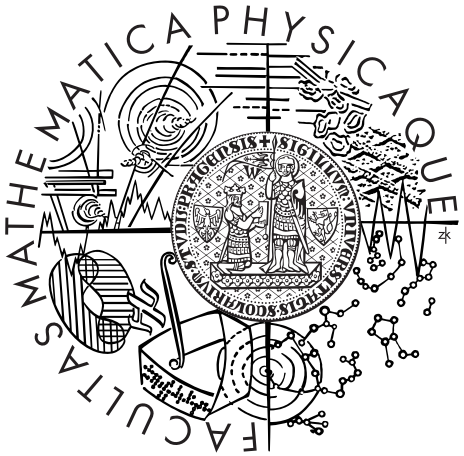
Charles University in Prague
Faculty of Mathematics and Physics

MASTER THESIS



Katarína Valentovičová

Claims reserving with copulae for multiple lines of business

Department of Probability and Mathematical Statistics

Supervisor of the master thesis: RNDr. Michal Pešta, Ph.D.
Study programme: Mathematics
Specialization: Financial and Insurance Mathematics

Prague 2015

In this place, I would like to express my gratefulness to the supervisor of my master thesis, RNDr. Michal Pešta, Ph.D., for his time, valuable suggestions and advice.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In Prague on 31st July 2015 Katarína Valentovičová

Název práce: Rezervování škod pomocí kopul pro více pojistných kmenů

Autor: Katarína Valentovičová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucíbakalářské práce: RNDr. Michal Pešta, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Stanovení rezerv a odhad předpokládaného škodního průběhu jsou jedním ze základních problémů pojišťovnictví. Celkové rezervy jsou obvykle stanovené za předpokladu nezávislosti pojistných kmenů. Nicméně, modelování závislosti mezi více pojistnými kmeni se stalo jednou z rozhodujících otázek stanovení rezerv. V této souvislosti, kopule představují užitečný nástroj pro vytváření modelů, které jsou schopné zachytit závislostní strukturu líp než modely založené na klasických mírách závislosti. Tato práce se zabývá výkladem modelu kopulové regrese, jeho vlastnostmi a využitím v praxi, přičemž se soustředí na jeho aplikaci v neživotním pojištění. Tenhle přístup spájí modelování marginálií pomoci zobecněných lineárních modelů a zachycen závislostní struktury pomocí kopule. V závěru práce aplikujeme teoretické postupy na reálná data.

Klíčová slova: stochastické rezervování škod, kopula, neživotní pojištění, více pojistných kmenů

Title: Claims reserving with copulae for multiple lines of business

Author: Katarína Valentovičová

Department: Department of Probability and Mathematical Statistics

Supervisor of the master thesis: RNDr. Michal Pešta, Ph.D., Department of Probability and Mathematical Statistics

Abstract: Claims reserving and claims process estimation present classical problems in general insurance. The overall reserves are often determined under the assumption of independence among the lines of business. Though, recently modelling of the dependence among multiple lines of business has become crucial issue of reserving process. In this context, copulae provide a useful tool to construct models which go beyond the classical ones in terms of dependence structure. This thesis deals, in particular, with the copula regression model, its properties and possible applications in general insurance. This approach combines GLM modelling of margins and then expressing the dependence structure using copula. The theoretical methods are illustrated on a real data set.

Keywords: stochastic claims reserving, copula, non-life insurance, multiple lines of business

Contents

Introduction

Claims reserving and claims process estimation present classical problems in general insurance. Several models have been developed in order to determine the claims reserves based on the historical observations of claims payments data. The primary goal of those models is to set an adequate reserve to cover losses that have been incurred but not yet reported and developed. The extensive overview of classical and commonly used methods can be found in monographs of England and Verrall (2002) and Wütrich and Merz (2008). Most of the methods introduced herein deal with claims reserving problem of a single line of business.

In practice, portfolio of almost every major insurance company consists of multiple lines of business. It is natural to assume, that the lines of business are somehow related to each other. In order to set the value of reserves for overall portfolio, it is necessary to capture, understand and properly express this relationship. The dependence structure might appear on several levels, among the losses as they develop in time within one accident year or on the other hand, within one development year among losses in different accident years. Some authors also consider the dependence with respect to the calendar years. The question of how to cope with dependent lines of business, to which an insurance company has to face, is surely of utmost importance.

In the presented thesis, the copula approach is exhibited in order to associate the claims form several run-off triangles. Thus the appearance of dependence structure among the cells of run-off triangles is supposed. Copulae provide a useful tool to construct models which go beyond the classical ones in terms of dependence structure. The focus lies on copula regression model which is based on combination of generalized linear models used to determine the distribution of margins and subsequently association of margins using copula.

The outline of the thesis is as follows. In the first chapter, the basic theory of copulae is introduced. The definition of the copula, basic properties as well as the most important results concerning copulae are described. We introduce examples of elliptical and Archimedean families of copulae which are later used to model the dependence structure of chosen lines of business

The claims reserving notation is summarized in Chapter 2. The two reserving methods are stated as well: the chain-ladder method, which can be considered as industry's benchmark; and generalized linear models in terms of claims reserving.

The focus of the third chapter lies on the construction of copula regression model. The modelling of margins using GLM approach is described and then set into copula framework. Special attention is paid to the estimation of the copula regression model as well as well as to the later review of suitability of the model via goodness-of-fit test.

