0.6	0.380 (0.463)	0.675 (0.363)	0.464 (0.346)	0.627 (0.299)	0.515 (0.278)	0.567 (0.262)

*: the root mean square error (RMSE)

Tables 3.6, 3.7, and 3.8 present the parameter estimates and the corresponding RMSE (values in parenthesis) using the true values in Case 2 for the number of subjects $n = 200, 500$, and 1000, respectively. We get quite similar results as those for Case 1. The estimates from both the MLE and LA methods are very close to the true values, especially for large number of subjects and longer follow-up times. For parameters α_0 and α_1 , the marginal likelihood method provides more accurate estimation results than MLE in terms of root mean square errors. However, MLE gives better estimation results for parameters ϕ and γ_1 than the marginal likelihood method. When comparing the results among the four parameters, both MLE and LA methods gave quite accurate estimates for the parameter ϕ . On the other hand,

Table 3.4: Estimates of AR-GARCH parameters for Case 1: $n = 500$.

parameters	true value	T=20		T=50		T=100	
		MLE	LA	MLE	LA	MLE	LA
ϕ	0.8	0.665 (0.292)	0.666 (0.270)	0.802 (0.092)	0.778 (0.096)	0.807 (0.070)	0.797 (0.067)
α_0	1	1.128 (1.093)	1.114 (0.965)	1.188 (0.977)	1.259 (0.965)	1.260 (0.998)	1.273 (1.018)
α_1	0.3	0.456 (1.223)	0.234 (0.424)	0.402 (0.291)	0.247 (0.271)	0.274 (0.201)	0.260 (0.186)
γ_1	0.6	0.421 (0.466)	0.642 (0.374)	0.488 (0.286)	0.598 (0.245)	0.561 (0.258)	0.589 (0.236)

Table 3.5: Estimates of AR-GARCH parameters for Case 1: $n = 1000$.

parameters	true value	T=20		T=50		T=100	
		MLE	LA	MLE	LA	MLE	LA
ϕ	0.8	0.723 (0.220)	0.722 (0.213)	0.741 (0.113)	0.774 (0.121)	0.786 (0.065)	0.782 (0.065)
α_0	1	1.199 (1.022)	1.143 (0.990)	1.431 (1.083)	1.258 (1.073)	1.397 (1.049)	1.318 (0.946)
α_1	0.3	0.455 (0.742)	0.233 (0.415)	0.238 (0.291)	0.277 (0.243)	0.304 (0.181)	0.294 (0.182)
γ_1	0.6	0.474 (0.408)	0.607 (0.395)	0.561 (0.298)	0.571 (0.282)	0.529 (0.238)	0.561 (0.206)

estimates of α_0 are the worst among the four parameter estimates, but they are still much closer to the true value of α_0 than those in Case 1. In summary, the parameter estimates from the marginal likelihood method are more reliable than those from the maximum likelihood method. Moreover if we increase the number of replications of simulation study we will get more accurate estimation results for parameters in the AR-GARCH model.

Table 3.6: Estimates of AR-GARCH parameters for Case 2: $n = 200$.

parameters	true value	T=20		T=50		T=100	
		MLE	LA	MLE	LA	MLE	LA
ϕ	0.8	0.756 (0.210)	0.726 (0.211)	0.819 (0.117)	0.754 (0.208)	0.819 (0.107)	0.768 (0.088)
α_0	1	0.640 (0.706)	0.880 (0.618)	0.621 (0.603)	0.838 (0.514)	0.652 (0.571)	0.964 (0.479)
α_1	0.2	1.022 (1.943)	0.204 (0.336)	0.346 (0.601)	0.139 (0.219)	0.305 (0.542)	0.157 (0.153)
γ_1	0.2	0.187 (0.320)	0.322 (0.379)	0.247 (0.343)	0.367 (0.410)	0.173 (0.283)	0.257 (0.336)

Table 3.7: Estimates of AR-GARCH parameters for Case 2: $n = 500$.

parameters	true value	T=20		T=50		T=100	
		MLE	LA	MLE	LA	MLE	LA
ϕ	0.8	0.717 (0.218)	0.707 (0.221)	0.783 (0.112)	0.759 (0.113)	0.787 (0.073)	0.759 (0.105)
α_0	1	0.711 (0.584)	0.711 (0.568)	0.837 (0.513)	0.838 (0.451)	0.849 (0.460)	0.900 (0.479)
α_1	0.2	0.261 (0.655)	0.114 (0.240)	0.364 (0.689)	0.175 (0.208)	0.182 (0.166)	0.195 (0.165)
γ_1	0.2	0.312 (0.369)	0.361 (0.385)	0.259 (0.335)	0.302 (0.343)	0.274 (0.310)	0.287 (0.311)

3.4 Summary and Discussion

There are two main purposes in the simulation study. The first one is to compare our proposed model with the widely used logistic regression model. Specifically, both the logistic regression model in (2.28) and our proposed model are fitted to the data generated from the model specified in Section 3.1, and the differences between the two models were evaluated. The second purpose of the study is to illustrate the performance of two estimation methods, the maximum likelihood estimation and the marginal likelihood approach with Laplace

Table 3.8: Estimates of AR-GARCH parameters for Case 2: $n = 1000$.

parameters	true value	T=20		T=50		T=100	
		MLE	LA	MLE	LA	MLE	LA
ϕ	0.8	0.744 (0.166)	0.719 (0.246)	0.751 (0.210)	0.748 (0.210)	0.775 (0.075)	0.771 (0.076)
α_0	1	0.753 (0.536)	0.772 (0.535)	0.853 (0.468)	1.133 (0.459)	0.847 (0.441)	0.847 (0.441)
α_1	0.2	0.263 (0.444)	0.220 (0.376)	0.248 (0.293)	0.209 (0.227)	0.175 (0.192)	0.165 (0.187)
γ_1	0.2	0.294 (0.363)	0.337 (0.384)	0.239 (0.322)	0.246 (0.326)	0.226 (0.301)	0.238 (0.310)

approximation. There are two parts of parameters in our proposed models (2.28) and (2.29), the time-varying coefficients β and the time-series parameters φ . The Kullback-Leibler (KL) divergence was employed to evaluate the performance of estimating β . The results based on the three approaches, GLM, MLE, and LA, were presented in Tables 3.1 and 3.2. The paired difference t -test between GLM and LA in the tables showed that the improvement for β by our proposed model with the AR-GARCH component is significant. The paired difference t -test between MLE and LA in the tables showed that the marginal likelihood approach with Laplace approximation gives better parameter estimates than the MLE method. To measure the accuracy of the estimation results of time-series parameters φ , the root mean square error (RMSE) was calculated for each estimated parameter based on both LA and MLE methods. The results showed that both methods provided very good estimation results. Overall, the marginal likelihood method gives slightly better parameter estimates than MLE in terms of RMSE, especially for α_1 and γ_1 .

In this simulation study, two cases based on different sets of the true parameters in the AR-GARCH model were considered to generate two groups of data. Similar results were obtained for both cases although we also found few differences among the parameter estimates between the two cases. For example, MLE gave better estimates for the time-series parameters in Case 2 than those in Case 1. To illustrate our proposed modeling procedure in the future, we may choose different sets of the true parameters in the AR-GARCH model

and see whether there are any scheme change in the simulation results as we discussed in this chapter for the two cases.

This simulation study compared the estimation results between our proposed model, logistic regression model with AR(1)-GARCH(1,1) structure, and logistic regression model based on simulated data in Section 3.1. In the future, we may consider the following model in comparison with our proposed model,

$$\begin{aligned}\log \frac{p_{it}}{1 - p_{it}} &= \beta_t x_{it} \\ \beta_t &= \phi \beta_{t-1} + \varepsilon_t,\end{aligned}\tag{3.4}$$

where $\varepsilon_t \sim N(0, 1)$. In model (3.4), we use an AR(1) model for time-varying coefficient $\{\beta_t\}$. Both maximum likelihood and marginal likelihood methods can be used to estimate the parameters for this model based on simulated data in Section 3.1. Then we can compare the KL divergence between the model (3.4) and our proposed model to see if our proposed model will show improved results after adding the GARCH component.

