1-26-2015

Multivariate Longitudinal Data Analysis for Actuarial Applications

Priyantha K. Hewa Katuwandeniyage
priyantha_hk@yahoo.com

Multivariate Longitudinal Data Analysis for Actuarial Applications

Priyantha K. Hewa Katuwandeniyage, Ph.D.

University of Connecticut, 2015

ABSTRACT

Analysis of longitudinal data has increased in popularity in recent years for several disciplines that is commonly used to understand the dynamic nature and the heterogeneity within and among subjects. There has been a much more rapid progress of longitudinal analysis for univariate data. However, there is a developing interest of extending the longitudinal framework to handle multivariate responses for obvious reasons: to capture dependence structure of the responses and thereby to increase the efficiency of the model. Actuarial applications in this area are very limited at the moment and it is our hope to contribute to this developing literature. Most work has focused on the assumption of multivariate normal for the joint responses; we propose a more flexible framework of using copula functions to integrate the dependence among responses and the classical random effects approach to identify intertemporal dependence within a subject and unobservable subject-specific heterogeneity among observations. Covariate information is taken into account for observable subject-specific effects through the regression model for the marginals.

For empirical illustration, we analyzed two datasets which are directly related with the insurance industry. Our first data set is used to understand the global

insurance demand in both life and non-life insurance. Simultaneously, we used the proposed models to understand the association between these two insurance lines. Loss triangles corresponding to four insurance lines have been considered under the second data set. We transformed loss triangle data into the longitudinal framework to apply the above mentioned new method. In both empirical studies, Archimedean and Elliptical family copulas are incorporated. To illustrate the flexibility of the proposed model, we have considered different skewed distributions, such as lognormal, GB2, and Weibull.

Multivariate Longitudinal Data Analysis for Actuarial Applications

Priyantha K. Hewa Katuwandeniyage

M.S., University of Connecticut, 2008

B.S., University of Sri Jayewardenepura, 2005

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2015

APPROVAL PAGE

Doctor of Philosophy Dissertation

Multivariate Longitudinal Data Analysis for Actuarial Applications

Presented by

Priyantha K. Hewa Katuwandeniyage, B.S., M.S.

Major Advisor _____
Professor Emiliano A. Valdez

Associate Advisor _____
Professor James G. Bridgeman

Associate Advisor _____
Professor Brian M. Hartman

University of Connecticut

2015

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to everyone who have contributed and extended support during the completion of this thesis.

Foremost, I would like to express the deepest appreciation and thanks to my major advisor, Professor Emiliano A. Valdez, for his continued and untiring support towards my Ph.D research. His excellent guidance, patience, enthusiasm, immense knowledge, and faith in me throughout this process have been extremely helpful to finish this work. Professor Valdez is a truly an inspiration for me.

Besides my major advisor, I would like to thank Professors James H. Bridgeman and Brian M. Hartman for being extraordinary committee members who were more than generous with their expertise and precious time. In addition, I could not be thankful enough, to Professor Jay Vadiveloo, who helped me to gain working experience through actuarial projects and guidance to relate my research work with actuarial applications.

Next on my list to thank is the Department of Mathematics of the University of Connecticut, together with its Quantitative Learning Center for generously providing me financial assistance throughout my graduate studies. Special thanks go to Professors Thomas Roby, Sarah Frey, David Gross, and Reed Solomon for arranging the financial support especially at times when I needed it most.

I am also deeply humbled by the help during the completion of my Ph.D. extended to me by several other people connected with the Department of Mathematics. Tops

on this list are the support and help by Monique Roy, Tammy Prentice, and Cara Light.

I am indebted to the many friendly and bright colleagues in the department and the university who also extended support in many indirect ways. Here I want to acknowledge the following people: Asiri, Shirani, Milanthi, Bernardo, Gao, Rozita, David, Gi, Sudath, Chandrika, Priyanga, Amali, and Januka.

A special thank goes to William J. Thompson, Stephen J. Kaczmarek, and Andrea Sheldon for giving me the opportunity to work at Milliman Hartford Health and to help secure my actuarial career.

I would like to thank members of my family for their endless love and support throughout my life, especially their understanding for having a limited time to lovingly share while completing my dissertation. Special thanks go to my mother Yasawathie Uwana Hweage and my father Somapala Hewa Katuwandeniyage, sister Nadeere Hewa Katuwandeniyage, brother-in-law Prasanna Katuwandeniya, niece Ruwindi Katuwandeniya, and nephew Omal Katuwandeniya as well as my in-laws.

Last but not least, I would like to thank my loving wife, Ushani Dias, for her remarkable support and sharing knowledge while we together go through this difficult process of completing the thesis.

Contents

Chapter 1

Introduction

The natural approach to model construction of cross-sectional data over time is the use of longitudinal analysis, which allows for understanding the dynamic relationship that evolves within a given subject while simultaneously exploring cross-sectional heterogeneity among the observations. Over the past several decades, longitudinal analysis has gained popularity not only within the statistics discipline, but also among several other disciplines including medical statistics, finance, and insurance where its widespread potential applications are apparent. The primary motivation for the statistical analysis of time sensitive data has been typically to understand changes that evolve over time. To illustrate, in many branches of insurance and financial institutes, companies continue to implement strategies to manage their portfolio of products in order to remain competitive and actuaries analyze the impact of these strategies in their claims over time to establish equitable pricing.

When the primary variable of interest is a single dimension, this is commonly referred to as *univariate longitudinal data* analysis. In the statistics literature, there

is a toolbox of classes of statistical models, including mixed-effects models, multi-level models, hierarchical models, to name a few, available in analyzing longitudinal data. Many of these statistical methodologies are very well developed for univariate longitudinal data. In the early development of univariate longitudinal data models, for example, a subject i for $i = 1, 2, \ldots, N$ is observed over a period of time t for $t = 1, 2, \ldots, T$ and the simple mixed-effects ANOVA model may be expressed as

$$y_{it} = b_i + \mathbf{X}'_{it}\beta + \varepsilon_{it}, \tag{1.0.1}$$

where y_{it} is the one-dimensional primary outcome of interest, $\mathbf{X}'_{it}$ is a set of observable covariates, β is a vector of regression coefficients. Here b_i is the random effects and ε_{it} is the disturbance term. In a typical assumption, $b_i \sim N(0, \sigma_b)$ which has the interpretation of explaining the presence of heterogeneity among the observations not measured by the presence of the covariates, and $\varepsilon_{it} \sim N(0, \sigma_\varepsilon)$. Due to the fact that above model can control the heterogeneity of individual subject within the data, the model is also known as heterogeneous model. See Fitzmaurice et al. (2008) for a comprehensive review of the origins of longitudinal data models. The book of Singer and Willett (2003) additionally provides for an appreciation of the applications of longitudinal data models in various disciplines.

In actuarial science, the main outcome of interest y_{it} represents claims (either number of times or amount, depending on the modeling situation), made by policy-holder i in a portfolio of N policies over a period of time t. Here the data may be unbalanced if the time period of observation is not the same for all observations. It is not uncommon for an insurer to take into account the history of claims of policy-holders to assess the premium for the following. Such is called *credibility ratemaking*

and the corresponding models, referred to as *credibility models*, lead to a premium that accounts for the variation due to the policyholder's experience and that due to the overall portfolio of policies. Frees, Young, and Luo (1999) demonstrated that the class of credibility models studied in actuarial science falls within the more general framework of longitudinal data models. In particular, Frees, Young, and Luo (1999) examined the link between the class of linear mixed-effects models expressed by

$$y_{it} = \mathbf{Z}'_{it}\alpha_i + \mathbf{X}'_{it}\beta + \varepsilon_{it} \tag{1.0.2}$$

and common actuarial credibility models. Here, the $\mathbf{Z}_{it}$ is a set of covariates with policyholder-specific random effects parameters α_i. With only an intercept in the random effects term, equation (1.0.1) is indeed a special case of equation (1.0.2). The linear mixed-effects model is undoubtedly the most widely used longitudinal data model and can be traced track to the original work of Laird and Ware (1982).

The disturbance component in the longitudinal data framework as expressed in (1.0.2) is typically assumed to follow a Gaussian multivariate distribution. However, such restriction can easily be relaxed using a more general multivariate distribution function based on copulas; Frees and Wang (2005) and Frees and Wang (2006) explored these extensions to the copula framework. Furthermore, Shi and Frees (2010) and Shi and Frees (2011) examined additional actuarial applications of longitudinal data models. The book by Frees (2004) provides for a comprehensive theory and applications of longitudinal data models. Author describes the longitudinal analysis as a combination of typical regression and time series analysis. This concept envision more actuarial related applications within the longitudinal framework as many actuarial related applications can be found on regression, and time series context. For

example, we can consider the demand of global insurance over countries as a cross sectional study, regression model, while dynamic change in insurance demand over time analyze with time series models. Clearly, it is worth to analyze global insurance demand using a longitudinal model, which could address both cross sectional and dynamic analysis together.

Today, it is not uncommon to find situations where the outcome of interest comes in the form of more than a single dimension. Such cross-sectional data across time with multiple response outcomes can be considered as an extension of univariate longitudinal data and have been known as *multivariate longitudinal data*. Consider the case where we have a set of N subjects observed over T time periods for a set of m responses. Here, the formulation is for the case of a balanced dataset, but the extension should be straightforward, although messy, to the unbalanced case. Now let y_{itk} denote the observation from i-th individual in t-th time period on the k-th response. Hence, for a given subject i, the matrix

$$\mathbf{Y}_i = \begin{pmatrix} y_{i11} & y_{i21} & \cdots & y_{iT1} \\ y_{i12} & y_{i22} & \cdots & y_{iT2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{i1m} & y_{i2m} & \cdots & y_{iTm} \end{pmatrix}$$

represents observations over T time periods corresponding to m number of response variables. By letting $\mathbf{y_{it}} = (y_{it1}, y_{it2}, \ldots, y_{itm})'$ for $t = 1, 2, \ldots, T$, we can express $\mathbf{Y}_i = (\mathbf{y_{i1}}, \mathbf{y_{i2}}, \ldots, \mathbf{y_{iT}})$, for $i = 1, 2, \ldots, N$.

Unlike the univariate case, multivariate longitudinal data allows the extension to examine the joint evolution of multiple responses over a period of time. For a fixed time, the responses may be correlated while simultaneously correlated within a given subject over time. Clearly, when longitudinal data is available, understanding the

heterogeneity among the observations and the dynamic relationship within a given subject that evolves over time helps improve the quality of the data analysis. In addition, understanding relationships among the response variables may dramatically increase model accuracy and efficiency thereby helping to improve possible predictive power of the model. Joe (1997) emphasizes the importance of the joint analysis of multivariate response variables in cross-sectional studies.

Understanding the relationship between different response variables within a longitudinal framework require exceptional techniques. In the literature, this has been accomplished by extending univariate longitudinal data analysis to a higher dimension and calling it multivariate longitudinal data analysis. When longitudinal data is available for multiple response variables, multivariate analysis can outperform univariate analysis. The multivariate longitudinal data analysis allows understanding the relationship between the response variables in addition to exploring the dynamic dependency and heterogeneity within subjects and among observations respectively. Also, joint analysis of multiple response variables can be used to compare different responses. Therefore, when modeling multivariate longitudinal data, the choice of multivariate longitudinal data analysis is more appealing than simply applying univariate longitudinal data analysis to each response variable separately.

Multivariate longitudinal data analysis is not surprisingly a stranger in the medical statistics discipline. As pointed out by Bandyopadhyay, Ganguli, and Chatterjee (2011), a 2007 special issue in the *Statistical Methods in Medical Research* journal was devoted to the various methods and applications of multivariate longitudinal data analysis. For example, using hearing data, Fieuws, Verbeke, and Molenberghs (2007) indicated that research questions with multivariate longitudinal data cannot

be fully addressed without taking the dependency among the response variables into account, and that the association structure of the responses may "be of direct scientific relevance." The scientific investigation considered in their paper has to do with the association of the various hearing thresholds determined at various frequencies, both for the right and left ears. Other scientific work considered in Fieuws and Verbeke (2009) include the study of transmission of psychiatric disorders among family members (father, mother, children) and the effect of renal graft transplantation from a deceased or living donor, These scientific investigations provide relevant indications that understanding the association among the responses may be an important factor to consider when analyzing multivariate longitudinal data.

Actuarial application in multivariate longitudinal context is very limited to date. Most of the research has restricted itself to the univariate longitudinal analysis of each response variable separately in multivariate longitudinal data. Even though univariate longitudinal analysis is a promising statistical tool for univariate data, it is not a robust technique to model data with multiple response variables. This is because analyzing each response variable separately does not capture the dependencies among the response variables. However, there is a developing interest has been observed in recent research by Shi (2012). It is our hope to contribute to this developing literature on multivariate longitudinal analysis within an actuarial context. In this thesis, we focus on the general framework of specifying the joint distribution of the multiple outcomes based on copulas which allows for the flexibility of modeling the dependence structure. In understanding the evolution of the joint outcome over time, we exploit the commonly used random effects models. This work is also motivated by the possible variety of actuarial and insurance applications of multivariate longitudinal data

analysis. For empirical illustration purposes, we examined two data sets which are directly related with the insurance industry.

The joint global demand between life and non-life insurance over the specific period of time with countries as our observations is the first application we considered. This analysis has been largely been motivated by what we have observed regarding the strong positive association between life and non-life insurance demand. Here we considered the demand variable to be the *insurance density*, defined as the amount of premiums per capita. Gross domestic product (GDP) per capita, urbanization and religious are few covariates we captured in our analysis to explain the demand for insurance lines. It is interesting to see the significant impact from religious factor on both insurance lines.

As a second application of our proposed technique and also something interesting as an actuarial application, we considered loss reserving of property and casualty insurance. Consider the case of a loss triangle often studied in insurance loss models. The amount of loss observed usually starts from the time of the accident and could evolve over a period of development years. The loss variable can be multivariate in structure in many sense: different lines of business (e.g. homeowners, automobile), various types of perils covered (e.g. hurricane, flood, theft) within the same line of business, or various types of losses (e.g. property damage, personal injury, theft) in the case of automobile insurance. In this thesis, we focus on the innovative use of multivariate longitudinal data analysis in loss reserving for a general insurer with multiple lines of business. We find that if we structure the correlated loss triangles for several lines of business as a longitudinal framework, we are better able to understand the underlying dependencies among the lines of insurance and at the same, capture

the dynamic emergence of the losses.

This thesis has been structured and organized as follows.

In Chapter 2, we discuss the literature related to longitudinal data analysis. Section 2.2 discuss the common methods for analyzing univariate longitudinal data that have appeared in the literature. Section 2.3 extends these methods to the case where the response variable is multivariate. We end this chapter by providing the model construction and specification of the multivariate longitudinal data method that was used all throughout the thesis.

Chapter 3 provides for an empirical application of our proposed multivariate construction motivated by the data used to analyze global insurance demand. Using data of a pair of insurance demands for each of 75 countries collected over a six-year period, we find strong evidence of dependence of demand between life and non-life insurance. Our model also allowed for identifying the heterogeneity that is commonly observed because of differences that is usually present among different countries.

In Chapter 4, we motivate our application in loss reserving for general insurance by examining various loss reserving methods that have appeared in the literature. As background, we reviewed and examined commonly used methods of loss reserving in the univariate case where the loss data comes in a triangular format. The bottom part of this triangle is unobserved, and therefore must be predicted to give the loss reserve estimates. Some methods extended to the case where we have a multivariate loss triangle have been done and these methods are discussed in Section 4.4.

Chapter 5 provides our second interesting empirical application of our proposed multivariate construction. Here, we are motivated by the idea of loss reserving for an insurance company with multiple, possible correlated, lines of business. We find that

our method can handle the estimation of the presence of the dependency among the various lines of insurance while at the same time, we are able to capture the dynamic nature of the loss data that occur over calendar years.

Finally, in Chapter 6, we summarize the findings in this thesis while at the same time, we provide some interesting sketches for possible future research.

Chapter 2

Literature and Model Construction

2.1 Introduction

Even though regression analysis is a popular statistical tool in modeling the collection of subjects across population, assumption of independence of the observations limits its ability to address the possible presence of dependency over time. Observations of a response along with corresponding covariates over time for a set of subjects create a longitudinal data framework. In contrast to cross-sectional data, longitudinal data allows us to further capture additional information about the data. The natural approach to analyzing such data with a single response is to use univariate longitudinal analysis which allows us to understand the dynamic relationship within a given subject, while simultaneously exploring the cross-sectional heterogeneity among the observations. Over the past few decades, clearly because of its importance, longitudinal data analysis has gained popularity in several disciplines such as medical statistics, finance, insurance, and actuarial science, e.g. Frees, Young, and Luo (1999) and Frees

and Shi (2010). There has been a much more rapid progress of longitudinal analysis for a univariate response data. See Diggle et al. (2013).

Increasingly becoming important these days where there is abundance of data is the extension of longitudinal data analysis from a single response to multiple responses. Such analysis is referred to in this thesis as *multivariate longitudinal data analysis* where we now observe several responses within a subject and it becomes important to capture the possible dependence among the responses. Multivariate longitudinal data analysis will continue to capture the nature of the dynamic relationship within the subjects and the presence of heterogeneity across the subjects in the data set. This type of an extension requires a more general model structure for analyzing the data.

In this chapter, we briefly discuss some longitudinal data methods for analyzing univariate data and later extend this to the limited literature on multivariate data. In the concluding sections of the chapter, we present our suggested methodology for handling multivariate responses using copula framework. Copula models allowed us to capture the presence of dependency among the responses with the flexibility of separating the effects of the marginals. The primary motivation, as later discussed, is to extend several existing literature that primarily relies on the Gaussian model structure and the use of limited distributions to model the marginal components.

2.2 Common methods for univariate longitudinal data

When analyzing a single response with cross-sectional data over time, methods used are those that help to distinguish difference among the subjects due to changes within

the subjects over time. Classical regression techniques do not always work in the context of longitudinal data because the observations within the subjects may be correlated. In the literature, some classical techniques such as the use of paired t-test and analysis of covariance have been used but there are usually disadvantages of using these methods.

Consider the vector of n repeated measurements for the i-th subject, denoted by

$$\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in})'.$$

A starting point of analysis is usually based on the general multivariate regression model where we assume the following specification

$$y_i = x_i\beta + \varepsilon_i,$$

where x_i is the design matrix, β is the corresponding set of regression parameters, and the random component ε_i usually assumes a multivariate normal distribution with a covariance structure. The mean structure is the usual classical linear regression model but the covariance assumption can be designed to allow for parsimonious structures. While the covariance matrix may be unstructured, there is also a large selection of covariance structure that can be used within the model that may lead to more efficient inferences. For example, the simplest covariance structure assumes independence while auto-regressive covariance structure allows for time dependence.

A natural extension commonly used is to capture time dependence with the use of

random effects. Here, the linear mixed-effects model has the following specification:

$$\mathbf{y}_i = x_i\beta + z_i\mathbf{b}_i + \varepsilon_i,$$

where we have the addition of the vector $\mathbf{b}$ often referred to as the random effects. These random effects typically assume also a multivariate normal distribution model, but a more general multivariate distribution can be specified. Additionally, the random effects together with the error components are all independent. In the case where we only have a single random effect denoted by α, the linear mixed-effects model reduces to what we sometimes call random effects model:

$$\mathbf{y}_i = x_i\beta + \alpha_i + \varepsilon_i.$$

See Fitzmaurice et al. (2008).

Frees and Wang (2006) indicated that longitudinal analysis significantly outperforms cross-section regression analysis because of the ability to account for historical trend informations with longitudinal data. It is a well known fact that a large portion of the total variation corresponding to the response variables can be explained by the subject-specific random effects rather than time-specific random effects. Also, Frees (2004) pointed out that dynamic dependency of subjects can be determined by introducing subject-specific random effects in the regression model for longitudinal data.

2.3 The extension with multivariate response

When modeling multiple response variables over time for the same set of subjects, multivariate longitudinal analysis is preferred in model estimation. Within the limited literature on multivariate longitudinal analysis for continuous response variables, an early study of Rochon (1996) introduced seemingly unrelated regression analysis for joint analysis of multiple response variables. He used the technique of generalized estimating equations (GEE) to estimate parameters and considered the generalized linear model to construct the relationship between a set of repeated measures and covariates. Finally, he applied seemingly unrelated regression models to combine GEE models.

The classical approach in multivariate longitudinal analysis is the random effects model. Reinsel (1982) extends the random effects model approach to analyze the balanced multivariate data when responses are from a multivariate normal distribution. He used multivariate random effects covariance structure in order to combine the multiple response variables. Shah, Laird, and Schoenfeld (1997) extended this approach to address the missing data in multivariate longitudinal analysis. Due to the complexity of covariance structure as the number of response variables increases, Fieuws and Verbeke (2006) introduced pair-wise fitting of mixed models for joint modeling of multivariate longitudinal models incorporating multivariate random effects approach.

Surprisingly, even far beyond a decade ago, a copula-based approach for multivariate longitudinal analysis was introduced by Lambert and Vandenhende (2002). They considered a copula function to fit univariate longitudinal model for each response variable separately and used another copula function to capture the time dependence for each response variable. The model was calibrated using medical-related data to

understand "hemodynamic effect of a new antidepressant" drug. Although the authors recognized the flexibility of specifying various copula models, they used, for simplicity, the Gaussian copula models, but allowing for different marginal distribution models.

Two excellent survey papers on the use of random effects model for multivariate repeated measures may be found in Fieuws, Verbeke, and Molenberghs (2007) and Bandyopadhyay, Ganguli, and Chatterjee (2011).

2.4 Our multivariate model construction

The new technique we propose for multivariate longitudinal analysis is a combination of classical and modern approaches in the literature. Most studies, that address multivariate longitudinal analysis, used the classical normal distribution assumption for response variables. Unfortunately, in practice, this assumption does not hold. Particularly, in actuarial science, skewed distributions of continuous random variables are very commonly applied in research. We can relax this assumption by introducing exponential family distributions for response variables. The proposed flexible framework requires copula functions to integrate the dependence among responses and the classical random effects to identify intertemporal dependence within a subject and unobservable subject-specific heterogeneity among observations. Covariate information is taken into account for observable subject-specific effects through the generalized linear regression model for the marginals. Our proposed method is also flexible with both balanced and unbalanced data, an important aspect of longitudinal investigation.

According to Bandyopadhyay, Ganguli, and Chatterjee (2011), the methods to analyze multivariate longitudinal data analysis may be broadly classified into three categories. The first approach is to express the multivariate outcome into a single summary measure so that the analysis therefore reduces to univariate longitudinal data analysis. The use of regression models without the explicit specification of the covariance structure is the second approach. This can be done by the use of separate regression models for each outcome and then combining the regression coefficients into one single global estimate. The third approach is to explicitly specify the association structure either through a covariance with a Gaussian multivariate structure or more broadly, through a more general multivariate distribution model for the multiple outcomes. The work of Gao et al. (2006) explores the covariance specification within a Gaussian multivariate structure using SAS as a computing tool. In some sense, the approach proposed in this thesis falls in this third category.

2.4.1 Notation and assumptions

Suppose we have a set of observations on n subjects collected over T time periods for a set of m response variables. Let $y_{it,k}$ denote the observation from i-th individual in t-th time period on k-th response. Hence, for a given subject response $\mathbf{Y_i}$ can be expressed as a matrix $\mathbf{Y_i} = (y_{it,k})_{T\times M}$. Considering column vectors of above matrix, we can re-write the matrix $\mathbf{Y_i}$ as a vector of column vectors $\mathbf{Y_i} = (\mathbf{y_{i1}}, \mathbf{y_{i2}}, ..., \mathbf{y_{iT}})$.

Similarly, for q set of predictor variables, $\mathbf{X_{it}} = (\mathbf{x_{it,1}}, \mathbf{x_{it,2}}, ..., \mathbf{x_{it,m}})$ indicates the matrix of predictor variables. We will be using α_{ik} to represent the random effects component corresponding to the i-th subject from the k-th response variable and $G(\alpha)$ for pre-specified distribution function of corresponding random effects.

In the model specification, we explicitly make the following assumptions:

- The observations $\{\mathbf{Y_i}\}$ are independent for a given time t and response k.
- Each response variable over time is assumed to have the same parametric distribution.
- $\{\mathbf{X_{it}}\}$ are non stochastic variables.
- The random effect components $\{\alpha_{ik}\}$ are identically and independent distributed.
- Random effects and covariates are independent.
- The same family of copula functions will be applicable over time.

2.4.2 The model structure

Our proposed methodology consists of the use of a random effect model to capture dynamic dependency and heterogeneity, and a copula function to incorporate dependency among the response variables. This extension naturally allows us to extend univariate to multivariate longitudinal data analysis.

There are several advantages to our proposed methodology. Marginal distribution of response variables are not restricted to normal distribution. Available covariate information are incorporated to determine some distribution parameter. Dependency structure among responses are modeled using copula functions. Intertemporal dependency within subjects and unobservable subject specific heterogeneity are captured utilizing random effect term. Parameter estimation of the proposed model are facilitated by MLE. Model construction has the flexibility to accommodate both balanced and unbalanced data.

Consider the situation where we observe m number of responses over T time periods for N different subjects. Observed data for subject i can therefore be expressed as

$$\{(y_{i11}, y_{i12}, \ldots, y_{i1m}), \ldots, (y_{iT1}, y_{iT2}, \ldots, y_{iTm})\}$$

so that

$$\mathbf{y}_{it} = (y_{it1}, y_{it2}, \ldots, y_{itm}) \text{ for } i = 1, 2, \ldots, N \text{ and } t = 1, 2, \ldots, T$$

is the ith observation in the tth time period corresponding to the various m responses. The joint distribution of the m responses over time can be expressed as

$$\mathrm{H}(\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{iT}) = P\left(\mathbf{Y}_{i1} \leq \mathbf{y}_{i1}, \mathbf{Y}_{i2} \leq \mathbf{y}_{i2}, \ldots, \mathbf{Y}_{iT} \leq \mathbf{y}_{iT}\right). \tag{2.4.1}$$

For an arbitrary m uniform random variables on the unit interval, the copula function, C, can be uniquely defined as

$$C(u_1, \ldots, u_m) = P(U_1 \leq u_1, \ldots, U_m \leq u_m). \tag{2.4.2}$$

For a random vector of dimension m expressed as $(y_1, y_2, \ldots, y_m)$, according to Sklar's Theorem, its joint multivariate distribution function can be expressed in terms of the marginals through the copula function as

$$F(y_1, y_2, \ldots, y_m) = C(F_1(y_1), F_2(y_2), \ldots, F_m(y_m)), \tag{2.4.3}$$

with $F_k(y_k)$ referring to the marginal distribution function of the k-th response. It is

also well-known that the corresponding joint density function has the form

$$f(y_1, y_2, \ldots, y_m) = c(F_1(y_1), F_2(y_2), \ldots, F_m(y_m)) \prod_{k=1}^{m} f_k(y_k), \tag{2.4.4}$$

where $f_k(y_k)$ are the marginal density functions and c is the density associated with the copula function C. Copulas have been for years dominated the literature on dependencies in various disciplines ranging from statistics, actuarial science, social science and insurance. It a rather flexible tool that allows for understanding a wide range of dependence structure while at the same time, is adaptable to separating the effects of the intrinsic characteristics of the marginal distributions such as skewness and thickness of tails.

If $\{\alpha_{ik}\}$ represent random effects with respect to the k-th response variable, the conditional joint distribution at time t is

$$H(\mathbf{y_{it}}|\alpha_{i1}, \ldots, \alpha_{im}) = C(F(y_{it,1}|\alpha_{i1}), \ldots, F(y_{it,m}|\alpha_{im})).$$

The corresponding conditional joint density at time t has the expression

$$h(\mathbf{y_{it}}|\alpha_{i1}, \ldots, \alpha_{im}) = c(F(y_{it,1}|\alpha_{i1}), \ldots, F(y_{it,m}|\alpha_{im})) \prod_{k=1}^{m} f(y_{it,k}|\alpha_{ik}),$$

where $F(y_{it,k}|\alpha_{ik})$ denotes the distribution function of k-th response variable at time t. The use of random effects is one of the several techniques to model longitudinal data that allows for controlling variables that may change over time; such is typical for accounting individual heterogeneity that may be present in the data.

We need to express the likelihood of the data in order to perform maximum

likelihood estimation. If ω represents the set of parameters in the model, the likelihood of the i-th subject is therefore given by

$$L(\boldsymbol{\omega}|(\mathbf{y_{i1}}, \ldots, \mathbf{y_{iT}})) = h(\mathbf{y_{i1}}, \ldots, \mathbf{y_{iT}}|\boldsymbol{\omega}). \tag{2.4.5}$$

We can write

$$h(\mathbf{y_{i1}}, \ldots, \mathbf{y_{iT}}|\boldsymbol{\omega}) = \int_{\alpha_{i1}} \cdots \int_{\alpha_{im}} h(\mathbf{y_{i1}}, \ldots, \mathbf{y_{iT}}|\alpha_{i1}, \ldots, \alpha_{im}) \, dG(\alpha_{i1}) \cdots dG(\alpha_{im}),$$

where, more specifically, $\boldsymbol{\omega} = \{\boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\alpha}\}$. The symbol $\boldsymbol{\eta}$ represents the systematic component, which determines the location parameter of the assumed distribution of the response variable, and $\boldsymbol{\tau}$ denotes the rest of the corresponding parameters in the underlying distribution. The set of random effects in the regression model is denoted by $\boldsymbol{\alpha}$.

Traditionally, regression models with empirical data incorporate linear relationships between location parameter of the assumed distribution for the response variable and independent variables (also known as explanatory variables or covariates). In our proposed model, we relax the linearity assumption and allow the systematic component to be a function of covariates including non-linear relationships. Therefore, in general, we can express the systematic component as follows:

$$\eta(\mu(x)) = X'\beta \tag{2.4.6}$$

Under independence over time for a given random effect:

$$h(\mathbf{y_{i1}}, \ldots, \mathbf{y_{iT}}|\alpha_{i1}, \ldots, \alpha_{im}) = \prod_{t=1}^{T} h(\mathbf{y_{it}}|\alpha_{i1}, \ldots, \alpha_{im})$$

$$= \int_{\alpha_{i1}} \ldots \int_{\alpha_{im}} \prod_{t=1}^{T} h(\mathbf{y_{it}}|\alpha_{i1}, \ldots, \alpha_{im}) dG(\alpha_{i1}) \cdots dG(\alpha_{im})$$

and from the previous slides, we have

$$= \int_{\alpha_{i1}} \ldots \int_{\alpha_{im}} \prod_{t=1}^{T} c(F(y_{it,1}|\alpha_{i1}), \ldots, F(y_{it,m}|\alpha_{im})) \prod_{k=1}^{m} f(y_{it,k}|\alpha_{ik}) dG(\alpha_{i1}) \cdots dG(\alpha_{im})$$

Then, we can write the log likelihood function as

$$\sum_i \log \Big\{ \int_{\alpha_{i1}} \ldots \int_{\alpha_{im}} \prod_{t=1}^{T} \prod_{k=1}^{m} c(F(y_{it,1}|\alpha_1), \ldots, F(y_{it,m}|\alpha_m)) \times f(y_{it,k}|\alpha_{ik}) dG(\alpha_{i1}) \cdots dG(\alpha_{im}) \Big\}$$

The likelihood function above is not typically in explicit (or closed) functional form. Hence, to evaluate the above integrals, we need to employ special numerical techniques. For our purposes and for simplicity, we used Monte Carlo integration techniques to evaluate this log likelihood function.

The concept of joint analysis of multivariate response variables rapidly improved in research studies after introducing copula function as a tool for relating several dimensions of an outcome. This flexible approach allows researchers to understand the joint evolution of response variables, without the restriction of a bivariate case, based on different families of

distributions while exploring dependencies among the response variables. For more theories and example about copula functions, one may refer Nelsen (2007), and Joe (1997). The usefulness of copula functions in joint analysis and their applications in actuarial related problems are explored by Frees and Valdez (1998). Elliptical copula functions were employed by Frees and Wang (2006) for both discrete and continuous longitudinal data to address the time dependencies.

2.4.3 Copula model diagnostic

There is a wide variety of families of copulas, e.g. Archimedean and Elliptical, that are available for model calibration. Hence the issue of choosing the correct family of copula for fitting data has been an interesting topic in the literature. As part of a preliminary investigation of the data, one can use scatter plots of the rank data to get an initial idea about possible candidate of copulas. Rank scatter plot is not usually a strong approach to justify the goodness-of-fit of selected copula functions. However, rank scatter plot allows us to visualize the dependency of response variables regardless of the marginal effect. One can also describe the rank scatter plot as graphical representation of Spearman's rho correlation.

There are few graphical tools for copula validation that has appeared in the literature. Fisher and Switzer (1985) introduced Chi-plots which is based on chi-square statistics. Frees and Valdez (1998) proposed qq plots to select appropriate bivariate Archimedean copula. Later, a rank-based approach, K-plot, which is inspired by the concept of qq-plot, proposed by Genest and Boies (2003). Another graphical tool, which is inspired by the univariate pp-plot, was introduced by Sun, Frees, and Rosenberg (2008). The so-called "Copula pp-plot" can be used to evaluate the goodness-of-fit of estimated copula functions in both Archimedean and Elliptical family. "Copula pp-plot" compares the probability values corresponding to the empirical and theoretical copula functions. Here we plot the probabilities of empirical copula against those of the theoretical copula to produce the

"copula pp-plot".