Finally, the fourth chapter is dedicated to the application of copula regression

model on the real data. The R software is used to model fitting and subsequently expressing the value of reserves of overall considered portfolio. The results of reserving process are presented and compared from statistical and numerical point of view.

Chapter 1

The theory of copulae

Copulae present very useful tool to construct models which go beyond the classical ones in terms of dependence structure. They provide the understanding of dependence at more detailed level. The notion copula is quite recent in statistic, although several results concerning copulae go back to the development of the theory of probabilistic metric spaces. Below, we follow the monograph of Nelsen (2006) and the publication of McNeil et al. (2005) which describes copulae in concept of risk management.

1.1 Definitions and basic properties

In this section we will deal with the notion of n-dimensional real function H and its certain properties which lead to definition of the copula. We will first briefly describe the notation we will use. Through this thesis we denote $\overline{\mathbb{R}}$ the extended real line given by $[-\infty, \infty]$. For a function H, we denote by Dom$H$ and Ran$H$ the domain, respectively the range of the function H. When dealing with copulae, it is very common to refer to function f as increasing (rather than non-decreasing) whenever $x \leq y$ implies that $f(x) \leq f(y)$. We will stay consistent with this notation. Finally, we define the generalized inverse of a function f as $f^{-1}(y) = \inf\{x \in \mathbb{R} \mid f(x) \geq y\}$ for all $y \in \text{Ran}f$, using the convention $\inf \emptyset = \infty$.

Consider n-dimensional real function H. Let the domain of H be given by the cartesian product of sets $S_1, \ldots, S_n$ where each set S_k has a smallest element a_k. The function H is said to be *grounded* if $H(\boldsymbol{t}) = 0$ for all $\boldsymbol{t}$ in Dom$H$ where $t_k = a_k$ for at least one k. If S_k is non-empty for all k and has a greatest element b_k, then *margins* of H can be defined. The one-dimensional margins (called just margins) of H are the functions H_k defined on the set S_k as $H_k(x) = H(b_1, \ldots, b_{k-1}, x, b_{k+1}, \ldots, b_n)$ for all $x \in S_k$. Similarly, higher-dimensional margins are defined.

Definition 1. *An n-dimensional distribution function is a function H with domain $\overline{\mathbb{R}}^n$ such that H is grounded, n-increasing and $H(\infty, \ldots, \infty) = 1$.*

Lemma 1. *Let H be grounded n-increasing function with domain $S_1 \times \cdots \times S_n$, where $S_1, \ldots, S_n$ are non-empty subsets of $\overline{\mathbb{R}}$. Then H is increasing in each argument, i.e., if $(t_1, \ldots, t_{k-1}, x, t_{k+1}, \ldots, t_n)$ and $(t_1, \ldots, t_{k-1}, y, t_{k+1}, \ldots, t_n)$ are*

in Dom$H$ *and* $x \leq y$, *then*

$$H(t_1, \ldots, t_{k-1}, x, t_{k+1}, \ldots, t_n) \leq H(t_1, \ldots, t_{k-1}, y, t_{k+1}, \ldots, t_n).$$

In addition, if H *has margins, then for points* $\boldsymbol{x} = (x_1, \ldots, x_n)$ *and* $\boldsymbol{y} = (y_1, \ldots, y_n)$ *in* Dom$H$

$$|H(\boldsymbol{x}) - H(\boldsymbol{y})| \leq \sum_{k=1}^{n} |H_k(x_k) - H_k(y_k)|.$$

Proof. For the proof, see Schweizer and Sklar (1983). ❑

As the consequence of the Lemma 1, it follows that the margins of n-dimensional distribution function are distribution functions. In the following text, we will denote the margins of H as $F_1, \ldots, F_n$.

Now we settled all the definitions and properties needed to introduce the notion of copula.

Definition 2. *A* n*-dimensional copula is a mapping* $C : [0,1]^n \to [0,1]$ *with the following properties:*

1. $C(u_1, \ldots, u_n)$ *is increasing in each component* $u_i \in [0,1]$, $i \in \{1, \ldots, n\}$;
2. $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$ *for all* $u_i \in [0,1]$ *and* $i \in \{1, \ldots, n\}$;
3. *for all* $(a_1, \ldots, a_n)$ *and* $(b_1, \ldots, b_n) \in [0,1]^n$ *where* $a_i \leq b_i$, *it holds*

$$\sum_{i_1=1}^{2} \cdots \sum_{i_n=1}^{2} (-1)^{i_1 + \cdots + i_n} C(u_{1i_1}, \ldots, u_{ni_n}) \geq 0,$$

where $u_{j1} = a_j$ *and* $u_{j2} = b_j$ *for all* j *in* $\{1, \ldots, n\}$.

Hence a copula C is a distribution function on $[0,1]^n$ with uniformly distributed margins on $[0,1]$. The following theorem is a direct consequence of the Lemma 1.

Theorem 2. *Let* C *be an* n*-copula. Then for every* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *in* $[0,1]^n$

$$|C(\boldsymbol{v}) - C(\boldsymbol{u})| \leq \sum_{k=1}^{n} |v_k - u_k|.$$

Thus C *is uniformly continuous on* $[0,1]^n$.

The important property of copulae is introduced by so-called *Fréchet-Hoeffding bounds* inequality.