There are some advantages and disadvantages for the two estimation methods we considered, the maximum likelihood estimation and the marginal likelihood approach with Laplace approximation. The MLE method is a very standard procedure in statistical modeling and the results in this simulation study show that MLE usually gives quite accurate parameter estimates, especially for large number of subjects. Another attraction is that the gradient of the likelihood function of our considered models (2.28) and (2.28) is rather easy to derive theoretically. Then the optimization problem is relatively fast to be solved using the L-BFGS-B algorithm in the R program. Comparing with the results obtained by using MLE, the marginal likelihood method with Laplace approximation gives more accurate estimates of the time-varying coefficients and the AR-GARCH parameters since the means of the KL divergences and RMSE by LA are smaller than those by MLE for the two cases in our simulation. However, there are some disadvantages using the marginal likelihood method. It is very difficult to calculate the marginal likelihood function in our proposed model because it involves high-dimension integral. In this study, the Laplace approximation was used to estimate this integral. There still exists some problems, such as, it is difficult to find the first gradient of the Laplace approximation since the Laplace approximation equation involves the determinant. Therefore, it becomes very slow to solve the optimization problem when the Laplace approximation is used.

CHAPTER 4

APPLICATION

In this Chapter we will examine the time-varying effects of potential risk factors, such as age, cigarettes per day (CSM), serum cholesterol (SCL), systolic blood pressure (SBP), and body mass index (BMI), on coronary heart disease (CHD) in the Framingham Heart Study (FHS). Section 4.1 briefly introduces the FHS data. The risk factors or covariates were measured every two years with a follow-up time of more than 40 years. For the first diagnosis of the covariates effects, a logistic regression model is fitted based on measurements of the covariates in a single exam, such as measurements at exam 4. The dependent variable is the occurrence or non-occurrence of the CHD at the end of a follow-up period. To incorporate all repeated observations, a pooled logistic regression model is used to describe the relationship between the risk factors and the development of coronary heart disease, which is discussed in Section 4.2. In Section 4.3, we will apply our proposed model to the FHS data to explore the time-varying covariates effects. First of all, we will study the structures of the time-varying coefficients of covariates and determine a specific ARMA-GARCH model for the estimated coefficients of each covariate. A two-stage method is used to specify the models. We will first fit a multiple logistic model for each two-year examination interval and obtain a time series of coefficients for each covariate. Six different models will be considered for these coefficients and the Bayesian information criterion (BIC) is employed for model selection. Finally, two estimation methods, the maximum likelihood estimation method and the marginal likelihood approach with Laplace approximation, are used to jointly estimate the parameters in the specified models, which are special models of our proposed model in equations (2.24) and (2.25). In Section 4.4, we summarize and compare the results which are obtained from the pooled logistic regression and the proposed model.

4.1 The Framingham Heart Study

Cardiovascular disease (CVD) is one of the leading causes of death and serious illness overall the world. The Framingham Heart Study was designed to focus on identifying major risk factors that contribute to arteriosclerotic (coronary heart disease, angina and stroke) and hypertensive cardiovascular disease. The study was started in 1948 under the direction of the National Heart Institute (now known as the National Heart, Lung, and Blood Institute). The participants had been examined every two years for detailed medical history, physical examination, and laboratory tests. In our study, the endpoint of interest is the initial incidence of coronary heart disease (CHD). At the beginning of the study, 5,209 subjects between 30 to 62 years of age (2873 women and 2336 men) were enrolled. People who had already developed a coronary heart disease were not included in the study. Potential risk factors under consideration in this study include age, cigarettes per day (CSM), serum cholesterol (SCL), systolic blood pressure (SBP), and body mass index (BMI, weight in kilograms/height in meters squared). Table 4.1 summarizes the variables provided in the data set. From this table, we can see that the range of measurements for SBP is from 82 to 300. We will change scale of SBP by dividing the measurements by 10. Scale of the CSM variable (ranging from 0 to 60) is processed in the same manner.

The measurements in the Framingham Heart Study were collected every two years and continued for about 40 years. In our study, measurements from the 20 exams are analyzed. Measurements of covariates made prior to exam 3 are not used since some stress effects were observed in the earlier determinations of some characteristics. Therefore, patients who entered into exam 4 are considered in our study and followed by 34 years from exam 4 to exam 20. There were a total of 4541 (2532 women and 2009 man) subjects at exam 4. After removing patients who got CHD before exam 4, the number of subjects is reduced to 4397. Table 4.2 displays some results at exam 4. Of the 4397 subjects, 1498 or 34.07 percent ultimately developed CHD at the end of the 34-years follow-up period. The percentages of people in the study ultimately developed CHD are 26.72% for female and 43.56% for male, respectively. Disease rate for men is almost twice as that for women, which implies that men are more likely to develop CHD than women.

Figure 4.1 shows boxplots of serum cholesterol (SCL), systolic blood pressure (SBP), age and body mass index (BMI) of patients at exam 4. Covariates measured in exam 4

Table 4.1: Variable Description

Variable	Description	Units	baseline measurements		
			Min	Max	Mean
sex	participant gender	1 = men; 2 = women			
age	age at exam	year	28	62	44.03
scl	serum cholesterol	mg/100ml	96	503	221.2
sbp	systolic blood pressure	mm Hg	82	300	136.9
csm	cigarettes per day		0	60	9.37
hgt	participant height	inches	51	76.5	64.81
wgt	participant weight	pounds	67	300	153.1
BMI	body mass index	kg/m ²	14.12	56.68	25.58
chd	the first exam CHD happened		1	21	11.21
chdfate	ultimately developed CHD	1= ultimately CHD; 0= otherwise.			

Table 4.2: Disease rate for both gender at the end of FHS

	Freq. (%)	CHD* (%)
female	2478 (56.36%)	662 (26.72%)
male	1919 (43.64%)	836 (43.56%)
total	4397 (100%)	1498 (34.07%)

*: patients ultimately developed CHD

are split into groups according to the grouping variable `chdfate` (1 = ultimately CHD, 0 = otherwise). The boxplots show that patients who develop CHD have higher SCL, SBP, age and BMI than others, which indicates that SCL, SBP, age and BMI might be potential CHD risk factors. Then, we fit a multiple logistic regression using measurements of the covariates at exam 4 to examine the effects of the potential risk factors. The dependent variable is the occurrence or non-occurrence of the CHD at the end of the 34-year-follow-up period. Table 4.3 presents the results of logistic regression using covariates' measurements at exam 4, which shows that sex, serum cholesterol, systolic blood pressure and body mass index are the significant variables of CHD incidence in the logistic model. Age is marginally significant with a p -value of 0.07, while CSM is not significant in this analysis. From the table, we can see that the coefficient of sex is negative, which means that males are more likely to get CHD than females, and the result is consistent with the disease rates listed in Table 4.2.

Table 4.3: Logistic Regression based on measurements at exam 4

Variables	Coef.	SE	z value	$\Pr(> z)$
Sex	-0.806	0.076	-10.623	<0.001
SCL	0.008	0.001	9.87	<0.001
SBP	0.095	0.017	5.684	<0.001
CSM	0.042	0.032	1.309	0.191
BMI	0.043	0.009	4.650	<0.001
Age	0.008	0.004	1.812	0.070

In the analysis described above we have used only the covariate measurements at exam 4 as the independent variables. In the following section, the pooled logistic regression method is employed to allow the use of all available measurements in the 20 exams.

4.2 The Pooled Logistic Regression Method

In the Framingham Heart Study, potential risk factors are recorded repeatedly for every two years during the follow-up period. To incorporate all available measurements, the

Figure 4.1: Boxplots of potential risk factors

pooling of repeated observations (PRO) method [34, 35] is used. In our study, each two-year examination interval is treated as a short-term follow-up study. At the beginning of each short-term follow-up study, the measurements of risk factors were recorded and used to predict the disease development at the end of the short-term study. The patients under consideration for each interval are free of the event at the beginning of the intervals. The PRO method pools all observations over multiple intervals into a single sample to evaluate the short-term development of the disease.

In this study, the dependent variable CHD is coded as the first exam at which the disease happened. For example, if chd of a patient is 10, that means the patient was diagnosed with CHD at exam 10, but not at previous exams. The outcome measurement of dependent variable is like a survival time. Figure 4.2 illustrates an example for this kind of data. At the beginning of the study, there are 2000 patients at risk of a disease with a total of 4 examinations. After the first examination interval, 50 patients develop the disease and 60 subjects are lost. Removing these 110 individuals from the population at risk, 1890 subjects followed to next examination interval. The measurements of risk factors among the 1890 subjects are examined at exam 2. Continue this procedure to the end of the study. The PRO method pools all subjects who are at risk in each interval to a single sample with $2000 + 1890 + 1735 = 5625$ subjects and $50 + 55 + 70 = 175$ events. Logistic regression can be employed to this pooled data, which is called the pooled logistic regression method.