The probabilities of empirical and theoretical copulas can be calculated using the residual values obtained by the marginal models. Departure from the line connecting $(0, 0)$ and $(1, 1)$ on the graph indicates the rejection of the underlying theoretical copula assumption. An empirical copula function can be formulated as

$$C_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1} \left(\frac{\tilde{X}}{n+1} \leq u, \frac{\tilde{Y}}{n+1} \leq v \right), \tag{2.4.7}$$

where $\tilde{X}$ and $\tilde{Y}$ are the rank values corresponding to random variables X and Y. $\mathbf{1}(\Lambda)$ denotes the indicator function of set Λ. In this thesis, we used the `R` package *copula* to calculate the copula functions and corresponding probabilities.

Chapter 3

Global Insurance Demand

3.1 Introduction

Some people are risk lovers. They believe that the more risk you take the more you can possibly gain. In contrast, some people are risk averse. However, in reality, people cannot avoid risk. Risk may occur as a result of human activities or action of nature. However, people can manage the financial risk resulting from these consequences. Individuals and businesses can manage small financial losses through their savings or investments. But, it is very unlikely that individuals or businesses can survive themselves with large financial losses. For example, the super storm Sandy in 2012 created damages over $70 billion, which implies high financial loss per individual, only in New Jersey.

Insurance companies facilitate their clients (individuals or businesses) by mitigating the financial risk due to the uncertainty . The insurance agreements they make with individuals or businesses provide financial assistance in the event of future financial loss in exchange of premiums they collect. Insurance companies render financial protection for large risk pools.

In simple terms, insurance companies pool their clients with similar risk characteristics and bear the financial risk by distributing the financial impact among everyone in the risk pool. Insurance companies have usually adequate financial capital to assist their clients who encounter large financial losses. Therefore, individuals and businesses do not need to set aside large capitals to protect their own future financial losses, but may use those capitals for investments. This could inspire people to assume more risk with their financial investments than they would do in the absence of insurance. This is expected to increase total investments within the economy and thereby improving economic growth.

In recent decades, instead of providing traditional financial protection for their clients, insurance companies started to move into the investment market and start playing a major role in the world economy. There has been a number of empirical research that found that the activities in the insurance industry provide engine to economic growth. For example, see Arena (2008). These research results signify the importance of the insurance industry in the global economy. Clearly, economies in most developed countries are very sensitive to their insurance industry. This is mainly, because, large amounts of financial assets are governed by the insurance industry. Furthermore, in most of these countries, the insurance industry provides a large number of direct and indirect employment opportunities through their services. The following Table 3.1.1 indicates the financial wealth of insurance industry around the world for the period 2008-2013.

Table 3.1.1: Total premiums collect by annual (USD billion)
Source: Annual Swiss Re sigma reports.

Year	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**
Life	2,490	2,332	2,520	2,627	2,621	2,608
Non-life	1,779	1,735	1,819	1,970	1,992	2,033

As we mentioned earlier, there are different types of insurance products that can be observed in the market: life, property, health, liability, and auto insurance, to name a

few. For our purposes, we categorized every product not considered life as non-life. In the early days, there is a greater demand for non-life insurance but the life insurance market alone has gained rapid growth and development over the years. This is especially true for countries with advanced economies. However, the annual reports from SwissRe indicates that total premiums collected over different regions have different level of growth in both life and non-life insurance.

In this chapter, our primary motivation is to understand the relationship between life and non-life insurance demand. To do so, we used data consisting of a pair of insurance demands for each of 75 different countries observed during the period 2004-2009. We also recognize the heterogeneous characteristics of each country that may affect insurance demand. For this type of problem, the use of multivariate longitudinal framework is most suitable. We have a pair of possibly dependent response variables (in the form of demand) with each observation exhibiting heterogeneity (through the form of observable covariates).

There are three measures that can be used as proxy variables to insurance demand.

- Insurance penetration.
- Insurance density.
- Insurance in force.

Gross domestic product (GDP) is the common measure for the economic performance in a country. The ratio of total insurance premiums collected to GDP is defined to be the insurance penetration. This relative measure of the premiums implies significance of the insurance market within the economy for a given country.

Insurance density, also known as insurance premiums per capita, is another proxy for insurance consumption. This variable refers to the total insurance premiums relative to the size of the population of a country. For example, the life insurance density in the United

States in year 2009 was $ 1,603.00. This was the average total premium for life insurance assessed for every living person of United States in year 2009.

Finally, the last variable we mentioned above, insurance in force, has appeared in some literature studies to measure life insurance demand. This measures the total amount of life insurance policies in force, plus usually the total dividends paid by the insurance companies. Other studies sometimes used relative terms, e.g. insurance in force relative to the GDP of the country.

In this study, insurance densities for both life and non-life insurance were utilized as proxy to insurance demand. A set of popular explanatory variables were incorporated as insurance determinants to explain the global insurance demand over time. In the following section, we discuss some research work that has appeared in the literature over the past years related to insurance demand. In Section 3.3,we provide a detailed discussion about some of the more commonly used and often accepted explanatory variables that have appeared in the literature. Here, we also discuss some of these research findings explaining their significance and impact on insurance demand. Finally in section 3.4, we examine our cases study involving global insurance demand where we implemented the proposed model in chapter 2. We provided calibration results and discuss interpretations of these results.

3.2 Some literature on demand

Understanding insurance demand is a classical research topic that has been covered over the years in the economic, insurance, and actuarial literature. Numerous research studies, both theoretical and empirical, that are focused on insurance demand indicate how important this topic is in understanding the consumer market. Furthermore, it is well-known that the insurance industry has a significant role on the global economy so that understanding consumption within the insurance market is undeniably important.

Most research studies on global insurance demand consider the work of Yarri (1965) as the first to provide the theoretical model for life insurance demand. Fischer (1973), Campbell (1980), and Lewis (1989) are a few other theoretical models that appeared in the literature. Beenstock, Dickinson, and Khajuria (1986) also developed a theoretical model that incorporates several economic and social factors in understanding life insurance premiums.

Besides these theoretical models, there are ample research studies with empirical illustrations used to calibrate these theoretical models as well as other possible statistical models. It is clear that for a given country, insurance demand can be affected by various factors. To illustrate, income level, dependency ratio, inflation, government regulation and intervention, underwriting cost, urbanization, life expectancy or death ratio, and education level are just a few popular factors examined in several of these studies. These factors have direct and indirect impact on insurance demand, and help us to understand the often observed deviation of insurance consumption between countries. Therefore, studies that involve understanding the nature of insurance consumption and its determinants using real observed data are always interesting to all parties within the insurance industry.

Within the large body of insurance demand literature, most research works are focused on life insurance demand. One good reason for this is the continuing growth of life insurance premiums especially in advanced economies. In addition to the need to cover premature death, people tend to purchase life insurance products as an investment or saving vehicle.

The rapid development in insurance industry around 1980 to 1990, especially life insurance, largely encouraged Browne and Kim (1993) to investigate the factors that affect life insurance demand. Based on data observed for 45 different countries, income, inflation, the price of insurance, dependency ratio, government spending on social security, and the country's predominant religion are significant factors that explain the deviation of life insurance consumption. In examining life insurance market in developing countries, Outreville (1996)

pointed out that life insurance as a savings or investment product is usually lower in demand for developing countries than for developed countries. As explained by the author, the main reason for this low demand is that "developing countries have a supply-leading causality pattern of development" as opposed to a demand-following causality pattern observed in developed countries. In addition, Beck and Webb (2003) conducted a research study incorporating 63 countries over the period of 1980 to 1996 and found that the development of the banking sector and inflation are more robust predictors of the insurance demand than income level.

Apart from the research focused on global insurance consumption, there are few studies to understand the insurance demand within an individual country. For example, the research study by citeHwang2003 explored that economic security, education level, and the change in social structure directly impact life insurance demand in China. Surprisingly however, this empirical study indicates no negative impact of inflation on life insurance demand during the study period. Lim and Haverman (2004) focused on the Malaysian life insurance industry. A novel statistical procedure by Lenten and Rulli (2006) unraveled the time series properties of life insurance demand in the Australian insurance market over the period between 1981 and 2003. Finally, two comprehensive literature survey that seek for factors influencing life insurance demand can be found in Zietz (2003) and Outreville (2013).

Outreville (1990) examined the relationship between property/casualty insurance and economic/financial development. For empirical illustrations, he incorporated 55 developing countries in his research work. Results of his research work indicate that developing countries have less demand for property and casualty insurance. Browne, Chung, and Frees (2000) investigate the insurance consumption of property and liability insurance in OCED (Organization for Economic Cooperation and Development) countries over a period of time covering 1987 through 1993. This could be the first research work to identify the insurance demand broken down by coverage specific level. Within non-life insurance, the study ex-

plored insurance consumption for individual motor vehicle insurance and general liability insurance, which is primarily purchased by businesses.

A recent study of Dragos (2014) incorporated a panel data of 17 countries with emerging economies over the period of years between 2001 and 2011. Using both random and fixed-effects longitudinal models with Asian and European countries considered separately, this study focused on the demand for both life and non-life insurance. However, in this study, the longitudinal models are constructed separately for both market. The results of the study did not indicate consistency among Asian and European countries.

3.3 Common determinants

Although we have ample research studies about insurance consumption and its determinants, outcomes of these research do not generally provide a clear and consistent picture about the effects of these determinants in explaining variation across countries. Therefore, numerous variables have been considered in the literature studies. For a start, Zietz (2003) and Outreville (2013) which are previously mentioned, provide detailed descriptions about common determinants of life insurance demand.

One explanation of possible inconsistencies may have to do with differences among the countries over an extended period of time. To illustrate, it is apparent that certain economic characteristics and lifestyles do indeed widely vary by country and even within a country, there are possible variations over a period of time. Because of such, it is therefore not surprising to see inconsistencies of the effects for different studies. A recent study from Brokesova, Pastorakova, and Ondruska (2014) indicates that some determinants of insurance demand in advanced economies have different impact than in transition economies. A literature review by Eling, Pradhan, and Schmit (2014), from year 2000 to 2014, of macro insurance demand identifies 12 different determinants that led to inconsistent feet towards

insurance demand.

Another possible explanation is the evolving nature of policy coverage for different types of insurance products. Today, for example, we find many different types and variations of insurance products available in the market. Within the life insurance sector alone, there are several types of term life insurance product with savings or investment vehicle. Demand for these different products can vary by country due to various economic and demographic factors. For example, young people tend to buy more traditional life insurance products covering mortality risk alone, while the elder population tends to purchase term life insurance products with savings component.

In this section, we discuss about the predictor variables we use in our research and their previous appearance in some literature studies. A bunch of factors that drive insurance demand up and down in different countries can be observed in various studies. These various studies may have used different factors, but careful comparison of some of these factors provide a common core. For example, personal income, household income, expected future income, gross domestic product (GDP), and price of insurance all represent affordability of insurance products. Age, life expectancy, and death ratios can all represent mortality or survival characteristics of individuals. Therefore, our choice of predictor variables may fall within the purview of: dependency ratio, income level, mortality characteristics (e.g. life expectancy), religion, and urbanization.

Dependency ratio

In general, the ratio of total number of dependents to the working population considered is called the dependency ratio. Lenten and Rulli (2006) describe this determinant as an indicator of the number of members dependent on the main income source per family. While most research studies indicate a positive relationship between dependency ratio to life insurance consumption, there are some empirical studies that indicate otherwise. For

example, Beenstock, Dickinson, and Khajuria (1986) and Browne and Kim (1993) found a positive relationship, Burnett and Palmer (1984) and Outreville (1996) found insignificant impact, and Li et al. (2007) found a negative relationship.

Beck and Webb (2003) and Lester, Rocha, and Feyen (2011) provide explanation to these inconsistencies. The nature of the dependence is often categorized in varying degrees. In some studies, children such as those under 18 years old are considered dependent while in other studies, people over age 65 are considered dependent. Hence conclusively, the nature of the impact of the dependency ratio on insurance demand can vary. This is because of the different degrees of needs and preferences of life insurance coverage by age; for example, life insurance products with a high savings component as well as annuity products tend to be favored by countries with dependency ratios that have a large proportion of elderly population.

Income level

It is intuitive that affordability of insurance products is not an issue with greater income level. People tend to purchase insurance products against possible financial losses when the income level is high. Almost all research studies, which use income level as a determinant, find positive relationship between insurance demand and income level. Browne and Kim (1993), Browne, Chung, and Frees (2000), Treerattanapun (2011), and Brokesova, Pastorakova, and Ondruska (2014) are to name a few research studies that indicate positive impact of income level towards insurance demand. However, Anderson and Nevin (1975) found a negative impact among middle-income families, but a positive relationship between low and high-income level families; families considered in this study included young marrieds. Furthermore, a recent study by Dragos (2014) indicates that income level is not a significant factor in the demand for non-life insurance in Asia.

Gross Domestic product also known as GDP is one of the popular proxy variable for

income level in many research areas. Outreville (1990) showed that GDP factor is not influenced by the country's currency factor. Hence, especially for global insurance demand, GDP factor is a much more suitable proxy for income level of a country.

Another possible proxy for income level is the education level which is commonly linked to risk aversion. However, numerous of these studies usually examine simultaneously the effect of income level and education level on the demand for insurance. It is therefore not surprising to find inconsistencies on the empirical results. For example, Hammond, Houston, and Melander (1967), Burnett and Palmer (1984), and Browne and Kim (1993) find that education has a positive effect on insurance demand, while Duker (1969), Anderson and Nevin (1975), and Auerbach and Kotlikoff (1991) find the opposite. Outreville (1996), Beck and Webb (2003), Nesterova (2008) , and Lester, Rocha, and Feyen (2011) are a few research studies to find non-significant impact of education on insurance demand. Finally, a recent study by Dragos (2014) finds that level of education has a significant influence to demand in non-life insurance but not to life insurance.

Mortality characteristics

In understanding mortality characteristics, several studies use either life expectancy, also known as life expectancy at birth, or death ratio. As per Outreville (2013), a country with higher life expectancy tends to have lower demand for life insurance that covers pure mortality risk. On the other hand, it is possible that higher life expectancy lowers the overall cost of insurance which increases affordability and hence, one could expect a positive relationship. Beenstock, Dickinson, and Khajuria (1986), Outreville (1996), Browne and Kim (1993), and Li et al. (2007), to name a few, found positive, but statistically weak relationship with life insurance demand.

As we did in our case study as explained in later sections, we used death ratio as a proxy to life expectancy at birth. It appears intuitive that life expectancy at birth and death ratio

of a country have an inverse relationship. As indicated by the studies above, relationship between death ratios and life insurance demand could go either direction.

Religion

It is believed that certain religious beliefs tend to oppose the concept of insurance, especially that of life insurance. It is therefore a common theme among studies of insurance demand to examine how the presence of certain religion could affect insurance demand. A number of these studies tend to examine the proportion of the population that are Muslims who follow Islamic religion. For example, the early research work of Browne and Kim (1993) indicates that countries with higher proportion of Muslims tend to have less demand for life insurance. In addition, Beck and Webb (2003) and Park and Lemaire (2011) considered percentages of many other types of religions (e.g. Christians, Buddhist), but found a similar relationship as in Browne and Kim (1993) only for the Muslim population. However, there is nowadays an increasing demand for Takaful insurance structured around the Islamic insurance concept.

Urbanization

A ratio of the urban population to the total population of a country can be considered as the urbanization factor. Earlier studies of insurance demand did not consider urbanization as an important determinant, but later studies of both life insurance and non-life insurance demand found that urbanization became an increasingly important determinant. Most research studies have found a positive relationship between urbanization and insurance consumption. Browne, Chung, and Frees (2000) and Esho et al. (2004) found a positive relationship of urbanization to non-life insurance demand; such studies incorporated urbanization as proxy to loss probability. Furthermore,]citeEsho2004 argued that most countries tend to have a higher rate of crime in urban areas. A study by Park and Lemaire (2011) about non-life insurance demand describes that development of urban areas in many coun-

tries is a result of industrialization and economic development. People living in urban areas generally tend to be exposed to a higher risk of financial loss compared to those living in rural areas. Therefore, demand for insurance products increase with development and growth of urban areas.

Beck and Webb (2003) argue that the cost of life insurance products in urban areas could be low as a result of high density of consumers in certain geographic areas. It is intuitive that low cost of products will increase the demand and hence, a positive relationship can be observed. A study involving East European and some former soviet union countries, conducted by Nesterova (2008), found that life insurance demand is not influenced by urbanization. On the other hand, studies from Hwang and Gao (2003), and Sen (2008) indicated positive relationship of urbanization with life insurance demand. A more recent study by Dragos (2014) and Brokesova, Pastorakova, and Ondruska (2014) found that urbanization has significant positive effect on life insurance demand in Asian insurance markets, but not significant for East European markets.

3.4 A case study

In this section, we examine the relationship between life and non-life insurance demand as a multivariate longitudinal data problem. Our primary response variable is a bivariate vector which consists of two components with one component describing life insurance demand and the other, non-life insurance demand. As a background, we note that insurance demand in the literature uses one of three possible variables of interest: (a) insurance density which is the premiums collected per capita, (b) insurance penetration which is the ratio of insurance premiums to the level of GDP, and (c) insurance in-force which is the outstanding face amount plus dividend. For simplicity and for our purposes, the first definition is what we used to describe insurance demand.

Why we want to study this pair of insurance demand is primarily motivated by Figure 3.4.1. This figure exhibits the observed closeness of the movement of real growth rates between life insurance and non-life insurance during the period between 1980 and 2010. In this figure, we are showing a strong positive co-movements of the two growth rates. This strong positive co-movement may provide information to insurance companies who wish to penetrate insurance markets in certain countries. For example, if an insurance company has a strong market share of life insurance in a particular country, that company may want to strongly consider penetrating the non-life insurance market in the area, after considering additional characteristics or determinants that can help to assess further the level of non-life insurance demand.

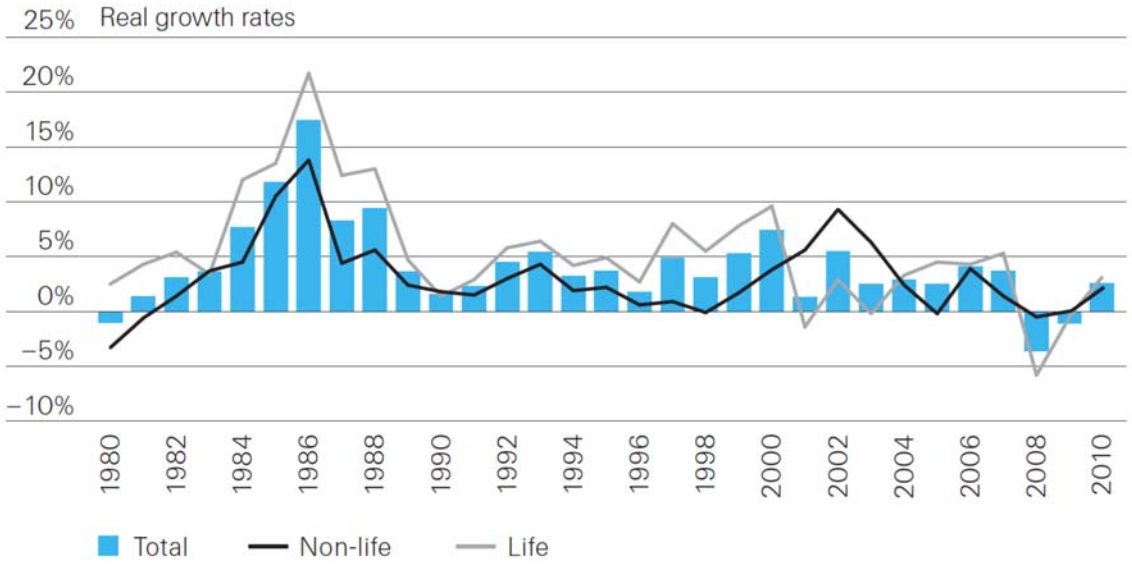

Figure 3.4.1: Real growth rates of life and non-life insurance: 1980-2010
Source: Swiss Re Economic Research & Consulting

3.4.1 The data and its sources

Our data consists of a pair of insurance demands for each of 75 countries collected during the period of 2004 to 2009, inclusive. This pair of demands was derived using the ratio of insurance premiums to the total population. Insurance premiums were collected from Swiss Re sigma reports done annually. The records of data also consisted of characteristics of different countries, which we call independent variables, as detailed below in the table. These characteristics were used as covariates in explaining the heterogeneity of insurance demand among the countries. The data explaining religion and Muslim population is collected from Kettani (2010). The database contains average percentage value of Muslim population in each country in decades. For our study, we used the average percentage values for the period years 2000 to 2010. The rest of the variables are collected from the World Bank database.

Table 3.4.2 provides summary statistics of the different variables in the dataset observed between years 2004 and 2009. The table provides the range of values for each statistic. For example, for the non-life insurance, we obtained a minimum premium density of 0.74 (in millions) and 1.26 (also in millions). For the life insurance, on the other hand, we obtained a minimum premium density of 0.49 (in millions) and 1.28 (in millions). Furthermore, the mean death rate observed for the different countries between years 2004 and 2009 are in the range of 7.87 and 9.00, expressed in terms of thousands. As suspected, our correlation between life insurance and non-life insurance demand is in the range of 0.75 and 0.80, an indication of a strong positive co-movement. A matrix of correlation between the independent variables is subsequently in the Table 3.4.3.

In preserving the notation used in the previous chapter when we discussed the model construction, we used a pair of insurance demand $(y_{it,1}, y_{it,2})$ where the first component refers to non-life insurance and the second, life insurance. This data is further disaggregated according to country i and over time t, denoting the respective calendar years 2004 through

Table 3.4.1: Variables in the study

Response variables	
Non-life density	Premiums per capita in non-life insurance
Life density	Premiums per capita in life insurance
Independent variables	
GDP per capita	Ratio of gross domestic product (current US dollars) to total population
Religious	Percentage of Muslim population
Urbanization	Percentage of urban population to total population
Death rate	Percentage of death
Dependency ratio	Ratio of population over 65 to working population

2009. The covariates for each country i and time t are described by the vector $\mathbf{x}_{it}$ and will serve as predictor variables in our analysis. For notational convenience, we can write the observable data consisting of the available information as follows:

$$\left\{(y_{it,1}, y_{it,2}), \mathbf{x}_{it},\ t = 1, ..., 6,\ i = 1, 2, \ldots, 75\right\}.$$

In other words, during the observation period, we have 75 countries for which each country is observed 6 (yearly) times. However, in the final calibration of our data, we dropped the observations from three different countries (The Netherlands, Ireland and the UK) because of unusually high insurance demand for these countries in comparison with the rest. In the final analysis then, only 72 countries were used. Furthermore, there are potentially 5 predictor variables which we considered for both life insurance and non-life insurance demand. However, not surprisingly, the death rates and dependency ratios did not affect non-life insurance demand.

Figure (3.4.2/3.4.3) display separately the non-life and life insurance premiums per capita over time in a longitudinal framework where each time series curve represents a country i. These two graphs demonstrate the presence of time dependence of the data. Figure (3.4.5) provide the scatter plots of the two response variables based on the raw

observations. While this figure tells us already how positively related the two responses for each time period observed, Figure (3.4.6), which displays the ranked responses for each time period, shows an even stronger positive relationship when the marginal effect is discarded.

Table 3.4.2: Summary statistics of variables in year 2004 to 2009.

Variable	**Minimum**	**Maximum**	**Mean**	**Correlation with Life insurance**	**Correlation with Non-life insurance**
Non-life insurance	(0.74, 1.26)	(2427.61, 2857.40)	(386.28, 516.99)	(0.75, 0.80)	-
Life insurance	(0.49, 1.28)	(3058.58, 3803.76)	(503.87, 697.39)	-	(0.75, 0.80)
GDP per capita	(375.20, 550.90)	(56311.50, 94567.90)	(13896.60, 20524.50)	(0.77, 0.82)	(0.90, 0.91)
Death rate	(1.50, 1.52)	(16.17, 17.11)	(7.87, 8.00)	(0.09, 0.11)	(0.06, 0.07)
Urbanization	(11.92, 13.56)	(100,100)	(64.90, 66.29)	(0.37, 0.42)	(0.45, 0.46)
Religious	(0.01,0.01)	(99.61, 99.61)	(22.12, 22.12)	(-0.30, -0.29)	(-0.30, -0.28)
Dependency ratio	(1.25, 1.39)	(29.31, 33.92)	(14.89, 15.55)	(0.57, 0.61)	(0.57, 0.60)

Table 3.4.3: Correlation matrix of covariates in year 2004 to 2009.

	GDP per capita	Death rate	Urbanization	Religious
Death rate	(0.01, 0.03)			
Urbanization	(0.49, 0.52)	(-0.16, -0.15)		
Religious	(-0.29, -0.25)	(-0.38, -0.34)	(-0.14, -0.13)	
Dependency ratio	(0.58, 0.62)	(0.53, 0.54)	(0.30, 0.32)	(-0.53, -0.52)

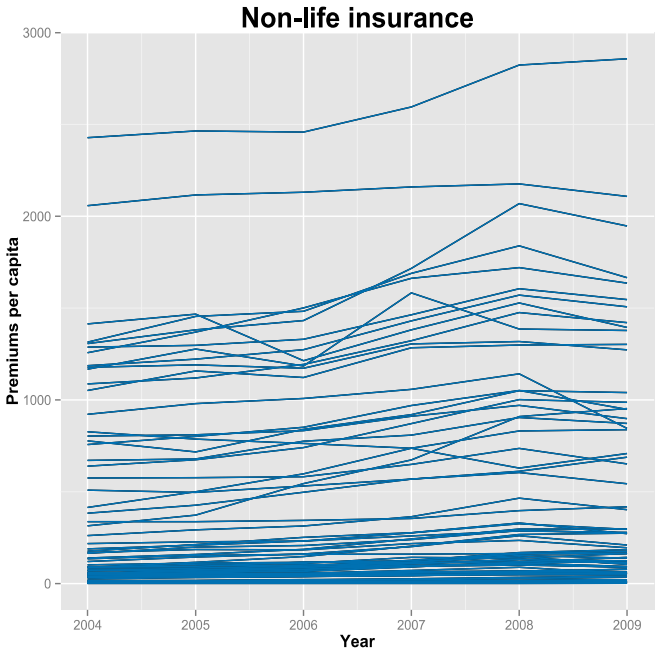

Figure 3.4.2: Non-life insurance premium per capita over time

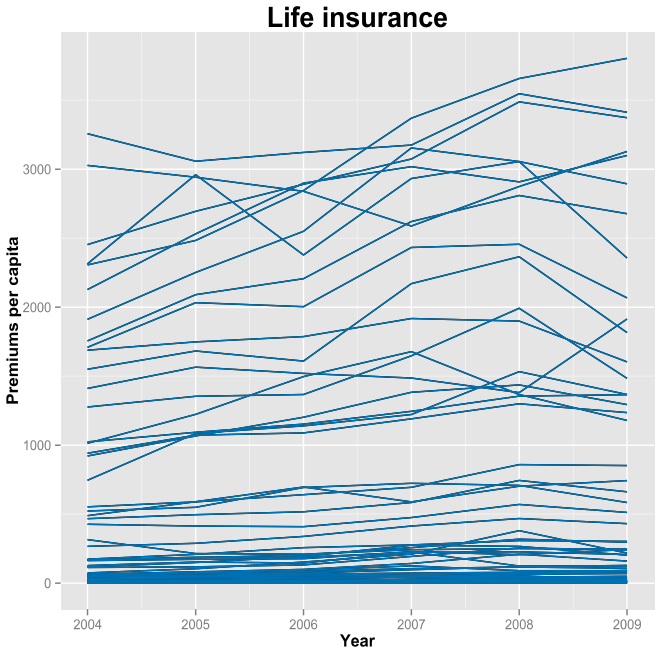

Figure 3.4.3: Life insurance premium per capita over time

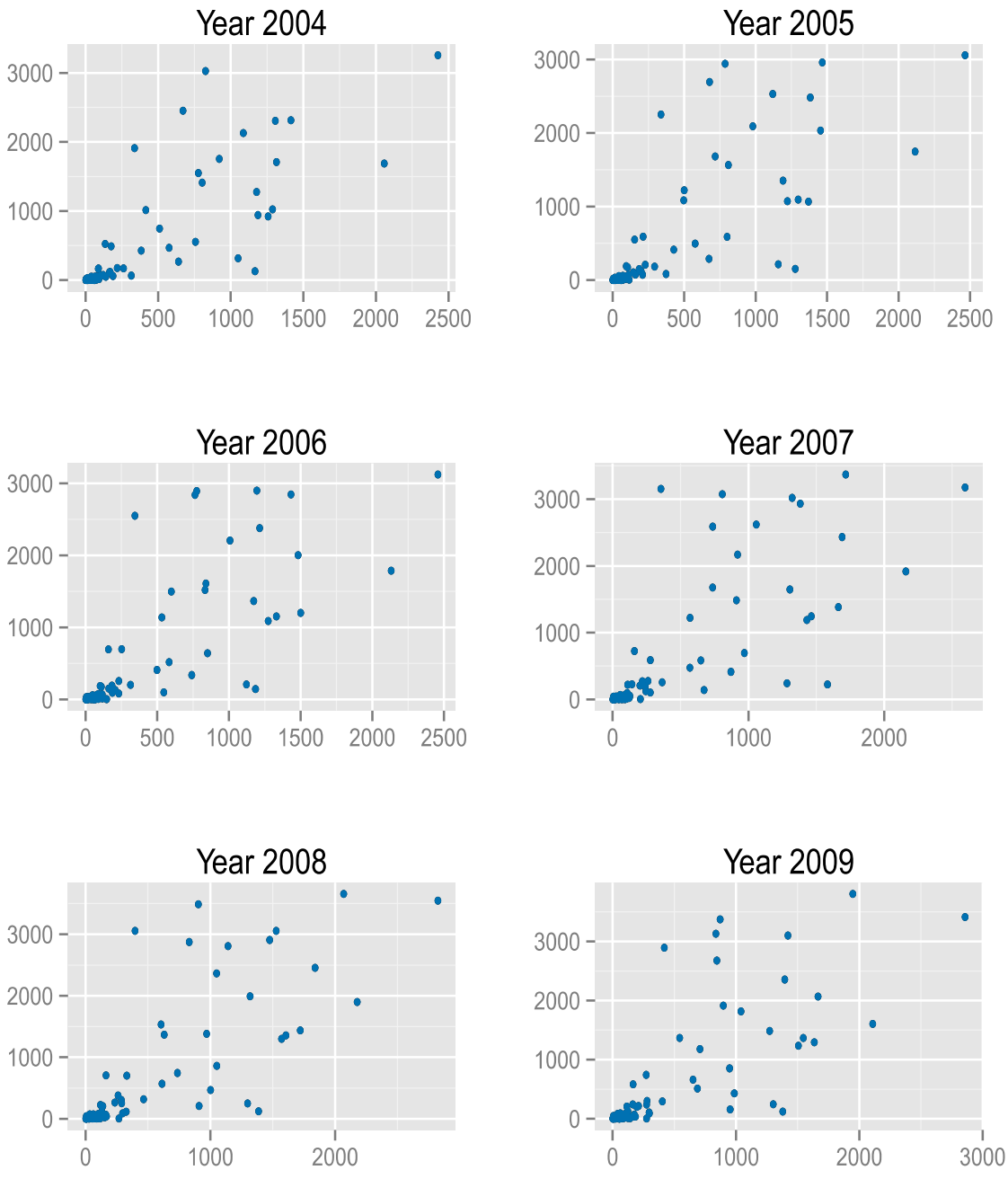

Figure 3.4.4: Scatter plots of the two response variables

x-axis: non-life insurance and y-axis: life insurance

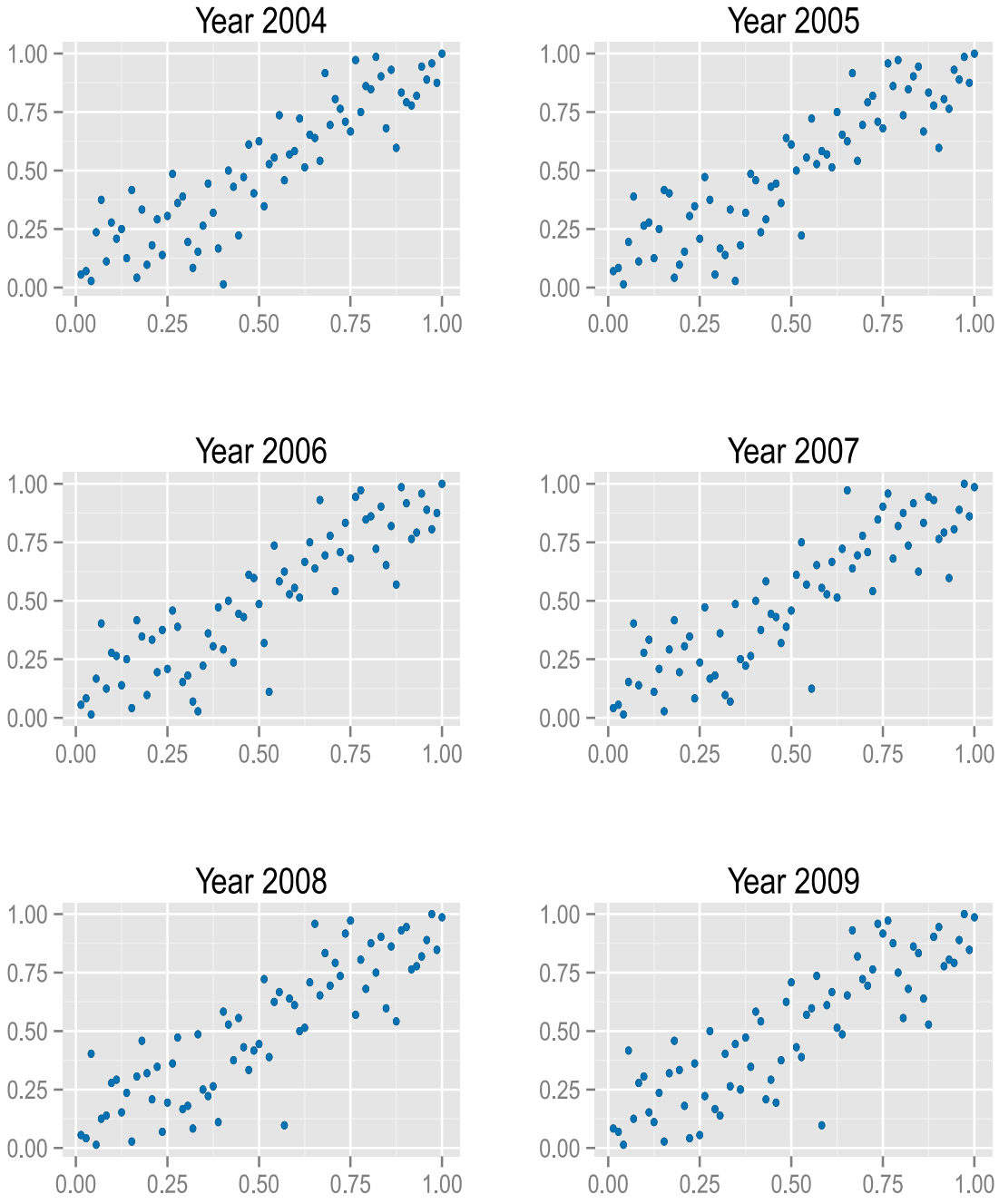

Figure 3.4.5: Scatter plots of the ranked response variables

x-axis: non-life insurance and y-axis: life insurance

3.4.2 Model calibration

Copulas provide a flexible way to construct multivariate distributions and are used in this section. By definition, a copula is simply a multivariate joint distribution defined on a d-dimensional cube $[0,1]^d$ such that every marginal follows uniform distribution on interval $[0,1]$. A copula captures both linear and nonlinear relationship and has been widely employed in multivariate analysis. The advantage is that it separates the modeling of marginal and dependence structure. In this application, we use the Gaussian family of copulas as specified below:

$$C(u_1, u_2; \rho) = \Phi_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \tag{3.4.1}$$

However, as explained later in the model calibration results, we additionally examined the quality of the fit of some Archimedean copulas as summarized in Table 3.4.4

Table 3.4.4: Some archimedean copulas

Family	**Copula function**	**Parameter**	**Kendall's τ**	**Spearman's ρ**
Clayton	$[u^{-a} + v^{-a} - 1]^{-1/a}$	$a \geq -1$	$\frac{a}{2+a}$	Complicated form
Frank	$\frac{1}{a}\ln\left[1 + \frac{(e^{au}-1)(e^{av}-1)}{e^a - 1}\right]$	$a \in \mathbb{R}$	$1 - \frac{4}{a}\{D_1(-a) - 1\}$	$1 - \frac{12}{a}\{D_2(-a) - D_1(-a)\}$
Gumbel	$\exp\left[-[(-\ln u)^a + (-\ln v)^a]^{1/a}\right]$	$a \geq 1$	$1 - a^-1$	No closed form

$D_1(.)$ and $D_2(.)$ indicate Debye functions.