Theorem 3. *Let* C *be an* n*-copula, then for every* $\boldsymbol{u}$ *in* $[0,1]^n$

$$W^n(\boldsymbol{u}) \leq C(\boldsymbol{u}) \leq M^n(\boldsymbol{u}),$$

where M^n *and* W^n *are functions on* $[0,1]^n$ *defined as follows*

$$M^n(\boldsymbol{u}) = \min(u_1, \ldots, u_n),$$
$$W^n(\boldsymbol{u}) = \max(u_1 + \cdots + u_n - n + 1, 0).$$

The function M^n is an n-copula for each $n \geq 2$, while W^n is not a copula for any $n > 2$. We refer to M^n as the *Fréchet-Hoeffding upper bound* and W^n as the *Fréchet-Hoeffding lower bound*. The third important copula is so-called *product copula* Π^n defined on $[0,1]^n$ as $\Pi^n(\boldsymbol{u}) = u_1 \cdots u_n$. We will further explore the importance of copulae M^n, W^2, Π^n in the Section 1.4.

1.2 Sklar's theorem

We introduced the definition of a copula in the previous section. From this definition it follows that a copula C is a distribution function on $[0,1]^n$ with uniformly distributed margins. Sklar's Theorem, one of the most important result in the theory of copulae, enlightens the position of copulae in the relationship between multivariate distribution functions and their univariate margins and represents the theoretical base of many applications. This theorem, as well as the word *copula* in a statistical and mathematical sense, was firstly introduced in Sklar (1959).

Theorem 4. *Let H be an n-dimensional distribution function with margins $F_1, \ldots, F_n$. Then there exists an n-copula C such that for all $\boldsymbol{x}$ in $\overline{\mathbb{R}}^n$,*

$$H(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)). \tag{1.1}$$

If $F_1, \ldots, F_n$ are continuous, then C is unique; otherwise C is uniquely determined on $\mathrm{Ran}F_1 \times \cdots \times \mathrm{Ran}F_n$. Conversely, if C is an n-copula and $F_1, \ldots, F_n$ are univariate distribution functions, then the function H defined in (1.1) is an n-dimensional distribution functions with margins $F_1, \ldots, F_n$.

Proof. The proof of the theorem for the bivariate copula can be seen in Nelsen (2006). The complete proof is stated in Schweizer and Sklar (1983). The important part of the proof is to show so-called *extension lemma*, which demonstrates that every n-subcopula (see Nelsen (2006), Definition 2.10.5.) can be extended to an n-copula. For this proof, see Sklar (1996).

❑

Corollary. Let H be an n-dimensional distribution function with continuous margins $F_1, \ldots, F_n$ and copula C satisfying (1.1). Then for any $\boldsymbol{u} \in [0,1]^n$

$$C(u_1, \ldots, u_n) = H\left(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)\right). \tag{1.2}$$

1.3 Copulae and random variables

In this section we will deal with copulae in the relationship with random variables. The results concerning dependency structure of random variables will be presented in terms of the theory of copulae.

Definition 3. *If the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ has joint distribution function H with continuous marginal distribution functions $F_1, \ldots, F_n$, then the copula of H (or $\boldsymbol{X}$) is the distribution function C of $(F_1(X_1), \ldots, F_n(X_n))$.*

The copula C of a random vector $\boldsymbol{X}$ with continuous margins is uniquely determined as it is stated in the Theorem 4. The standard representation of the distribution function of a random vector is then

$$H(x_1, \ldots, x_n) = \mathsf{P}\{X_1 \leq x_1, \ldots, X_n \leq x_n\} = C(F_1(x_1), \ldots, F_n(x_n)).$$

One of the most useful properties of copulae is that copulae of random vectors are invariant in terms of strictly increasing transformations of random vectors. When considering strictly monotonous transformations, copulae are even invariant or change in a certain way depending on the monotony.

Theorem 5. *Let $(X_1, \ldots, X_n)^T$ be a random vector with continuous margins and copula C. If $T_1, \ldots, T_n$ are strictly increasing function on $\mathrm{Ran}X_1, \ldots, \mathrm{Ran}X_n$, then $(T_1(X_1), \ldots, T_n(X_n))^T$ has also copula C.*

Proof. For the proof, see McNeil et al. (2005). ❑

As it follows from the Definition 2, copula C presents a distribution function on $[0,1]^n$ with uniformly distributed margins on $[0,1]$. Moreover, there are several situations, such as random variate generation from copula or estimation of parameters, in which the *copula density* comes into focus. If it holds that

$$C(u_1, \ldots, u_n) = \int_0^{u_1} \cdots \int_0^{u_n} \frac{\partial^n}{\partial s_1 \ldots \partial s_n} C(s_1, \ldots, s_n)\, \mathrm{d}s_1 \ldots \mathrm{d}s_n,$$

then C is said to be absolutely continuous and C admits a density $c(u_1, \ldots, u_n) = \frac{\partial^n}{\partial u_1 \ldots \partial u_n} C(u_1, \ldots, u_n)$.

1.4 Dependence structure

In probability and statistics, dependence structure belongs to one of the most studied aspects and properties of random variables. Copulae represent very powerful tool which allows us to isolate in a certain sense the dependence relations among the components of a random vector. In this section we will explore how can copulae be used in order to study, express and measure dependence. For the detailed overview of indices of dependence, we refer to Schweizer and Wolff (1981), Wolff (1980) and Lancaster (1982).