Figure 4.2: Pooling of repeated observations

The logistic regression model is used for prediction of the probability of an event occurring in an interval of observations. The pooled logistic regression method is a short-term approach and examine the two-year prediction for disease in our study. However, the logistic regression analysis in Section 3.1 is used for a long-term follow-up study since the patients are followed more than 40 years to record the initial incidence of a disease.

Table 4.4 presents the results of the pooled logistic regression method applied to the Framingham Heart Study. All measurements of the risk factors from exam 3 to exam 19 are used in the regression analysis with the first diagnosis of CHD from exam 4 to exam 20. Hence, there are 17 grouping intervals. Table 4.4 shows that all risk variables under consideration are statistically significant in the pooled logistic regression model. Recalling the results in Table 4.3 in Section 4.1, cigarettes per day (CSM) and age measured at exam 4 are not statistically significant in the logistic model. The results suggest that current age and CSM are more important predictors of the CHD incidence than age and CSM measured at exam 4. Current serum cholesterol, systolic blood pressure and body mass index are also significant predictors of CHD, which is consistent with the results in Table 4.3. The results imply that these risk factors are important covariates for both long-term and short-term study.

Table 4.4: Pooled Logistic Regression from exam 3 to exam 19 for both male and female

Variables	Coef.	SE	z value	$\Pr(> z)$
Sex	-0.742	0.111	-6.686	<0.001
SCL	0.006	0.001	5.232	<0.001
SBP	0.099	0.023	4.290	<0.001
CSM	0.153	0.044	3.467	<0.001
BMI	0.049	0.012	4.119	<0.001
Age	0.049	0.006	8.492	<0.001

To examine if risk factors have the same effects for male and female on the development of CHD, we fit pooled logistic regression model for men and women separately. Table 4.5 and Table 4.6 display the results for male and female, respectively. All risk factors considered

in our study are statistically significant in the pooled logistic model for male, which is consistent with the results for all male and female. However, risk variable CSM is not statistically significant with a p -value of 0.205 in the pooled logistic model for female. The results suggest that the effect of cigarettes per day to the development of CHD is different for male and female.

Table 4.5: Pooled Logistic Regression from exam 3 to exam 19 for male

Variables	Coef.	SE	z value	Pr(> z)
SCL	0.007	0.002	4.092	<0.001
SBP	0.124	0.032	3.897	<0.001
CSM	0.162	0.051	3.163	0.002
BMI	0.042	0.019	2.182	0.029
Age	0.045	0.008	5.795	<0.001

Table 4.6: Pooled Logistic Regression from exam 3 to exam 19 for female

Variables	Coef.	SE	z value	Pr(> z)
SCL	0.005	0.002	3.230	0.001
SBP	0.071	0.034	2.080	0.037
CSM	0.118	0.093	1.267	0.205
BMI	0.055	0.015	3.533	<0.001
Age	0.055	0.008	6.226	<0.001

The PRO method can be applied to the data with repeated measurements on independent variables but with only one measurement on the dependent variable such as death or initial incidence of an interested event. In our study, the dependent variable is the examination time at which the first CHD event occurs. Using the PRO method, all available measurements of

the covariates can be considered to evaluate the relationship between risk factors and short-term disease development. Logistic model and Cox proportional hazard model can be used in the PRO method. In our study, pooled logistic regression model is employed and provide short-term prediction of the probability of CHD incidence using potential risk factors, SCL, SBP, CSM, sex and age. There are some limitations in the PRO method. First, observation intervals should be the same during the follow-up period. Secondly, the effects of risk factors on dependent variable are assumed not changing over time. However, literature reviews show evidence that the effects of covariates on CHD change over time [3, 11]. To capture the time-varying effects of covariates, our proposed model will be applied to the Framingham data.

4.3 Time-varying Coefficient Model with ARMA-GARCH Structure

In this section, our proposed model will be applied to the Framingham Heart Study. In our study, the measurements of covariates from exam 3 to exam 19 are included to predict the events from exam 4 to exam 20. The proposed model in Section 2.3 is a very general model which includes generalized linear models and ARMA-GARCH models. A two-stage method will be used to specify time series models for the coefficients of each considered covariate. In the first stage, a multiple logistic regression model is used to estimate the probability of CHD incidence for each two-year examination interval. The coefficients in the logistic models are allowed to change for different examination intervals. These time-varying coefficients are recorded. In the second stage, different time series models for the time-varying coefficients of each covariate are compared and selected by BIC. After the models are specified, the maximum likelihood and the marginal likelihood estimation methods, discussed in Section 2.3.1 and 2.3.2, are used to estimate the parameters in the proposed model jointly.

4.3.1 Model Specification

Time-varying coefficient models have been used to study CHD events in FHS in the past. Literature [3, 11] shows that the effects of covariates are different for male and female. In the following study, the Framingham data for men and women will be analyzed separately.

Effects of Risk Factors for Male

The result in Table 4.5 showed that the covariates, serum cholesterol (SCL), systolic blood pressure (SBP), cigarettes per day (CSM), Body mass index (BMI) and age, are statistically significant in the pooled logistic regression model for male. However, SCL will not be considered for the time-varying coefficient models in this study because there are no measurements of this variable at exams 11, 12 and 16-19. To apply our proposed time series model, we need to have measurements for each examination interval. Four risk factors, SBP, CSM, BMI, and age will be used to evaluate the incidence of CHD for male in this study.

There were 1919 male subjects aged 34 to 69 at exam 4. They were free of disease at exam 3 and were followed for thirty-two years to exam 20. The measurements of risk factors from exam 3 to exam 19 are used in this study. The multiple logistic regression model will be fitted to examine the effects of the considered covariates to the CHD incidence for each two-year examination interval. The coefficients of covariates are allowed to change for each two-year examination interval. Individuals with missing measurements will be dropped in this study. The coefficients of covariates are recorded and ARMA-GARCH models are identified for these coefficients.

The multiple logistic regression model for male has the form:

$$\log \frac{p_{it}}{1 - p_{it}} = \beta_{0t} + \beta_{1t}x_{1i,t-1} + \beta_{2t}x_{2i,t-1} + \beta_{3t}x_{3i,t-1} + \beta_{4t}x_{4i,t-1}, \quad (4.1)$$

where p_{it} is the conditional probability of observing an event at time t given that the individual is event free at time $(t - 1)$; $x_{1i,t-1}$, $x_{2i,t-1}$, $x_{3i,t-1}$, and $x_{4i,t-1}$ are measurements at time $(t - 1)$ for age, CSM, SBP, and BMI, respectively.

Based on the results from the multiple logistic regression models, we will examine the time series structures for the intercept β_{0t} and coefficients of the four covariates. The following models will be considered for the coefficients and BIC is used for model selection.

Model 1: trend + AR(1)-GARCH(1,1)

$$\begin{aligned} \beta_{kt} &= w_k + \rho_k t + \xi_{kt}, & k &= 0, 1, 2, 3, 4 \\ \xi_{kt} &= \phi_k \xi_{k,t-1} + \varepsilon_{kt}, \\ \varepsilon_{kt} &= \sigma_{kt} z_{kt}, & z_{kt} &\sim N(0, 1), \\ \sigma_{kt}^2 &= \alpha_{k0} + \alpha_{k1} \varepsilon_{k,t-1}^2 + \gamma_{k1} \sigma_{k,t-1}^2. \end{aligned} \quad (4.2)$$

Model 2: trend + AR(1): In model (4.2), set $\alpha_{k1} = 0$ and $\gamma_{k1} = 0$.

Model 3: trend + white noise: In model (4.2), set $\alpha_{k1} = 0$, $\gamma_{k1} = 0$, and $\phi_k = 0$.

Model 4: AR(1)-GARCH(1,1): In model (4.2), set $\rho_k = 0$.

Model 5: AR(1): In model (4.2), set $\rho_k = 0$, $\alpha_{k1} = 0$ and $\gamma_{k1} = 0$.

Model 6: white noise: In model (4.2), set $\rho_k = 0$, $\alpha_{k1} = 0$, $\gamma_{k1} = 0$, and $\phi_k = 0$.