The model specification is flexible enough to accommodate any marginals; however, for our purposes, we chose the class of Generalized Beta of the Second Kind (GB2) distributions. This class of marginal distributions is extremely flexible to accommodate highly-skewed distributions as we observe in our data. See Figure (3.4.4) and Figure (3.4.5) for histograms of non-life and life insurance premiums per capita. Both figures exhibit very long tails of the distribution. The GB2 distributions fit well for this type of data. For $Y \sim \text{GB2}(a, b, p, q)$

with $a \neq 0, b, p, q > 0$, the density function can be expressed as

$$f_y(y) = \frac{|a|\, y^{ap-1} b^{aq}}{B(p,q)(b^a + y^a)^{(p+q)}} \tag{3.4.2}$$

where $B(\cdot,\cdot)$ is the usual Beta function. Its distribution function can be expressed as

$$F_y(y) = B\left(\frac{(y/b)^a}{1+(y/b)^a}; p, q\right) \tag{3.4.3}$$

where $B(\cdot;\cdot,\cdot)$ is the incomplete Beta function. Using simple probability concepts, it is also straightforward to derive the following expression for the mean of a GB2 random variable:

$$\mathrm{E}(Y) = b\, \frac{B(p+1/a, q-1/a)}{B(p,q)}.$$

The regression is introduced into the GB2 marginals through the scale parameter. Suppose $\mathbf{x}$ is a vector of known covariates. We have: $Y|\mathbf{x} \sim \mathrm{GB2}(a, b(\mathbf{x}), p, q)$, where

$$b(\mathbf{x}) = \alpha + \beta'\mathbf{x}$$

Define residuals $\varepsilon_i = Y_i e^{-(\alpha_i + \beta'\mathbf{x}_i)}$ so that

$$\log Y_i = \alpha_i + \beta'\mathbf{x}_i + \log \varepsilon_i$$

where $\varepsilon_i \sim \mathrm{GB2}(a, 1, p, q))$. Construction of pp-plots arising from these residuals can then be used for diagnostics as later demonstrated for our data.

Finally, a natural assumption for the random effects term for the k-th response, for $k = 1, 2$, is based on the normal distribution as specified below:

$$\alpha_{ik} \sim N\left(0, \sigma_k^2\right)$$

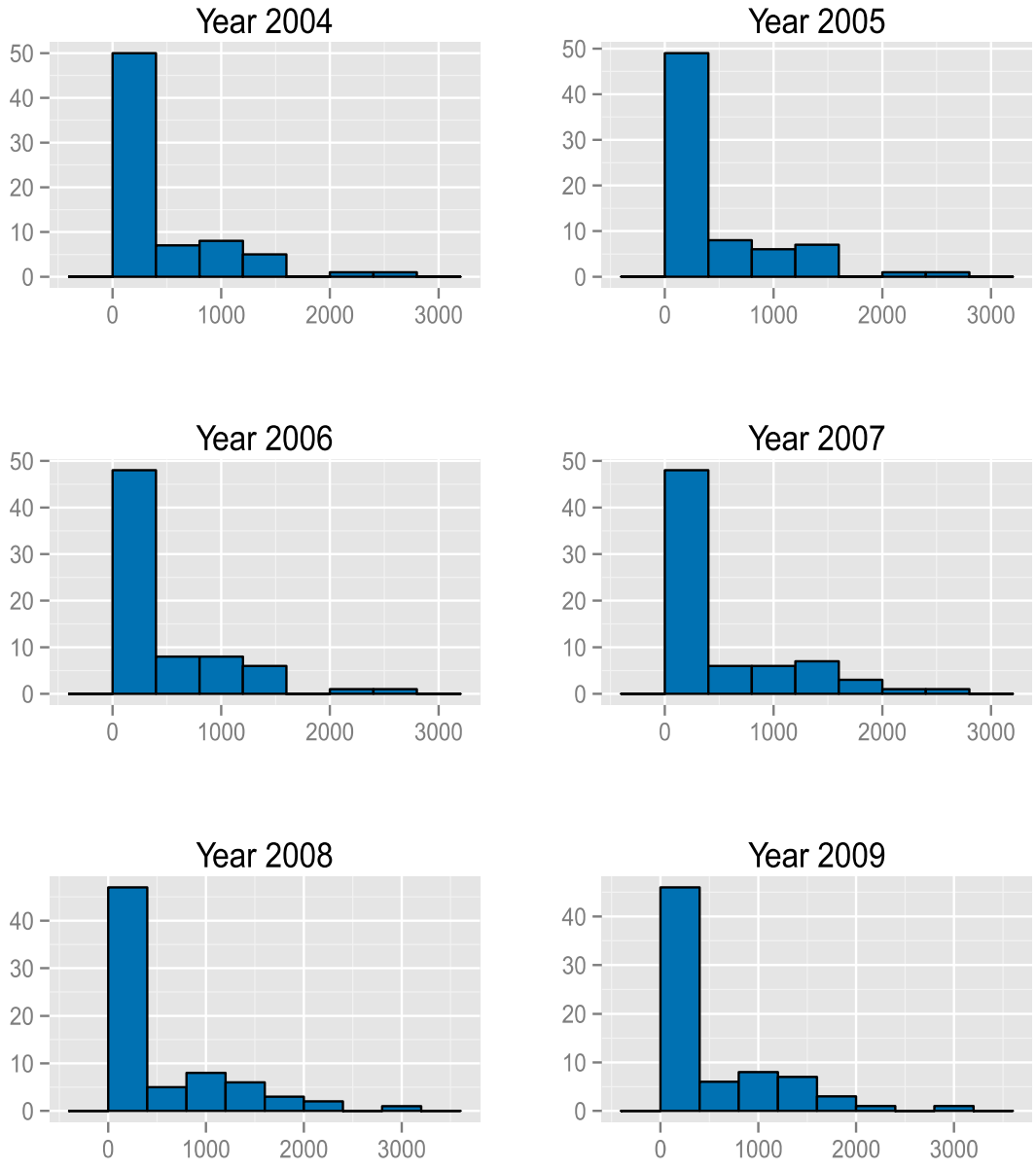

Figure 3.4.6: Non-life insurance premium per capita

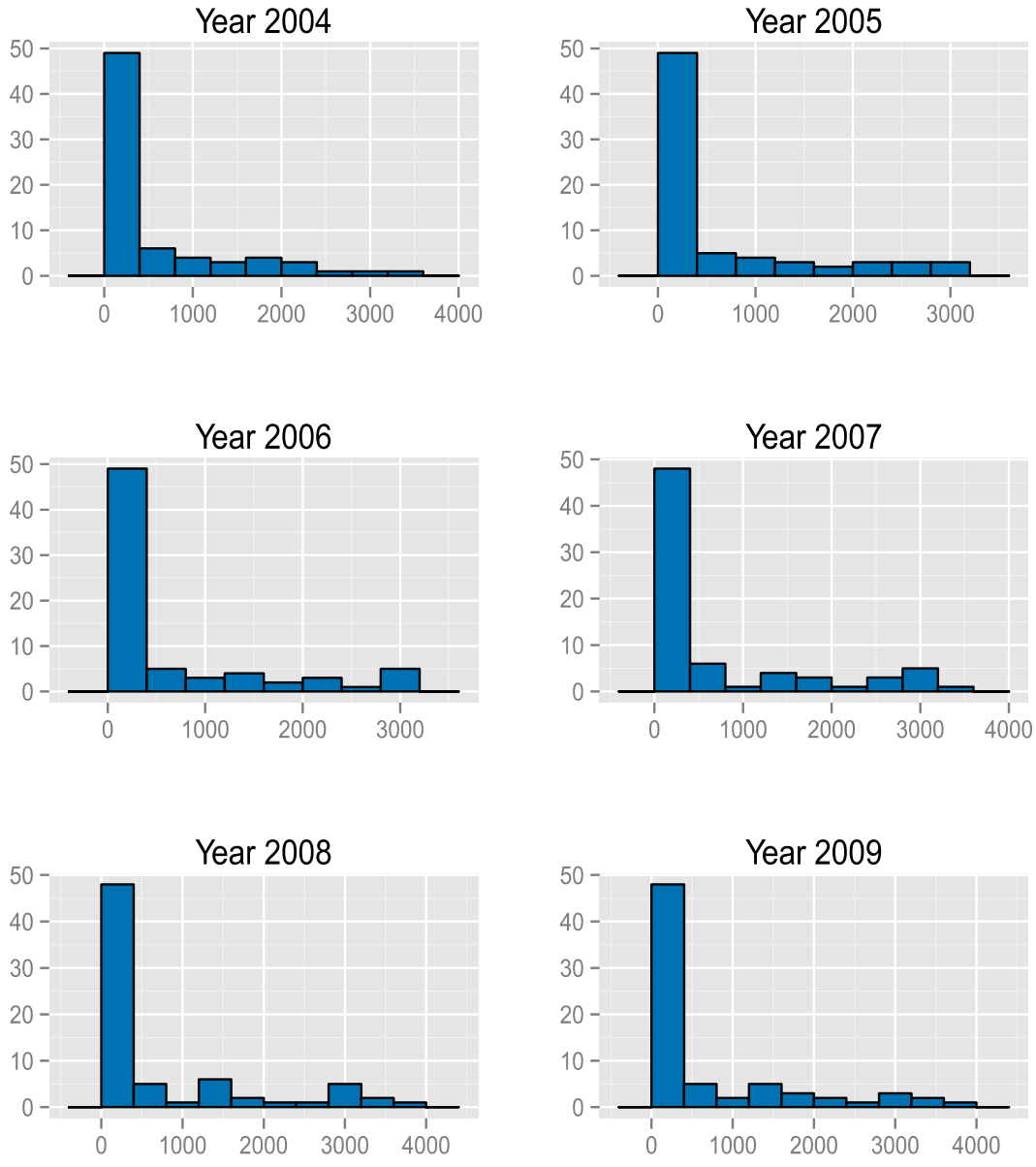

Figure 3.4.7: Life insurance premium per capita

3.4.3 Model calibration results with diagnostics

Maximum likelihood estimates together with respective standard error of these estimates are summarized in Tables 3.4.5 and 3.4.6. The estimates of the parameters in the non-regression component of the GB2 marginals are statistically significant. For the non-life insurance part, the regression parameter estimates demonstrate that all three predictor variables used (GDP per capita, Religious, Urbanization) are statistically significant. The signs of these coefficients tell us the following stories: (a) GPD per capita has a positive effect on demand, (b) there is an inverse relationship between the percentage of Muslims in the population and demand, and (c) urbanization has a positive effect on demand. For the life insurance part, this same set of predictor variables has the same effect on life insurance demand. Unexpectedly, the additional variables (Death rate, Dependency ratio) do not statistically help explain life insurance demand. Finally, the standard deviation estimate is significantly different from zero which explains the presence of a random effect. Hence this tells us the presence of a dynamic relationship of the data.

For marginal diagnostics, we present PP plots in Figures 3.4.8 and 3.4.9. These figures are enough to convince us of the quality of the fit of the GB2 marginals. However, we see that we may be able to further improve the quality of the fit by using additional covariates. This will be subject for further studies.

In the table, the estimate for the correlation parameter in the Gaussian copula is 0.5174 and is statistically significantly different from zero. This indicates a very strong dependence between the responses. This is not at all surprising according to our preliminary investigation of the scatter plots of the ranked responses.

Table 3.4.5: Fitted models

	Univariate fitted model for insurance demand					
	Non-life insurance			Life insurance		
Parameter	**Estimate**	**Std Error**	**p-val**	**Estimate**	**Std Error**	**p-val**
Covariates						
GDP per capita	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000
Religious	-0.0085	0.0023	0.0000	-0.0231	0.0040	0.0000
Urbanization	0.0567	0.0022	0.0000	0.0279	0.0061	0.0000
Death rate				0.0035	0.0333	0.9164
Dependency ratio				-0.0440	0.0297	0.1390
GB2 Marginals						
a	2.5636	0.1397	0.0000	1.0427	0.0611	0.0000
p	1.3957	0.1356	0.0000	3.7321	0.5371	0.0000
q	0.5369	0.0364	0.0000	0.5081	0.0330	0.0000
Random effect						
$Sigma_\alpha$	0.6471	0.0535	0.0000	0.8507	0.1088	0.0000

Table 3.4.6: Gaussian copula estimate

Parameter	**Estimate**	**Std Error**	**p-val**
ρ	0.5174	0.0315	0.0000

We also examined the quality of the fit of other Archimedean copulas and the estimated parameters are presented in Table 3.4.7. All three parameter estimates indicate again the presence of strong positive dependence. The Frank copula seems to perform slightly better than the Gaussian copula because it has marginally lower AIC and BIC statistics. This can also be seen in the copula PP-plots as exhibited Figure 5.4.7.

Table 3.4.7: Estimated copula functions

Copula function	**Estimated parameter**	**Standard error**	**P-value**	**AIC**	**BIC**
Clayton	0.4016	0.0592	0.0000	5689.501	5693.569
Frank	3.8477	0.3329	0.0000	5627.828	5631.896
Gumbel	1.5820	0.0605	0.0000	5629.81	5633.879
Normal	0.5174	0.0315	0.0000	5630.651	5634.719

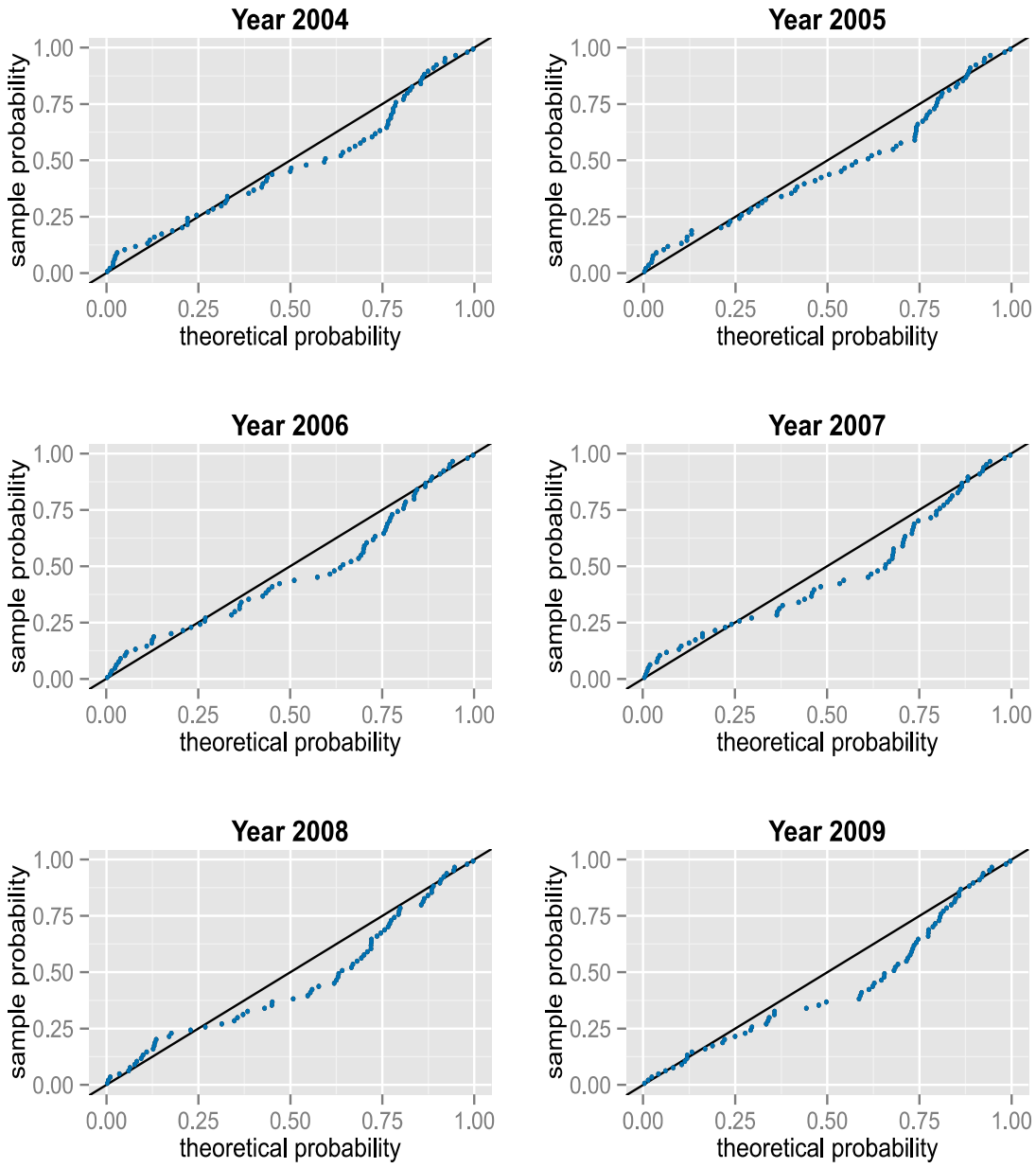

Figure 3.4.8: PP plots of the residuals for marginal diagnostics: non-life insurance

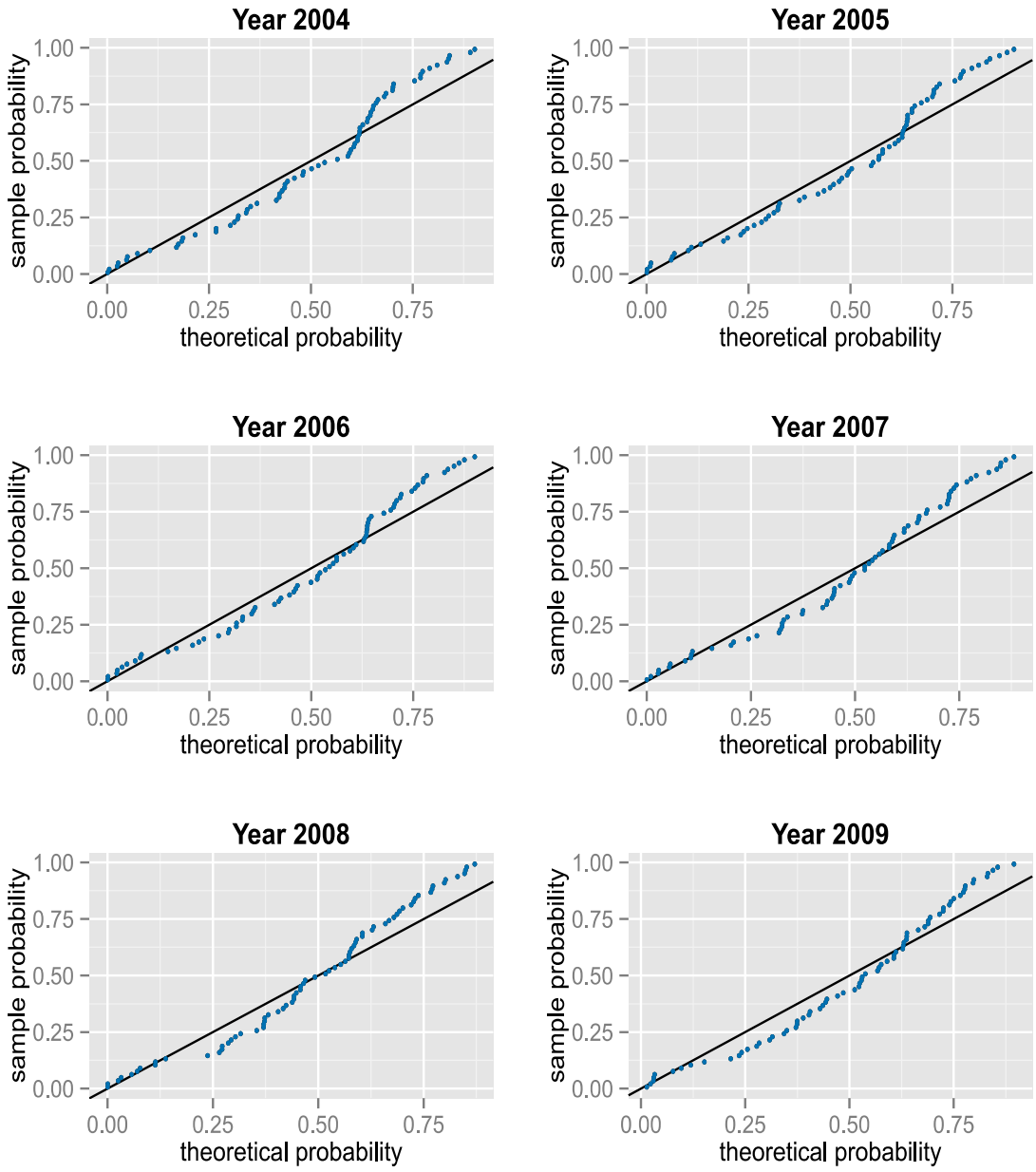

Figure 3.4.9: PP plots of the residuals for marginal diagnostics: life insurance

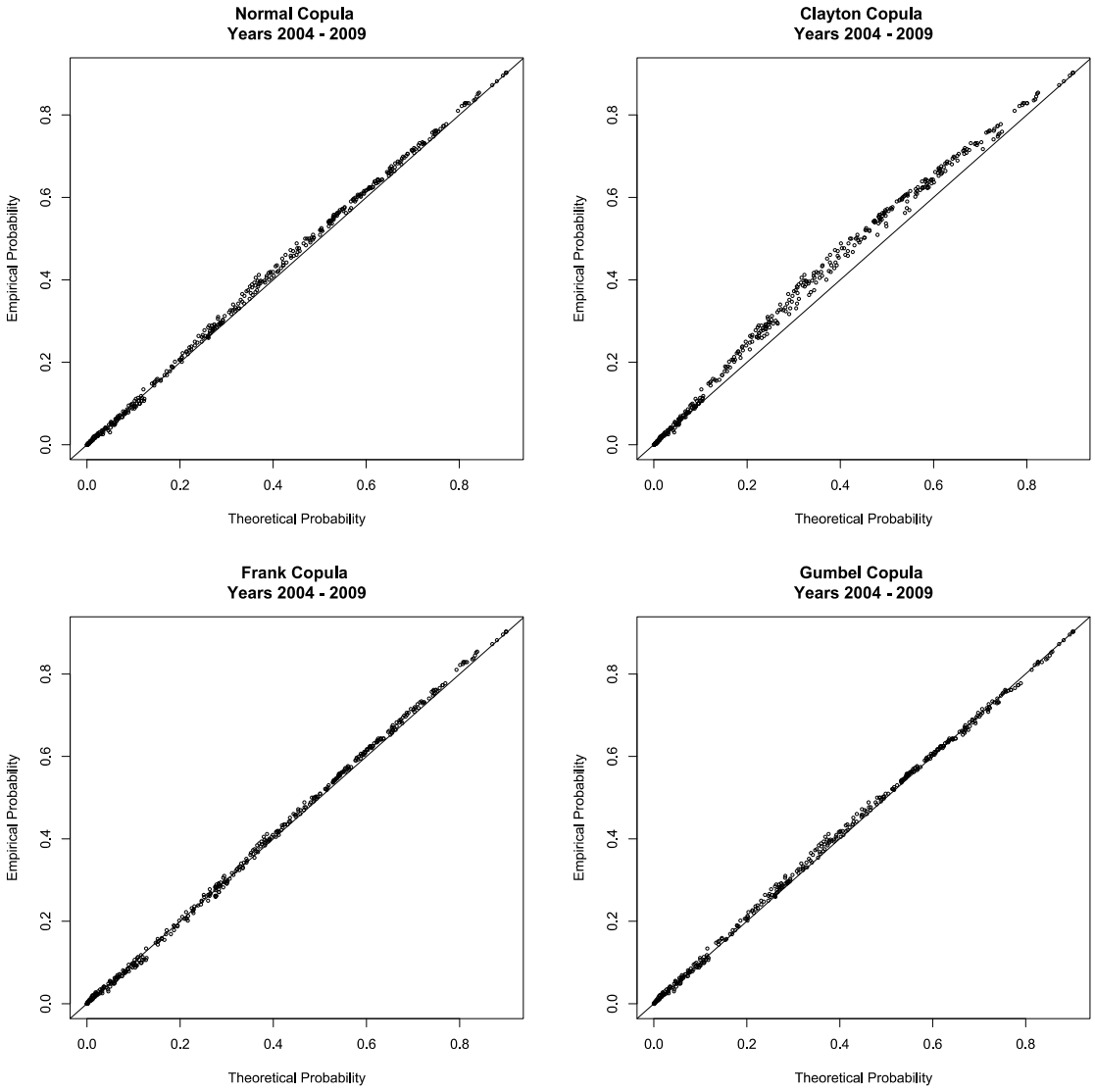

Figure 3.4.10: Copula diagnostics
PP - plots for normal, Clayton, Frank and Gumbel copula for year 2004 - 2009

Chapter 4

Loss Reserving in General Insurance

4.1 Introduction

Insurance companies collect premiums in order to bear the risk according to a contract agreement. In exchange, the insurance company pays the insured the promised benefit in the event of a covered financial loss. However, in many instances, insurers do not know the actual financial obligation of the insurance agreement they offered. Therefore, estimating expected future liabilities is vital to the company's financial assessments, financial decisions, strategic planning, capital requirements, and ratemaking. An accurate estimate of future obligation allows senior management to set aside enough amount of money, as a reserve, to help them alleviate the financial risk associated with the insurance products they sell. In many insurance companies, reserve amount corresponding to claims carry a considerable portion of total liability

of the company. Hence, the reserve figure is an important component of the ongoing viability of an insurance institution. Determining its optimal value together with management of these reserves is increasingly becoming integral part of every insurance enterprise risk management process. Especially, in property and casualty insurance, actuarial reserve is encountered extensively. Establishing the correct amount of reserve is aimed at helping to reduce the financial burden especially under extreme conditions. These loss reserves are undoubtedly one of the most substantial liability items on the insurance companys balance sheet; accurate calculations and reporting of these reserve estimates are therefore crucial to the company and other interested parties such as shareholders, auditors, tax authorities and regulators, to name a few. A systematic over-estimation of these provisions generally leads to increase in premiums for which affordability of coverage then becomes an issue. On the other hand, should these provisions become too low, this could affect the financial position of the insurer for which solvency then becomes an issue. Therefore, failure to calculate reserves with precision for outstanding and incurred but not reported (IBNR) claims will adversely affect a company's future financial development.

Historically, heavy-tailed (or long tailed) distributions can be observed for losses associated with general insurance. Even though most non-life insurance policies are available for a short period of time, the settlement process of claims can take a long time interval. For several types of insurance products such as professional liability insurance, the possible claims from accidents originating in a particular year often cannot be finalized, or realized, in the same year the accidents happen. There are many possible reasons for this: delay of loss reporting to the insurer, delay caused by difficulty of assessing the severity of the loss, long legal procedures to settle claims,

just to name a few. However, these claims have to be related to the same years for which the premiums were actually paid, henceforth, each year, the insurer has to set aside sufficient amount of reserves to cover losses for accidents that have already occurred, but either have not been reported or settled with the insurer. Figure 4.1.1 indicates the time line of a typical claim process in non-life insurance.

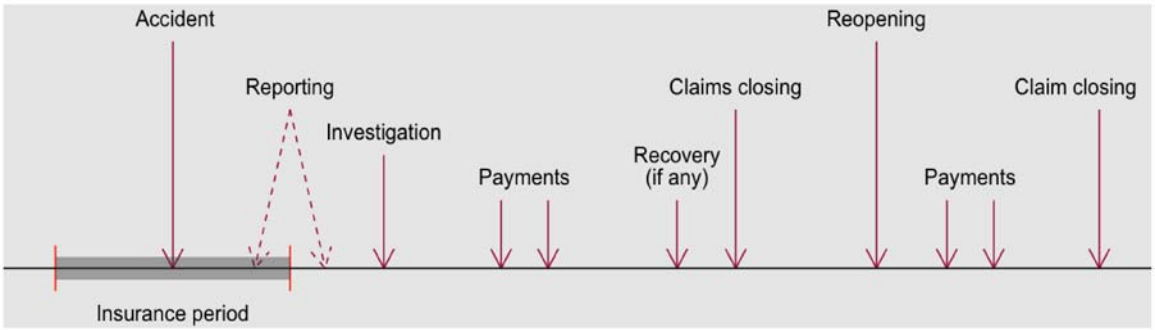

Figure 4.1.1: Time line of a non-life insurance claim

Source: "Stochastic claims reserving methods in insurance," by M.V. Wüthrich and M. Merz, The Wiley Finance Series, 2008, page 2.

Loss reserving is one of the most important topic in actuarial science, and hence, there is extensive research work in the area. This extensive body of literature includes simple loss reserving approach, like chain ladder, to more complex methods as generalized linear models (GLM) for loss reserving. As the insurance industry expands, more insurance lines and insurance products were introduced to the insurance market. As new insurance products become available to industry, complexity of claims handling as well as estimating loss reserving receive more attention in recent history. Therefore, estimating loss reserves for multiple lines of business with the consideration of dependencies among different insurance lines became more important. Due to this reason, there is an increasing interest that we are now seeing in both research

and practice on multivariate loss triangles. The demand for methods of loss reserving producing high accuracy of results in the insurance market compels the continued development in loss reserving over the years. Mack (1993), Taylor (2000), England and Verrall (2002), Schmidt (2006), Merz and Wuthrich (2008b) and Frees and Shi and Frees (2011) are to name a few research work, which explain different approach and endless progress in loss reserving research. In this chapter, we discuss many loss reserving approaches that have appeared both in research and practice.

In this chapter, we explore different reserving techniques that have appeared in the loss reserving literature. We started with univariate methods. Simple techniques, like Chain ladder method and Additive method, as well as stochastic methods are covered. Before we discuss about multivariate methods, we also discussed parametric models under the univariate case.