First we will develop the meaning of Fréchet-Hoeffding bounds M^n and W^n and product copula Π^n in terms of dependence structure of random variables. Since W^n is copula only for $n = 2$, in this case we will deal only with dependence relationship between two random variables.

Definition 4. *The random variables $X_1, \ldots, X_n$ with copula C are said to be comonotonic if C is Fréchet-Hoeffding upper bound, i.e. $C = M^n$.*
The pair of random variables X_1 and X_2 with copula C is countermonotonic if they have as copula the Fréchet-Hoeffding lower bound, i.e. $C = W^2$.

The properties of comonotonicity and countermonotonicity express perfect dependence among random variables as it will be shown in following proposition.

Proposition 6. *Random variables $X_1, \ldots, X_n$ are comonotonic if and only if*

$$(X_1, \ldots, X_n) \stackrel{d}{=} (T_1(Z), \ldots, T_n(Z))$$

for some random variable Z and increasing functions $T_1, \ldots, T_n$.
Random variables X_1 and X_2 are countermonotonic if and only if

$$(X_1, X_2) \stackrel{d}{=} (T_1(Z), T_2(Z))$$

for some random variable Z with T_1 increasing and T_2 decreasing, or vice versa.

Proof. For the proof, see McNeil et al. (2005). ❑

Corollary. Let $X_1, \ldots, X_n$ be random variables with continuous distribution functions. They are comonotonic if and only if for every pair (i, j) we have $X_j = T_{ji}(X_i)$ almost surely for some increasing transformation T_{ji}.

Proof. For the proof, see McNeil et al. (2005). ❑

Random variables $X_1, \ldots, X_n$ with distribution functions $F_1, \ldots, F_n$ are independent if and only if their joint distribution function H satisfies $H(x_1, \ldots, x_n) = F_1(x_1) \cdots F_n(x_n)$ for all $x_1, \ldots, x_n \in \overline{\mathbb{R}}^2$. Hence the following result is direct consequence of Theorem 4.

Theorem 7. *Let $(X_1, \ldots, X_n)^T$ be a vector of continuous random variables with copula C, then $X_1, \ldots, X_n$ are independent if and only if $C = \Pi^n$.*

With respect to the previous theorem, Π^n is called an independent copula. Moreover, based on the relationships between dependence structure of random variables and their copulae expressed in Definition 4, M^n is referred to as a copula of comonotonicity and W^2 is called a copula of countermonotonicity.

1.4.1 Kendall's tau and Spearman's rho

Kendall's tau and Spearman's rho are very important measures of dependence between two random variables, which present an alternative to the to the linear correlation in cases when linear correlation can lead to inappropriate results since it is not a copula-based measure of dependence.

Before stating the definition of Kendall's tau and Spearman's rho, it is important to introduce the notion of *concordance.* Let $(X, Y)^T$ be random vector of continuous random variables and let $(x, y)^T$ and $(\widetilde{x}, \widetilde{y})^T$ be two observations of this vector. Then we say that $(x, y)^T$ and $(\widetilde{x}, \widetilde{y})^T$ are *concordant* if $(x - y)(\widetilde{x} - \widetilde{y}) > 0$; and *discordant* if $(x - y)(\widetilde{x} - \widetilde{y}) < 0$.

Theorem 8. *Let $(X, Y)^T$ and $(\widetilde{X}, \widetilde{Y})^T$ be independent vectors of continuous random variables with joint distribution functions H and $\widetilde{H}$, with margins F of X and*

$\widetilde{X}$ *and* G *of* Y *and* $\widetilde{Y}$*. Further, let* C *and* $\widetilde{C}$ *be copulae of* $(X,Y)^T$ *and* $(\widetilde{X},\widetilde{Y})^T$*, respectively, i.e.* $H(x,y) = C(F(x),G(y))$ *and* $\widetilde{H}(x,y) = \widetilde{C}(F(x),G(y))$*. Let* Q *denote the difference between probability of concordance and discordance of* $H(x,y) = C(F(x),G(y))$*, i.e.*

$$Q = P\left\{\left(X-\widetilde{X}\right)\left(Y-\widetilde{Y}\right) > 0\right\} - P\left\{\left(X-\widetilde{X}\right)\left(Y-\widetilde{Y}\right) < 0\right\}.$$

Then, it holds

$$Q = Q\left(C,\widetilde{C}\right) = 4\iint_{[0,1]^2} \widetilde{C}\,(u,v)\,\mathrm{d}C\,(u,v) - 1.$$

Proof. For the proof, see Nelsen (2006). ❑

Definition 5. *Let* $(X,Y)^T$ *and* $(\widetilde{X},\widetilde{Y})^T$ *be independent and identically distributed random vectors with joint distribution functions* H*. Then the Kendall's tau, denoted as* τ *(*$\tau_{X,Y}$ *respectively), is defined as the probability of concordance minus the probability of discordance:*

$$\tau = \tau_{X,Y} = P\left\{\left(X-\widetilde{X}\right)\left(Y-\widetilde{Y}\right) > 0\right\} - P\left\{\left(X-\widetilde{X}\right)\left(Y-\widetilde{Y}\right) < 0\right\}.$$

Following result is direct consequence of Definition 5 and Theorem 8.