First, let's look at the time series structures for the intercept β_{0t} . Figure 4.3(a) is the time series plot for the estimated intercepts. We can see that there is an obvious upward linear trend. The least square method can be used to estimate this linear trend which is shown in Figure 4.3(b) (dashed line). Figure 4.3(c) is the time series plot after removing the linear trend. Since we just have 17 observations in the time series, and sample size is too small to perform model identification by the usual time series analysis techniques, we will fit the six models described above and compare them by BIC. Table 4.7 shows the values of BIC for these models. The "Trend + white noise" model will be used for the intercept β_{0t} since it has the smallest BIC value of 82.383. In Table 4.7, estimated maximized log-likelihood value ($\log(L)$) for each model is also presented. We can see that estimated maximized log-likelihood values for the models have order: trend + AR(1)-GARCH(1,1) > trend + AR(1) > trend + White noise and AR(1)-GARCH(1,1) > AR(1) > White noise. These results are expected to be seen since the white noise model is nested in the AR(1) model and the AR(1) model is nested in the AR(1)-GARCH(1,1) model.

Table 4.7: Model selection for the intercept β_{0t}

Model	BIC	$\log(L)$
Trend+White Noise	82.383	-36.942
Trend+AR(1)	84.131	-36.399
Trend+AR(1)-GARCH(1,1)	89.076	-36.038
White Noise	123.061	-58.697
AR(1)	111.542	-51.521
AR(1)-GARCH(1,1)	100.484	-43.159

Using the same analysis method above, we evaluate the effect of age over time. There

is an obvious decreasing linear trend in the time series plot of coefficients in Figure 4.4(a). Dashed line in Figure 4.4(b) is the fitted linear trend. After removing the linear trend, a time series plot of the coefficients of age is presented in Figure 4.4(c). Similarly to the model selection procedure for the intercept β_{0t} , values of BIC for different models are listed in Table 4.8, which suggests that the “trend + AR(1)” model has the smallest BIC value of -79.144 thus will be selected.

Table 4.8: Model selection for the age effect

Model	BIC	log(L)
Trend+White Noise	-78.280	43.389
Trend+AR(1)	-79.144	45.238
Trend+AR(1)-GARCH(1,1)	-73.478	45.239
White Noise	-66.346	36.006
AR(1)	-64.346	36.423
AR(1)-GARCH(1,1)	-58.680	36.424

Figure 4.5 presents the effects of CSM for the development of coronary heart disease over time. There are no measurements of this covariate at exams 2, 3, 6 and 16. The measurements in previous exams will be used for the missed values. Specifically, measurements of CSM at exam 1 will be used for those at exams 2 and 3 and measurements at exam 5 will be used for those at exam 6 and so on. Similarly to the analysis of the age-effect, we fit multiple logistic regression model (4.1) and record the time-varying coefficients of CSM. A time series plot of the time-varying coefficients is presented in Figure 4.5(a) which shows that the effect of CSM stays almost the same in the first 30 years and decreases thereafter. The BIC values in Table 4.9 show that the AR(1) model is more appropriate for the coefficients than other models.

Then let’s discuss the effects of SBP on CHD for male. A time series plot of the time-varying coefficients of SBP is presented in Figure 4.6(a). We can see that the coefficient values are almost always above zero, which means that patients with high systolic blood pressure are more likely to develop coronary heart disease. Table 4.10 shows that the “trend

Table 4.9: Model selection for the CSM effect

Model	BIC	log(L)
Trend+White Noise	-11.649	10.074
Trend+AR(1)	-11.717	11.525
Trend+AR(1)-GARCH(1,1)	-6.576	11.788
White Noise	-5.862	5.764
AR(1)	-14.180	11.339
AR(1)-GARCH(1,1)	-8.550	11.358

+ white noise” model has the smallest BIC value comparing to other five models.

Table 4.10: Model selection for the SBP effect

Model	BIC	log(L)
Trend+White Noise	-28.810	18.655
Trend+AR(1)	-25.987	18.660
Trend+AR(1)-GARCH(1,1)	-20.342	18.671
White Noise	-28.437	17.052
AR(1)	-25.647	17.073
AR(1)-GARCH(1,1)	-20.351	17.259

Finally let’s discuss the effects of BMI on CHD. Figure 4.7(a) shows the time series plot of the time-varying coefficients of BMI. in which we can see that the coefficient values are mainly above zero. It means that patients with high body mass index are more likely to develop coronary heart disease. Table 4.11 shows that the “trend + white noise” model is the best model to be used for the BMI effects among the models displayed in the table.

Table 4.11: Model selection for the BMI effect

Model	BIC	log(L)
Trend+White Noise	-49.528	29.014
Trend+AR(1)	-46.763	29.048
Trend+AR(1)-GARCH(1,1)	-42.146	29.573
White Noise	-48.196	26.931
AR(1)	-45.376	26.938
AR(1)-GARCH(1,1)	-40.023	27.095

In summary, the model used for male is

$$\log \frac{p_{it}}{1 - p_{it}} = \beta_{0t} + \beta_{1t}x_{1i,t-1} + \beta_{2t}x_{2i,t-1} + \beta_{3t}x_{3i,t-1} + \beta_{4t}x_{4i,t-1},$$

where

$$\begin{aligned}
\beta_{0t} &= \alpha_0 + \rho_0 t + \varepsilon_{0t}, & \varepsilon_{0t} &\sim N(0, \sigma_0^2), \\
\beta_{1t} &= \alpha_1 + \rho_1 t + \xi_{1t}, & \xi_{1t} &= \phi_{11}\xi_{1,t-1} + \varepsilon_{1t}, & \varepsilon_{1t} &\sim N(0, \sigma_1^2), \\
\beta_{2t} &= \alpha_2 + \xi_{2t} & \xi_{2t} &= \phi_{21}\xi_{2,t-1} + \varepsilon_{2t}, & \varepsilon_{2t} &\sim N(0, \sigma_2^2), \\
\beta_{3t} &= \alpha_3 + \rho_3 t + \varepsilon_{3t}, & \varepsilon_{3t} &\sim N(0, \sigma_3^2), \\
\beta_{4t} &= \alpha_4 + \rho_4 t + \varepsilon_{4t}, & \varepsilon_{4t} &\sim N(0, \sigma_4^2),
\end{aligned} \tag{4.3}$$

where p_{it} is the event probability at time t given that the individual is event free at time $(t-1)$; $x_{1i,t-1}$, $x_{2i,t-1}$, $x_{3i,t-1}$, and $x_{4i,t-1}$ are measurements at time $(t-1)$ for age, CSM, SBP, and BMI, respectively.

Effects of Risk Factors for Female

The results in Table 4.6 in Section 4.2 show that four covariates, serum cholesterol (SCL), systolic blood pressure (SBP), Body mass index (BMI) and age, are statistically significant in the pooled logistic regression model for female. Cigarettes per day (CSM) is not statistically significant and will not be included in the following analysis. In addition, SCL is also not considered in the analysis due to too many missing measurements.

There were 2478 female subjects aged 34 to 68 at exam 4. Similarly to the analysis for male, we first fit multiple logistic regression model using age, BMI, and SBP as covariates for each two-year examination interval. Then we analyze the structure of time-varying coefficients and choose a proper time series model using BIC for the coefficients of each covariate.

The multiple logistic regression model with time-varying coefficients for female is

$$\log \frac{p_{it}}{1 - p_{it}} = \beta_{0t} + \beta_{1t}x_{1i,t-1} + \beta_{2t}x_{2i,t-1} + \beta_{3t}x_{3i,t-1}, \quad (4.4)$$

where p_{it} is the conditional probability of observing an event at time t given that the individual is event free at time $(t - 1)$; $x_{1i,t-1}$, $x_{2i,t-1}$, and $x_{3i,t-1}$ are measurements at time $(t - 1)$ for age, SBP and BMI, respectively.

After we fit the logistic regression models, time-varying coefficients of covariates are recorded. Based on the estimated coefficients, we identify proper models by comparing BIC values of the six fitted models described in the previous section. We first look at the time series structure of the time-varying intercepts β_{0t} . A time series plot in Figure 4.8(a) shows that there is an increasing linear trend for the intercepts. Therefore, a linear regression model may be used for the time-varying coefficients. After removing the linear trend, the de-trended time series is presented in Figure 4.8(c). Six models, “trend + white noise” model, “trend + AR(1)”, “trend + AR(1)-GARCH(1,1)”, white noise model, AR(1), and AR(1)-GARCH(1,1), are considered for this series. BIC is employed for model selection. Table 4.12 presents the BIC values for the six fitted models. Since the “trend + white noise” model has the smallest BIC value, this model will be used for the intercept β_{0t} in this study.