4.2 Common univariate methods

Most loss reserving methods have largely been based on the observable claims data from a single line of business that may be represented by a (univariate) random variable $C_{i,j}$, for cumulative claim payments, or $X_{i,j}$, for incremental payments, where the suffix i refers to the year of the accident and the suffix j to the development year. Therefore, calendar year of the claim is represented by $i + j$. In reality, claims are not paid in full and insurance companies make series of payments over the years to make final settlement of a claim. This series of payments indicates the development of claims over the years and in loss reserving literature we call development year payments.

Under the classical assumption, for each accident year, claims are settled either in the accident year or within a fixed number of years, n so that we set both $i, j \in \{0, 1, 2, \ldots, n\}$. Usually, observed losses can be arranged in a triangular format which produces the so-called loss run-off triangle also known as loss triangles. In this setup, the upper part of the triangle represents observable claims for calendar years $i+j \leq n$, while non-observable claims $i+j > n$ are represented by the lower part of the triangle. Table 4.2.1 illustrates a run-off triangle which represents the observable cumulative losses. Under the assumption that all claims corresponding to each accident year are settled in a fixed number of n development years, ultimate claim amounts of accident year i can be denoted by $C_{i,n}$. Hence, the outstanding claim reserve can be expressed as

$$R_i = C_{i,n} - C_{i,n-i}$$

for accident year $i = 1, 2, \ldots, n$.

One can easily transform cumulative claim payments into incremental claim payments by taking the difference of consecutive cumulative claim payments for a given accident year i.

$$X_{i,j} = C_{i,j} - C_{i,j-1}$$

for $j \in \{1, 2, \ldots, n\}$ and $i \in \{0, 1, \ldots, n\}$. It is intuitively appealing to think that modeling loss reserves using cumulative claims is equivalent to modeling loss reserves using incremental claims. In this regards, in the literature, we can observe some loss reserving models that utilize incremental claims for reserving. run-off triangle for incremental losses can be setup as table 4.2.2.

Predicting the lower part of the triangle (i.e. claims for $i + j > n$) in both cumu-

Table 4.2.1: Univariate cumulative loss triangle

	Development year								
Accident year	0	1	...	j	...	$n-i$	...	$n-1$	n
0	$C_{0,0}$	$C_{0,1}$	...	$C_{0,j}$	...	$C_{0,n-i}$	...	$C_{0,n-1}$	$C_{0,n}$
1	$C_{1,0}$	$C_{1,1}$	...	$C_{1,j}$	...	$C_{1,n-i}$	...	$C_{1,n-1}$	
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$			
i	$C_{i,0}$	$C_{i,1}$	...	$C_{i,j}$	...	$C_{i,n-i}$			
$\vdots$	$\vdots$	$\vdots$		$\vdots$					
$n-j$	$C_{n-j,0}$	$C_{n-j,1}$	...	$C_{n-j,j}$					
$\vdots$	$\vdots$	$\vdots$							
$n-1$	$C_{n-1,0}$	$C_{n-1,1}$							
n	$C_{n,0}$								

Table 4.2.2: Univariate incremental loss triangle

	Development year								
Accident year	0	1	...	j	...	$n-i$	...	$n-1$	n
0	$X_{0,0}$	$X_{0,1}$	...	$X_{0,j}$	...	$X_{0,n-i}$	...	$X_{0,n-1}$	$X_{0,n}$
1	$X_{1,0}$	$X_{1,1}$	...	$X_{1,j}$	...	$X_{1,n-i}$	...	$X_{1,n-1}$	
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$			
i	$X_{i,0}$	$X_{i,1}$	...	$X_{i,j}$	...	$X_{i,n-i}$			
$\vdots$	$\vdots$	$\vdots$		$\vdots$					
$n-j$	$X_{n-j,0}$	$X_{n-j,1}$	...	$X_{n-j,j}$					
$\vdots$	$\vdots$	$\vdots$							
$n-1$	$X_{n-1,0}$	$X_{n-1,1}$							
n	$X_{n,0}$								

lative and incremental setup provides estimates for accident year reserve or calendar year reserve. Forecasted claims of a particular year accident represent the accident year reserve, and calendar year reserve represents the prediction of the claims for a particular calendar year. Extensive research on accident year reserve has appeared in loss reserve literature. See, for example, Taylor (2000) for a comprehensive treatment of existing loss triangle models used in theory and practice. However, for accounting,

risk management and strategic decision purpose, calendar year reserves can be very useful for insurance companies.

4.2.1 Chain ladder method

Even though origins of chain ladder approach are not altogether clear, it is the most famous method in research as well as in practice among all techniques for loss reserving. This is mainly because of its simplicity. Due to the enormous utilization of chain ladder approach, there are many research that have been done and well documented in the literature. Kremer (1982), Taylor and Ashe (1983), Renshaw (1989), Mack (1993), to name a few research work, which made substantial developments in the understanding of chain ladder technique. One can define chain ladder approach as a non-parametric method for loss reserving. However, a number of stochastic modeling that involved chain ladder approach can also be found in theory. See, for example, Verrall (1989) and England and Verrall (2002).

Suppose cumulative claim amount $C_{i,j}$ of accident year i and development year j is strictly positive. Then, individual development factors corresponding to each accident and development year can be estimated as

$$\widehat{\psi}_{i,j} = \frac{C_{i,j}}{C_{i,j-1}}$$

for $i \in \{0, 1, \ldots, n-j\}$ and $j \in \{1, 2, \ldots, n\}$.

Model assumptions

The basic assumptions of the chain ladder method include

1. Cumulative claims $C_{i,j}$ are mutually independent for all accident years i.

2. There exists a set of positive deterministic factors ψ_j for each development year $j \in \{1, 2, \ldots, n\}$ and these factors are mutually independent for all accident years.

Estimates of these age-to-age development factors in the chain ladder method satisfy following equation.

$$\widehat{\psi}_j = \frac{\sum_{i=0}^{n-j} C_{i,j}}{\sum_{i=0}^{n-j} C_{i,j-1}} \tag{4.2.1}$$

Formulation of these age-to-age development factors can be rewritten as weighted average of above mentioned individual development factor.

$$\widehat{\psi}_j = \frac{\sum_{i=0}^{n-j} C_{i,j-1} \cdot \widehat{\psi}_{i,j}}{\sum_{i=0}^{n-j} C_{i,j-1}}$$

Therefore, chain ladder prediction for $i + j > n$ is

$$\widehat{C}_{i,t} = C_{i,n-i} \cdot \prod_{k=n-i+1}^{t} \widehat{\psi}_j$$

for all $i \in \{0, 1, \ldots, n\}$ and $t \in \{n - i + 1, n - i + 2, \ldots, n\}$.

Based on the above definitions from chain ladder method, one can express the predicted ultimate cumulative claims amount for each accident year i as

$$\widehat{C}_{i,n} = C_{i,n-i} \cdot \widehat{\psi}_{n-i+1} \cdot \ldots \cdot \widehat{\psi}_n$$

and the corresponding accident year predicted reserve

$$\widehat{R}_i = \widehat{C}_{i,n} - C_{i,n-i} = C_{i,n-i} \left(\widehat{\psi}_{n-i+1} \cdot \ldots \cdot \widehat{\psi}_n - 1 \right) \tag{4.2.2}$$

for $i \in \{1, \ldots, n\}$.

Despite numerous research under this method and its vast popularity in practice, it is well documented that there are several disadvantages in the chain ladder. See Klugman, Panjer, and Willmot (2012), for example.

4.2.2 Additive method

Additive loss reserving, also known as incremental loss ratio method, is one of the most popular and simple approach which is based on incremental losses $X_{i,j}$. The additive loss reserving method utilizes additional information other than loss amounts from the run-off triangles. To capture the loss ratio corresponding to run-off data, incremental claims are divided by an exposure variable which can be used to measure the volume of the business. Accident year premiums, number of policies, or number of claims are to name a few common exposure variables used in research and practice within the context of additive model. More importantly, unlike the chain ladder approach, predictions from the additive method does not depend only on the last observation on the diagonal. This helps to overcome the effect of outliers on the diagonal for predictions. See Merz and Wuthrich (2009a), for brief summarization of the differences between the chain ladder and the additive methods.

Model assumptions

1. A set of parameters $\theta_0, \theta_1, \ldots, \theta_n$ indicates observed exposure variable over accident years.

2. There exists unknown parameters $\kappa_1, \kappa_2, \ldots, \kappa_n$ which satisfy

$$\kappa_j = E\left[\frac{X_{i,j}}{\theta_i}\right] \text{ for all } i, j \in \{0, 1, 2, \ldots, n\}.$$

The parameter κ_j indicates the development patterns for incremental loss ratio over the accident years. The additive prediction of the incremental losses from the lower part of the triangle can be estimated as

$$\widehat{X}_{i,j} = \theta_i \widehat{\kappa}_j$$

where

$$\widehat{\kappa}_j = \frac{\sum_{k=0}^{n-j} X_{k,j}}{\sum_{k=0}^{n-j} \theta_k}$$

for $i \in \{0, 1, \ldots, n\}$ and $j \in \{n-i+1, \ldots, n\}$. Hence, the predicted cumulative losses, $\widehat{C}_{i,j}$, with $i + j > n$ are defined as

$$\widehat{C}_{i,j} = C_{i,n-i} + \theta_i \sum_{l=n-i+1}^{j} \widehat{\kappa}_l$$

and the estimated ultimate reserve can be expressed as

$$\widehat{R}_i = \widehat{C}_{i,n} - C_{i,n-i} = \theta_i \sum_{l=n-i+1}^{n} \widehat{\kappa}_l \tag{4.2.3}$$

for $i \in \{1, \ldots, n\}$.

4.2.3 Bornhuetter-Ferguson method

A unifying survey of some methods of loss reserving based on run-off triangles can be found from Schmidt and Zocher (2007). In this survey, a general class of loss reserving method called the Bornhuetter-Ferguson method is re-examined. Most of the common loss reserve techniques, including chain ladder approach, are derived as special cases of the general Bornhutter-Ferguson method exploiting the development pattern for cumulative quotas.

Based on the assumptions that there exist two sets of parameters $\phi_0, \phi_1, \ldots, \phi_n$ and $\lambda_0, \lambda_1, \ldots, \lambda_n$ with $\lambda_n = 1$, the Bornhuetter-Ferguson method can be defined as

$$\widehat{C}_{i,j} = C_{i,n-i} + \left(\widehat{\lambda}_j - \widehat{\lambda}_{n-i}\right) \widehat{\phi}_i \tag{4.2.4}$$

where $\widehat{\phi}_0, \widehat{\phi}_1, \ldots, \widehat{\phi}_n$ and $\widehat{\lambda}_0, \widehat{\lambda}_1, \ldots, \widehat{\lambda}_n$ represent prior estimators of the expected ultimate cumulative losses, $E[C_{i,n}]$, and the development patterns, respectively. These cumulative predictions highly rely on prior estimators which are largely based on actuarial judgment using both internal and external information. See Schmidt and Zocher (2007).

4.3 Stochastic methods

Despite the simplicity of traditional loss reserving techniques discussed in the previous section, common disadvantage is their inability to provide more information regarding future obligations other than a single estimate of the loss reserve. Stochastic modeling can be used to overcome this issue in loss reserving. In addition to the best

estimate, stochastic loss reserving techniques can be used to estimate the variability of claims reserve. It is obvious that none of the techniques can predict the actual future obligations. Therefore, understanding the uncertainty of the prediction could lead an insurance company to have a better picture of all possible scenarios of future obligations and allow to make a more informed financial decision, for example, to set extra reserve for catastrophic events. In the statistical literature, estimating the standard deviation resulting from the desired statistical model is a common approach that is used to assess the uncertainty. In this regard, two strands of research works can be found in the loss reserve literature:

1. Defining the first two moments without specifying the underlying distribution.

2. Specifying the distribution for the underlying data.

4.3.1 Mack model

A prominent work of Mack (1993) introduced distribution-free method to estimate the standard error of reserve estimates computed by the chain ladder approach. He defined first two moments of cumulative losses without specifying the underline distribution of cumulative losses.

Suppose $A_{i,j} = \{C_{i,k} | 0 \leq k \leq j\}$ be the set of all observed data up to the $j + 1^{th}$ development year. Then, according to the Mack model, the first two moments of loss distribution are defined as

$$E\left(C_{i,j+1}|A_{i,j}\right) = \psi_j C_{i,j} \tag{4.3.1}$$

$$Var\left(C_{i,j+1}|A_{i,j}\right) = \sigma_j^2 C_{i,j} \tag{4.3.2}$$

for all $i \in \{0, \ldots, n\}$ and $j \in \{0, \ldots, n-1\}$. Unknown parameter ψ can be calculated

by the equation 4.2.1, which is an unbiased estimate. An unbiased estimator of σ_j^2 is given by

$$\widehat{\sigma}_j^2 = \frac{1}{n-j-1}\sum_{i=0}^{n-j} C_{i,j}\left(\frac{C_{i,j+1}}{C_{i,j}} - \widehat{\psi}_j\right)^2$$

for all $i \in \{0,\ldots,n\}$ and $j \in \{1,\ldots,n-1\}$. The model proposed the use of conditional mean squared error value in order to understand the variability of the estimated loss reserve. Therefore,

$$MSE\left(\widehat{R}_i\right) = E\left(\left(\widehat{R}_i - R_i\right)^2 | B\right)$$

where $B = \{C_{i,j} | i+j < n+1\}$ is the set of all observed data, which is corresponding to upper triangle data in run-off triangle. It is also possible to show that

$$E\left(\left(\widehat{R}_i - R_i\right)^2 | B\right) = E\left(\left(\widehat{C}_{i,n} - C_{i,n}\right)^2 | B\right)$$

and hence $MSE\left(\widehat{R}_i\right) = MSE\left(\widehat{C}_{i,n}\right)$. See Mack (1993) for more details.

Following the Mack's model for loss reserving plus estimating variability of predicted reserve, a number of stochastic loss reserving methods has appeared in the literature. For example, Schmidt and Schnaus (1996) extended Mack's model that allows to characterize the optimality of chain ladder factors.

4.3.2 GLM models

The parametric methods based on distributional families gained much attention in loss reserving as these methods could derive the corresponding predictive distribution of unpaid losses. The use of GLM models within this context has also been described

extensively in the actuarial literature; see Renshaw (1989) and Kass et al. (2008). Over-dispersed Poisson model, which can reproduce the chain ladder estimation, is one of the popular reserving models within the GLM context. Under the assumption that incremental claims, $X_{i,j}$, are independently distributed, the first two moments can be defined as

$$E\left(X_{i,j}\right) = \mu_{i,j} \text{ and } Var\left(X_{i,j}\right) = \phi\mu_{i,j} \tag{4.3.3}$$

In this model, the chain ladder type linear predictor has the mean parameter of the Poisson distribution via a log link function.

$$\log\left(\mu_{i,j}\right) = c + \alpha_i + \beta_j \tag{4.3.4}$$

where α_i and β_j stand for accident years and development years, respectively, with initial values $\alpha_1 = \beta_1 = 0$. Verrall (2000) derived the negative binomial model as an extension of the poisson model.

The gamma distribution is another assumption for loss data, especially for skewed and medium-tailed data, commonly used in the actuarial literature. Mack (1991) pioneered the gamma distribution for loss triangle data. Renshaw and Verrall (1998) applied gamma distribution for incremental claims within the GLM framework. The approach is very similar to over-dispersed Poisson model. The only difference is $Var\left(X_{i,j}\right) = \phi\mu_{i,j}^2$ instead of $Var\left(X_{i,j}\right) = \phi\mu_{i,j}$ in equation 4.3.3. The log link function incorporated has the linear predictor with mean parameter as in equation 4.3.4. Again, this model can provide similar results as the chain ladder approach but not guaranteed for all data set.

4.3.3 Lognormal model

Early work on stochastic claim reserving has been done by Kremer (1982). He is the first to introduce lognormal model for loss reserving with incremental claim amounts. Reserve from lognormal model can well address the data that display heavy tail distribution. Logarithmic transformation of incremental claims defines a new variable $Y_{i,j} = \log(X_{i,j})$ and based on lognormal model, we have

$$Y_{i,j} = \mu_{i,j} + \varepsilon_{i,j}$$

which satisfies

$$Y_{i,j} \sim LN\left(\mu_{i,j}, \sigma^2\right) \text{ and } \varepsilon_{i,j} \sim N\left(0, \sigma^2\right)$$

where the mean parameter, $\mu_{i,j}$, can be estimated using the chain ladder type of structure as

$$\mu_{i,j} = c + \alpha_i + \beta_j.$$

Here, α_i and β_j represent accident year i and development year j, respectively.

Via the introduction of a loglinear model with parameters to allow for trends in a run-off triangle in three directions (horizontally, vertically and diagonally), the actuarial group from the Katholieke-Universiteit of Leuven has suggested an important step in the development of stochastic reserving techniques. Hereby the (i, j)th element in a run-off triangle is modeled by means of three parameters, namely $log(X_{i,j}) = \alpha_i + \beta_j + \gamma_{i+j} + \varepsilon_{i,j}$, see, for example, see Goovaerts et al. (1990) and Barnett and Zehnwirth (2000). The parameters α_i and β_j describe respectively the effects of accident and development year, and the additional parameter γ_{i+j} describes the calendar year effect (to model e.g. the effects of economic inflation and changing

legislations). Techniques to estimate these parameters are well-known. Many of these research works suggested to use the lognormal models which allow for computing the least squares estimates of the parameters.

England and Verrall (2002) reviewed a number of of stochastic models in the literature. They discussed the connection between each approach while explaining how to implement each method in practice. Wuthrich and Merz (2008) is another example that provides a comprehensive review of stochastic claim reserving in non-life insurance.

4.4 Multivariate methods

For multiple lines of business, the collection of univariate loss triangles derived from each lines of business can be used to determine the aggregate loss reserves corresponding to the total portfolio of loss triangles. Each of these univariate loss triangles may be interpreted as a subportfolio of an aggregate portfolio. Analogously, for an industry based study, each loss triangle comes from a different insurance company. See working paper from Shi (2013), for example. This type of study can be used for several reasons. For example, it can be used for understanding claim trends, settlement patterns, and loss ratio changes over time. On the other hand, a company based study, which includes multiple loss triangles from the same company but from different lines of business is common in practice. Simple aggregation of loss reserving corresponding to multiple loss triangles does not provide much accurate results because of the possible dependency among individual loss triangles. Therefore, multivariate loss reserving techniques has gained increasing interest over the last decades

in the literature on loss reserving.

Suppose an insurance company has m number of subportfolios corresponding to m number of different insurance lines with the same number of development years. All claims are settled either in the accident year or within fixed number of years, n. The incremental paid claims vector of m dimension can be denoted by

$$\mathbf{X_{i,j}} = (X_{i,j,1}, \ldots, X_{i,j,m}) \tag{4.4.1}$$

and cumulative paid claims by

$$\mathbf{C_{i,j}} = (C_{i,j,1}, \ldots, C_{i,j,m}) \tag{4.4.2}$$

where

$$C_{i,j,k} = \sum_{p=0}^{j} X_{i,p,k}$$

for $i, j \in \{0, 1, \ldots, n\}, k \in \{1, 2, \ldots, m\}$ and $X_{i,p,k}$ indicates claim amount of $i+1$-th accident year $j+1$-th development period from k-th subportfolio. Assuming data available for the same set of accident years throughout all subportfolios, the cumulative claim amounts corresponding to k-th subportfolio can be denoted by

$$\mathbb{D}_k = \{C_{i,j,k} | i + j \leq n\}$$

and cumulative claims over all subportfolios by

$$\mathbb{D} = \bigcup_{k=1}^{m} \mathbb{D}_k.$$

The goal is to predict the claim amount belonging to complement of the above set. One can represent the multivariate loss triangle as in Table 4.4.1 with cumulative paid claim variables as we discussed in this section. Similar to the univariate case, multivariate loss triangle for incremental paid claims can be derived by taking the difference of the corresponding cumulative payments in each subportfolio.

Table 4.4.1: Multivariate incremental loss triangle

	Development year								
Accident year	0	1	...	j	...	$n-i$	...	$n-1$	n
0	$\mathbf{C_{0,0}}$	$\mathbf{C_{0,1}}$	...	$\mathbf{C_{0,j}}$	...	$\mathbf{C_{0,n-i}}$	...	$\mathbf{C_{0,n-1}}$	$\mathbf{C_{0,n}}$
1	$\mathbf{C_{1,0}}$	$\mathbf{C_{1,1}}$	...	$\mathbf{C_{1,j}}$	...	$\mathbf{C_{1,n-i}}$	...	$\mathbf{C_{1,n-1}}$	
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$			
i	$\mathbf{C_{i,0}}$	$\mathbf{C_{i,1}}$	...	$\mathbf{C_{i,j}}$	...	$\mathbf{C_{i,n-i}}$			
$\vdots$	$\vdots$	$\vdots$		$\vdots$					
$n-j$	$\mathbf{C_{n-j,0}}$	$\mathbf{C_{n-j,1}}$	...	$\mathbf{C_{n-j,j}}$					
$\vdots$	$\vdots$	$\vdots$							
$n-1$	$\mathbf{C_{n-1,0}}$	$\mathbf{C_{n-1,1}}$							
n	$\mathbf{C_{n,0}}$								

The concept of distribution-free method on univariate loss reserving by Mack (1993) has been extended for the multivariate setup. Braun (2004) introduced a bivariate stochastic model which is an extended version of the Mack's model to estimate the prediction error of the chain ladder method for a portfolio of several correlated run-off triangles. To capture the correlation of subportfolios, fixed correlation structure is assumed between individual development factors of the two corresponding development years. He developed correlation coefficients assuming independence across accident years, but dependence across development years. In order to approximate the lower and upper bounds for total unpaid losses of a portfolio, Hürlimann (2005) proposed a linear approximate estimation of the bivariate chain ladder factors. Pöhl

and Schmidt (2005) introduced multivariate chain ladder approach which is an extension of Schmidt and Schnaus (1996) and Braun (2004). In the same year Kremer (2005) also proposed multivariate extension of the chain ladder approach. However, due to the complexity of the parameter estimation and failure to address the issues of additivity of dependent loss triangles, the=is proposed method did not get enough attention. Schmidt (2006) discussed an extended version of two univariate methods, additive and chain ladder, within a multivariate setup.

4.4.1 Multivariate chain ladder (MCL)

Pöhl and Schmidt (2005) extended the multivariate version of Schmidt and Schnaus (1996) model and bivariate model of Braun (2004). This new model they named as multivariate chain ladder approach for portfolios of loss triangles with different lines of business. This model assumed fixed conditional correlation coefficients between the individual development factors of a fixed development period over the subportfolios. This is also equivalent to developing correlation coefficients between aggregate claims of a fixed development year over the portfolio. Unlike the model from Braun (2004), MCL discusses multivariate predictions as well as prediction error.

We can use matrices to simplify the complexity of the multivariate chain ladder formulation. Suppose

$$\mathbf{diag}(\mathbf{a}) = \begin{pmatrix} a_1 & & 0 \\ & a_2 & \\ & \cdot & \\ 0 & & a_n \end{pmatrix} \text{ and } \mathbf{diag}(\mathbf{a})^p = \begin{pmatrix} a_1^p & & 0 \\ & a_2^p & \\ & \cdot & \\ 0 & & a_n^p \end{pmatrix}$$

indicate n by n diagonal matrices where $\mathbf{a} = \{a_1, \ldots, a_n\}$. The individual devel-

opment factors corresponding to the kth subportfolio will be denoted by

$$\psi_{i,j,k} = \frac{C_{i,j,k}}{C_{i,j-1,k}} \text{ and } \psi_{\mathbf{i,j}} = \{\psi_{i,j,1}, \ldots, \psi_{i,j,m}\}$$

and we can write

$$\mathbf{C_{i,j}} = \mathbf{diag}(\mathbf{C_{i,j-1}}) \cdot \psi_{\mathbf{i,j}} = \mathbf{diag}(\psi_{\mathbf{i,j}}) \cdot \mathbf{C_{i,j-1}}$$

for $i \in \{0, 1, \ldots, n\}$, $j \epsilon \{1, 2, \ldots, n\}$ and $k \epsilon \{1, 2, \ldots, m\}$.

Model assumption:

Model assumptions for the multivariate chain ladder method is very similar to the univariate case. In addition to basic chain ladder assumptions, the multivariate extension incorporates correlation among different subportfolios in the estimation of deterministic constant corresponding to the development factors.

1. Cumulative claims $\mathbf{C_{i,j}}$ are independent over accident years i.

2. There exists an m-dimensional vector for each development year j which consists of positive deterministic factors

$$\psi_{\mathbf{j}} = (\psi_{i,j,1}, \ldots, \psi_{i,j,m})^T$$

and symmetric positive definite m by m matrix Σ_j for $j \in \{1, \ldots, n\}$.

Under the above assumptions, the multivariate chain ladder model can be defined as

$$\mathbf{E}\left[\mathbf{C_{i,j}}|\mathbf{C_{i,j-i}}\right] = \mathbf{diag}\left(\psi_{\mathbf{j}}\right) \cdot \mathbf{C_{i,j-1}} \tag{4.4.3}$$

$$\mathbf{Cov}\left(\mathbf{C_{i,j}}, \mathbf{C_{i,j}}|\mathbf{C_{i,j-1}}\right) = \mathbf{diag}\left(\mathbf{C_{i,j-1}}\right)^{\frac{1}{2}} \cdot \mathbf{\Sigma_j} \cdot \mathbf{diag}\left(\mathbf{C_{i,j-1}}\right)^{\frac{1}{2}}. \tag{4.4.4}$$

Therefore, ultimate claim amounts by accident year is represented by

$$\mathbf{E}\left[\mathbf{C_{i,n}}|\mathbb{D}\right] = \prod_{\mathbf{j=n-i+1}}^{\mathbf{n}} \mathbf{diag}\left(\psi_{\mathbf{j}}\right) \cdot \mathbf{C_{i,n-i}} \tag{4.4.5}$$

and estimation can be completed using

$$\widehat{\mathbf{C}}_{\mathbf{i,n}} = \prod_{\mathbf{j=n-i+1}}^{\mathbf{n}} \mathbf{diag}\left(\widehat{\psi}_{\mathbf{j}}\right) \cdot \mathbf{C_{i,n-i}} \tag{4.4.6}$$

where

$$\widehat{\psi}_{\mathbf{j}} = \left(\widehat{\psi}_{i,j,1}, \ldots, \widehat{\psi}_{i,j,m}\right)$$

$$= \left(\sum_{\mathbf{i=0}}^{\mathbf{n-j}} \mathbf{diag}\left(\mathbf{C_{i,j-1}}\right)^{\frac{1}{2}} \mathbf{\Sigma_j^{-1}} \mathbf{diag}\left(\mathbf{C_{i,j-1}}\right)^{\frac{1}{2}}\right)^{-1} \mathbf{diag}\left(\mathbf{C_{i,j-1}}\right)^{\frac{1}{2}} \mathbf{\Sigma_j^{-1}} \mathbf{diag}\left(\mathbf{C_{i,j-1}}\right)^{\frac{1}{2}} \cdot \psi_{\mathbf{i,j}}$$

There are some developments on multivariate chain ladder method that has appeared in the literature. Schmidt (2006), Merz and Wuthrich (2008a), Merz and Wuthrich (2008b) and Zhang (2010) provide brief highlights. Schmidt (2006) compared the results from reserves of multivariate loss triangles using different methods. First, they applied univariate chain ladder method for an aggregate portfolio, and later applied univariate chain ladder method on loss triangles from each subportfolio separately. Finally, multivariate chain ladder method was applied to the same set of subportfolios. Even though, multivariate chain ladder method provides optimal results among other approaches considered in the paper, the authors pointed out

that results also may be dependent on various factors related to the data. Merz and Wuthrich (2008a) also developed multivariate chain ladder method and later, Merz and Wuthrich (2008b) extended this work using the mean square error prediction (MSEP) on several correlated run-off portfolios. Extended version of multivariate chain ladder method is introduced by Zhang (2010). This new model is named as the General Multivariate Chain Ladder (GMCL) model.

4.4.2 Multivariate additive model

Multivariate additive method for loss reserving with correlated run-off triangles was introduced by Hess, Schmidt, and Zocher (2006). Under the model assumptions, non-observable incremental claims, the lower part of the triangle, are predicted incorporating Gauss-Markov linear predictors. Therefore, this method is constrained within certain type of correlated run-off data due to its linear model setup.

Model assumptions:

1. There exist positive definite symmetric matrices $\mathbf{\Theta_0}, \mathbf{\Theta_1}, \ldots, \mathbf{\Theta_n}$ and $\mathbf{\Sigma_0}, \mathbf{\Sigma_1}, \ldots, \mathbf{\Sigma_n}$ along with vectors of unknown parameters $\kappa_\mathbf{0}, \kappa_\mathbf{1}, \ldots, \kappa_\mathbf{n}$ which satisfy

$$\kappa_\mathbf{j} = \mathbf{E}\left[\frac{\mathbf{X_{i,j}}}{\mathbf{\Theta_i}}\right].$$

2. The covariance structure satisfies the following:

$$\mathbf{Cov}\left[\mathbf{X_{i,j}}, \mathbf{X_{k,l}}\right] = \begin{cases} \mathbf{\Theta_i^{1/2}\Sigma_j\Theta_i^{1/2}} & \text{if } i = k \text{ and } j = l \\ 0 & \text{Otherwise.} \end{cases}$$

Based on the above assumptions, Hess, Schmidt, and Zocher (2006) provided following formula for the Gauss-Markov predictors of $X_{i,j}$ when $i + j > n$:

$$\widehat{\mathbf{X}}_{\mathbf{i,j}} = \mathbf{\Theta_i} \left(\sum_{\mathbf{k=0}}^{\mathbf{n-j}} \mathbf{\Theta_k^{1/2} \Sigma_j^{-1} \Theta_k^{1/2}} \right)^{\mathbf{-1}} \sum_{\mathbf{k=0}}^{\mathbf{n-j}} \left(\mathbf{\Theta_k^{1/2} \Sigma_j^{-1} \Theta_k^{1/2}} \right) \mathbf{\Theta_k^{-1} X_{k,j}}$$

Furthermore, it has been shown that the Gauss-Markov predictors for any sum of incremental claims provides the same results as the sums of the Gauss-Markov predictors for the single incremental claims. However, this does not mean that univariate additive method can be used for aggregate portfolio of run-off triangles for loss reserving. See Schmidt (2006). Based on theHess, Schmidt, and Zocher (2006) model, Merz and Wuthrich (2009b) introduced a stochastic model which allows to estimate conditional MSEP for the ultimate claims of a total portfolio using Gauss-Markov predictors. The combined model of chain ladder and additive model under the multivariate context was introduced by Merz and Wuthrich (2009a) for loss reserving. They are the first to discuss the MSEP of the ultimate total claim of a portfolio using a combined model in subportfolios. Ludwig and Schmidt (2010) applied the multivariate additive model to predict the calendar year reserve.

Loss reserving methods we discussed above, mainly, can be categorized as non-parametric or distribution-free approaches. It is understandable that actuaries who mainly work on loss reserves are more interested about the variability of the loss prediction than merely point estimation of future claim obligations. Parametric approach on loss reserving can easily address the prediction uncertainty based on experience loss data. Furthermore, this allows the construction of the prediction distribution and therefore, parametric approach has gained gain much more attention in recent research. Actuaries can get more information about possible future claim scenarios

and the nature of the claims distribution when parametric models are used. Various parametric methods have been proposed in the literature within the multivariate context. In the following sections, we discuss some parametric approach for multivariate loss triangles as we have seen in the literature.

4.4.3 General multivariate chain ladder (GMCL) model

Zhang (2010) introduced the general chain ladder approach within the multivariate context. He explained the multivariate chain ladder model, introduced by Pöhl and Schmidt (2005), as a special case of this stochastic reserving model. The proposed model utilized error terms from each accident year to address the dynamic correlation and structural dependency among the loss triangles corresponding to different lines of business. Parameter estimation for this proposed model used seemingly unrelated regression (SUR) method. The author introduced a base model to explain the model set up and for comparison with existing multivariate approach.

Model assumption:

1. $\mathbf{E}\left[\epsilon_{\mathbf{i,j}}|\mathbf{X_{i,0}}, \mathbf{X_{i,1}}, \ldots, \mathbf{X_{i,j-1}}\right] = \mathbf{0}$.
2. $\mathbf{Cov}\left[\epsilon_{\mathbf{i,j}}|\mathbf{X_{i,0}}, \mathbf{X_{i,1}}, \ldots, \mathbf{X_{i,j-1}}\right] = \mathbf{diag}\left(\mathbf{X_{i,j}}\right)^{\mathbf{1/2}} \cdot \mathbf{\Sigma_j} \cdot \mathbf{diag}\left(\mathbf{X_{i,j}}\right)^{\mathbf{1/2}}$.
3. Losses corresponding to different accident years are independent.
4. Error terms, $\epsilon_{\mathbf{i,j}}$, are symmetrically distributed.

Under the above assumption, the base model of the GMCL model defines

$$\mathbf{X_{i,j+1}} = \mathbf{B_j} \cdot \mathbf{X_{i,j}} + \epsilon_{\mathbf{i,j}}$$

where $\mathbf{B_j}$ is an m by m square matrix with coefficients for each development period. Coefficients corresponding to each row represent different loss triangles.