Theorem 9. *Kendall's tau of continuous random variables* X *and* Y *with copula* C *is given by*

$$\tau_{X,Y} = \tau_C = Q(C,C) = 4\iint_{[0,1]^2} C\,(u,v)\,\mathrm{d}C\,(u,v) - 1.$$

Another important measure of dependence among random variables based on concordance is Spearman's rho.

Definition 6. *Let* $(X_1,Y_1)^T, (X_2,Y_2)^T$ *and* $(X_3,Y_3)^T$ *be independent and identically distributed random vectors with joint distribution functions* H*. Then the Spearman's rho, denoted as* $\rho_{X,Y}$ *, is defined as follows:*

$$\rho_{X,Y} = 3\left(P[(X_1-X_2)(Y_1-Y_3) > 0] - P[(X_1-X_2)(Y_1-Y_3) < 0]\right).$$

Theorem 10. *Spearman's rho of continuous random variables* X *and* Y *with copula* C *is given by*

$$\begin{aligned}\rho_{X,Y} = \rho_C =& 3Q(C,\Pi)\\ =& 12\iint_{[0,1]^2} uv\mathrm{d}C(u,v) - 3\\ =& 12\iint_{[0,1]^2} C(u,v)\mathrm{d}u\mathrm{v} - 3.\end{aligned} \tag{1.3}$$

Proof. The formula (1.3) follows from Theorem (8). For the detailed proof, see Nelsen (2006). ❑

1.4.2 Tail dependence

Tail dependence is a concept that describes the limiting proportion that the value of one random variable exceeds a certain threshold conditioned that the other random variable has already exceeded given threshold. As it will be shown in this section, tail dependence is a copula property, which means that tail dependence of random variables X and Y is invariant under strictly increasing transformations X and Y. Other results and properties related to tail dependence are listed in Juri and Wütrich (2003) and Joe et al. (2010).

Definition 7. *Let X and Y be continuous random variables with distribution functions F and G. The upper tail dependence parameter λ_U is defined as follows:*

$$\lambda_U = lim_{t\to 1^-}\, \mathsf{P}\left[Y > G^{-1}(t) \mid X > F^{-1}(t)\right],$$

if the limit exists. Similarly, the lower tail dependence parameter λ_L is the limit (if it exists)

$$\lambda_L = lim_{t\to 0^+}\, \mathsf{P}\left[Y \leq G^{-1}(t) \mid X \leq F^{-1}(t)\right].$$

If $\lambda_U = 0$ $(\lambda_L = 0)$, we say that random variables X and Y are asymptotically independent in upper (lower) tail.

The representation of tail dependence parameters stated in the following theorem demonstrates the copula property of this measure.

Theorem 11. *Let X and Y be continuous random variables, whose copula is C, with distribution functions F and G and let λ_U and λ_L be the parameters of tail dependence given in Definition 7. If the limits λ_U and λ_L exist, then*

$$\lambda_U = 2 - lim_{t\to 1^-} \frac{1 - C(t,t)}{1-t}$$

and

$$\lambda_L = lim_{t\to 0^+} \frac{C(t,t)}{t}.$$

Proof. For the proof, see Nelsen (2006). ❑

Further it can be shown that the parameter of upper tail dependence can be expressed as a limit

$$\lambda_U = lim_{t\to 0^+} \frac{\widehat{C}(t,t)}{t},$$

where $\widehat{C}$ denotes the survival copula of $(X_1, \ldots, X_n)^T$, i.e.

$$\overline{H}(x_1, \ldots, x_n) = \mathsf{P}\left\{X_1 > x_1, \ldots, X_n > x_n\right\} = \widehat{C}\left(\overline{F}_1(x_1), \ldots, \overline{F}_n(x_n)\right).$$

Note that the copula C and survival copula $\widehat{C}$ of random vector $\boldsymbol{X}$ are equalled for each $\boldsymbol{X}$ that is *radially symmetric*. Random vector $\boldsymbol{X}$ is radially symmetric about $\boldsymbol{a}$ if $\boldsymbol{X} - \boldsymbol{a} \overset{d}{=} \boldsymbol{a} - \boldsymbol{X}$. Hence for the copula C of radially symmetric random vector it holds that $\lambda_U = \lambda_L$.

1.5 Copula faimilies

Copulae can be divided into several classes based on their common characteristics and properties. The most widely-known copula families are probably *elliptical* and *Archimedean* copulae. In the following section, we deal with these two classes, their definitions and basic properties are introduced.

1.5.1 Elliptical copulae

Elliptical copulae present the group of copulae of *elliptical distributions*. The class of elliptical distributions consists of wide range of multivariate distributions which share some important properties with multivariate normal distribution. We will limit only on the basic definition in order to define and work with elliptical class of copulae. For further details on elliptical distributions, see Cambanis et al. (1981).