The time-varying effect of age for female is very similar to that for male. There is a decreasing linear trend in the time series plot 4.9(a) and the fitted linear line is presented in Figure 4.9(b). After removing the linear trend, we choose a proper model for the de-trended series 4.9(c). Table 4.13 lists BIC values for the six fitted models. We can see that “trend + white noise” has the smallest BIC value thus is considered as the time series model for the age effect.

For the effect of SBP, Figure 4.10 presents structures of the time-varying coefficients of SBP obtained from the multiple logistic regression model (4.4). From the time series plot of the estimated coefficients in Figure 4.10(a), we can see that almost all values of coefficients are positive, which means that higher systolic blood pressure results in higher risk for female

Table 4.12: Model selection for the intercept β_{0t}

Model	BIC	log(L)
Trend+White Noise	75.194	-33.348
Trend+AR(1)	78.028	-33.347
Trend+AR(1)-GARCH(1,1)	83.489	-33.245
White Noise	129.039	-61.686
AR(1)	114.997	-53.249
AR(1)-GARCH(1,1)	112.422	-49.128

Table 4.13: Model selection for the age effect

Model	BIC	log(L)
Trend+White Noise	-70.386	39.443
Trend+AR(1)	-67.556	39.444
Trend+AR(1)-GARCH(1,1)	-62.331	39.665
White Noise	-60.308	32.987
AR(1)	-60.974	34.737
AR(1)-GARCH(1,1)	-55.560	34.863

patients to get coronary heart disease. Table 4.14 gives the BIC values for different models. The white noise model has the smallest BIC value among these six models listed in Table 4.14, therefore is chosen as the model for the SBP effect.

In the end, let's discuss the time-varying effect of BMI for female. Figure 4.11(a) is the time series plot of coefficients of BMI. From this figure, we can see that the coefficients are mostly positive over time, which means that one with high BMI will be more likely to develop CHD. BIC values in Table 4.15 suggest that white noise model may be used to examine BMI effect on the development of CHD.

Table 4.14: Model selection for the SBP effect

Model	BIC	log(L)
Trend+White Noise	-33.633	21.066
Trend+AR(1)	-32.676	22.005
Trend+AR(1)-GARCH(1,1)	-28.235	22.617
White Noise	-34.823	20.244
AR(1)	-32.885	20.692
AR(1)-GARCH(1,1)	-27.240	20.703

Table 4.15: Model selection for the BMI effect

Model	BIC	log(L)
Trend+White Noise	-60.649	34.575
Trend+AR(1)	-58.082	34.708
Trend+AR(1)-GARCH(1,1)	-52.422	34.711
White Noise	-63.279	34.473
AR(1)	-60.611	34.556
AR(1)-GARCH(1,1)	-54.955	34.561

In summary, the model used for female is

$$\log \frac{p_{it}}{1 - p_{it}} = \beta_{0t} + \beta_{1t}x_{1i,t-1} + \beta_{2t}x_{2i,t-1} + \beta_{3t}x_{3i,t-1},$$

where

$$\begin{aligned}
\beta_{0t} &= \alpha_0 + \rho_0 t + \varepsilon_{0t}, & \varepsilon_{0t} &\sim N(0, \sigma_0^2), \\
\beta_{1t} &= \alpha_1 + \rho_1 t + \varepsilon_{1t}, & \varepsilon_{1t} &\sim N(0, \sigma_1^2), \\
\beta_{2t} &= \alpha_2 + \varepsilon_{2t}, & \varepsilon_{2t} &\sim N(0, \sigma_2^2), \\
\beta_{3t} &= \alpha_3 + \varepsilon_{3t}, & \varepsilon_{3t} &\sim N(0, \sigma_3^2),
\end{aligned} \tag{4.5}$$

where p_{it} is the event probability at time t given that the individual is event free at time $(t - 1)$; $x_{1i,t-1}$, $x_{2i,t-1}$, and $x_{3i,t-1}$ are measurements at time $(t - 1)$ for age, SBP and BMI, respectively.

4.3.2 Model Estimation

In the previous section, multiple logistic regression model and analysis of structures of time-varying coefficients are used to examine the time-varying effects of covariates on the CHD incidence, and the models are specified in equations (4.3) and (4.5) for male and female, respectively. These two models are special cases of our proposed model in equations (2.24) and (2.25). In this section, we will compare our specified models (4.3) and (4.5) to other models with different time series components and perform the model selection based on BIC criteria. Then the maximum likelihood and the marginal likelihood methods discussed in Section 2.3.1 and 2.3.2 will be used to estimate the parameters in the selected models.

Model Estimation for male

First, we draw a comparison between our specified model (4.3) and the following models based on the FHS data:

Model (a): the multiple logistic regression model with “linear trend + AR(1)” model for intercept and each covariate.

Model (b): the multiple logistic regression model with “linear trend + AR(1)-GARCH(1,1)” model for intercept and each covariate.

Model (c): the multiple logistic regression model with “linear trend + white noise” structure for intercept and each covariate.

Model (d): the multiple logistic regression model with AR(1) structure for intercept and each covariate.

BIC values and estimated log-likelihood functions based on the MLE method are calculated to compare these different models. The results are displayed in Table 4.16, which shows that BIC for the specified model is the smallest one among the five models. The estimated log-likelihood values have the order: Specified model < Model (a) < Model (b), as would be expected since Specified model is nested in Model (a) and Model (a) is nested in Model (b).

The above model comparison shows that our specified model (4.3) in Section 4.3.1 performs the best among all the considered models. Now we will do the parameter estimation

Table 4.16: Model comparison for male

Model	BIC	log(L)
Specified model (4.3)	6165.809	-2584.822
Model a	6197.072	-2580.727
Model b	6289.215	-2577.483
Model c	6351.815	-2682.765
Model d	6946.620	-3004.816

using the two methods described before, the maximum likelihood and the marginal likelihood methods. Both of them rely on the joint likelihood function of our considered model. In model (4.3), p_{it} is a conditional probability given that the individual is event free at the time $(t - 1)$, that is, population under consideration at exam t should be event free at exam $(t - 1)$. Based on this condition, the joint log-likelihood function for male has the form,

$$\begin{aligned}
& l(\{\beta_{0t}\}, \{\beta_{1t}\}, \{\beta_{2t}\}, \{\beta_{3t}\}, \{\beta_{4t}\}, \alpha_0, \rho_0, \alpha_1, \rho_1, \phi_{11}, \alpha_2, \phi_{21}, \alpha_3, \rho_3, \alpha_4, \rho_4, \sigma_0^2, \sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) \\
&= \sum_{R_t} \sum_{t=1}^T [y_{it}(\beta_{0t} + \beta_{1t}x_{1i,t-1} + \beta_{2t}x_{2i,t-1} + \beta_{3t}x_{3i,t-1} + \beta_{4t}x_{4i,t-1}) \\
&\quad - \log(1 + e^{(\beta_{0t} + \beta_{1t}x_{1i,t-1} + \beta_{2t}x_{2i,t-1} + \beta_{3t}x_{3i,t-1} + \beta_{4t}x_{4i,t-1})})] - \sum_{k=0}^4 \sum_{t=1}^T \left[\frac{\varepsilon_{kt}^2}{2\sigma_k^2} + \frac{\log \sigma_k^2}{2} \right] + \text{Constant},
\end{aligned} \tag{4.6}$$

where R_t denotes the population at risk at time $(t - 1)$, that is, the individual in risk set R_t is event free at time $(t - 1)$; $T = 17$ in our study; the random error term ε_{kt} is defined in equation (4.3). Based on the log-likelihood function (4.6), parameters in model (4.3) can be estimated by both maximum likelihood and marginal likelihood methods.

In our model, there are two parts of parameters, $\beta = (\{\beta_{0t}\}, \{\beta_{1t}\}, \{\beta_{2t}\}, \{\beta_{3t}\}, \{\beta_{4t}\})$ and $\varphi = (\alpha_0, \rho_0, \alpha_1, \rho_1, \phi_{11}, \alpha_2, \phi_{21}, \alpha_3, \rho_3, \alpha_4, \rho_4, \sigma_0^2, \sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)$. The maximum likelihood method estimates these two parts of parameters together. The marginal likelihood method will first estimate the time series parameters φ , then β can be estimated based on the fitted

parameters $\hat{\varphi}$.