Zhang (2010) emphasized that the idea of overestimating large values and underestimating low values from the univariate chain ladder method, by Barnett and Zehnwirth (2000), is even correct on the multivariate chain ladder approach. According to the author, one can avoid this issue by introducing intercept terms in the model. He provided the following general multivariate model with the addition of intercept terms for the above mention base model under the same set of model assumptions:

$$\mathbf{X_{i,j+1}} = \mathbf{A_j} + \mathbf{B_j} \cdot \mathbf{X_{i,j}} + \epsilon_{\mathbf{i,j}}$$

where $\mathbf{A_j}$ represents the vector of intercepts for each loss triangle. Numerical example of paid and incurred loss triangles which were modeled simultaneously have been considered to illustrate the proposed model. Also, MCL models proposed by Pöhl and Schmidt (2005) and Merz and Wuthrich (2008b) were discussed as special cases of the general multivariate chain ladder model. See Zhang (2010) for more details.

4.4.4 Bayesian models

Shi, Basu, and Meyers (2012) considered multivariate lognormal model for loss reserving of multiple loss triangles within the Bayesian context. This work mainly focused on dependencies among the multiple loss triangles from different insurance lines. Normalized incremental claims, $\mathbf{y_{i,j}}$, is assumed to have the multivariate lognormal distribution

$$f(y_{i,j}|\mu_{\mathbf{i,j}}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{m/2}\,|\Sigma|^{1/2}\left(\prod_{k=1}^{m} y_{i,j}^{k}\right)}\mathbf{exp}\left(\frac{1}{2}(\mathbf{log}\ y_{i,j} - \mu_{i,j})^{T}\Sigma^{-1}(\mathbf{log}\ y_{i,j} - \mu_{i,j})\right)$$

with $\mathbf{y_{i,j}} = (\mathbf{y_{i,j,1}}, \mathbf{y_{i,j,2}}, \ldots, \mathbf{y_{i,j,m}})^{\mathbf{T}}$, location vector, $\mu_{\mathbf{i,j}} = (\mu_{\mathbf{i,j,1}}, \mu_{\mathbf{i,j,2}}, \ldots, \mu_{\mathbf{i,j,m}})^{\mathbf{T}}$, and scale matrix, $\mathbf{\Sigma}$ which is an m by m matrix with each component indicating the pairwise covariance of the different loss triangles.

The authors argued that dependencies among the different insurance lines can be observed through the calendar year effect. Hence, the model addresses the correlation among the claims across and within loss triangles by introducing random effect terms in the mean structure of the multivariate lognormal distribution. The results from the corresponding work showed that consideration of calendar year effect in addition to the correlation among the lines of business alone improved the model fitting considerably. Incorporating identical random effects among the lines of business, the mean structure of the distribution can be expressed as

$$\mu_{i,j}^{k} = \mu^{k} + A_{i}^{k} + B_{j}^{k} + \eta_{i+j}$$

where η_{i+j} indicates the random effect term and A_{i}^{k}, B_{j}^{k} denote the accident year and development year effects corresponding to the k-th line of business. In their proposed model setup, the authors assumed a common calendar year effect among the loss triangles from different lines. However, this could be arguably expressed as a disadvantage because one could observe different random effects for each lines of business.

4.4.5 Copula models

Copula applications on loss reserving are not completely new in property and casualty insurance. The construction of the multivariate distribution model using copula functions allows to address the dependency among different lines of business and gives more flexibility on the marginal models for individual lines. Brehm (2002) used a normal copula to estimate the pairwise correlation among calendar year inflations for different lines of business with lognormal assumption for the marginal distributions of claims for each insurance line. Besides the accident year and development year parameters in the regression equation (see 4.4.7) for location parameters, the model utilized a set of parameters to address the calendar year effect as well:

$$y_{i,j,k} = A_{i,k} + \sum_{p=0}^{j} B_{p,k} + \sum_{t=0}^{i+j} C_{t,k} + \epsilon_{i,j,k} \tag{4.4.7}$$

where $A_{i,k}$, $B_{p,k}$ and $C_{t,k}$ denote the kth insurance line regression parameters corresponding to accident year, development year, and calendar year, respectively.

Another copula application can be found in the work by Zhao and Zhou (2010) on individual claim loss reserving methods. In their work, copula functions were incorporated to model the dependence structure of the claim occurrence time with delays in the individual loss model. Shi and Frees (2011) extend copula based approach on multivariate loss triangles addressing the possibility of expanding the marginal distributions for claims data among lines of business. The lognormal and gamma distributions were assumed to represent normalized incremental claims corresponding to two different lines of insurance. The parameters corresponding to these distributions were estimated through regression equations with standard covariates such as

accident years and development years. Under copula regression, one large correlation matrix, $\mathbf{A}$, was introduced in the model to capture the pairwise correlation among the insurance lines as well as correlation within each loss triangle simultaneously:

$$\mathbf{A} = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \dots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \dots & 1 \end{pmatrix} \tag{4.4.8}$$

The model proposed by De Jong (2012) improved the pairwise correlation to higher dimension. Gaussian copula with normal marginals are incorporated together with factor models to address various types of time dependencies within the data.

As we discussed previously, risk margins of total loss reserving and characteristics associated with corresponding distributions can be valuable for loss reserving. Shi and Frees (2011), De Jong (2012) and Shi, Basu, and Meyers (2012) are to name a few recent research work which incorporate parametric distribution assumptions in estimating loss reserves. Understanding different correlations are critical in total loss reserving. Modeling the correlations due to the loss development over the years and dependencies among different insurance lines of business are very important. Some parametric models in loss reserving based on loss triangle data assume that claims in development years are independent. This is a major drawback since more often this assumption is violated in practice. Shi and Frees (2011) used a correlation matrix (see 4.4.7) to overcome this issue by capturing association among and within loss triangles simultaneously. De Jong (2012) and Shi, Basu, and Meyers (2012) considered calendar year effect within loss triangles in the model construction. Both methods utilized random effect approach to understand the calendar year effect but one can observe some variation in the model setup. De Jong (2012) discussed a

much more flexible dependency structure among loss triangles with typical normal assumption for marginals and normal copula function.

Chapter 5

Correlated Loss Triangles for Multiple Lines of Business

5.1 Introduction

As demand for insurance products increase, most insurance companies expand their business on multiple lines of insurance. Different insurance products offered by the same insurance company is becoming even more popular as policyholders find it more convenient to shop for insurance products to protect their assets against financial losses. Today, it is rare to find an insurance company that renders insurance product focused on a single line of business in the insurance market.

Due to the different characteristics of different lines of insurance, companies with insurance products in several lines generally subdivide its portfolio into several sub-portfolios by line of business. Hence, individual sub-portfolio for each insurance line is incorporated for claims handling. Loss development factors and other correspond-

ing characteristics may be assumed to be homogeneous within these sub-portfolios. Actuaries evaluate the unpaid losses and uncertainty associated with each insurance line based on the observed claims for different accident years and development years. However, insurance companies as well as other interested parties like investors, regulators, and shareholders are sometimes more interested in the aggregated reserve amount for all lines of business which provides the company's total obligations in future years.

In this regard, there are two simple approaches that has been observed in practice. One can aggregate total claims across the lines of businesses for a given insurance company and try to estimate the total reserve. In this approach, one major assumption to make is homogeneity among the lines of business. It is clear that under this strong assumption, which is not realistic in general, we could distort estimates of the aggregate loss reserve thereby sacrificing the accuracy. Merz and Wuthrich (2009a) provide a combined framework of chain ladder and additive loss reserving which can be used with homogeneous loss triangle data. The second approach assumed non-homogeneity among different lines of insurance but used the assumption of independence among the different lines. The total loss reserve for a given insurance company is then estimated using simple aggregation of reserve from each line. Unfortunately, the assumption of independence among sub-portfolios is not realistic in today's complex insurance market. There can be some observable reasons for dependencies in claim development over the years for different lines of business. This may be attributable to the operation of the company under the same administration using similar strategic and common risk pooling management.

As pointed out by Ajne (1994), simple aggregation of loss triangles from multiple

lines of business ignores the dependency among those sub-portfolios. Therefore, calculating loss reserve for each insurance line individually, without taking into account of dependency among different lines, could lead to an inaccurate total loss reserve. This issue in loss reserving is known as the additivity arising from a simple addition for aggregation. Depending on the covariance among the different lines of insurance, the simple aggregation of reserve from each line could decrease (with negative covariance) or increase (with positive covariance) the company's total reserve value. Therefore, aggregation of loss reserving of each sub-portfolio does not equal to the loss reserve estimate for the aggregate claims of portfolio when dependency among different lines of insurance take into account.

Different types of dependencies can be observed within and among sub-portfolios of losses for a given insurance company. The following are possible dependencies involved with loss triangles listed by Holmberg (1994) and Schmidt (2006):

1. Dependence within accident years.

2. Dependence among accident years.

3. Dependence between different line of business.

It is important to account for these dependencies to improve the accuracy of the total aggregate reserve. Understanding the effect from each sub-portfolio on total loss reserve and the effect of one sub-portfolio on other sub-portfolios could assist management of insurance companies to make strategic decisions. Additionally, there has been some research work that accounts for the possible dependencies through calendar years. Clearly, it is important for insurance companies to determine the reserve more efficiently by taking into account these various type of dependencies

associated with insurance claims.

In this chapter, we demonstrate the modeling of multivariate loss triangles using the proposed multivariate longitudinal model developed in this thesis. The use of a longitudinal framework for loss triangle data is new in the literature and we hope that this new framework could help in the progress of work on loss reserving. First, we discuss the multivariate framework for loss triangle data. Next, we provide additional details about the model construction before we present the empirical work performed. Finally, we present the results of the estimated models together with residual diagnostic tests to evaluate the reliability of the fitted model.

5.2 The multivariate longitudinal framework

Claim payments in loss triangles contain multiple accident years with their corresponding loss development over the calendar years. This characteristic of loss triangle data is also similar to observations over time across a collection of subjects from a given population. Therefore, clearly, one can organize longitudinal framework for the losses arising from loss triangles. In addition to, when considering loss triangles corresponding to multiple lines of insurance, one can obtain a multivariate longitudinal framework for loss triangle data. In order to get the multivariate longitudinal framework, we re-arranged each loss triangle data. According to this new structure, we consider each accident year as a subject while calendar years (diagonals of the loss triangle) are used to denote the time variable. Figure 5.2.1 illustrates the longitudinal framework of loss triangle data after rearrangement according to the above description. Unobservable losses corresponding to development of claims by accident are

represented by the "Unobservable Data" section. This is equivalent to the lower part of the loss triangle. Observed data is indicated by the "Observed Data" section in the figure. This new structure of the loss triangle data provides for a highly unbalanced longitudinal data.

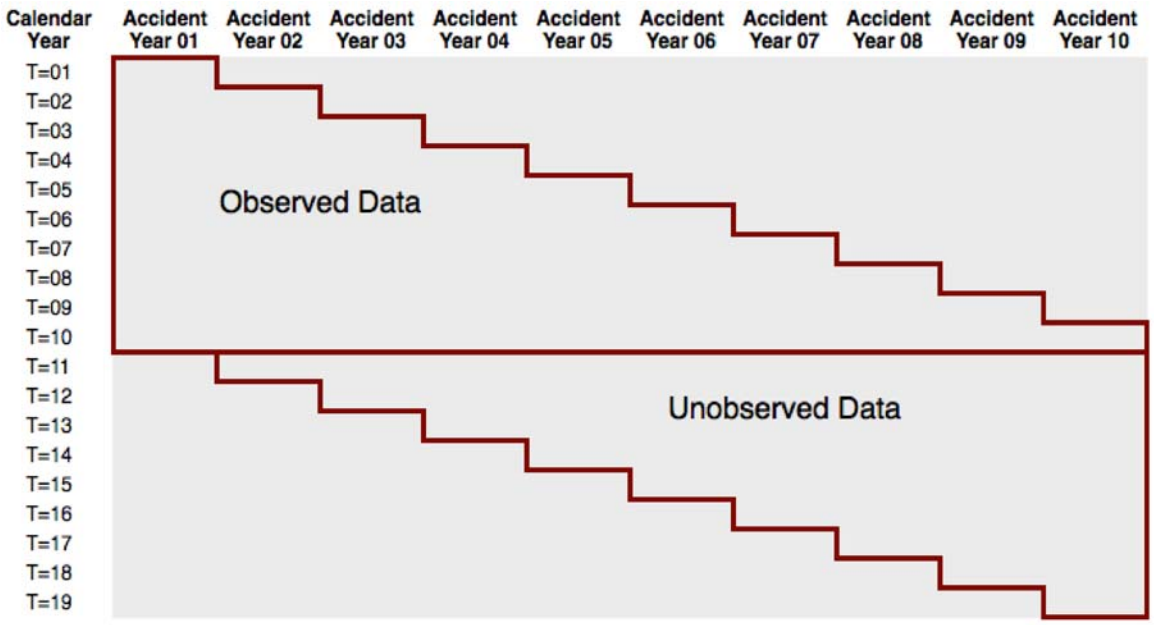

Figure 5.2.1: Longitudinal framework of loss triangle data

Longitudinal data provides for a higher level of understanding the underlying data. It is clear that losses across the calendar year are not homogeneous due to various reasons. When incorporating longitudinal framework, one can easily capture the heterogeneity among the losses from different accident years within the same calendar year. In addition to understanding heterogeneity among accident years, one can also capture the dynamic relationships of the losses within an accident year over several calendar years. Such a flexibility in reserve estimation is uncommon and atypical among the methods which have appeared in the literature.

To the best of our knowledge, longitudinal framework for loss triangle data is still new in the literature. It is our intension to explore the possibility of longitudinal model structure for loss triangle data.

5.3 Model construction

Our innovative approach is different from existing methods as we described in previous chapter for many reasons. The proposed model is able to address dependency among different lines of insurance, dynamic correlation of losses within an accident year over the calendar years, and also the heterogeneity among the accident year losses in the same calendar year. Unlike some of the parametric approaches that have appeared in the literature, our proposed model is not constrained by the independence assumption among losses in the upper part of the loss triangle. Such an assumption is clearly not realistic in practice. it is intuitive that losses corresponding to the same accident year over different calendar years exhibit a dynamic relationship. Longitudinal framework allows us to incorporate accident year losses without restricting to the independence assumption. We utilized the accident year specific random effect terms to get the conditionally independent losses over calendar years.

It has been shown that development years are generally more statistically significant than accident years in loss prediction models for many parametric studies. Therefore, development years are selected as the only covariate in our proposed model. As usual, development years are considered to be categorical variables in the regression equation (see Table 5.3.1). Therefore, the model is required to estimate coefficients for each corresponding development year. However, leaving accident years out from

the set of covariates in the model, we are able to significantly reduce the number of parameters to estimate in the model fitting process. Almost all parametric approach for multivariate loss reserving in the literature consume a lot of parameters to estimate and because of the relatively small sample size of observations in a loss triangle, parametric distribution methods often encounter the additional issue of model overfitting. Clearly, because of the manner we set up the framework, our proposed model overcomes this issue because of the the significant reduction in the number of coefficients arising in the model.

Table 5.3.1: Categorical variable: development year

	D2	D3	D4	D5	D6	D7	D8	D9	D10
Dev.Year 1	0	0	0	0	0	0	0	0	0
Dev.Year 2	1	0	0	0	0	0	0	0	0
Dev.Year 3	0	1	0	0	0	0	0	0	0
Dev.Year 4	0	0	1	0	0	0	0	0	0
Dev.Year 5	0	0	0	1	0	0	0	0	0
Dev.Year 6	0	0	0	0	1	0	0	0	0
Dev.Year 7	0	0	0	0	0	1	0	0	0
Dev.Year 8	0	0	0	0	0	0	1	0	0
Dev.Year 9	0	0	0	0	0	0	0	1	0
Dev.Year 10	0	0	0	0	0	0	0	0	1

The proposed approach incorporated some typical assumptions of losses in loss triangles, such as accident year claims are independently distributed among different accident years. Just as in many studies appearing in the literature, we did not restrict the distribution assumption about the random effect variable to be either fixed effect or having the same distribution over different lines of insurance. We relaxed the accident year specific random effects assumption by allowing independence, but different

distribution within a Gaussian framework for each line of insurance. In the absence of availability of more information, it is reasonable enough to assume Gaussian distributions for random effects. However, one could assume alternate distributions such as lognormal or gamma, which could increase the number of parameters use in the model. Again, with limited number of data, overfitting could be an issue when predicting future payments. Furthermore, model complexity can be more pronounced when other distributions are used for random effects. We observed that the mixed effect models in the statistic literature often incorporate Gaussian distribution for the random effect terms. Since our approach significantly minimizes the number of parameters to be estimated, the resulting predicted loss reserves are much more reliable for the lower part of the loss triangle. Our study assumes that each insurance line has been observed for same set of accident years and data are available for all lines of insurance throughout the same set of calendar years.

In our proposed model, we used incremental paid claims for loss reserving. However, as we described in the previous chapter, loss reserving based on cumulative paid claims are equivalent to loss reserving based on incremental paid claims. This is because one can always transform cumulative paid claims into incremental paid claims by taking the difference between the cumulative paid claims in two consecutive years. In loss reserving, especially when considering the dependency among multiple lines of insurance , the use of normalized claims data is not uncommon. It is intuitive that the best way to address the dependency among different lines of insurance is through the losses corresponding to each loss triangle. However, the direct magnitude of the losses corresponding to each triangle could significantly vary by the nature of the line of insurance. There are many reasons for this difference. For example, one can

observe high cost claims in private passenger auto insurance, while observing low cost claims in commercial multiple peril. Furthermore, high risk pool and large member portfolios can lead to different claims distributions for each line of insurance. These variations for different lines of insurance could distort or hide the presence of the nature of the dependency among these various lines. Henceforth, clearly, direct paid claims are not suitable for our purposes.

By normalizing incremental paid claims, one can get incremental payment per unit of exposure. These normalized paid claims can be used to compare paid claims from each line of insurance regardless of the magnitude of claim amounts. The number of insurance policies and the total earned premiums are to name a few common exposure variables that have been used in practice. In this study, we incorporate earned net premiums to calculate the normalized loss ratio based on the incremental paid claims from our observed data and this is developed in the following.

Suppose vector $\mathbf{Y_{ijk}}$ and $\mathbf{X_{ijk}}$ denote the normalized loss ratio and incremental paid claims respectively in i-th accident year and j-th development year for k-th line of insurance.

$$\mathbf{Y_{ijk}} = \frac{\mathbf{X_{ijk}}}{\omega_{\mathbf{ik}}} \tag{5.3.1}$$

where $\omega_{\mathbf{ik}}$ represents the vector of exposure in i-th accident year for k-th line of insurance.

One important advantage with parametric approach in model fitting is the ability to incorporate maximum likelihood estimation, which can provide reliable estimates of our unknown parameters. We used simulation approach to address the parameter uncertainty within the model. Since we have used a parametric setup in our model,

we are able to construct the entire predictive distribution for each line of insurance as well as for the aggregate portfolio.

We added more flexibility to our model by relaxing the distribution assumption on normalized loss ratios and allowed for skewed and heavy-tailed distributions for the marginals. Lognormal and gamma distributions are alternative marginals to the typical normal distribution that is sometimes used in loss reserving. Kremer (1982), Mack (1991), Renshaw and Verrall (1998), and England and Verrall (2002) are to name a few research works which have used these distributions for loss reserving. However, our proposed model is not restricted only for these distributions. In our study, we investigated a number of different distribution assumptions including generalized beta of the second kind (GB2), generalized gamma, gamma, lognormal, and Weibull distributions for each loss triangle.

Both Archimedean and Elliptical copulas are integrated to develop the joint distribution of multiple lines of insurance. Even though bivariate distributions are available for some family of distributions, extensions to higher dimension are not quite natural. Copulas can add much more power in the model construction as it can facilitate to accommodate higher dimensions. More importantly, copulas are flexible and separate the effects of the marginals, and because of these, we can freely assume marginals from different families of distributions. This is particularly important with empirical analysis. For example, preliminary analysis of claims data in loss triangles corresponding to four different lines of lines of insurance indicate different behavior of skewness. Furthermore, one can develop the marginals separately from the dependency structure of joint distribution so much so that model evaluation for marginals and joint distribution can be conducted separately.

In contrast to our work, Shi, Basu, and Meyers (2012) utilized calendar year specific random effect terms in the regression models to accommodate the dependency among different loss triangles. However, assumption of common calendar year effect across different lines of business is not completely realistic. Depending on the nature of each line of business, it is intuitive that we can observe different calendar year effect.

Even though our proposed model is not motivated by the work of Shi and Frees (2011), the approach we proposed here can be considered as extension to their work. We have added more flexibility in our approach. Instead of simply focusing on the dependency between different lines of insurance, our proposed model capture the dependency within accident years as well as subject specific heterogeneity incorporating random effects.

5.4 Empirical analysis

For empirical illustration of our proposed model, we used loss triangles which are reported in Schedule P of the annual statements required by the National Association of Insurance Commissioners (NAIC) for one of the licensed property and casualty insurance company in the United States. The following sub-sections discuss the model development of these data based on our proposed new methodology. Also, model evaluation and predicted loss reserves can be found in more details towards the end of this chapter. Insurance companies which offer coverage for multiple lines of insurance maintain several loss triangles corresponding to each line of insurance or a single triangle for homogeneous lines of insurance. For simplicity, we will use the same set

of notations we described in Section 4.4.

5.4.1 Data set

Data contains cumulative paid losses and net premiums collected over accident years from year 1988 to 1997. Even though most of the literature so far about multivariate loss reserving incorporate only a maximum of two different lines of insurance, here we used four different lines of insurance from the selected firm. We wanted to demonstrate that our approach is flexible enough to accommodate more than two lines of lines of insurance in theory as well as in practice. The following are the four lines of insurance considered in our study.

- Private passenger auto liability and medical
- Homeowners and farm owners
- Commercial multiple peril, and
- Other liability - Occurrence

For simplicity and reference purposes, we re-label these response variables after calculating the normalized loss ratios as in Table 5.4.1

Table 5.4.1: Response variables

ILRH	Incremental Loss Ratio for Homeowners/ farm owners
ILRP	Incremental Loss Ratio for Private passenger auto liability/medical
ILRC	Incremental Loss Ratio for Commercial multiple peril
ILRO	Incremental Loss Ratio for Other liability (Occurrence)

Cumulative paid losses and net premiums earned in each accident year for the four lines of business are displayed in Tables 5.4.3, 5.4.2, 5.4.4, and 5.4.5. Also, Table 5.4.6 indicates loss ratios corresponding to the incremental paid claims in the loss triangle data for these four lines of insurance. We considered the classical assumption in loss reserving with loss triangles data. All observed claims will be paid in full by the end of a fixed number of years. The time series plots of loss ratios over the development years in Figure 5.4.1 confirms that our assumption is reasonable for the data on hand. Each line corresponding to an accident year in a time series plot represents development pattern of loss ratio of paid claims. The decreasing pattern of these lines indicate that claims will be closed in a fixed time in the future. These graphs also show that each line of insurance has different decreasing pattern which indicates different correlation over time among the different lines of insurance.

Table 5.4.2: Cumulative paid losses Homeowners and farm owners

Homeowners/Farm owners insurance											
		Development year									
Accident year	Premiums	0	1	2	3	4	5	6	7	8	9
1988	94,070	35,605	52,161	54,137	55,539	56,476	57,403	57,815	58,104	58,225	58,396
1989	95,508	44,730	63,955	65,957	69,086	67,497	69,923	70,510	70,676	70,704	
1990	92,420	36,486	50,508	53,424	55,501	56,295	56,790	57,116	57,243		
1991	101,766	48,418	64,347	68,343	69,696	70,595	70,847	71,209			
1992	112,464	96,567	68,343	135,037	136,941	137,795	138,297				
1993	128,460	61,010	80,471	84,079	85,744	86,502					
1994	143,295	78,147	95,470	100,343	103,247						
1995	150,882	67,096	83,911	87,414							
1996	121,487	75,116	90,978								
1997	32,694	10,779									

Table 5.4.3: Cumulative paid losses for Private passenger auto liability and medical

Private passenger - auto liability/medical insurance											
		Development year									
Accident year	Premiums	0	1	2	3	4	5	6	7	8	9
1988	906,236	224,230	444,587	550,263	614,499	646,555	661,208	670,101	674,655	676,514	679,176
1989	964,751	243,325	479,993	597,425	662,098	697,533	710,619	716,832	719,696	721,908	
1990	1,015,900	256,357	510,765	634,461	705,322	738,529	752,960	761,574	767,622		
1991	1,117,065	276,302	549,534	687,614	759,545	792,495	807,148	815,227			
1992	1,238,859	318,085	638,439	790,579	863,181	904,970	922,784				
1993	1,362,581	361,131	719,340	871,564	949,251	991,851					
1994	1,522,338	413,286	802,548	968,688	1049,053						
1995	1,704,342	463,972	876,510	1047,437							
1996	1,901,566	509,094	945,440								
1997	2,161,063	584,107									

Table 5.4.4: Cumulative paid losses for commercial multiple peril insurance

Cumulative paid losses for commercial multiple peril insurance											
		Development year									
Accident year	Premiums	1	2	3	4	5	6	7	8	9	10
1988	79,592	11,378	21,163	27,999	33,247	37,077	38,843	39,830	40,875	41,611	42,040
1989	56,342	10,927	17,559	21,762	24,773	26,611	28,758	29,484	29,900	30,232	
1990	13,188	2,649	3,990	4,351	4,803	4,931	5,115	5,177	5,323		
1991	12,188	3,797	5,345	5,989	6,217	6,353	6,410	6,431			
1992	11,871	5,039	7,025	7,330	7,684	8,102	8,330				
1993	9,358	2,769	3,856	4,180	4,304	4,382					
1994	9,346	2,432	3,182	3,418	3,549						
1995	9,989	1,903	2,717	3,165							
1996	9,835	2,827	4,095								
1997	10,792	2,408									

Table 5.4.5: Cumulative paid losses for Other liability - Occurrence

Cumulative paid losses for other liability (Occurrence) insurance											
		Development year									
Accident year	Premiums	1	2	3	4	5	6	7	8	9	10
1988	51,288	1,371	3,332	6,592	9,270	10,621	12,651	14,460	14,723	15,130	15,283
1989	28,924	1,384	3,007	4,965	7,304	9,103	9,863	9,712	9,845	10,170	
1990	11,051	710	1,583	2,190	3,422	3,696	4,201	4,587	5,390		
1991	10,340	1,341	1,908	3,132	3,908	4,398	4,783	4,946			
1992	11,783	2,429	3,296	4,532	5,171	5,468	5,609				
1993	17,332	4,732	11,445	13,131	13,745	14,862					
1994	73,041	10,211	16,432	18,835	19,299						
1995	83,136	11,348	18,953	21,474							
1996	90,914	13,058	20,482								
1997	101,791	10,169									

Table 5.4.6: Incremental loss ratio of different lines of insurance

	Development lag									
Accident year	0	1	2	3	4	5	6	7	8	9
1988	0.3785	0.1760	0.0210	0.0149	0.0100	0.0099	0.0044	0.0031	0.0013	0.0018
	0.2474	0.2432	0.1166	0.0709	0.0354	0.0162	0.0098	0.0050	0.0021	0.0029
	0.1430	0.1229	0.0859	0.0659	0.0481	0.0222	0.0124	0.0131	0.0092	0.0054
	0.0267	0.0382	0.0636	0.0522	0.0263	0.0396	0.0353	0.0051	0.0079	0.0030
1989	0.4683	0.2013	0.0210	0.0161	0.0166	0.0088	0.0061	0.0017	0.0003	
	0.2522	0.2453	0.1217	0.0670	0.0367	0.0136	0.0064	0.0030	0.0023	
	0.1939	0.1177	0.0746	0.0534	0.0326	0.0381	0.0129	0.0074	0.0059	
	0.0478	0.0561	0.0677	0.0809	0.0622	0.0263	0.0001	0.0046	0.0112	
1990	0.3948	0.1517	0.0316	0.0225	0.0086	0.0054	0.0035	0.0014		
	0.2523	0.2504	0.1218	0.0698	0.0327	0.0142	0.0085	0.0060		
	0.2009	0.1017	0.0274	0.0343	0.0097	0.0140	0.0047	0.0111		
	0.0642	0.0790	0.0549	0.1115	0.0248	0.0457	0.0349	0.0727		
1991	0.4758	0.1565	0.0393	0.0133	0.0088	0.0025	0.0036			
	0.2473	0.2446	0.1236	0.0644	0.0295	0.0131	0.0072			
	0.3115	0.1270	0.0528	0.0187	0.0112	0.0047	0.0017			
	0.1297	0.0548	0.1184	0.0750	0.0474	0.0372	0.0158			
1992	0.8586	0.2998	0.0422	0.0169	0.0076	0.0045				
	0.2568	0.2586	0.1228	0.0586	0.0337	0.0144				
	0.4245	0.1673	0.0257	0.0298	0.0352	0.0192				
	0.2061	0.0736	0.1049	0.0542	0.0252	0.0120				
1993	0.4749	0.1515	0.0281	0.0130	0.0059					
	0.2650	0.2629	0.1117	0.0570	0.0313					
	0.2959	0.1162	0.0346	0.0133	0.0083					
	0.2730	0.3873	0.0973	0.0354	0.0644					
1994	0.5454	0.1209	0.0340	0.0203						
	0.2715	0.2557	0.1091	0.0528						
	0.2602	0.0802	0.0253	0.0140						
	0.1398	0.0852	0.0329	0.0064						
1995	0.4447	0.1114	0.0232							
	0.2722	0.2421	0.1003							
	0.1905	0.0815	0.0448							
	0.1365	0.0915	0.0303							
1996	0.6183	0.1306								
	0.2677	0.2295								
	0.2874	0.1289								
	0.1436	0.0817								
1997	0.3297									
	0.2703									
	0.2231									
	0.0999									

Pairwise association between corresponding cells from multiple loss triangles is the most common approach to understand the dependency among different lines of business. It is also common that typical approaches ignore the calendar year influence while restricting the model under the assumption of independence among the accident years.

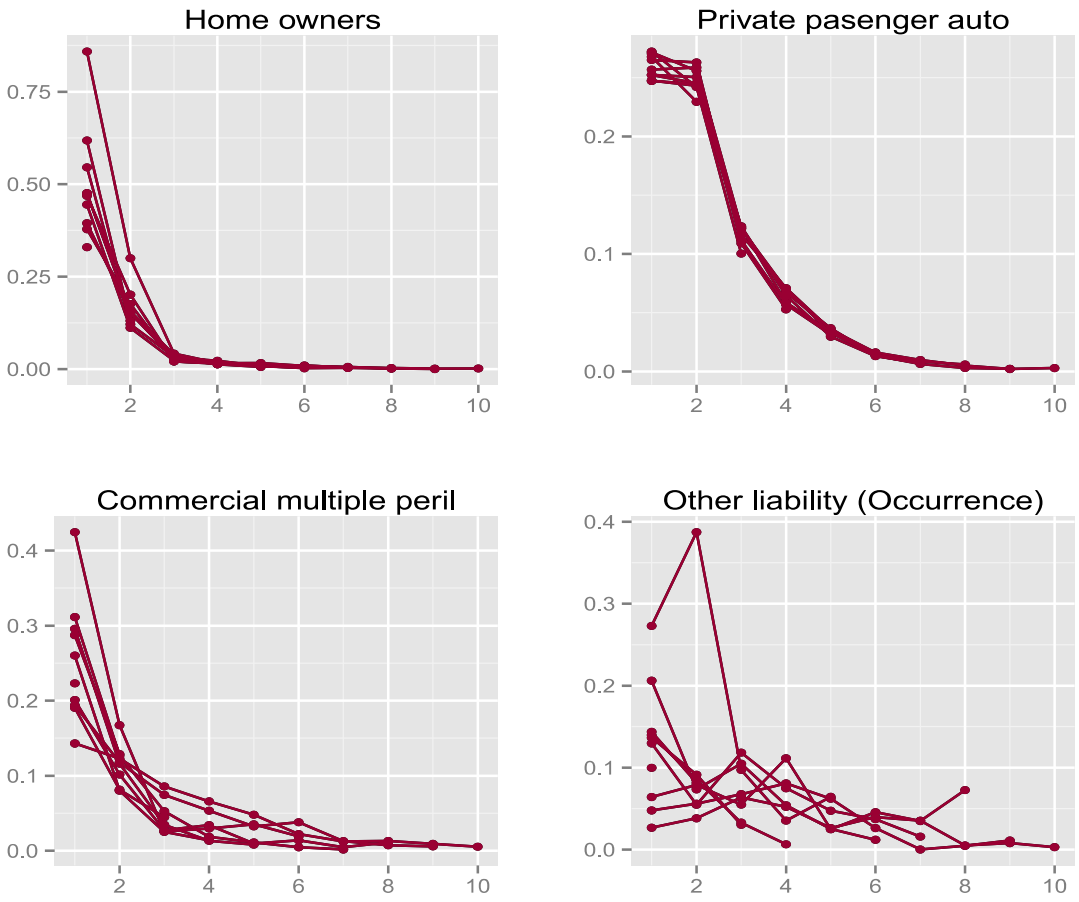

Figure 5.4.1: Multivariate time series plots of incremental loss ratio

5.4.2 Marginal distributions

Parametric approach in loss reserving is generally limited to few parametric distributions. As we mentioned earlier, the lognormal and gamma distributions are alternatives to the typical Gaussian assumptions for many parametric approaches. This is because these distributions allow for some flexibility in terms of skewness and the tail. Please refer to section 4.3.2 and 4.3.3 for more details. In this study, we wanted to explore the possibility of a few other parametric distributions in additions to lognormal and gamma. Due to the skewed nature of the data generally exhibited in loss triangles, we focused on GB2 (Generalized Beta of the Second kind), GG (Generalized Gamma), and Weibull distributions. However, both GB2 and GG distributions failed to converged when estimating the model parameters using the maximum likelihood method. One sensible explanation could be that these distributions are required to estimate more parameters and loss triangle data do always have small sample of data. However, we found that the Weibull distribution is much suitable for loss triangle of Other liability (Occurrence) losses . To the best of our knowledge, this is the first time we see that the Weibull distribution is being used for loss reserving methods. We believe that this is an appealing finding for loss reserving via parametric distributions.