Definition 8. *The n-dimensional random vector* $\boldsymbol{X}$ *is said to have an elliptical distribution with parameters* $\boldsymbol{\mu}, \Sigma$ *and* ϕ*, if for some* $\boldsymbol{\mu} \in \mathbb{R}^n$ *and some* $n \times n$ *non-negative definite matrix* Σ*, the characteristic function* $\varphi_{\boldsymbol{X}-\boldsymbol{\mu}}(\boldsymbol{t})$ *of* $\boldsymbol{X} - \boldsymbol{\mu}$ *is a function of the quadratic form* $\boldsymbol{t}^T\Sigma\boldsymbol{t}$*, i.e.* $\varphi_{\boldsymbol{X}-\boldsymbol{\mu}}(\boldsymbol{t}) = \phi\left(\boldsymbol{t}^T\Sigma\boldsymbol{t}\right)$*. We write that* $\boldsymbol{X} \sim E_n(\boldsymbol{\mu}, \Sigma, \phi)$*.*

The function ϕ from the Definition 8 is referred to as a *characteristic generator*. When $\phi(u) = \exp(-u/2), E_n(\boldsymbol{\mu}, \Sigma, \phi)$ is the multivariate normal distribution $N_n(\boldsymbol{\mu}, \Sigma)$; and for $n = 1$ the class of elliptical distributions correspond to the class of one-dimensional symmetric distributions.

Gaussian copula

One of the most important representative of the elliptical class of copulae is *Gaussian copula*. The Gaussian copula is a copula of the n-variate normal distribution with linear correlation matrix R with the following form:

$$C_R^{Ga}(\boldsymbol{u}) = \Phi_R^n\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)\right),$$

where Φ^{-1} denotes the inverse of the univariate standard normal distribution and Φ_R^n stands for the joint distribution function of the n-variate standard normal distribution with linear correlation matrix R. From the properties of multivariate normal distribution, the bivariate Gaussian copula can be expressed as follows:

$$C_R^{Ga}(u, v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left\{-\frac{s^2 - 2\rho st + t^2}{2(1-\rho^2)}\right\} \mathrm{d}s\mathrm{d}t,$$

where ρ is the coefficient of linear correlation of the corresponding normal distributions.

The copula of comonotonicity and the independent copula are special cases of Gaussian copula. If the matrix of linear correlations $R = \boldsymbol{I}_{n\times n}$, then it can be shown that $C_R^{Ga} = \Pi^n$. In the case that $R = \boldsymbol{J}_{n\times n}$, we get $C_R^{Ga} = M^n$. Note that $I_{n\times n}$ and $J_{n\times n}$ are matrices of dimension $n \times n$, where $I_{n\times n}$ has ones on its diagonal and $J_{n\times n}$ denotes the matrix whose each element is equalled to one.

One of the properties of random vectors with elliptical distribution is radial symmetry, thus the Gaussian copula equals to its corresponding survival copula. Due to this equality, the parameters of lower and upper tail dependence of random variables X and Y with bivariate Gaussian copula are equalled as well. As it is shown in McNeil et al. (2005), this parameter has zero value, hence the Gaussian copula is asymptotically independent in both tails.

1.5.2 Archimedean copulae

In this section we discuss an important class of copulae called *Archimedean copulae*. They have several advantages which make them very useful in practice. We have seen that elliptical copulae are restricted to have radial symmetry. In case of Archimedean family, the copulae are not necessarily radially symmetric, so they allow to better approximate asymmetry in dependence structure. Unlike the elliptical copulae many of these copulae have closed form expressions. In general, Archimedean copulae allow for a big variety of different dependence structures. For further results concerning Archimedean copulae, we refer to Ballerini (1994) and Müller and Scarsini (2005).

Let Φ denote the set of functions $\varphi : [0, 1,] \to [0, \infty]$ which are continuous, strictly decreasing, convex such that $\varphi(0) = \infty$ and $\varphi(1) = 0$. Each $\varphi \in \Phi$ has an inverse $\varphi^{-1} : [0, \infty] \to [0, 1,]$ with the same properties, except $\varphi^{-1}(0) = 1$ and $\varphi^{-1}(\infty) = 0$. Each member of Φ generates a copula C given by

$$C(u, v) = \varphi^{-1}\left(\varphi(u) + \varphi(v)\right). \tag{1.4}$$

Copulae of the above form are called Archimedean copulae and the function φ is referred to as a *generator* of C.

Note that, in the definition of Φ, it is not necessary for $\varphi(0)$ to be infinite in order to generate a copula. When $\varphi(0)$ is finite, the Archimedean copula is given by

$$C(u, v) = \varphi^{[-1]}\left(\varphi(u) + \varphi(v)\right),$$

where $\varphi^{[-1]}$ denotes so-called *pseudo-inverse* function given by

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t), & 0 \le t \le \varphi(0), \\ 0, & \varphi(0) \le t \le \infty. \end{cases}$$

If $\varphi(0) = \infty$, the generator function φ is said to be *strict* and the associated Archimedean copula is called a *strict copula.*

One of the useful properties of Archimedean copulae is that Kendall's tau of random variables X and Y with an Archimedean copula C generated by function φ is given by

$$\tau_C = 1 + 4 \int_0^1 \frac{\varphi(t)}{\varphi'(t)} \mathrm{d}t. \tag{1.5}$$

Further we will discuss the three members of Archimedean family, namely *Gumbel, Clayton* and *Frank* copula.

Gumbel copula

The generator function of Gumbel copula is given by

$$\varphi(t) = (-\ln t)^{\theta}, \ \theta \geq 1.$$

Then from (1.4) we get

$$C_{\theta}(u,v) = \exp\left\{-\left[(-\ln u)^{\theta} + (-\ln v)^{\theta}\right]^{1/\theta}\right\}.$$

Furthermore, it holds that $C_{\theta} = \Pi^2$ for $\theta = 1$; and $\lim_{\theta\to\infty} C_{\theta} = M^2$, hence the independent copula and the copula of comonotonicity are the special cases of Gumbel copula.