Table 4.17 presents the estimates of time-series parameters based on the maximum likelihood method (MLE). We can see that all parameter estimates are statistically significant at the 95% level except for α_2 , ρ_3 , ρ_4 , σ_1^2 , σ_2^2 and σ_4^2 . The estimate for ϕ_{21} is a borderline case with a p -value of 0.0501.

Table 4.17: Estimates of time-series parameters in the proposed model for male by MLE

Parameters	Estimates	SE	t value	p value
α_0	-11.9980	1.1199	-10.7136	< 0.0001
ρ_0	0.4898	0.1290	3.7975	0.0001
α_1	0.0718	0.0114	6.3045	< 0.0001
ρ_1	-0.0045	0.0013	-3.5157	0.0004
ϕ_{11}	0.4818	0.1124	4.2856	< 0.0001
α_2	0.0699	0.0845	0.8274	0.4080
ϕ_{21}	0.5189	0.2649	1.9589	0.0501
α_3	0.1874	0.0445	4.2132	< 0.0001
ρ_3	-0.0073	0.0047	-1.5332	0.1253
α_4	0.0751	0.0363	2.0685	0.0386
ρ_4	-0.0047	0.0035	-1.3414	0.1798
σ_0^2	0.9368	0.1792	5.2268	< 0.0001
σ_1^2	0.0003	0.0004	0.9741	0.3300
σ_2^2	0.0154	0.0086	1.7771	0.0756
σ_3^2	0.0031	0.0011	2.8951	0.0038
σ_4^2	0.0023	0.0029	0.7874	0.4311
BIC	6165.809	log(L)	-2584.822	

Table 4.18 presents the estimates of time-series parameters based on the marginal likelihood method with Laplace approximation (LA). We get similar results as those in Table 4.17. The estimate for ρ_4 is still not statistically significant at the 95% level. Based on the

simulation results in Section 3.3, marginal likelihood method with Laplace approximation gives more accurate estimates. For the final model and disease prediction, we will use the estimates from LA.

Table 4.18: Estimates of time-series parameters in the proposed model for male by LA

Parameters	Estimates	SE	t value	p value
α_0	-10.3041	0.1533	-67.2153	< 0.0001
ρ_0	0.3333	0.0821	4.0582	< 0.0001
α_1	0.0695	0.0108	6.4081	< 0.0001
ρ_1	-0.0041	0.0012	-3.4142	0.0006
ϕ_{11}	0.4998	0.1664	3.0036	0.0026
α_2	0.1276	0.0351	3.6388	0.0003
ϕ_{21}	0.4941	0.1959	2.5227	0.0116
α_3	0.1826	0.0376	4.8564	< 0.0001
ρ_3	-0.0102	0.0037	-1.9962	0.0459
α_4	0.0753	0.0357	2.1092	0.0349
ρ_4	-0.0029	0.0018	-1.5895	0.1112
σ_0^2	1.2644	0.2615	4.8356	< 0.0001
σ_1^2	0.0001	0.0001	1.7271	0.0842
σ_2^2	0.0519	0.0248	2.0918	0.0365
σ_3^2	0.0137	0.0082	1.6663	0.0957
σ_4^2	0.0016	0.0011	1.4394	0.1501
BIC	6103.224	log(L)	-2972.708	

Figure 4.12 shows the time-varying effects of covariates estimated from the two-stage approach (solid line), the maximum likelihood method (dotted line), and the marginal likelihood method with Laplace approximation (dashed line) based on model (4.3). From these plots, we can see that the overall shapes obtained from these three methods are very similar for each covariate in the considered model, especially for SBP and BMI. However, for

age and CSM, the estimates of the time-varying coefficients by MLE and LA are not very close to those by the two-stage method, but the trends by the three methods have similar patterns. In addition, the estimates of time-varying coefficients from MLE and LA are very similar and are almost identical as showed in the plots. These results are expected since the simulation study in Section 3.3 shows that the estimates of coefficients by MLE and LA are very close to each other for large sample size of subjects in our proposed model. In our study, the number of male subjects is over 1,000. Therefore, we get very similar results of time-varying effects of covariates from both the MLE and LA methods.

Model Estimation for female

Similarly to the model estimation for male, we first perform the model comparison for female. Models (a), (b), (c), and (d), which are described in the previous section, are fitted. BIC values and estimated log-likelihood functions based on the MLE method are computed and listed in Table 4.19. From the table, we can see that model (4.5) has the smallest BIC value among the five models and the estimated log-likelihood values have the order: Specified model < Model (a) < Model (b).

Table 4.19: Model selection for female

Model	BIC	log(L)
Specified model (4.5)	5413.194	-2307.066
Model a	5448.343	-2293.908
Model b	5471.247	-2264.382
Model c	5531.563	-2356.006
Model d	6199.746	-2690.098

Our specified model (4.5) will be used for the following analysis. Using the same

estimation procedure for male, the log-likelihood function based on model (4.5) is

$$\begin{aligned}
& l(\{\beta_{0t}\}, \{\beta_{1t}\}, \{\beta_{2t}\}, \{\beta_{3t}\}, \alpha_0, \rho_0, \alpha_1, \rho_1, \alpha_2, \alpha_3, \sigma_0^2, \sigma_1^2, \sigma_2^2, \sigma_3^2) \\
&= \sum_{R_t} \sum_{t=1}^T [y_{it}(\beta_{0t} + \beta_{1t}x_{1i,t-1} + \beta_{2t}x_{2i,t-1} + \beta_{3t}x_{3i,t-1}) \\
&\quad - \log(1 + e^{(\beta_{0t} + \beta_{1t}x_{1i,t-1} + \beta_{2t}x_{2i,t-1} + \beta_{3t}x_{3i,t-1})})] - \sum_{k=0}^3 \sum_{t=1}^T \left[\frac{\varepsilon_{kt}^2}{2\sigma_k^2} + \frac{\log \sigma_k^2}{2} \right] + \text{Constant},
\end{aligned} \tag{4.7}$$

where R_t denotes the population at risk at time $(t - 1)$, that is, the individual in risk set R_t is event free at time $(t - 1)$; $T = 17$ in our study; the random error term ε_{kt} is defined in equation (4.5). Based on the log-likelihood function (4.7), parameters in model (4.5) can be estimated by both maximum likelihood and marginal likelihood methods. The parameters that need to be estimated include $\boldsymbol{\beta} = (\{\beta_{0t}\}, \{\beta_{1t}\}, \{\beta_{2t}\}, \{\beta_{3t}\})$ and $\boldsymbol{\varphi} = (\alpha_0, \rho_0, \alpha_1, \rho_1, \alpha_2, \alpha_3, \sigma_0^2, \sigma_1^2, \sigma_2^2, \sigma_3^2)$.

The estimates of time-series parameters based on the maximum likelihood method (MLE) are presented in Table 4.20. From the table, we can see that the estimates for α_2 , σ_1^2 and σ_3^2 are not statistically significant at the 95% level with p -values of 0.0882, 0.1300 and 0.0983, respectively. The estimates for other parameters are statistically significant at the 95% level.

Table 4.21 presents the estimates of time-series parameters based on the marginal likelihood method with Laplace approximation (LA). The results are similar to those obtained by MLE. However, from the simulation results, estimates from the LA method will be used for final model and disease prediction.

Figure 4.13 shows the plots of the estimated time-varying coefficients obtained from the three methods, the two-stage approach (solid line), the maximum likelihood method (dotted line), and the marginal likelihood method with Laplace approximation (dashed line), using covariates in model (4.5). From these plots, we can see that the three methods give very similar coefficient estimates for the intercept and all considered covariates. Coefficients estimates by MLE and LA are almost the same, which is consistent with the results for male in Figure 4.12.