In Table 5.4.7, we also provide the regression equation setup for location parameters in each of the distributions we used together with their corresponding residual calculations. $\mathbf{x}$ and β represent the vector of covariates and its corresponding vector of coefficients, respectively. In regression analysis, residuals play an important role. Covariates are used to determine the effect on dependent variables. One can explain residuals as proxy to pure value of corresponding dependent variable after removing the effect of the covariates. Therefore, residuals can be used to evaluate the good-

ness of fit of each marginals we modeled. In loss triangle studies, typical covariates used are accident years and development years, but this requires the strong assumption of claims within accident years being independently distributed. It is intuitive that claims within accident years can have association because those claims could have calendar year effect. Based on the model setup we used, we only considered the development year as covariate in the regression equation, but we accounted for the effect of calendar years by the addition of a random effect term to capture the dynamic relationships over calendar years. In both gamma and lognormal models, the regression equation is integrated to estimate the location parameter while in the Weibull model, the regression equation is integrated through the scale parameter. Except for lognormal model, both gamma and Weibull models are considered to have nonlinear relationship with their covariates.

Table 5.4.7: Marginal distributions

Marginals	**Density** $f(y)$	**Covariates**	**Residuals** R_i	**Line of business**
Gamma	$\frac{1}{\Gamma(\nu)y}\left(\frac{y\nu}{\mu}\right)^{\nu} e^{(-y\nu/\mu)}$	$\log \mu_i(\mathbf{x}) = \alpha_i + \beta'\mathbf{x}$	$\frac{Y_i}{\mu_i(\mathbf{x})}$	ILRA
Lognormal	$\frac{1}{\sigma\sqrt{2\pi}y} \exp\left[-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right]$	$\mu_i(\mathbf{x}) = \alpha_i + \beta'\mathbf{x}$	$\frac{\log(Y_i) - \mu_i(\mathbf{x})}{\sigma}$	ILRB, ILRE
Weibull	$\frac{\kappa}{\lambda}\left(\frac{y}{\lambda}\right)^{\kappa-1} e^{-(y/\kappa)^{\kappa}}$	$\log \lambda_i(\mathbf{x}) = \alpha_i + \beta'\mathbf{x}$	$\frac{Y_i}{\lambda_i(x)}$	ILRH1

5.4.3 Output of fitting the marginals

In finding the marginal distributions for each line of insurance, we analyzed the incremental loss ratios applicable to each line separately. After some investigation of fitting the marginals, as we already alluded in the previous section, we find that three families of distributions are most suitable for our four lines of business. In particular, we saw the gamma and Weibull distributions are much suitable for homeowners and other liability insurance, while both private passenger and commercial multiple peril insurance are best modeled with the lognormal distribution.

Figures 5.4.2, 5.4.3, 5.4.4, and 5.4.5 show the results of performing residual analysis corresponding to each of these marginal models for assessing goodness of fit. For diagnostic purposes, we show in these figures the quality of the fit of the marginals by examining the residuals in terms of its histograms, the empirical cumulative distribution functions (ECDFs), as well as their QQ and PP plots. All are displayed for model evaluation of the marginals. These figures support evidence of our choices for the marginal distributions. However, we want to emphasize how good the quality of the Weibull distribution is for the Other liability (Occurrence) insurance line. In addition to these graphical approaches, we also present the results of calculating the Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC) for each individual model. Our procedure of selecting the most reasonable model is based on an examination of these distribution graphs as well as a comparison of the numerical values of the AIC and BIC.

Table 5.4.8 provides the maximum likelihood estimates of the regression coefficients and other parameters in the respective marginal distributions fitted to each line of insurance. According to these estimates summarized in the table, all develop-

ment year factors are statistically significant with the exception for development year 2 in private passenger and other liability (occurrence) lines of insurance. The rest of the parameters in the marginal distributions for all lines of insurance indicate statistical significance. Surprisingly however, the standard deviations of the random effect distribution for the homeowner and other liability (occurrence) lines of insurance are not statistically significant.

Table 5.4.8: Fitted marginals for each line of insurance

	Lines of Business											
	ILRH			ILRP			ILRC			ILRO		
	Gamma distribution			Lognormal distribution			Lognormal distribution			Weibull distribution		
Parameter	Estimate	Std Error	p-val	Estimate	Std Error	p-val	Estimate	Std Error	p-val	Estimate	Std Error	p-val
Covariates												
Dev.Year 1	-	-	-	-	-	-	-	-	-	-	-	-
Dev.Year 2	-1.0997	0.1435	0.0000	-0.0479	0.0451	0.2936	-0.7626	0.1947	0.0003	-0.1771	0.2835	0.5356
Dev.Year 3	-2.8077	0.1480	0.0000	-0.8103	0.0466	0.0000	-1.7508	0.2023	0.0000	-0.5871	0.2801	0.0420
Dev.Year 4	-3.3975	0.1540	0.0000	-1.4253	0.0486	0.0000	-2.1746	0.2113	0.0000	-0.7229	0.2983	0.0196
Dev.Year 5	-3.9561	0.1621	0.0000	-2.0632	0.0511	0.0000	-2.5806	0.2229	0.0000	-1.1480	0.3095	0.0006
Dev.Year 6	-4.3969	0.1760	0.0000	-2.9075	0.0544	0.0000	-2.8066	0.2383	0.0000	-1.3550	0.3331	0.0002
Dev.Year 7	-4.7331	0.1859	0.0000	-3.5022	0.0589	0.0000	-3.7751	0.2579	0.0000	-1.6428	0.3670	0.0001
Dev.Year 8	-5.4997	0.2112	0.0000	-4.0707	0.0663	0.0000	-3.3521	0.2900	0.0000	-1.3261	0.4192	0.0029
Dev.Year 9	-6.4610	0.2564	0.0000	-4.7769	0.0772	0.0000	-3.8281	0.3451	0.0000	-2.5145	0.4739	0.0000
Dev.Year 10	-5.6530	0.3411	0.0000	-4.5198	0.1074	0.0000	-4.1785	0.4649	0.0000	-3.6613	0.6379	0.0000
Intercept	-0.6969	0.1002	0.0000	-1.3467	0.0320	0.0000	-1.4181	0.1573	0.0000	-1.9998	0.1980	0.0000
Marginals												
ν	6.3380	1.2835	0.0000	-	-	-	-	-	-	-	-	-
σ	-	-	-	0.0980	0.0100	0.0000	0.4207	0.0436	0.0000	-	-	-
κ	-	-	-	-	-	-	-	-	-	1.7299	0.2081	0.0000
Random effect												
σ_α	0.0561	0.0967	0.5646	0.2651	0.0870	0.0039	0.2651	0.0870	0.0039	0.2042	0.1158	0.0850
AIC		-338.2382			-388.4891			-260.1648			-201.8147	
BIC		-314.1502			-364.4011			-236.0768			-177.7267	

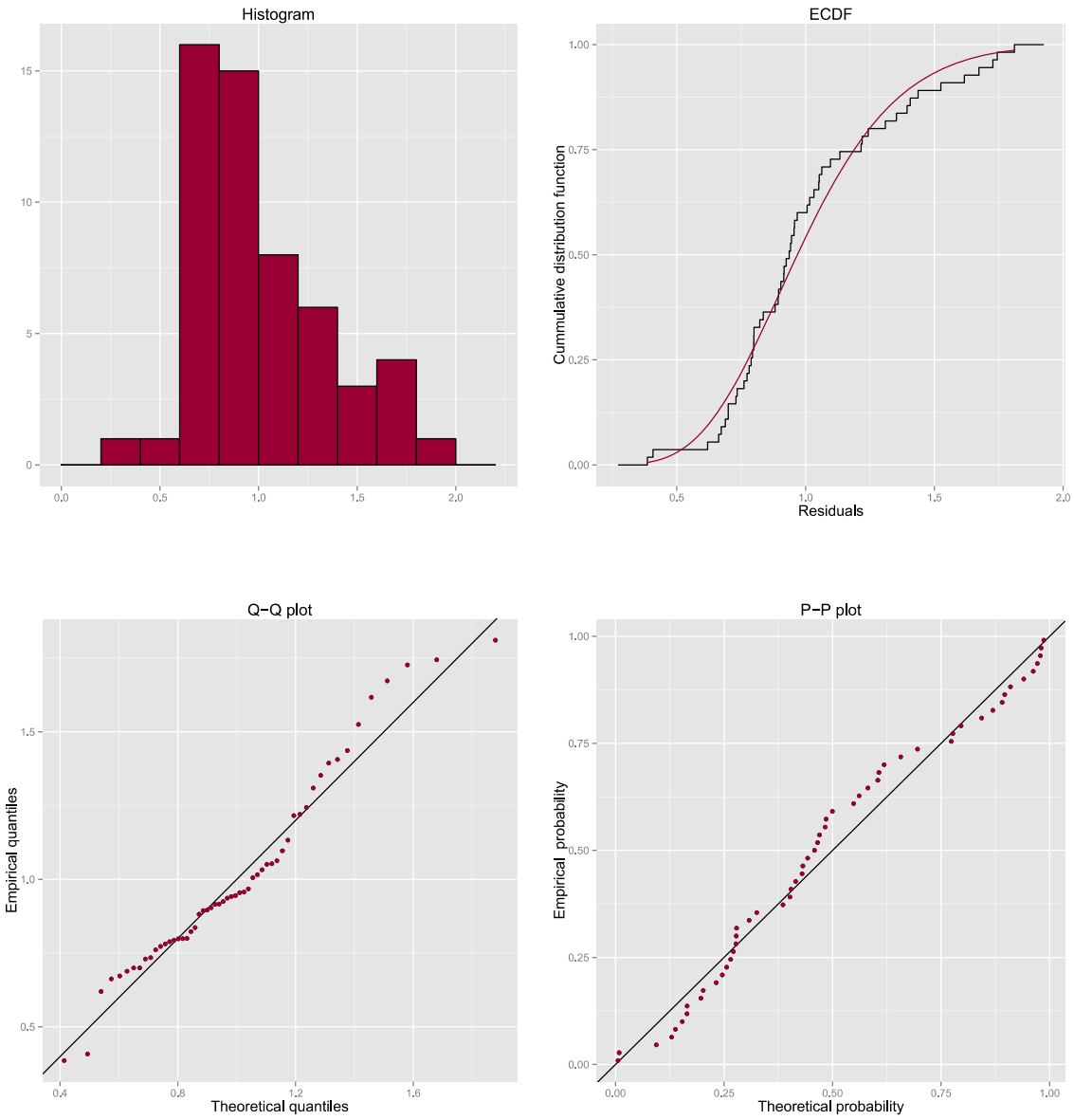

Figure 5.4.2: Residual diagnostics for variable ILRH

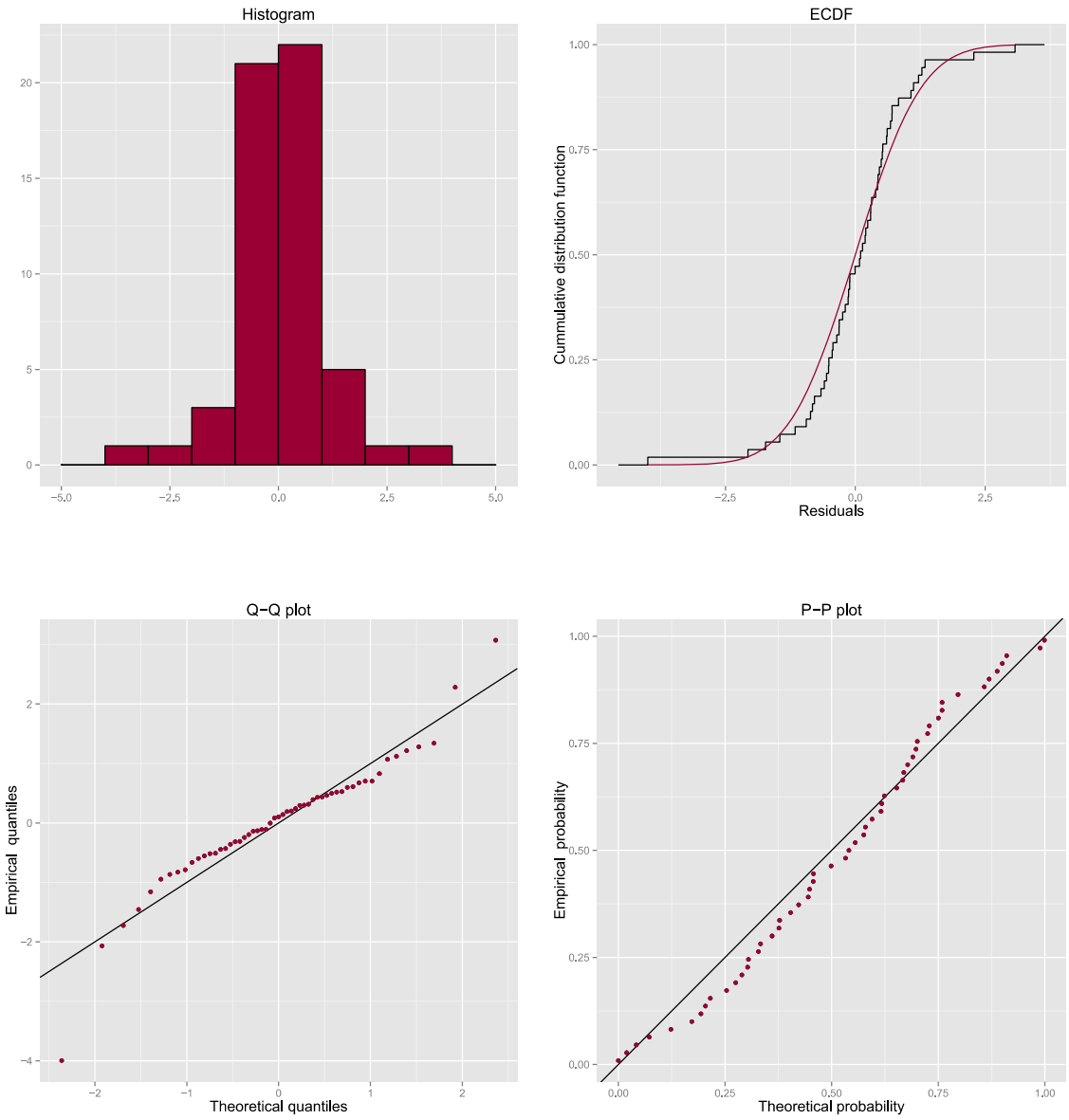

Figure 5.4.3: Residual diagnostics for variable ILRP

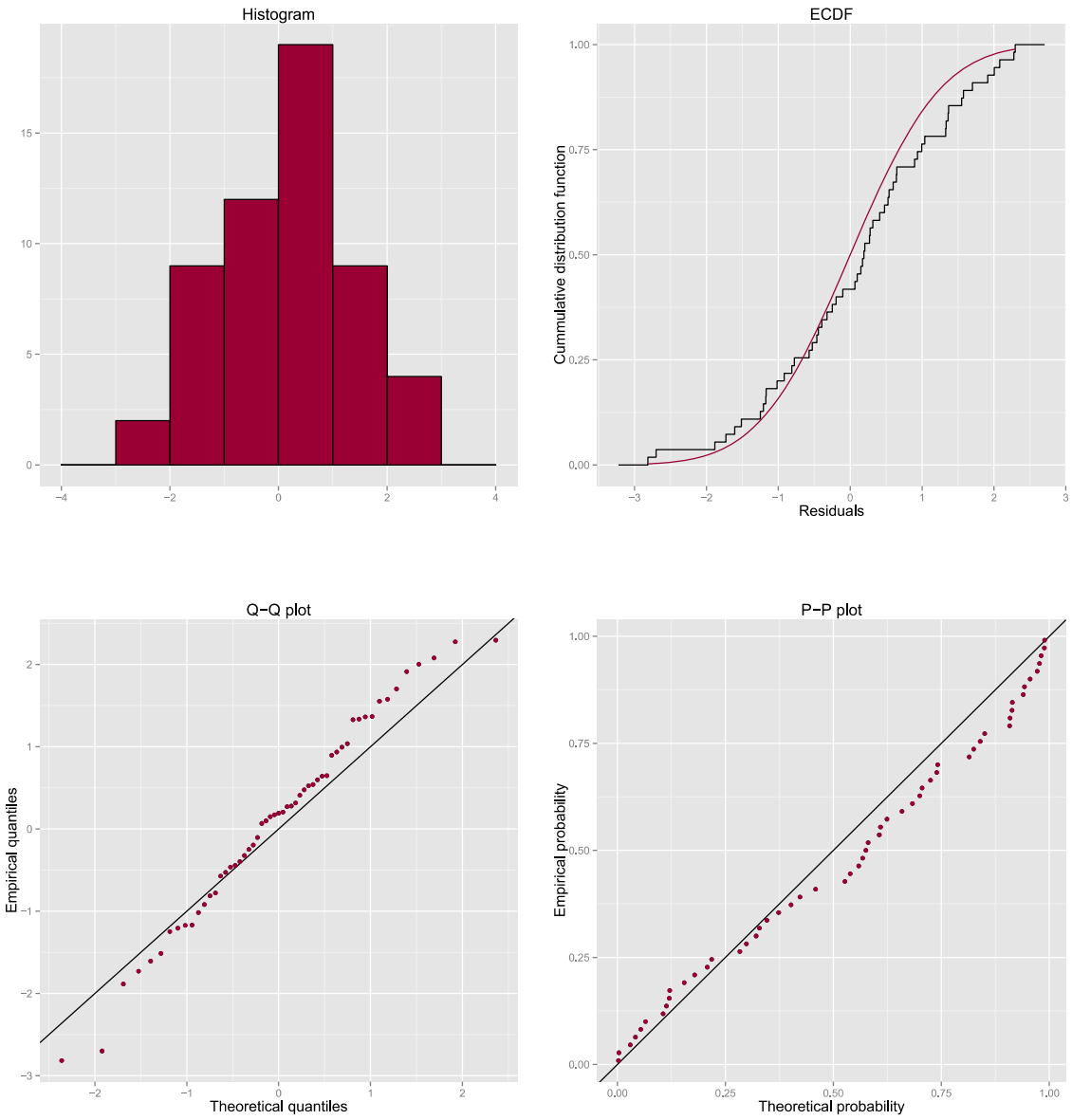

Figure 5.4.4: Residual diagnostics for variable ILRC

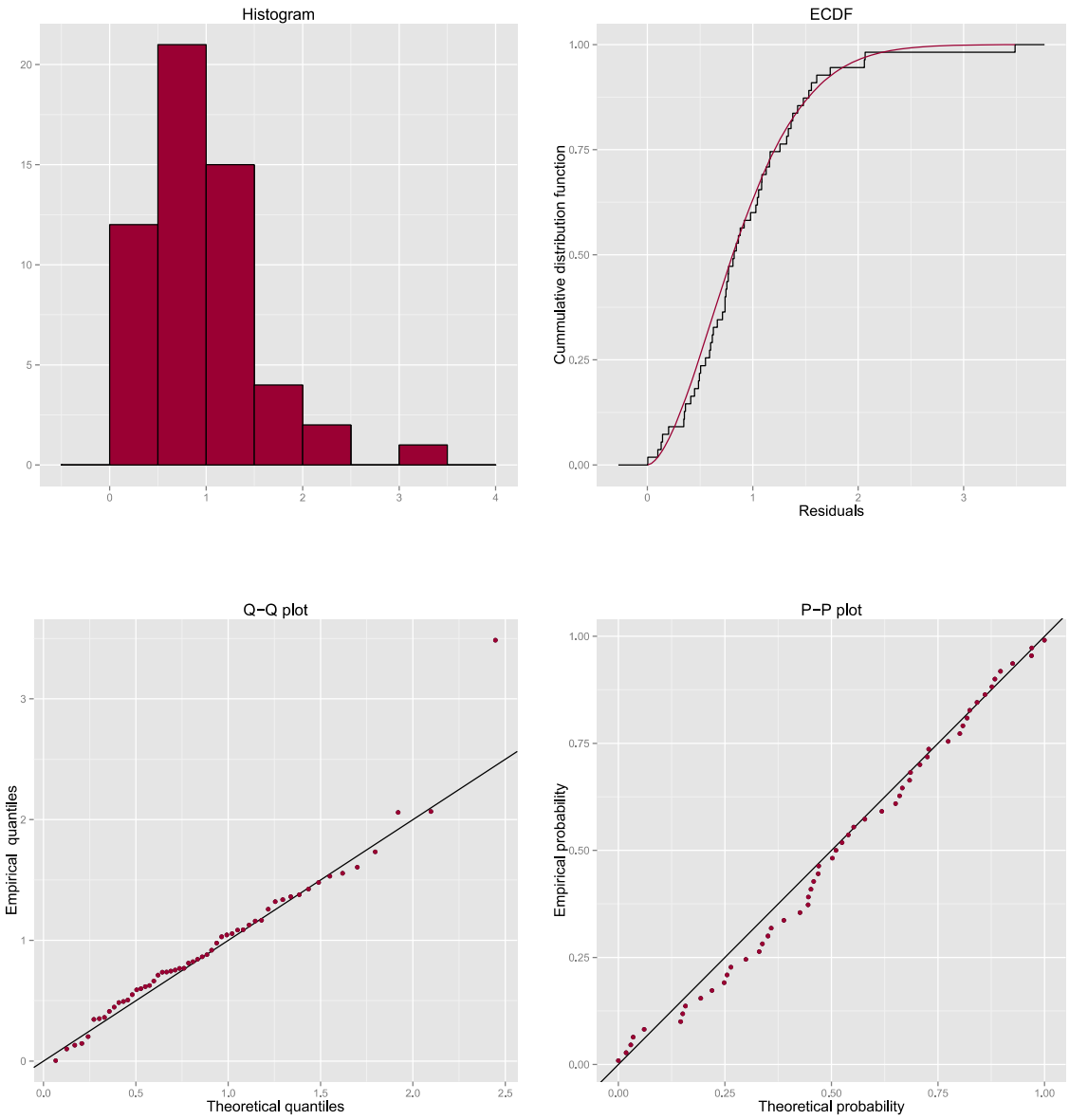

Figure 5.4.5: Residual diagnostics for variable ILRO

5.4.4 Preliminary investigation of dependence

One very important aspect of our model is its ability to measure and understand the presence of dependencies not only among different lines of insurance, but also the dynamic correlation within accident years. This dynamic correlation is usually measured in terms of a serial correlation. As explained in Chapter 2 where we developed the general multivariate framework, we modeled to capture this dynamic correlation for each individual line of insurance by incorporating accident year specific random effect terms in the regression equation. As n initial investigation of understanding this dependence, we simply applied basic statistical procedures of directly measuring the serial correlation for a given line of insurance as implied by the data.

To illustrate the serial correlation of a given response variable in two consecutive years, we considered y_{i1k} and y_{i2k}. Simple linear correlation can be expressed as:

$$\text{Corr}(y_{i1k}, y_{i2k}) = \frac{\text{Cov}(y_{i1k}, y_{i2k})}{\sqrt{\text{Var}(y_{i1k})\text{Var}(y_{i2k})}} \tag{5.4.1}$$

where we have

$$\text{Cov}(y_{i1k}, y_{i2k}) = \text{E}(\text{Cov}(y_{i1k}, y_{i2k})|\alpha_{ik}) + \text{Cov}(\text{E}(y_{i1k}|\alpha_{ik}), \text{E}(y_{i2k}|\alpha_{ik})). \tag{5.4.2}$$

Under the conditional independence,

$$\text{Cov}(y_{i1k}, y_{i2k}) = \text{Cov}(\text{E}(y_{i1k}|\alpha_{ik}), \text{E}(y_{i2k}|\alpha_{ik})) \tag{5.4.3}$$

where $\text{Var}(y_{ijk})$ can be expressed as,

$$\text{Var}(y_{ijk}) = \text{E}(\text{Var}(y_{ijk}|\alpha_{ik})) + \text{Var}(\text{E}(y_{ijk}|\alpha_{ik})). \tag{5.4.4}$$

Due to the small sample size of losses in loss triangles and its highly unbalanced structure, calculating serial correlation is limited to a few calendar years. For example, when calculating serial correlation between calendar years 9 and 10, we only consider 9 observed values form both calendar years. In other words, we only incorporate losses corresponding to the first 9 accident years. Even though comparatively small sample has been used for correlation calculations, based on the nature of the loss triangle data, we believe that this calculation indicates the importance of having the dynamic association in model estimation. Table 5.4.9 provides serial correlations of each individual insurance line for the last four calendar years. These results convincingly demonstrate a clear departure from the assumption of independence for losses within accident years; this has a potential severe effect of calculating unbiased and inaccurate reserve estimates.

Table 5.4.9: Calendar year correlation

ILRH

	CY 6	CY 7	CY 8	CY 9
CY 10	0.788	0.825	0.845	0.888
CY 9	0.817	0.841	0.888	
CY 8	0.839	0.883		
CY 7	0.879			

ILRP

	CY 6	CY 7	CY 8	CY 9
CY 10	0.855	0.912	0.928	0.909
CY 9	0.906	0.923	0.904	
CY 8	0.914	0.896		
CY 7	0.885			

ILRC

	CY 6	CY 7	CY 8	CY 9
CY 10	0.630	0.669	0.703	0.773
CY 9	0.642	0.692	0.764	
CY 8	0.667	0.757		
CY 7	0.744			

ILRO

	CY 6	CY 7	CY 8	CY 9
CY 10	0.110	0.123	0.108	0.161
CY 9	0.081	0.101	0.099	
CY 8	0.069	0.095		
CY 7	0.099			

In addition to the serial correlation we discussed above, we also wanted to understand the association between different lines of insurance. Before we incorporated the copula functions to investigate the association structure of the various response variables, we also did some basic preliminary analysis to visualize the dependency. The upper portion of the loss triangle as displayed in Figure 5.4.6 provides the scatter plots between pairs of lines of insurance. We can clearly visualize from these graphs that there is no strong evidence of a linear correlation among the response variables, but there is strong evidence of some form of nonlinear dependence between each other. This encouraged us to exploit the advantages of using copula functions to capture many forms of dependence, as we have shown on these graphs based on our loss triangle data for different lines of insurance.

5.4.5 Selection and estimation of the copulas

Copula application is not entirely new to the literature on insurance loss reserving, for example. see Section 4.4.5 for additional details. As we explored more possibilities for marginals, we also considered different copula functions from both Archimedean and Elliptical families. These two families of copulas have been used in many applications partly because of the seemingly straightforward procedure of fitting these copulas to data. Clayton, Frank, and Gumbel copula functions are considered under

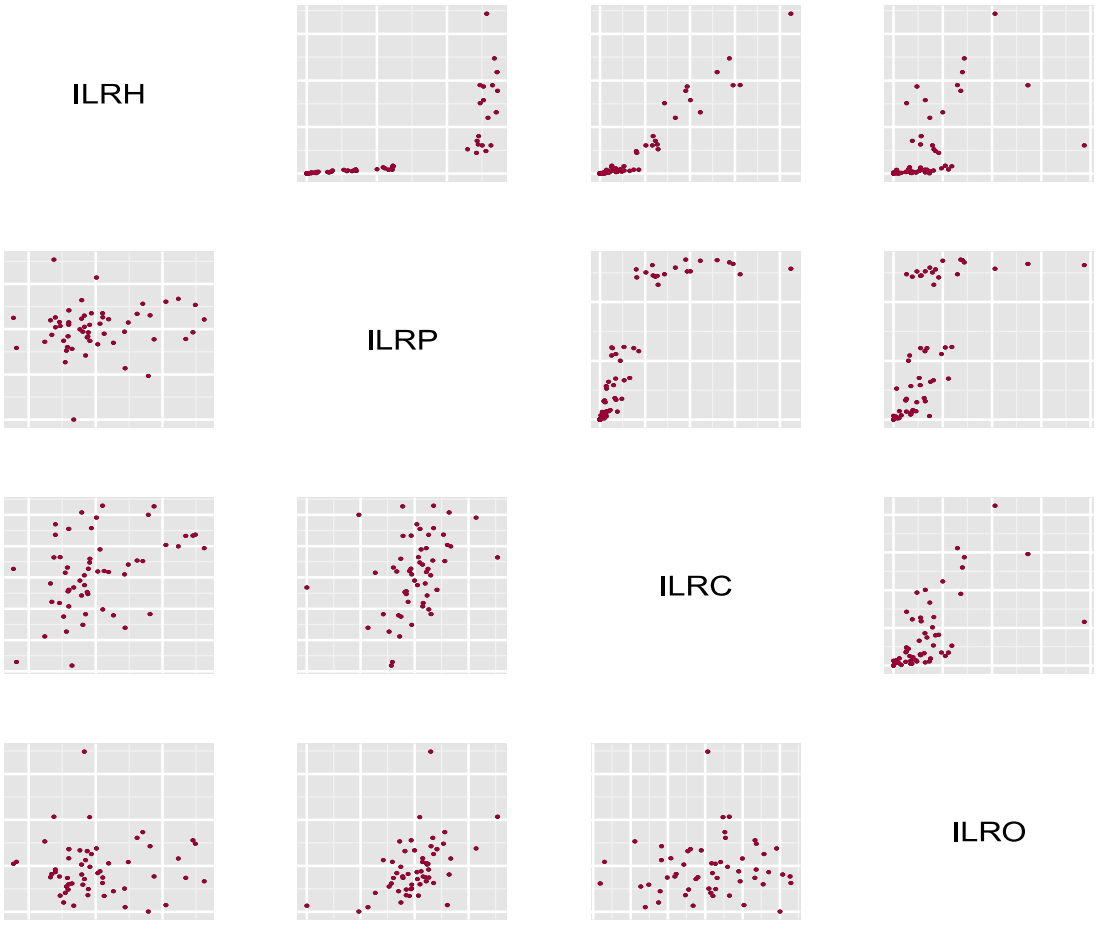

Figure 5.4.6: Correlation matrix
The upper part gives the loss ratios from raw data, while the lower part as the residuals after fitting marginals.

the Archimedean family. Both the Gaussian copula and t-copula functions fall under the Elliptical family, and for these copulas, we also considered the unstructured association matrix for different lines of insurance. Formulation of these Archimedean copulas and Gaussian copula can be found under the model calibration in section 3.4.2.

When comparing the two families of copulas, the Elliptical family of copula has the more advantage of allowing for capturing possible pairwise association. As such, this usually leads to a better and more intuitive interpretation of the resulting dependence parameters.

Table 5.4.10 summarizes the estimation results of fitting the three Archimedean copula functions: Clayton, Frank, and Gumbel. Not surprisingly as already indicated in our preliminary investigation, all three copulas provide statistical evidence of strong dependence by a simple examination of the p-values in the table. These p-values can be used to test whether the corresponding parameters are any different from the case of independence. Next, we compared these three copula models to choose the best of them based on the AIC and the BIC values. According to statistical theory, the lowest AIC and BIC values are generally the preferred model and in this case, clearly, Frank copula provides the best of these criterion values.

To investigate copula functions from the Elliptical family, results of model estimates are conspicuously demonstrated and summarized in Table 5.4.11. For these types of copulas, a major advantage is their ability to provide us a better understanding of the presence of dependence because they allow us to model pair-wise association through the correlation structure directly built into the copulas. For both the Gaussian and the t-copula, we see strong statistical evidence of some pair-wise

dependence, and not coincidentally, both copulas give the same pairs that are statistically significant. In this case, Lines 1 and 3 are, Lines 2 and 3 are, and so with Lines 2 and 4. Please refer to the table for which number refers to which line of insurance. Furthermore, when we compared the AIC and BIC values between the two copulas, we find that the Gaussian copula slightly performed better than the t-copula. This is also not at all surprising when in the t-copula, where we have an additional parameter of degrees of freedom, our estimates here give a value very very large. It is widely known that the t-copula converges to the Gaussian copula in the case of very large degrees of freedom.

Table 5.4.10: Estimated Archimedean copula functions

Copulas	**Parameter estimates**	**Standard error**	**p-value**	**AIC**	**BIC**
Clayton	0.1499	0.0566	0.0106	447.8156	449.8229
Frank	1.4907	0.4284	0.0010	441.6951	443.7024
Gumbel	1.1335	0.0526	0.0141	447.8317	449.8390

In addition to the AIC and BIC values for comparing different non-nested models, we also utilized the so-called "copula-PP" functions. Please refer to Chapter 2 for a discussion of these function in our copula diagnostic section. Figure 5.4.7 provides copula-PP plots for all three Archimedean copula functions and the Gaussian copula function. As we already concluded that the Gaussian copula is slightly more suitable than the t-copula, in order not to overwhelm the reader, we purposely ignored the construction of the "copula-PP" for the t-copula.