Using the relationship (1.5) it can be shown that the Kendall's tau of this copula is $1 - 1/\theta$. In the case of Gumbel copula, the marginal distributions are asymptotically independent in lower tail, i.e. $\lambda_L = 0$, but have upper tail dependence expressed by $\lambda_U = 2 - 2^{1/\theta}$.

Clayton copula

In the contrary with the Gumbel copula, Clayton copula is specific in capturing the dependence of the random variables on lower tails. The Clayton copula is defined as follows

$$C_{\theta}(u,v) = \max\left(\left[u^{-\theta} + v^{-\theta} - 1\right]^{-1/\theta}, 0\right),$$

where the corresponded generator function is $\varphi(t) = \left(t^{-\theta} - 1\right)/\theta$ for the parameter $\theta \in [-1,\infty) \setminus \{0\}$. For $\theta > 0$ the copulae are strict and the expression simplifies to

$$C_{\theta}(u,v) = \left[u^{-\theta} + v^{-\theta} - 1\right]^{-1/\theta}. \tag{1.6}$$

It the literature, the Clayton copula is commonly defined by (1.6) because several limit and other properties are derived from this expression. Similarly as for Gumbel copula, the independent copula and the copula of comonotonicity are the special cases of Clayton copula, in addition the copula of countermonotonicity belongs also to Clayton family. The associated equalities are as follows: $C_{-1} = W^2, \lim_{\theta\to\infty} C_{\theta} = M^2$ and $\lim_{\theta\to 0} C_{\theta} = \Pi^2$.

Kendall's tau of Clayton copula is $\theta/(\theta+2)$. The parameter of upper tail dependence is equalled to zero, so the random variables are asymptotically independent in upper tail. Contrariwise, the parameter of lower tail dependence is

$$\lambda_L = \begin{cases} 2^{-1/\theta}, & \theta > 0, \\ 0, & \theta \leq 0. \end{cases}$$

Frank copula

The Frank copulae are strict Archimedean copulae given by

$$C_{\theta}(u,v) = -\frac{1}{\theta}\ln\left(1 + \frac{(\mathrm{e}^{-\theta u} - 1)(\mathrm{e}^{-\theta v} - 1)}{\mathrm{e}^{-\theta} - 1}\right),$$

with the generator function in the following form

$$\varphi(t) = -\ln\frac{\mathrm{e}^{-\theta t}-1}{\mathrm{e}^{-\theta}-1},\ \theta \in \mathbb{R}\setminus\{0\}\,.$$

The special limit cases of Frank copulae are given as follows: $\lim_{\theta\to-\infty} C_\theta = W^2$, $\lim_{\theta\to\infty} C_\theta = M^2$ and $\lim_{\theta\to 0} C_\theta = \Pi^2$. Furthermore, members of Frank family are the only Archimedean copulae which satisfy the equation for radial symmetry.

The Frank copula has neither lower nor upper tail dependence, i.e. $\lambda_L = \lambda_U = 0$. Thus it is suitable for modelling the data characterized by weak tail dependence.

Chapter 2

Reserving theory

Claims reserving is one of the most important and crucial issues in insurance. In this thesis we deal with reserving for the group of products called *non-life insurance* (in UK known as *General Insurance* and in USA refereed to as *Property and Casualty Insurance*). The non-life insurance is a general title of all types of insurance products not covered by life insurance. As it is stated in Wütrich and Merz (2008) the non-life insurance considers the following lines of business: *motor/car insurance, property insurance, liability insurance, accident insurance, health insurance, marine insurance* and *other insurance.* Each of these lines of business can be further divided into several subclasses.

Insurance industry distinguishes between two main claim groups. The first one, known as RBNS (Reported but not settled), refers to the claims that have occurred and have been already reported to the insurer. The other group, IBNR (Incurred but not reported), is formed by the group of claims events that have occurred but have not been reported yet. Both reserves include an estimation of the costs incurred during the settlement of claims. In the case of IBNR reserve, the insurer knows neither the number of claims events that have occurred, nor the severity of each claim, hence, statistical methods based on historical development of claims need to be applied to estimate the amount of money to be paid for IBNR claims. The quality of this estimate depends on the chosen reserving method and its overall complexity.

In this chapter we will briefly introduce reserving procedure and mainly we will impose the notation used in order to describe specific reserving methods with copulae.

2.1 Claims reserving notation

In this section, we introduce classical notation used within the mathematical framework for claims reserving. To model a portfolio of insurance policies by *incremental payments* for claims, we consider a family of random variables $\{X_{i,j}\}$ for $i \in \{0, \ldots, I\}$ and $j \in \{0, \ldots, J\}$. The random variable $X_{i,j}$ is interpreted as claim amounts of *accident year* i which is settled with a delay of j years, thus in *development year* j. The last year of occurrence of claim event is I and the maximal value of development years is denoted by J. We assume that the observations of $X_{i,j}$ are available for all $i + j \leq I$ and that for $i + j > I$ the claim amounts $X_{i,j}$ are not observable. Typically, outstanding loss liabilities are pre-