Table 4.20: Estimates of time-series parameters in the proposed model for female by MLE

Parameters	Estimates	SE	t value	p value
α_0	-12.9756	1.0433	-12.4371	< 0.0001
ρ_0	0.3873	0.1079	3.5883	0.0003
α_1	0.0935	0.0172	5.4319	< 0.0001
ρ_1	-0.0042	0.0017	-2.4995	0.0124
α_2	0.0832	0.0488	1.7052	0.0882
α_3	0.0219	0.0090	2.4419	0.0146
σ_0^2	1.0514	0.2613	4.0232	< 0.0001
σ_1^2	0.0003	0.0002	1.5140	0.1300
σ_2^2	0.0407	0.0157	2.5893	0.0096
σ_3^2	0.0002	0.0001	1.6532	0.0983
BIC	5413.194	log(L)	-2307.066	

4.4 Summary and Discussion

To study the effects of potential risk factors on the CHD development in the Framingham Heart Study data, we first fitted a multiple logistic regression model using a single-measures of covariates at exam 4 in Section 4.1. The results showed that sex, serum cholesterol (SCL), systolic blood pressure (SBP), and body mass index (BMI) are statistically significant risk factors. In FHS, measurements of risk factors were recorded repeatedly for every two years and followed more than 40 years. To incorporate all available measurements, a pooled logistic regression model was employed in Section 4.2. This model is a short-term prediction model treating each observation interval as a mini-follow-up study. In the model, the effects of risk factors are assumed to be the same for every observation interval in the pooled logistic regression model.

To examine the time-varying effects of risk factors, our proposed model was applied, which is a very general model. In our study, BIC was used for model selection among

Table 4.21: Estimates of time-series parameters in the proposed model for female by LA

Parameters	Estimates	SE	t value	p value
α_0	-12.9680	0.1599	-81.1007	< 0.0001
ρ_0	0.3326	0.0092	36.1522	< 0.0001
α_1	0.0989	0.0078	12.6795	< 0.0001
ρ_1	-0.0044	0.0003	-14.6667	< 0.0001
α_2	0.1344	0.0187	7.1872	< 0.0001
α_3	0.0399	0.0073	5.4657	< 0.0001
σ_0^2	0.9944	0.0960	10.3583	< 0.0001
σ_1^2	0.0002	0.0002	0.8834	0.3770
σ_2^2	0.0176	0.0084	1.7024	0.0887
σ_3^2	0.0005	0.0002	2.0599	0.0394
BIC	5355.472	log(L)	-2626.514	

different time series models with time-varying effects of risk factors. First, multiple logistic regression model was fitted and time-varying coefficients were recorded. Then we analyzed the structures of coefficients to choose proper time series models. We compared six models, linear trend + white noise, linear trend + AR(1), linear trend + AR(1)-GARCH(1,1) model, white noise, AR(1), and AR(1)-GARCH(1,1) model. The model with the smallest BIC value was selected for final model. In the future, we may fit more models and compare the results to get more accurate prediction for disease development.

After we specified models described in models (4.3) and (4.5) for male and female, BIC was used to compare these models to other four different models with different time series components. The results showed that our specified models (4.3) and (4.5) performed the best among these five considered models. Then two estimation methods, the maximum likelihood method and the marginal likelihood method with Laplace approximation, were used to estimate the parameters in our specified models. The results in Section 4.3.2 showed that both methods gave very similar parameter estimates. However, estimates from the

marginal likelihood method were used to predict disease incidence since simulation results suggest that the marginal likelihood method give more accurate estimates.

In this study, we evaluated the effects of risk factors on CHD incidence for male and female separately. The effects of risk variables for male and female were showed in Figures [4.12](#) and [4.13](#). The effects of risk factors on CHD are different for male and female. In addition, we can see that the estimated coefficients of SBP and BMI are mostly positive for both male and female, which implies that high SBP and BMI result in high risk of CHD for both male and female.

Figure 4.3: Time series structure of intercept for male

Figure 4.4: Time series structure of the coefficients of age for male

Figure 4.5: Time series structure of the coefficients of CSM for male

Figure 4.6: Time series structure of the coefficients of SBP for male

Figure 4.7: Time series structure of the coefficients of BMI for male

Figure 4.8: Time series structure of intercept for female

Figure 4.9: Time series structure of the coefficients of age for female

Figure 4.10: Time series structure of the coefficients of SBP for female

Figure 4.11: Time series structure of the coefficients of BMI for female

Figure 4.12: Estimates of time-varying coefficients based on the proposed model for male

Figure 4.13: Estimates of time-varying coefficients based on the proposed model for female

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

In this dissertation, we developed time-varying coefficient models with ARMA-GARCH structures to examine the time-varying effects of covariates on dependent variables. In Chapter 1, we gave introduction and motivation of our study. In Chapter 2, we reviewed varying-coefficient models and time series models, and discussed their properties and estimation procedures. Then we proposed time-varying coefficient models with ARMA-GARCH structures and discussed two estimation methods, the maximum likelihood method and the marginal likelihood approach with Laplace approximation, based on our proposed model. Since our proposed model involves a lot of parameters, we provided a simple example to illustrate the estimation procedures, which consists of logistic regression model with a single covariate:

$$\log \frac{p_{it}}{1 - p_{it}} = \beta_t x_{it}, \quad (5.1)$$

where the response variable y_{it} follows Bernoulli distribution with success probability p_{it} . There is only one covariate x_{it} considered in this study. The coefficient β_t is allowed to change with time, which is more flexible than the fixed coefficient models. Furthermore, an AR(1)-GARCH(1,1) model is used for the time-varying coefficient β_t ,

$$\begin{aligned} \beta_t &= \phi \beta_{t-1} + \varepsilon_t \\ \varepsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 \sigma_{t-1}^2, \end{aligned} \quad (5.2)$$

where $z_t \sim N(0, 1)$. Simulation study is conducted in Chapter 3 based on models (5.1) and (5.2). In the simulation study, we investigated and compared the performance of two estimation methods, the maximum likelihood and the marginal likelihood methods. Two cases were considered based on different true parameters in model 5.2. Then two data sets

were generated according to these two cases. The results for both cases showed that marginal likelihood approach usually gave more accurate estimates of time-varying coefficients $\{\beta_t\}$ and time-series parameters $\{\phi, \alpha_0, \alpha_1, \gamma_1\}$. In Chapter 4, our proposed model was applied to the Framingham Heart Study (FHS) to evaluate the time-varying effects of potential risk factors, such as age, cigarettes per day (CSM), serum cholesterol (SCL), systolic blood pressure (SBP), and body mass index (BMI), on coronary heart disease (CHD). The results showed that the effects of risk factors on CHD are different for male and female. For male, a linear regression with AR(1) model was used for time-varying coefficients of age. To examine the time-varying effects of CSM, an AR(1) model was used. The effects of SBP and BMI followed a linear trend model. For female, a linear regression with an AR(1) model was also used to evaluate the effect of age. For the effects of SBP and BMI, the white noise model was used.

In my future research, the following topics will be considered. The first issue is to improve the efficiency of computation algorithms. In Chapter 2, two estimation methods were discussed to estimate the parameters in our proposed model: the maximum likelihood estimation and the marginal likelihood approach with Laplace approximation. Both methods are implemented in R. However, the program for marginal likelihood method runs very slowly. We need to find a way to expedite the program, such as calculating the gradient of the Laplace approximation equation (2.32). After speeding up the program, we will be able to increase the number of replications to obtain more accurate results.

Secondly, simulation will be conducted based on more examples of our proposed model (2.24) and (2.25) to illustrate model performance. In our simulation study, models (5.1) and (5.2) with only one covariate was studied to examine the time-varying relationships between covariates and response variable. In the future, more covariates will be considered, such as the model with two covariates,

$$\log \frac{p_{it}}{1 - p_{it}} = \beta_{1t}x_{1it} + \beta_{2t}x_{2it} \quad (p_{it} = \frac{e^{\beta_{1t}x_{1it} + \beta_{2t}x_{2it}}}{1 + e^{\beta_{1t}x_{1it} + \beta_{2t}x_{2it}}}), \quad (5.3)$$

where,

$$\begin{cases} \beta_{1t} = \phi_1\beta_{1,t-1} + \varepsilon_{1t} \\ \varepsilon_{1t} = \sigma_{1t}z_t \\ \sigma_{1t}^2 = \alpha_{10} + \alpha_{11}\varepsilon_{1,t-1}^2 + \gamma_{11}\sigma_{1,t-1}^2, \\ \beta_{2t} = \phi_2\beta_{2,t-1} + \varepsilon_{2t} \\ \varepsilon_{2t} = \sigma_{2t}z_t \\ \sigma_{2t}^2 = \alpha_{20} + \alpha_{21}\varepsilon_{2,t-1}^2 + \gamma_{21}\sigma_{2,t-1}^2. \end{cases}$$

The likelihood function based on this model is complicated, but the parameters can still be estimated by the same methods discussed in Section 2.3.1 and 2.3.2. In another direction, AR(1)-GARCH(1,1) model in (5.2) can be extended to more flexible ARMA-GARCH models. Also in our current study, the random variable z_t in model (5.2) was assumed to be normally distributed with mean zero and variance one. We may relax this assumption with variance σ^2 instead of one. Through our simulation studies, σ^2 can be estimated based on the two estimation methods, the maximum likelihood estimation and the marginal likelihood approach with Laplace approximation.