These figures, together with the AIC and BIC values, show that the results coming from the Frank and the Gaussian copulas indicate that they are more superior than

Table 5.4.11: Estimated Elliptical copula functions

	Gaussian copula			t-copula		
Parameter	**Estimate**	**Std Error**	**p-val**	**Estimate**	**Std Error**	**p-val**
r_{12}	0.0746	0.1289	0.5656	0.0797	0.1326	0.5506
r_{13}	0.3472	0.0963	0.0007	0.3429	0.0984	0.0011
r_{14}	-0.0563	0.1182	0.6362	-0.0439	0.1227	0.7219
r_{23}	0.3126	0.0969	0.0022	0.3201	0.0990	0.0022
r_{24}	0.5309	0.0785	0.0000	0.5290	0.0816	0.0000
r_{34}	0.0282	0.1005	0.7801	0.0404	0.1041	0.6998
df	-	-	-	75.9600	80.5460	0.3504
AIC		426.9003			429.5532	
BIC		438.9443			443.6046	

Here, r_{ij} represent the correlation between ith and jth lines of insurance.
1: ILRH, 2: ILRP, 3: ILRC and 4: ILRO

the other copula models. Here for purposes of being complete, we hereby present the functional forms of the Frank as well as the Gaussian copulas.

The Frank copula is derived using the generator $\psi(t) = -\log\left(\frac{e^{-\alpha t}-1}{e^{-\alpha}-1}\right)$ so that its four-dimensional copula representation is:

$$C(u_1, u_2, u_3, u_4) = -\frac{1}{\alpha}\log\left[1 + \frac{\prod_{i=1}^{4}(e^{-\alpha u_i} - 1)}{(e^{-\alpha} - 1)^3}\right]. \tag{5.4.5}$$

It can be shown that the inverse of the generator can be expressed as

$$\psi^{-1}(s) = -\frac{1}{\alpha}\log\left[1 + e^{\alpha s}\left(e^{-\alpha} - 1\right)\right]$$

and is completely monotonic for $\alpha > 0$. This complete monotonicity helps ensure a legitimate copula when extending to higher than two dimensions. See Frank (1979) and Genest (1987) for details of the characteristics of tis copula.

The Gaussian copula generated by a four-dimensional normal distribution with

linear correlation matrix Σ is given by

$$C\left(u_1, u_2, u_3, u_4\right) = H\left(\Phi^{-1}\left(u_1\right), \Phi^{-1}\left(u_2\right), \Phi^{-1}\left(u_3\right), \Phi^{-1}\left(u_4\right)\right) \tag{5.4.6}$$

where H is the joint distribution function of a standard normal random vector expressed as

$$H\left(x_1, x_2, x_3, x_4\right) = \int_{-\infty}^{x_4}\int_{-\infty}^{x_3}\int_{-\infty}^{x_2}\int_{-\infty}^{x_1} \frac{1}{\sqrt{(2\pi)^n\left|\Sigma\right|}} \exp\left(-\frac{1}{2}\mathbf{z^T\Sigma^{-1}z}\right) dz_1 dz_2 dz_3 dz_4 \tag{5.4.7}$$

and $\psi^{-1}\left(.\right)$ is the inverse of a standard normal distribution and

$$\psi\left(z\right) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} dw.$$

It is important to note that the Gaussian copula has zero tail dependence. See Embrechts, Lindskog, and McNeil (2003) for a proof of this result.

5.5 Application to reserve estimation

In this section we provide reserve estimation for individual lines of insurance as well as the combined portfolio. Due to the parametric nature of our model construction, we could easily incorporate Monte Carlo simulation techniques to estimate the reserve. For greater accuracy of the reserve calculations, we computed a large enough sample of reserve values based on multiple simulations. The final estimated reserve value is then based on the average of these simulated values. As we are able to estimate the reserve values for each accident year in every calendar year, our model allows

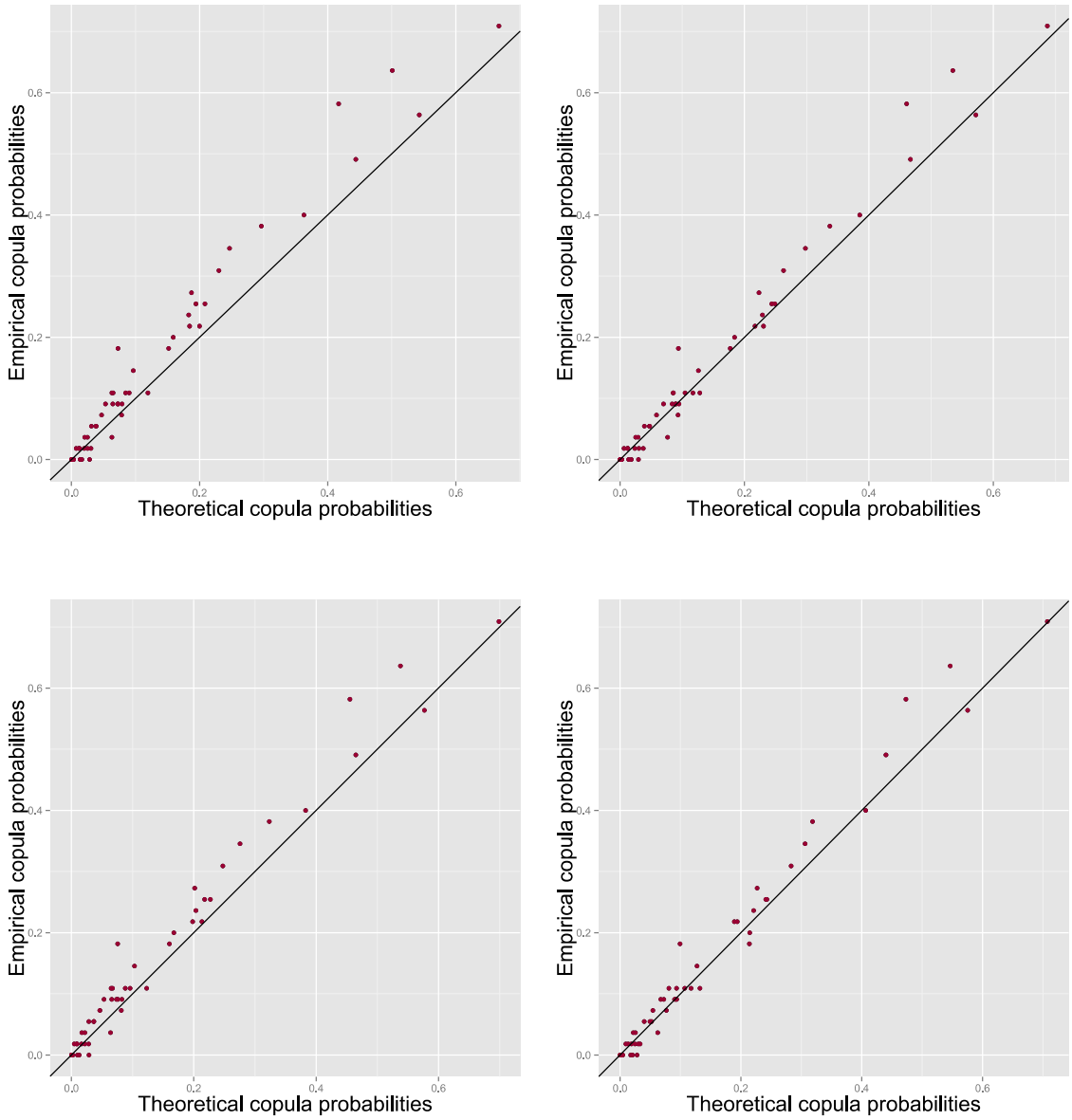

Figure 5.4.7: Copula PP-plots
From top left to bottom right: Clayton, Frank, Gumbel and Gaussian

us to produce both calendar year and accident year reserve estimates. This may be not possible, because the calculations are usually not as straightforward, with all available reserve estimate methods that have appeared in the literature. In this

section, we present results based on the calendar year, but we present the results based on accident year in the appendix.

Because our model estimates included calculations of the chosen copula functions, we needed to simulate values from these copulas. For our purposes, we present results based on both the Frank and the Gaussian copulas; both these copulas were the preferred choices. For the Frank copula, we relied on the `copula` R package to perform the simulations. For the Gaussian copula, we developed the simulation from fundamental principles. The following steps have been used to simulate values arising from a Gaussian copula with high dimensions.

- Construct the lower triangular matrix B so that the covariance matrix $V = BB^T$ using Choleski's decomposition;
- Generate a column vector of independent standard normal random variables $Z = (Z_1, Z_2, \ldots, Z_n)^T$;
- Take the matrix product of B and Z, i.e $Y = BZ$;
- Set $U_i = \Phi(Y_i)$ for $i = 1, 2, \ldots, n$;
- Set $X_i = F_{X_i}^{-1}(U_i)$ for $i = 1, 2, \ldots, n$;

A quick inspection of Tables 5.5.1 and 5.5.3 shows that the reserve estimates between the two copulas (Frank and Gaussian) do not vary too much. This is not at all surprising because based on our previous estimation, these two copulas are minimally different. Now just looking at the results of the reserve estimates based on the Gaussian copula, we find that reserve estimates are generally largest for the ILRP line of business while smallest for the ILRC line of business. For example, for

calendar year 1999, the reserve estimates are 471,009 and 1,483 for ILRP and ILRC, respectively. This has to do with the size exposure of this company we selected; this company has more policies ILRP and least policies in ILRC.

Another advantage of using parametric distributions for loss reserving is being able to estimate the range of reserve estimates rather than just a point estimates for the final reserve value. In Tables 5.5.1 and 5.5.3, we present the 5th percentile (lower bound) and 95th percentile (upper bound) of the reserve estimates for each calendar year for each individual line of insurance as well as reserve estimate for the entire portfolio. These percentiles were calculated based on the entire distribution of the simulated reserve values. For illustration purpose, again just looking at the Gaussian copula, to examine the range of estimate for the entire portfolio in calendar year 1999, we find that the 5th percentile is 414,413 and the 95th percentile is 591,979. For comparison purpose, we also present the range of values assuming normal distribution by adding and deducting 2 standard deviations away from the predicted value. This range of values provides a better guidance to the actuary who is usually faced with finding the right reserves to hold within a reasonable confidence level.

Table 5.5.1: Reserve estimates by calendar year - Gaussian copula

Calendar Year	ILRA			ILRB			ILRE			ILRH1		
	Lower Bound	Predicted Value	Upper Bound	Lower Bound	Predicted Value	Upper Bound	Lower Bound	Predicted Value	Upper Bound	Lower Bound	Predicted Value	Upper Bound
1998	8,035	14,736	22,975	808,695	957,024	1,121,485	1,324	3,015	5,659	5,236	25,980	55,660
1999	3,480	6,381	9,962	398,117	471,009	551,854	655	1,483	2,759	3,813	18,952	40,554
2000	2,071	3,798	5,929	209,116	247,448	289,926	431	976	1,814	2,976	14,769	31,655
2001	1,263	2,315	3,611	106,255	125,744	147,372	284	645	1,201	2,314	11,471	24,546
2002	754	1,382	2,158	51,317	60,703	71,130	188	431	807	1,790	8,919	19,102
2003	415	761	1,188	28,418	33,612	39,375	114	261	488	1,296	6,472	13,906
2004	232	426	666	15,791	18,691	21,897	88	204	385	891	4,509	9,813
2005	129	239	375	8,595	10,179	11,933	44	106	203	256	1,305	2,849
2006	31	57	90	5,184	6,157	7,236	18	45	91	62	323	719

Table 5.5.2: Combined reserve estimates by calendar year - Gaussian copula

Calendar Year	Avg - 2×StdDev	Lower Bound	Predicted Value	Upper Bound	Avg + 2×StdDev
1998	789,551	837,799	1,000,756	1,183,022	1,211,962
1999	389,146	414,413	497,826	591,979	606,506
2000	205,677	220,316	266,991	320,489	328,305
2001	105,336	113,926	140,176	170,864	175,015
2002	51,181	56,472	71,435	89,506	91,689
2003	28,209	31,757	41,107	52,750	54,004
2004	15,535	17,951	23,830	31,380	32,125
2005	8,566	9,425	11,828	14,726	15,091
2006	5,038	5,416	6,582	7,933	8,126

Table 5.5.3: Reserve estimates by calendar year - Frank copula

Calendar Year	ILRA			ILRB			ILRE			ILRH1		
	Lower Bound	Predicted Value	Upper Bound	Lower Bound	Predicted Value	Upper Bound	Lower Bound	Predicted Value	Upper Bound	Lower Bound	Predicted Value	Upper Bound
1998	8,046	14,740	23,167	806,041	956,724	1,124,442	1,168	3,017	6,033	4,955	25,938	57,449
1999	3,483	6,383	10,035	396,712	470,862	553,425	575	1,484	2,970	3,617	18,921	41,907
2000	2,073	3,798	5,972	208,415	247,369	290,741	378	977	1,953	2,817	14,744	32,656
2001	1,263	2,316	3,641	105,910	125,705	147,745	250	645	1,293	2,191	11,453	25,382
2002	754	1,382	2,172	51,126	60,683	71,324	167	432	863	1,701	8,907	19,731
2003	415	761	1,197	28,315	33,602	39,493	101	261	523	1,235	6,462	14,316
2004	233	426	670	15,742	18,685	21,964	79	204	409	861	4,500	9,964
2005	130	239	375	8,575	10,175	11,959	41	106	212	249	1,303	2,885
2006	31	57	90	5,184	6,155	7,234	17	45	91	62	322	715

Table 5.5.4: Combined reserve estimates by calendar year - Frank copula

Calendar Year	Avg - 2×StdDev	Lower Bound	Predicted Value	Upper Bound	Avg + 2×StdDev
1998	790,669	838,459	1,000,418	1,180,686	1,210,167
1999	390,955	415,529	497,650	589,682	604,344
2000	207,388	221,519	266,889	318,568	326,390
2001	106,584	114,935	140,118	169,513	173,652
2002	51,946	57,203	71,404	88,709	90,861
2003	28,713	32,259	41,086	52,250	53,459
2004	15,908	18,284	23,817	31,027	31,725
2005	8,729	9,549	11,822	14,567	14,915
2006	5,129	5,468	6,580	7,836	8,030

Based on our parametric model specification, we are also able to provide predictive distributions of the total reserve for the entire portfolio. In order to have a clearer understanding of these total reserves, we normalize the reserve estimates by dividing the values with the total earned premiums. This is a typical exercise in practice where companies are usually interested in the loss ratio; this ratio provides the estimated reserves per dollar of premium collected. These predictive distributions are summarized in Figure 5.5.1. This predictive distributions provide us a better picture of the variability present in the reserve estimates for the entire portfolio. The actuary who is faced with a difficult decision of making the correct reserve value to hold can rely on these figures. If he wants to hold a certain level of reserves in his company's portfolio, he can use this predictive distribution to allow him to estimate the level of confidence given the level of reserves.

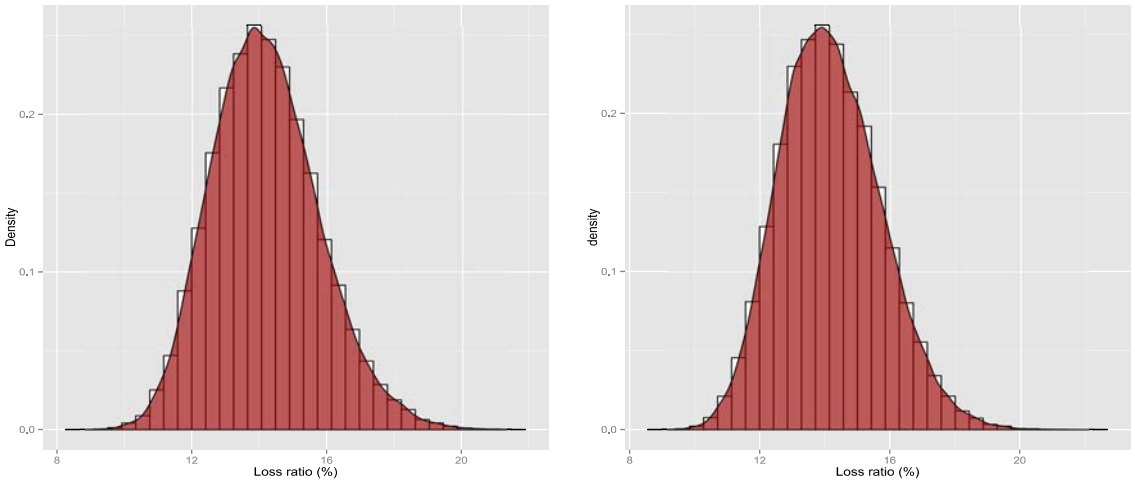

Figure 5.5.1: Total Unpaid Reserves per dollar premium exposure
Left panel: Gaussian copula, Right panel: Frank copula

Finally, we also compared the results of our model with the commonly used multivariate chain ladder method. This is summarized in Table 5.5.5, and although there

are some variations in the reserve estimates among the different models, we do not find very large deviations among these estimates. In some sense, we feel very confident about the estimates arising from our chosen model. Whether any one of these models provide better estimate, that is a question for additional work which goes beyond our purpose.

Table 5.5.5: Model comparison

Accident Year	**Multivariate Chain Ladder**	**Gaussian copula**	**Frank copula**
1989	3,463	3,245	3,243
1990	5,858	5,646	5,644
1991	11,978	11,811	11,808
1992	25,938	23,693	23,687
1993	50,797	47,289	47,259
1994	112,001	113,262	113,197
1995	234,878	241,633	241,617
1996	483,958	498,869	498,655
1997	1,129,869	1,115,084	1,114,673

Chapter 6

Concluding Remarks and Possible Further Work

In this thesis, we proposed one particular technique to model multivariate longitudinal data. We find in the statistics literature on multivariate longitudinal data analysis that the convention of assuming normality distributions for the response variables is commonplace. This is not at all surprising because the extension of the use of mixed effects models within this framework provides the ease of both interpretation and estimation. Several built-in packages provide these natural extensions. Furthermore, we find a huge potential of the uses of multivariate longitudinal data analysis in the field of actuarial science, insurance, and finance. We demonstrate this usefulness with some empirical investigation relating to global insurance demand and loss reserving in general insurance.

As a departure from the traditional normality assumption for response variables, we constructed our models to accommodate a more flexible approach so that other types of dependence models are used and that a larger family of marginal distributions

can be utilized. For example, we suggested several possible families of copula models as well as a wide variety of skewed distributions for single-valued random variables. The model specifies a structure that not only allows for adopting the dependence among the several response variables, but at the same time, the heterogeneity of the observables in the form of covariates and the dynamic nature within the longitudinal set up. The proposed model is based around the idea of copulas, a concept that has grown in rapid importance within several disciplines including actuarial science, insurance, and finance. The flexibility of copula functions, as well as its all-encompassing accommodation of wide variety of dependence, is what makes this widespread these days.

Copula functions are based on a uniform transformation of the marginals, and as such, the choice and modeling of the marginals are usually separately done from the dependence or multivariate structure. This allows the modeler to investigate the marginals separately from the dependence structure, and this gives flexibility to the choice of marginals as well. Even though there are some multivariate distributions available in the literature that can be directly applied to model multivariate data, their limitations to restricting the marginals present some problems when it comes to fitting empirical data. For example, multivariate normal or multivariate gamma distributions are forced to have all response variables from the same distribution family. Copulas do not have such similar problems.

There are other several advantages especially for our purposes. It is straightforward to accommodate more than two dimensions. So long as one can easily develop the likelihood of the data observed, the use of maximum likelihood estimation method to estimate the parameters becomes also straightforward. Of course, in several sit-

uations, the likelihood can be looking very complex and sometimes involve multiple integrations. As such, one can view this as a drawback, but numerical techniques including the use of simulation as implemented in this thesis work just fine. In the future, we hope that we can refine these numerical techniques in order to arrive at the estimates more efficiently and even more accurately, leading to model estimates that are much more reliable. However, despite this limitation, we find our technique work suitably well with our data and that we are able to rely on large sample properties of maximum likelihood estimates to perform additional statistical inference.

For each application in our analysis, we have suggestions and ideas to further advance the work we have accomplished in this thesis.

In our first application relating to global insurance demand, our study indicates that the insurance density or insurance premium, which is a proxy for insurance demand, in both life and non-life insurance are heavy-tailed distributions. This clearly indicates deviation from conventional normality assumptions. In our work, we were able to successfully incorporate skewed distributions for dependent variables such as GB2 distributions for our marginals.

Another important aspect of our research is understanding the relationship between life and non-life insurance demand. To the best of our knowledge, we are not aware any other study that have appeared in the literature involving understanding the possible relationship that exists between these two demands. Preliminary analysis as well as concluding results of our study implies that there is a strong statistical evidence of a positive relationship. This relationship between these two major lines of insurance could be used by insurance companies which may be seeking to enter into new market expansions into international operations. Furthermore, for a given

country, knowing the relationship of the demand between life and non-life insurance could help existing insurance companies to decide possible expansion into both lines of insurance.

The results in our study will possibly be very different from other studies that examine factors influencing insurance demand, although we suspect that the behavior of the relationship would be marginally different. One reason for this difference is that different set of explanatory or predictor variables may be used for different research studies. To illustrate, according to the annual sigma reports from Swiss Re, one can observe that there is noticeable difference in total premiums collected in different regions around the world. Clearly, this is an indication that insurance demand can vary by regions. We can further examine regional differences, for example, to extend our study where we distinguish total insurance premiums around the world according to regions as classified in Table 6.0.1. Although we were well aware of the impact of these regional differences on premiums, we did not account for these differences because this will complicate the estimation with the addition of several parameters in the model. This is especially true where such regions will have to fall as categorical variable. The purpose of this thesis is to simply demonstrate the usefulness of our model specification and their effectiveness in understanding relationships of the responses. It is however interesting to improve our model by incorporating other significant predictor variables such as regional differences.

In our definition for insurance demand, we utilized heavily the level of premiums collected to measure it. However, this can be argued in many respects. For example, the price of an insurance product may heavily depend on the nature of the product design or on the nature of the insurance product itself. Some life insurance products

Table 6.0.1: Regions to distinguish insurance premiums

North America
Western Europe
Advance Asia
South Asia
Oceania
Emerging Asia
Latin America
Central and Eastern Europe
Middle east and Central Asia
Africa

are designed to cover only the mortality component while in many more advanced countries, the investment component may be more pronounced within the product. To illustrate, term life insurance generally cheaper than whole life insurance products. Within the life insurance line of business alone, the variation of the product design can impact demand. Furthermore, Treerattanapun (2011) observed that auto insurance dominates the non-life insurance market, especially in developing countries, which suggests that consumption of insurance product vary even within a line of insurance. Auto insurance are generally less expensive when compared to some other non-life insurance products like homeowners insurance or liability insurance. For further investigation in this arena, it is also recommended to find a more fitting measure of insurance demand.

There are several rooms for improvement in our work on the loss reserving multivariate models. Some of these problems are well known in both theory and practice. To illustrate, in some insurance data, it is possible to have negative cumulative or incremental losses. These types of values are possible due to a number of reasons that generally arise during the claims handling process. Generally, an insured would claim

directly from its insurer regardless of who is at fault. In many instances, the insurer may first settle these claims but later investigate so that a possible salvage may occur when later determined another party is at fault. This gives rise to a correction in claims thereby leading to negative claims. In addition, reinsurance is a practice used by insurance companies to insure its losses beyond their capacity. Recovery arising from reinsurance generally appear on their books as negative claims. In our analysis of correlated loss triangles, we generally ignored the issue of negative increments but modeling this is not a difficult exercise. Therefore, extension to handle negative incremental claims in loss triangle should be straightforward. One possible approach is described in England and Verrall (2002) using normal approximations to the data. However, they pointed out that normal approximation would not be used in all situation unless available data satisfy the normality assumptions. Another possible approach is to separately model the negative incremental losses from the positive ones, and incorporating an additional parameter of the probability that it is negative (or its complement which is the probability it is positive). Certainly, there are other aspects of loss reserving that we ignored but can be handled without much additional complication in our model specification.

Bibliography

Ajne, B. 1994. "Additivity of chain-ladder projections." *ASTIN Bulletin* 24: 313-318.

Anderson, D. R., and J. R Nevin. 1975. "Determinants of young marrieds' life insurance purchasing behavior: An empirical investigation." *Journal of Risk and Insurance* 42: 375-387.

Arena, M. 2008. "Does Insurance Market Activity Promote Economic Growth? A Cross-Country Study for Industrialized and Developing Countries." *Journal of Risk and Insurance* 75 (4): 921-946.

Auerbach, A. T., and L. J. Kotlikoff. 1991. "How Rational is the Purchase of Life Insurance?" National Bureau of Economic Research, Working Paper No. w3063.

Bandyopadhyay, S., B. Ganguli, and A. Chatterjee. 2011. "A review of multivariate longitudinal data analysis." *Statistical Methods in Medical Research* 20 (4): 299-330.

Barnett, G., and B. Zehnwirth. 2000. "Best estimates for reserves." 87 (167): 245-321. Proceedings of the Casualty Actuarial Society

Beck, T., and I. Webb. 2003. "Economic, demographic, and institutional determinants of life insurance consumption across countries." *The World Bank Economic Review* 17 (1): 51-88.

Beenstock, M., G. Dickinson, and S. Khajuria. 1986. "The Determinant of Life Premiums: An International Cross-Sectional Analysis 1970-1981." *Insurance: Mathematics and Economics* 5: 261-270.

Braun, C. 2004. "The prediction error of the chain ladder method applied to correlated run-off triangles." *Astin Bulletin* 34 (2): 399-424.

Brehm, P. 2002. "Correlation and the aggregation of unpaid loss distributions." *CAS Forum* 2: 1-23.

Brokesova, Z., E. Pastorakova, and T. Ondruska. 2014. "Determinant of insurance industry development in transition economies: Empirical analysis of Visegrad group data." *The Geneva Papers on Risk and Insurance-Issues and Practice* 39: 471-492.

Browne, M. J., J. W. Chung, and E. W. Frees. 2000. "International Property-liability Insurance Consumption." *Journal of Risk and Insurance* 67: 73-90.

Browne, M. J., and K. Kim. 1993. "An International analysis of life insurance demand." *Journal of Risk and Insurance* 60: 616-634.

Burnett, M. J., and B. A Palmer. 1984. "Examining life insurance ownership through demographic and psychographic characteristics." *Journal of Risk and Insurance* 51: 453-467.

Campbell, R. A. 1980. "The demand for life Insurance: An application of the economics of university." *The Journal of Finance* 35: 1155-1172.

De Jong, P. 2012. "Modeling dependence between loss triangles." *North American Actuarial Journal* 16 (1): 74-86.

Diggle, P., P. Heagerty, Liang K. Y., and S. Zeger. 2013. *Analysis of Longitudinal Data*. Oxford University Press.

Dragos, S. L. 2014. "Life and non-life insurance demand: the different effects of influence factors in emerging countries from Europe and Asia." *Economic Research-Ekonomska Istraživanja* 27 (1): 169-180.

Duker, J. M. 1969. "Expenditures for life Insurance among working-wife families." *Journal of Risk and Insurance* 36: 525-533.

Eling, M., S. Pradhan, and J. T. Schmit. 2014. "Determinants of microinsurance demand." *The Geneva Papers on Risk and Insurance-Issues and Practice* 39 (2): 224-263.

Embrechts, P., F. Lindskog, and A. McNeil. 2003. "Modelling dependence with copulas and applications to risk management." 8 (1): 329-384.

England, P. D., and R. J. Verrall. 2002. "Stochastic claims reserving in general insurance." *British Actuarial Journal* 8 (3): 443-518.

Esho, N., A. Kirievsky, D. Ward, and R. Zurbruegg. 2004. "Law and the Determinants of Property-Casualty Insurance." *Journal of Risk and Insurance* 71: 265-283.

Fieuws, F., and G. Verbeke. 2009. "Joint models for high-dimensional longitudinal data." *Longitudinal data analysis* 16: 367-391.

Fieuws, S., and G. Verbeke. 2006. "Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles." *Biometrics* 62: 424-431.

Fieuws, S., G Verbeke, and G. Molenberghs. 2007. "Random-effects models for multivariate repeated measures." *Statistical Methods in Medical Research* 16: 387-397.

Fischer, S. 1973. "A life cycle model of life insurance purchases." *International Economic Review* 14 (1): 132-152.

Fisher, N. I., and P. Switzer. 1985. "Chi-plots for assessing dependence." *Biometrika* 72 (2): 253-265.

Fitzmaurice, G., M. Davidian, G. Verbeke, and G. Molenberghs. 2008. *Longitudinal Data Analysis*. CRC Press.

Frank, M. J. 1979. "On the simultaneous associativity ofF (x, y) and x+y- F (x, y)." *Aequationes Mathematicae* 19 (1): 194-226.

Frees, E. W. 2004. *Longitudinal and Panel Data: Analysis and Applications for the Social Sciences*. Cambridge University Press.

Frees, E. W., and E. A. Valdez. 1998. "Understanding relationships using copulas." *North American Actuarial Journal* 2: 1-25.

Frees, E. W., and P. Shi. 2010. "Long-tail longitudinal modeling of insurance company expenses." *Insurance: Mathematics and Economics* 47: 303-314.

Frees, E. W., and P. Wang. 2005. "Credibility using copulas." *North American Actuarial Journal* 2: 1-18.

Frees, E. W., and P. Wang. 2006. "Copula credibility for aggregate loss models." *Insurance: Mathematics and Economics* 38: 360-373.

Frees, E. W., Virginia R. Young, and Y. Luo. 1999. "A longitudinal data analysis interpretation of credibility models." *Insurance: Mathematics and Economics* 24: 229-247.

Gao, F., P. Thompson, C. Xiong, and J. P. Miller. 2006. "Analyzing multivariate longitudinal data using SAS." Paper 187-31 of SUGI 31, SAS Institute Inc.

Genest, C. 1987. "Frank's family of bivariate distributions." *Biometrika* 74 (3): 549-555.

Genest, C., and J. Boies. 2003. "Detecting dependence with Kendall plots." *The American Statistician* 57 (4).

Goovaerts, M. J., R. Kaas, A. E. Van Heerwaarden, and T. Bauwelinckx. 1990. "Effective Actuarial Methods." North-Holland, Amsterdam.

Hammond, J. D., D. B. Houston, and E. R. Melander. 1967. "Determinants of household life insurance premium expenditure: An empirical investigation." *Journal of Risk and Insurance* 34: 397-408.

Hess, K. Th., K. D. Schmidt, and M. Zocher. 2006. "Multivariate loss prediction in the multivariate additive model." *Insurance: Mathematics and Economics* 39 (2): 185-191.

Holmberg, R. D. 1994. "Correlation and the measurement of loss reserve variability." In: Casualty Actuarial Society (CAS) Forum Spring, pp. 247-278.

Hürlimann, W. 2005. "Approximate bounds for the IBNR claims reserves based on the bivariate chain-ladder model." *Belgian Actuarial Bulletin* 5 (1): 46-51.

Hwang, T., and S. Gao. 2003. "The determinants of the demand for life insurance in an emerging economy-the case of China." *Managerial Finance* 29: 82-96.

Joe, H. 1997. *Multivariate Models and Dependence Concepts.* Boca Raton, FL: Chapman & Hall/CRC.

Kass, R., M. Goovaerts, J. Dhaene, and M. Denuit. 2008. *Modern Actuarial Risk Theory using R.* Springer, Verlag Berlin Heidelberg.

Kettani, H. 2010. "World Muslim Population: 1950–2020." *International Journal of Environmental Science and Development* 1 (2): 127-70.

Klugman, S. A., H. H. Panjer, and G. E. Willmot. 2012. *Loss Models: from Data to Decisions.* John Wiley & Sons.

Kremer, E. 1982. "IBNR-claims and the two-way model of ANOVA." *Scandinavian Actuarial Journal* 1982 (1): 47-55.

Kremer, E. 2005. "The correlated chain-ladder method for reserving in case of correlated claims developments." *Blätter der DGVFM* 27 (2): 315-322.

Laird, N. M., and J. H. Ware. 1982. "Random effects models for longitudinal data." *Biometrics* 34: 69-76.

Lambert, P., and F. Vandenhende. 2002. "A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant." *Statistics in Medicine* 21: 3197-3217.

Lenten, L. J. A., and D. N. Rulli. 2006. "A Time-Series Analysis of the Demand for Life Insurance Companies in Australia: An Unobserved Components Approach." *Australian Journal of Management* 31: 41-66.

Lester, R., R. Rocha, and E. Feyen. 2011. *What drives the development of the insurance sector? An empirical analysis based on a panel of developed and developing countries.* The World Bank.

Lewis, F. D. 1989. "Dependents and the demand of life insurance." *American Economic Review* 79: 452 - 466.

Li, D., F. Moshirian, P. Nguyen, and T. Wee. 2007. "The demand for life insurance in OECD countries." *Journal of Risk and Insurance* 74: 637 - 652.

Lim, C. C, and S. Haverman. 2004. "Modelling life insurance demand from a macroeconomic perspective: The Malaysian case." Research paper: The 8th International Congress on Insurance, Mathematics and Economics, Rome.

Ludwig, A., and K. D. Schmidt. 2010. *Calendar Year Reserves in The Multivariate Additive Model.* Techn. Univ., Inst. für Mathematische Stochastik.