Finally, more time series models will be considered for time-varying coefficients of covariates in the application part. In Section 4.3, six models, linear trend + white noise, linear trend + AR(1) model, linear trend + AR(1)-GARCH(1,1) model, white noise, AR(1) model, and AR(1)-GARCH(1,1) model, were considered to evaluate the time-varying effects of covariates on CHD incidence. We may consider more time series models, such as AR(2) model, AR(2)-GARCH(1,1) model and so on, and select one of them by using BIC. Furthermore, more applications using our proposed model will be performed. In this dissertation, our proposed model was applied to the Framingham Heart Study (FHS) for evaluating time-varying effects of risk factors on incidence of coronary heart disease (CHD). In the future, we will examine whether there are time-varying relationships between risk factors and death, or other specified disease incidence, such as Cardiovascular Disease (CVD). Furthermore, we may apply our proposed model on other studies, such as, Multiple Risk Factor Intervention Trial (MRFIT) [36], The Strong Heart Study [37], and Chicago Heart Association Detection Project [38, 39] etc.

REFERENCES

- [1] D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society*, 34:187–220, 1972. [1](#), [2.1.1](#)
- [2] S.H. Walker and D.B. Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54:164–179, 1976. [1](#)
- [3] M. Wu and J.H. Ware. On the use of repeated measurements in regression analysis with dichotomous responses. *Biometrics*, 35(2):513–521, 1979. [1](#), [1](#), [4.2](#), [4.3.1](#)
- [4] T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796, 1993. [1](#), [2.1](#), [2.1.1](#)
- [5] R. S. Tsay. *Analysis of Financial Time Series*. Second edition, 2005. [1](#), [1](#)
- [6] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods*. Springer, second edition, 1991. [1](#), [2.2.1](#)
- [7] G. Box G.M. Jenkins and G. Reinsel. *Time Series Analysis: Forecasting and Control*. Third edition, 1994. [1](#), [2.2.1](#)
- [8] R.F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflations. *Econometrica*, 50:987–1007, 1982. [1](#), [2.2.2](#)
- [9] T. Bollerslev. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31:307–327, 1991. [1](#), [2.2.2](#), [2.2.2](#)
- [10] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Second edition, 1997. [1](#)
- [11] J. He. Bayesian dynamic survival models for longitudinal aging data. 2007. [1](#), [4.2](#), [4.3.1](#)
- [12] W. P. Castelli K. M. Anderson and D. Levy. Cholesterol and mortality: 30 years of follow-up from the framingham study. *JAMA*, 257(16):2176–2180, 1987. [1](#)
- [13] F. Canova. Modelling and forecasting exchange rates with a bayesian time-varying coefficient model. *Journal of Economic Dynamic and Control*, 17:233–261, 1993. [1](#)
- [14] Y. Chang and J. Phillips. Nonlinear econimetric models with cointegrated and deterministically trending regressors. *Journal of Econometrics*, 4:1–36, 2001. [1](#)

- [15] Z. Cai. Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics*, 136:163–188, 2007. [1](#)
- [16] M. C. Jensen F. Black and M. Scholes. The capital asset pricing model: Some empirical tests. *M. Jensen ed., Studies in the Theory of Capital Markets*, pages 79–121, 1972. [1](#)
- [17] E. Ghysels. On stable factor structures in the pricing of risk: do time varying betas help or hurt. *Journal of Finance*, 53:549–574, 1998. [1](#)
- [18] M. Reyes. Size time-varying beta and conditional heteroscedasticity in uk stock return. *Review of Financial Economics*, 8:1–10, 1999. [1](#)
- [19] K. Wang. Asset pricing with conditioning information: a new test. *Journal of Finance*, 58:161–196, 2003. [1](#)
- [20] Y. Chang and E. Martinez-Chombo. Electricity demand analysis using cointegration and error-correction models with time varying parameters: the mexican case. Working Paper, Department of Economics, Rice University, 2003. [1](#)
- [21] C.O. Wu D.R. Hoover, J.A. Rice and L. Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822, 2003. [2.1](#)
- [22] R. Chen and R.S. Tsay. Functional-coefficient autoregressive models. *ournal of the American Statistical Association*, 88:298–308, 1993. [2.1](#)
- [23] Y. Hong and T. Lee. Inference and forecast of exchange rates via generalized spectrum and nonlinear times series models. *Manuscript*, 1999. [2.1](#)
- [24] R.L. Eubank. Spline smoothing and nonparametric regression. *Journal of the American Statistical Association*, 84(406), 1989. [2.1.1](#)
- [25] D.J. Sargent. A flexible approach to time-varying coefficients in the cox regression setting. *Journal of the American Statistical Association*, 3:13–25, 1996. [2.1.1](#)
- [26] A. Agresti. *Categorical data analysis*. Wiley, second edition, 2002. [2.1.2](#)
- [27] D.B. Nelson. Stationary and persistence in garch(1,1) model. *Econometric Theory*, 6:318–334, 1991. [2.2.2](#)
- [28] P. Bougerol and N. Picard. Strict stationary of generalized autoregressive process. *The annals of probability*, 20:1714–1730, 1992a. [2.2.2](#)
- [29] P. Bougerol and N. Picard. Stationarity of garch processes and of some nonnegative time series. *J. Econometrics*, 52:115–127, 1992b. [2.2.2](#)
- [30] C. Francq and J. Zakoian. Maximum likelihood estimation of pure garch and arma-garch processes. *Bernoulli*, 10(4):605–637, 2004. [2.2.2](#)

- [31] L. Horvath I. Berkes and P. Kokoszka. Garch processes: structure and estimation. *Bernoulli*, 9(2):201–227, 2003. [2.2.2](#)
- [32] H. Skaug and D. Fournier. Automatic approximation of the marginal likelihood in non-gaussian hierarchical models. *Computational Statistics and Data Analysis*, 51:699–709, 2006. [2.3.2](#)
- [33] S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. [3.2](#)
- [34] K. Anderson L.A. Cupples, R.B. D’Agostino and W.B. Kannel. Comparison of baseline and repeated measure covariate techniques in the framingham heart study. *Statistics in Medicine*, 7:205–218, 1988. [4.2](#)
- [35] M. Lee R.B. D’Agostino and A.J. Belanger. Relation of pooled logistic regression to time dependent cox regression analysis: the famingham heart study. *Statistics in Medicine*, 9:1501–1515, 1990. [4.2](#)
- [36] J.D. Neaton and D. Wentworth. Serum cholesterol, blood pressure, cigarette smoking, and death from coronary heart disease: overall findings and differences by age for 316,099 white men. *Arch Intern Med*, 152:56–64, 1992. [5](#)
- [37] R.R. Fabsitz R. Devereux L. Best T.K. Welty W. Wang, E.T. Lee and B.V. Howard. A longitudinal study of hypertension risk factors and their relation to cardiovascular disease: The strong heart study. *Journal of The American Heart Association*, 47:403–409, 2006. [5](#)
- [38] J. Stamler A.R. Dyer and R.B. Shekelle. Serum cholesterol and mortality from coronary heart disease in yonger, middle-aged, and older men and women. *Annals of Epidemiology*, 2:51–57, 1992. [5](#)
- [39] C. Beam E.L. Navas-Nacher, L. Colangelo and P. Greenland. Risk factors for coronary heart disease in men 18 to 39 years of age. *Ann Intern Med*, 134(6):433–439, 2001. [5](#)

BIOGRAPHICAL SKETCH

Haiyan Zhao

Haiyan Zhao was born in the north of China, and received her BS degrees from the University of Science and Technology of China. She came to US in 2003 and obtained her MS degree in mathematics from the State University of New York at Stony Brook in 2005. In Sep. 2005, she enrolled into the Ph.D. program in the Department of Statistics of the Florida State University. She began working on her dissertation under the guidance of Professor Xufeng Niu and Professor Fred Huffer in 2006.

Haiyan's research interests include time-varying coefficient models, time series analysis, survival analysis, longitudinal data analysis.