Mack, T. 1991. "A simple parametric model for rating automobile insurance or estimating IBNR claims reserves." *ASTIN Bulletin* 21 (1): 93-109.

Mack, T. 1993. "Distribution-free calculation of the standard error of chain-ladder reserve estimates." *ASTIN Bulletin* 23 (2): 213-225.

Merz, M., and M. Wuthrich. 2008a. "Prediction error of the chain ladder reserving method applied to correlated run-off trapezoids." *Annals of Actuarial Science* 2 (1): 25-50.

Merz, M., and M. Wuthrich. 2008b. "Prediction error of the multivariate chain ladder reserving method." *North American Actuarial Journal* 12 (2): 175-197.

Merz, M., and M. Wuthrich. 2009a. "Combining chain-ladder and additive loss reserving methods for dependent lines of business." *Variance* 3 (2): 270-291.

Merz, M., and M. Wuthrich. 2009b. "Prediction error of the multivariate additive loss reserving method for dependent lines of business." *Variance* 3 (1): 131-151.

Nelsen, R. B. 2007. *An Introduction to Copulas.* Springer.

Nesterova, D. 2008. "Determinants of the demand for life insurance: Evidence from selected CIS and CEE countries." National University Kyiv-Mohyla Academy.

Outreville, J. F. 1990. "The economic significance of insurance markets in developing countries." *Journal of Risk and Insurance* 57: 487-498.

Outreville, J. F. 1996. "Life insurance market in developing countries." *Journal of Risk and Insurance* 63 (2): 263-278.

Outreville, J. F. 2013. "The relationship between insurance growth and economic development: 85 empirical papers for a review of the literature." *Risk Management and Insurance Review* 16: 71-122.

Park, S. C., and J. Lemaire. 2011. "The impact of culture on the demand for non-life insurance." University of Pennsylvania, Wharton School Working Paper IRM 2011-02.

Pöhl, C., and K. D. Schmidt. 2005. "Multivariate chain ladder." 36th International ASTIN Colloquium.

Reinsel, G. 1982. "Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure." *Journal of the American Statistical Association* 77: 190 - 195.

Renshaw, A. 1989. "Chain-ladder and interactive modeling (Claims reserving and GLIM)." *Journal of the Institute of Actuaries.* 116 (3): 559-587.

Renshaw, A., and R. Verrall. 1998. "A stochastic model underlying the chain-ladder technique." *British Actuarial Journal* 4: 903-923.

Rochon, J. 1996. "Analyzing bivariate repeated measures for discrete and continuous outcome variable." *Biometrics* 52: 740-750.

Schmidt, K. D. 2006. "Optimal and additive loss reserving for dependent lines of business." In: Casualty Actuarial Society (CAS) Forum Fall, pp. 319-351.

Schmidt, K. D., and A. Schnaus. 1996. "An extension of Mack's model for the chain ladder method." *Astin Bulletin* 26 (2): 247-262.

Schmidt, K. D., and M. Zocher. 2007. "The Bornhuetter-Ferguson principle." *Variance* 2 (1): 85-110.

Sen, S. 2008. *An Analysis of Life Insurance Demand Determinants for Selected Asian Economies and India.* Madras School of Economics.

Shah, A., N. M Laird, and D. Schoenfeld. 1997. "A random effects model with multiple characteristics with possibly missing data." *Journal of the American Statistical Association* 92: 775-779.

Shi, P. 2012. "Multivariate longitudinal modeling of insurance company expenses." *Insurance: Mathematics and Economics* 51: 204-215.

Shi, P. 2013. "A Multivariate analysis of intercompany loss triangles." https://sites.google.com/a/wisc.edu/peng-shi/.

Shi, P., and E. W. Frees. 2010. "Long-tail longitudinal modeling of insurance company expenses." *Insurance: Mathematics and Economics* 47: 303-314.

Shi, P., and E. W. Frees. 2011. "Dependent loss reserving using copulas." *ASTIN Bulletin* 41: 449-486.

Shi, P., S. Basu, and G. G. Meyers. 2012. "A Bayesian log-normal model for multivariate loss reserving." *North America Actuarial Journal* 16 (1): 29-51.

Singer, J. D., and J. B. Willett. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* Oxford University Press, Inc.

Sun, J., E. W. Frees, and M. A. Rosenberg. 2008. "Heavy-tailed longitudinal data modeling using copulas." *Insurance: Mathematics and Economics* 42 (2): 817-830.

Taylor, G. 2000. *Loss Reserving: An Actuarial Perspective.* Kluwer Academic Publishers.

Taylor, G. C., and F. R. Ashe. 1983. "Second moments of estimates of outstanding claims." *Journal of Econometrics* 23 (1): 37-61.

Treerattanapun, A. 2011. "The impact of culture on non-life insurance consumption." Paper presented at Wharton Research Scholars Project.

Verrall, R. J. 1989. "A State Space Representation of the Chain-ladder Linear Model." *Journal of the Institute of Actuaries* 116 (3): 589-609.

Verrall, R. J. 2000. "An investigation into stochastic claims reserving models and the chain-ladder technique." *Insurance: Mathematics and Economics* 26 (1): 91-99.

Wuthrich, M. V., and M. Merz. 2008. *Stochastic Claims Reserving Methods in Insurance.* Wiley.

Yarri, M. E. 1965. "Uncertain lifetime, life insurance, and the theory of the consumer." *Review of Economic Studies* 32: 137-150.

Zhang, Y. 2010. "A general multivariate chain ladder model." *Insurance: Mathematics and Economics* 46: 588-599.

Zhao, X., and X. Zhou. 2010. "Applying copula models to individual claim loss reserving methods." *Insurance: Mathematics and Economics* 46 (2): 290-299.

Zietz, E. N. 2003. "An examination of the demand for life insurance." *Risk Management and Insurance Review* 6 (2): 159-191.

Appendix A

A.1 Additional plots for global insurance demand

Originally, we considered 75 countries in our data set for the global insurance demand. We observed unusual trend of insurance per capita over time in three countries (Ireland, Switzerland, and United Kingdom). Preliminary statistics indicates that these countries could distort the dynamic dependency and hence we excluded them from the data set that we used in the final estimation of the model. Following Figure A.1.1 and A.1.2 show multivariate time series plots of original data set.

Under the copula framework, we tested copula functions in both Archimedean and Elliptical family. In Chapter 3, we display the so-called "copula pp-plot" for diagnostic of copula functions used in our model. Even though originally observed data are time dependent (see multivariate time series plots), regression models incorporated random effect terms to allow for the dynamic dependency within subjects. Henceforth, the resulting residuals from the estimated regression models can be assumed to be independent over time. We included all these residuals together for our copula

validation in Chapter 3. However, here we present the "copula pp-plot" by different years for different copula functions we considered.

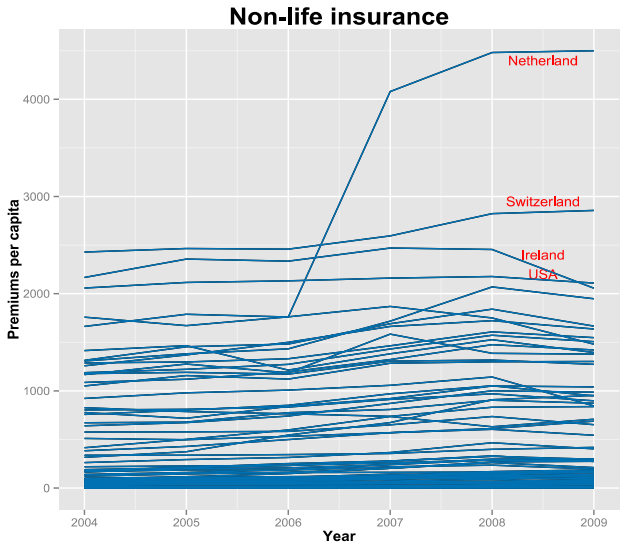

Figure A.1.1: Multivariate time series plot of non-life insurance per capita

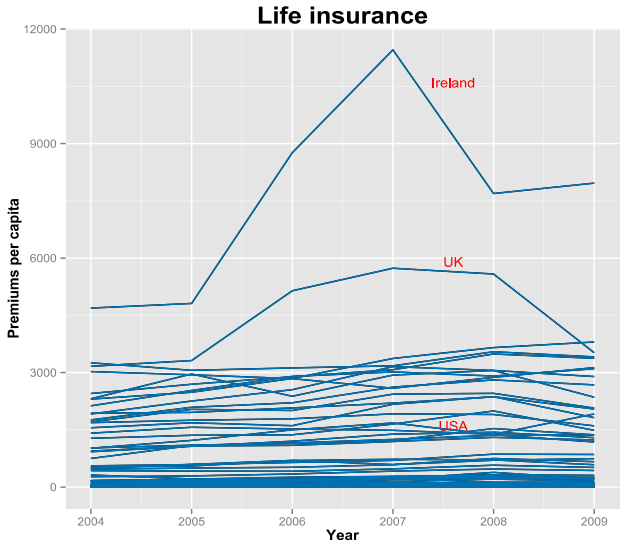

Figure A.1.2: Multivariate time series plot of life insurance per capita

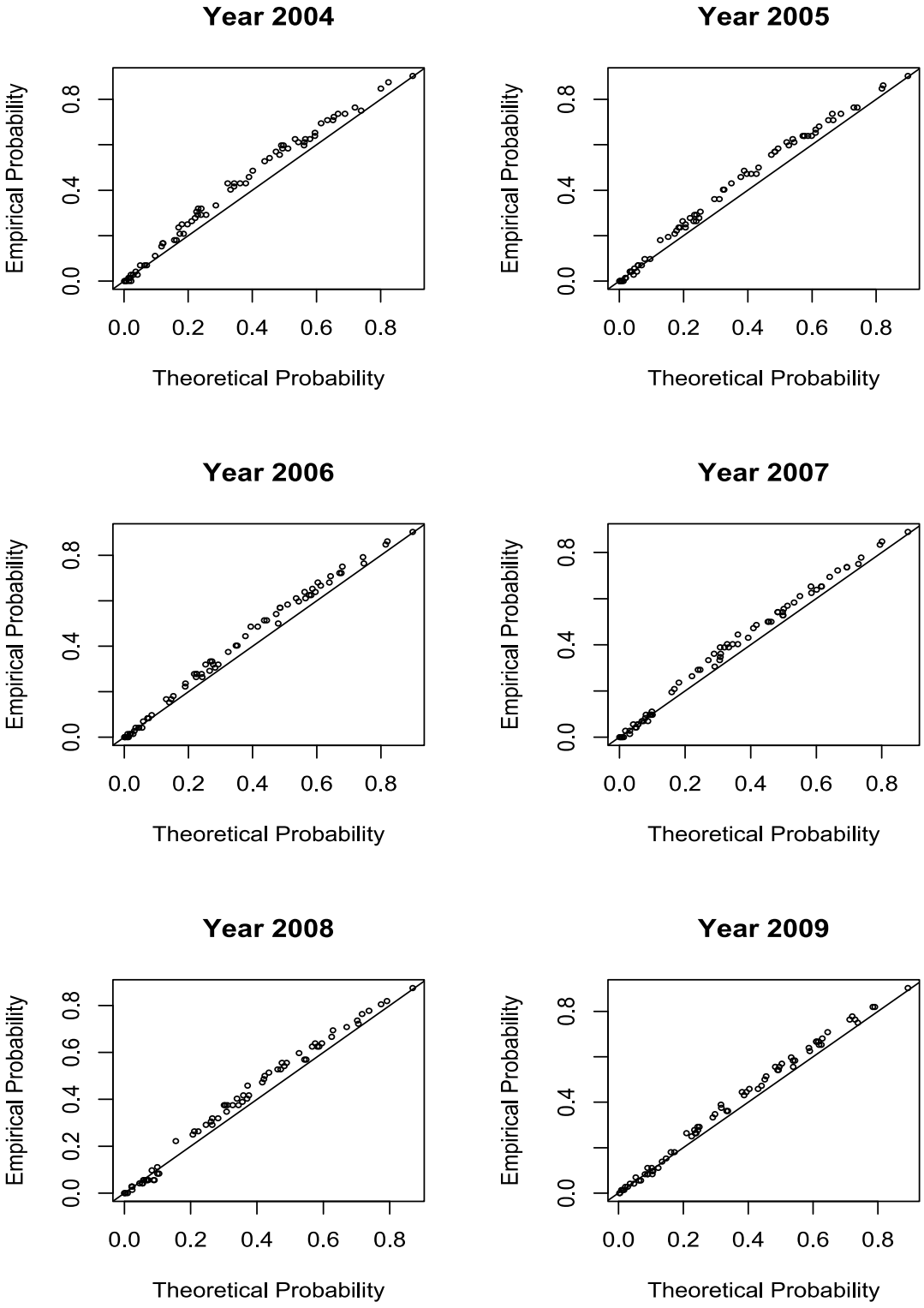

Figure A.1.3: Copula pp-plot for Clayton copula function

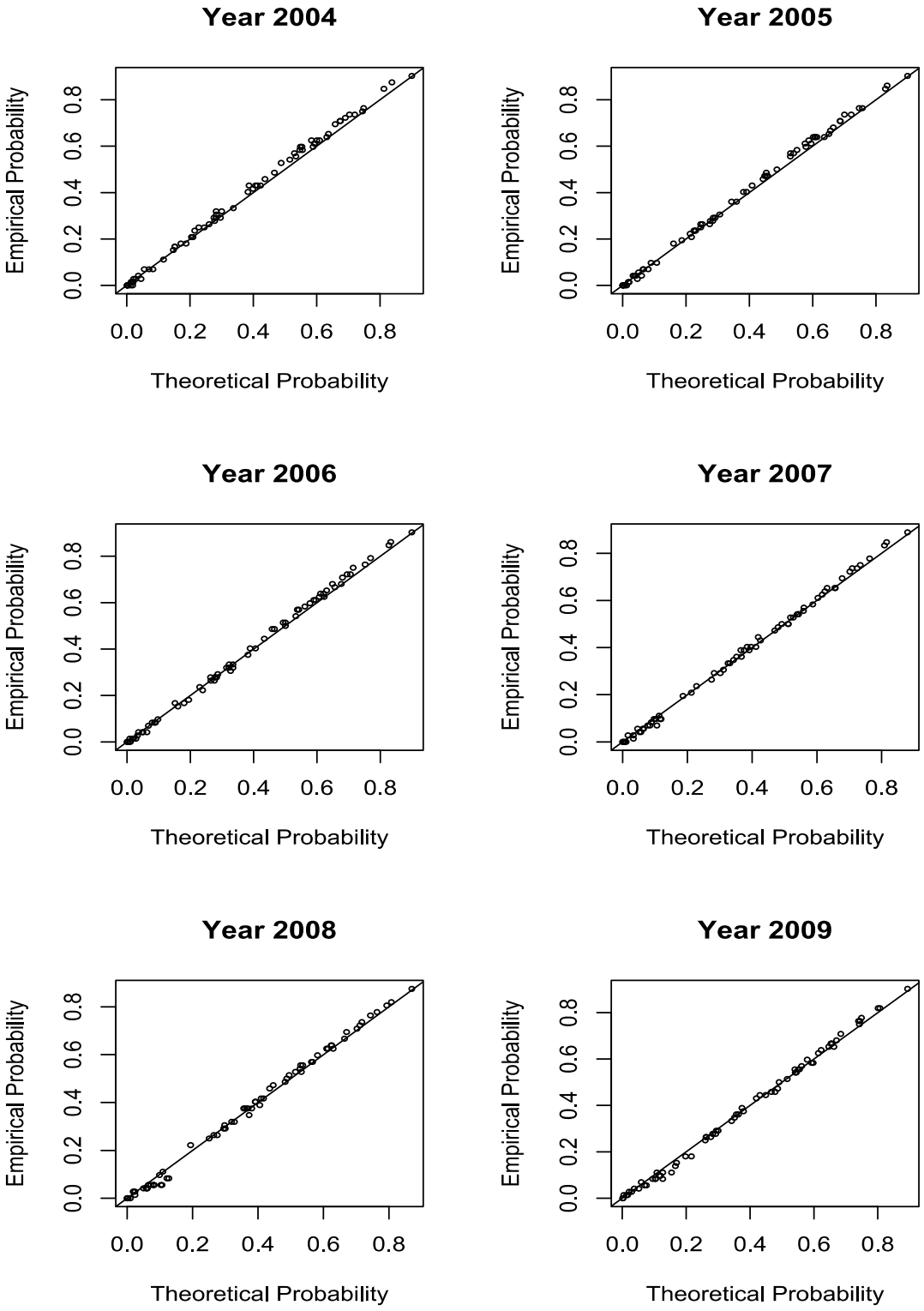

Figure A.1.4: Copula pp-plot for Frank copula function

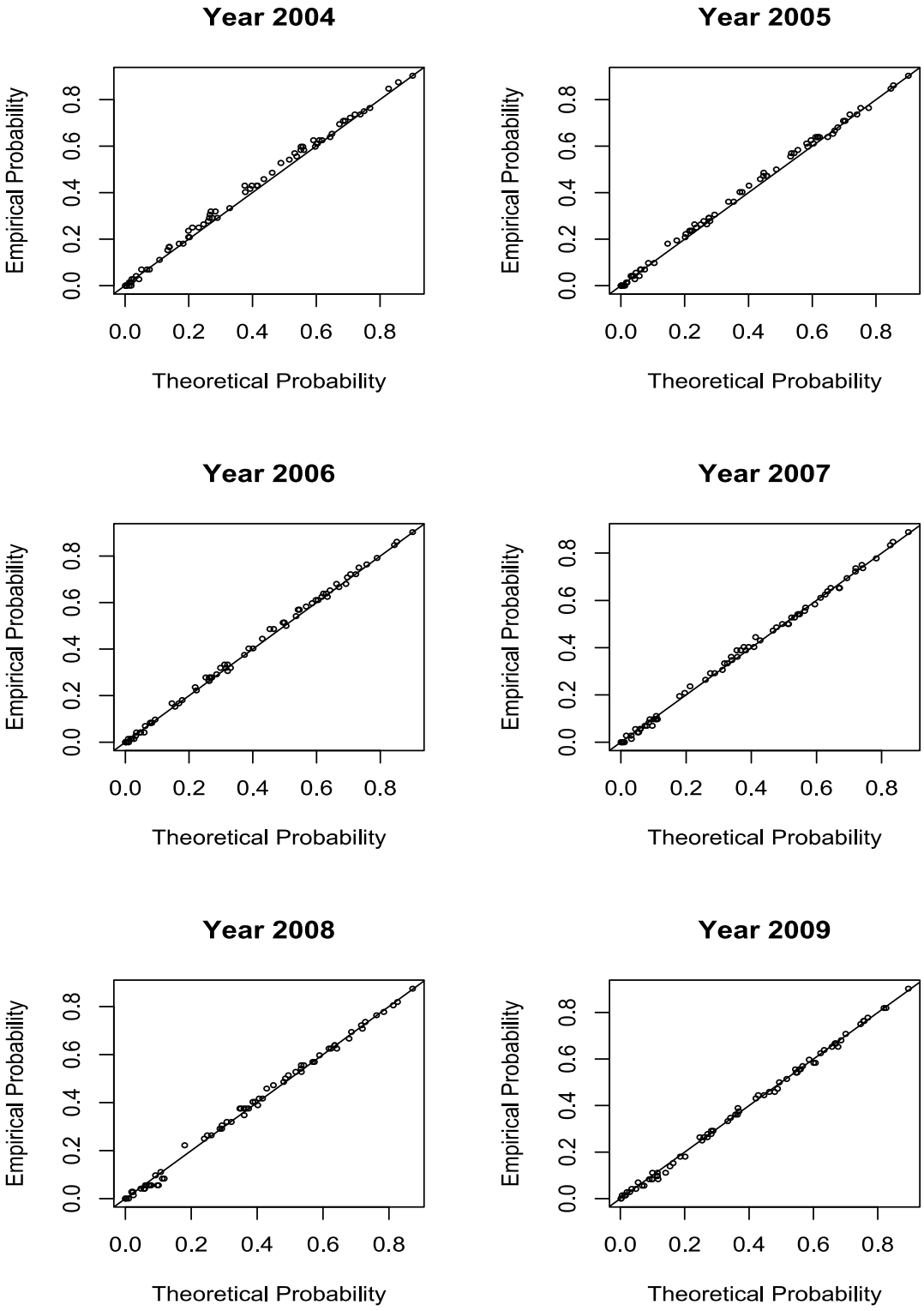

Figure A.1.5: Copula pp-plot for Gumble copula function

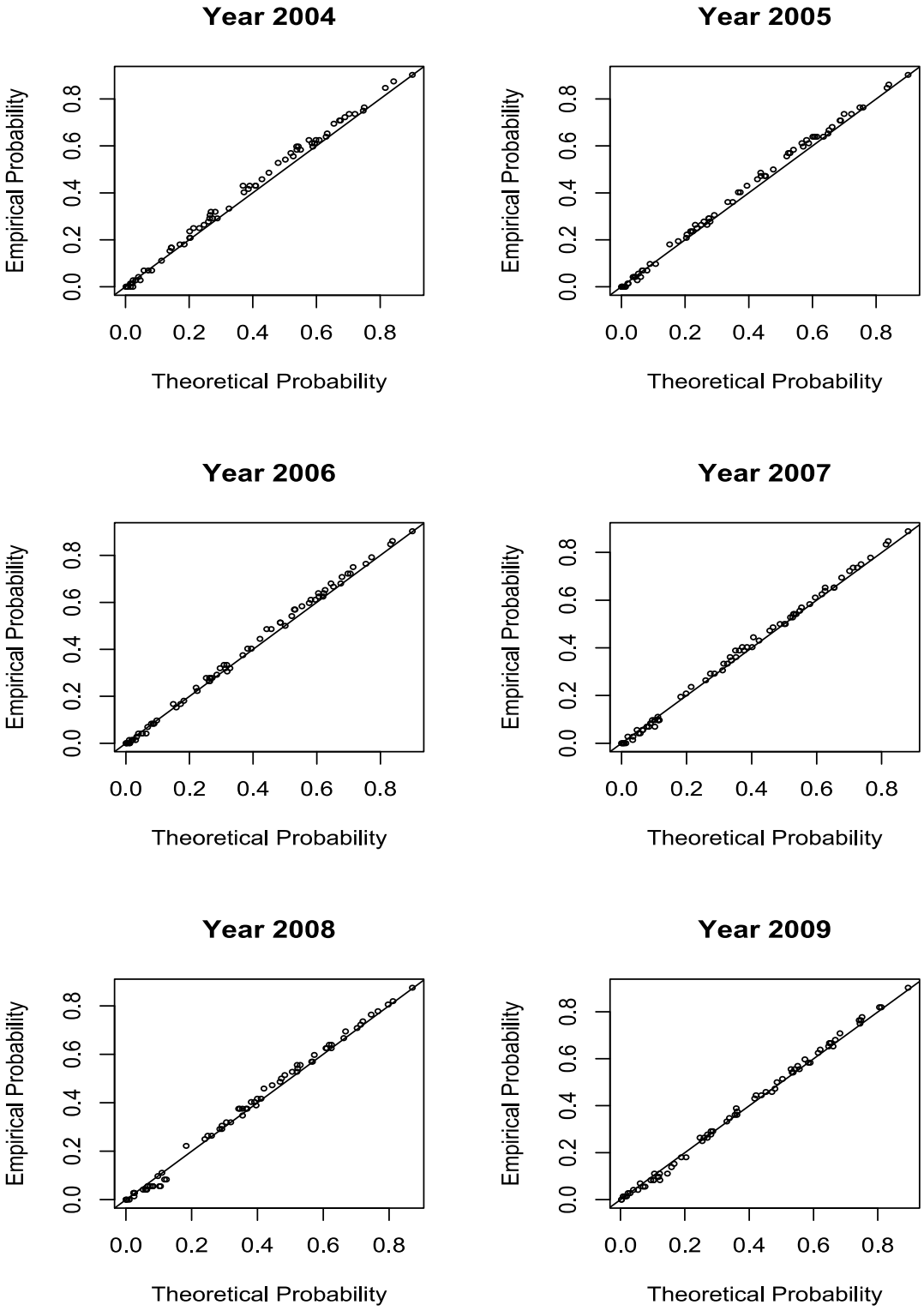

Figure A.1.6: Copula pp-plot for Normal copula function

A.2 Reserve estimates by accident year

Estimated reserves by accident year for the total portfolio from both the Gaussian and Frank copula models are provided in the following two tables. The corresponding reserve estimates for each individual lines of insurance are shown in Tables A.2.3 and A.2.4.

Table A.2.1: Combined reserve estimates by accident year - Gaussian copula

Calendar Year	Avg - 2×StdDev	Lower Bound	Predicted Value	Upper Bound	Avg + 2×StdDev
1989	2,458	2,651	3,245	3,938	4,032
1990	4,401	4,695	5,646	6,732	6,891
1991	9,103	9,749	11,811	14,183	14,520
1992	18,475	19,692	23,693	28,254	28,911
1993	36,732	39,167	47,289	56,537	57,845
1994	83,316	91,068	113,262	139,929	143,209
1995	181,821	196,740	241,633	294,272	301,445
1996	382,850	410,864	498,869	600,658	614,888
1997	866,081	924,983	1,115,084	1,331,428	1,364,087

Table A.2.2: Combined reserve estimates by accident year - Frank copula

Accident Year	Avg - 2×StdDev	Lower Bound	Predicted Value	Upper Bound	Avg + 2×StdDev
1989	2,474	2,670	3,243	3,916	4,012
1990	4,426	4,712	5,644	6,700	6,863
1990	9,198	9,823	11,808	14,071	14,417
1990	18,552	19,755	23,687	28,127	28,821
1990	37,021	39,430	47,259	56,117	57,497
1990	85,213	92,386	113,197	137,827	141,181
1990	185,643	199,355	241,617	290,439	297,590
1990	388,967	414,736	498,655	593,709	608,344
1990	877,550	931,686	1,114,673	1,318,771	1,351,797

Table A.2.3: Reserve estimates by accident year - Gaussian copula

Accident Year	ILRA			ILRB			ILRE			ILRH1		
	Lower Bound	Predicted Value	Upper Bound	Lower Bound	Predicted Value	Upper Bound	Lower Bound	Predicted Value	Upper Bound	Lower Bound	Predicted Value	Upper Bound
1989	91	167	263	2,316	2,749	3,232	93	237	474	18	92	205
1990	127	234	368	4,324	5,132	6,036	52	134	269	28	145	323
1991	252	466	732	8,945	10,629	12,505	94	241	480	92	475	1,056
1992	549	1,010	1,578	18,170	21,552	25,315	126	309	603	160	823	1,828
1993	1,053	1,945	3,054	36,334	43,180	50,754	155	400	794	343	1,764	3,903
1994	1,917	3,541	5,557	83,277	98,835	116,148	231	593	1,182	1,992	10,294	22,857
1995	3,388	6,248	9,803	183,253	217,781	255,976	366	942	1,884	3,211	16,662	37,105
1996	4,703	8,701	13,674	390,883	464,283	545,702	545	1,399	2,793	4,737	24,486	54,173
1997	4,222	7,783	12,180	898,051	1,066,427	1,253,355	1,129	2,913	5,825	7,338	37,960	84,451

Table A.2.4: Reserve estimates by accident year - Frank copula

Accident Year	ILRA			ILRB			ILRE			ILRH1		
	Lower Bound	Predicted Value	Upper Bound	Lower Bound	Predicted Value	Upper Bound	Lower Bound	Predicted Value	Upper Bound	Lower Bound	Predicted Value	Upper Bound
1989	91	167	263	2,315	2,748	3,230	91	237	473	17	92	203
1990	128	234	368	4,323	5,131	6,030	52	134	268	28	145	322
1991	254	466	733	8,953	10,626	12,489	93	241	482	91	474	1,052
1992	551	1,010	1,588	18,157	21,546	25,323	120	309	619	156	821	1,822
1993	1,061	1,945	3,056	36,355	43,157	50,734	154	399	798	336	1,759	3,896
1994	1,931	3,541	5,563	83,228	98,794	116,129	230	593	1,187	1,956	10,270	22,750
1995	3,411	6,253	9,830	183,425	217,759	255,954	365	945	1,887	3,192	16,661	36,954
1996	4,745	8,701	13,671	391,010	464,106	545,569	542	1,399	2,799	4,662	24,450	54,297
1997	4,251	7,785	12,237	897,944	1,066,094	1,253,066	1,124	2,915	5,842	7,234	37,878	83,976

A.3 Some R commands

The following are few sample R codes used to estimate the marginals and copulas in our model related to the global insurance demand and correlated loss triangles.

```
rm(list=ls())

WIPD <- read.csv("/Users/Research/GlobalDemand/BalanceData2.csv")
WIPD$LIFE.DENSITY<-WIPD$LifePremsM*1000000/WIPD$Population
WIPD$NON_LIFE.DENSITY<-WIPD$NonlifePremsM*1000000/WIPD$Population

vars<-c("LIFE.DENSITY","NON_LIFE.DENSITY","Country","Year",
        "Deathrate","GDP_PER_CAPITA","Urbanpopulation",
        "Religious","Dependency_ratio_O")

WIPD<-WIPD[vars]
WIPD<-na.omit(WIPD)
attach(WIPD)

library(GB2)
set.seed(2)
u1<-runif(720000)

 Loglikelihood<-function(parm,x1){

#GB2 parameters
 a<-parm[1]
 p<-parm[2]
 q<-parm[3]

#Sigma parameter for normal distribution of random effect
 rsigma<-parm[4]

#Independent variable coefficients
 G1<-parm[5]
 R1<-parm[6]
 U1<-parm[7]
 D<-parm[8]
```

```
 DEP<-parm[9]


#Reading variables from the data set
 gdp<-WIPD[,5]     # GDP_PER_CAPITA
 rel<-WIPD[,8]     # Religious (Muslim population)
 urb<-WIPD[,7]     # Urbanpopulation
 deth<-WIPD[,6]    # Death rate
 dpr<-WIPD[,9]     # Dependency ratio

#Normaly distributed random effects
 W<-qnorm(u1, mean=0,sd=rsigma)
 WR<-matrix(W,nrow=10000)
 fun<-matrix(nrow=10000,ncol=72); fun[,]<-0

 for(j in 1:10000){
 w<-c(WR[j,],WR[j,],WR[j,],WR[j,],WR[j,],WR[j,])
 ML<-(w+G1*gdp+R1*rel+U1*urb+D*deth+DEP*dpr) # Regression function
 D1<-dgb2(x1,a,exp(ML),p,q)
 M<-matrix(D1,nrow=72)
 fun[j,]<-M[,1]*M[,2]*M[,3]*M[,4]*M[,5]*M[,6]
 }

 Integral<-colMeans(fun)
 return(-sum(log(Integral)))
 }
       init.es<-c(0.01,2,1,2,0.001,0.001,0.001,0.001,0.001)
       fit.GB2<-optim(init.es, Loglikelihood,NULL,
       control=list(maxit=5000),x1=LIFE.DENSITY)
       parm.hat<-fit.GB2$par

 library(nlme)
 Hess<-fdHess(parm.hat, Loglikelihood, x1=LIFE.DENSITY)
 inv.Hess<-solve(Hess$Hessian)
 parm.SE<-sqrt(abs(diag(inv.Hess)))
 d.f<-length(NON_LIFE.DENSITY)-length(parm.hat)
 t_ratio<-parm.hat/parm.SE
 p_val<-pf(t_ratio*t_ratio,df1=1,df2=d.f,lower.tail=FALSE)

 out_putR<-cbind(parm.hat,parm.SE,t_ratio,p_val)
```

```
  out_putR<-round(out_putR,digits=4)
  out_putR


#Gaussian copula
library(mvtnorm)
"cn" <- function(r,u1,u2,u3,u4)
{
    rmat <- matrix(c(1,r[1],r[2],r[3],r[1],1,r[4],r[5],r[2],r[4],1,
                  r[6],r[3],r[5],r[6],1),nrow=4,ncol=4,byrow=TRUE)
    temp1 <- qnorm(u1)
    temp2 <- qnorm(u2)
    temp3 <- qnorm(u3)
    temp4 <- qnorm(u4)
    xvect <- matrix(c(temp1,temp2,temp3,temp4),nrow=length(temp1),
                                                        ncol=4)
    temp5 <- dmvnorm(xvect,sigma=rmat)
    return(temp5/(dnorm(temp1)*dnorm(temp2)*dnorm(temp3)
                                              *dnorm(temp4)))
}

# t-copula
library(mvtnorm)
"ct" <- function(r,u1,u2,u3,u4)
{
    rmat <- matrix(c(1,r[1],r[2],r[3],r[1],1,r[4],r[5],r[2],r[4],1,
                  r[6],r[3],r[5],r[6],1),nrow=4,ncol=4,byrow=TRUE)
    d<-r[7]
    temp1 <- qt(u1,df=d)
    temp2 <- qt(u2,df=d)
    temp3 <- qt(u3,df=d)
    temp4 <- qt(u4,df=d)
    xvect <- matrix(c(temp1,temp2,temp3,temp4),nrow=length(temp1),
                                                        ncol=4)
    temp5 <- dmvt(xvect,sigma=rmat,df=d,log=FALSE)
    return(temp5/(dt(temp1,df=d)*dt(temp2,df=d)*dt(temp3,df=d)
                                            *dt(temp4,df=d)))
}
```