Accident year	Development year						
	0	1	...	...	...	$J-1$	J
0	$X_{0,0}$	$X_{0,1}$	...	...	...	$X_{0,J-1}$	$X_{0,J}$
1	$X_{1,0}$	$X_{1,1}$	...	...	...	$X_{1,J-1}$	$X_{1,J}$
$\vdots$	$\vdots$	$\vdots$	Observations of r.v. $X_{i,j}$ $(i+j \leq I)$			$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$					
$\vdots$	$\vdots$	$\vdots$					
$\vdots$	$\vdots$	$\vdots$					
$I-1$	$X_{I-1,0}$	$X_{I-1,1}$					
I	$X_{I,0}$						

Table 2.1: Run-off triangle for incremental claim amounts $X_{i,j}$.

sented in so-called *run-off triangles* (see Table 2.1) through which the separation of insurance claims in two time axes can be captured. When dealing with run-off triangles of multiple lines of business, it is usual to work with standardized data which are expressed by the ratio $X_{i,j}/\omega_i$, where ω_i represents the exposure variable of the particular line of business in the i^{th} accident year measuring the volume of the business. The exposure variable can be the number of policies, the number of open claims or the earned premium.

The other approach to model a portfolio is by using *cumulative payments* given by $Y_{i,j} = \sum_{k=0}^{j} X_{i,k}$. We interpret $Y_{i,j}$ as the claim amounts up to development year j with accident year i. Hence, $Y_{i,j}$ is a random variable whose observations are available if $i+j \leq I$. The cumulative claim amounts can be represented by the (cumulative) run-off triangle as well.

The main goal is to predict the *ultimate claim amounts* $Y_{i,J}$ and subsequently the *outstanding claim reserve* for accident year i given by

$$R_i = Y_{i,J} - Y_{i,I-i} = \sum_{k=I-i+1}^{J} X_{i,k}$$

for all $i = 1, \ldots, I$. Moreover, usually it is important to estimate the whole distribution of the reserves in order to derive other distributional quantities.

For the purposes of this thesis, we will further suppose that the last year of occurrence of claim event I and the maximum development years observed J are equalled (i.e. $I = J$) and that $X_{i,j} = 0$ for all $j > J$. This assumption means that all claims will be closed in J years.

2.2 Basic claims reserving methods and reserving process overview

In this section we will briefly introduce two methods of claims reserving: the *chain-ladder* method which is often understood as straightforward algorithmic method working mainly with cumulative data; and the *generalized linear model*

(GLM) approach to model the incremental claims. Since the large number of various methods has been invented, we refer to Wütrich and Merz (2008) and England and Verrall (2002) for the complex overview of the claims reserving methods.

2.2.1 Chain-ladder method

The chain-ladder (CL) method is probably one of the most popular claims reserving method. This method provides a computational algorithm to set claims reserves and is widely used in practice due to its simplicity. However, the lack of underlying stochastic model on which the algorithm can be based results into uncertainty of such prediction, since it cannot be determined. Later, several appropriate stochastic models that justify the CL method were derived.

The distribution-free CL method, developed and described in Mack (1993), operates with cumulative data and derives so-called *development factors* (also known as *link ratios, CL factors* or *age-to-age factors*). It is based on the following assumptions:

- Cumulative claims $Y_{i,j}$ are independent random variables for different accident years i.
- There exist development factors $f_0, \dots, f_{J-1} > 0$ such that we have

$$\mathsf{E}\left[Y_{i,j} \mid Y_{i,0}, \dots, Y_{i,j-1}\right] = \mathsf{E}\left[Y_{i,j} \mid Y_{i,j-1}\right] = f_{j-1} Y_{i,j-1}$$

  for all $0 \leq i \leq I$ and $1 \leq j \leq J$.

Under the model assumptions it can be shown (for the proof, see Wütrich and Merz (2008)) that the following holds for conditionally expected claims:

$$\mathsf{E}\left[Y_{i,J} \mid \mathcal{D}_I\right] = \mathsf{E}\left[Y_{i,J} \mid Y_{i,I-i}\right] = Y_{i,I-i} f_{I-i} \cdots f_{J-1}, \tag{2.1}$$

where $\mathcal{D}_I = \{Y_{i,j}; i + j \leq I, 0 \leq j \leq J\}$ stands for the set of observations available at time I.

The relationship stated by (2.1) provides an algorithm to estimate the ultimate claim $Y_{i,J}$ based on the observations $\mathcal{D}_I$. If the development factors f_j are known, then the outstanding claims reserve is given by

$$\mathsf{E}\left[Y_{i,J} \mid \mathcal{D}_I\right] - Y_{i,I-i} = Y_{i,I-i}\left(f_{I-i} \cdots f_{J-1} - 1\right). \tag{2.2}$$

In practice, the development factors are often unknown, hence they need to be estimated. The estimation of f_j for all $j \in \{0, \dots, J-1\}$ is expressed as follows:

$$\widehat{f}_j = \frac{\sum_{i=0}^{I-j-1} Y_{i,j+1}}{\sum_{i=0}^{I-j-1} Y_{i,j}}. \tag{2.3}$$

Allying the formulas (2.1) and (2.3), the CL estimate of the ultimate claim $Y_{i,J}$ is then given by (note that we assume $I = J$)

$$\widehat{Y_{i,J}}^{CL} = \widehat{\mathsf{E}}\left[Y_{i,J} \mid \mathcal{D}_I\right] = Y_{i,I-i} \widehat{f}_{I-i} \cdots \widehat{f}_{I-1},$$

hence the estimate of claims reserve for the accident year i is

$$\widehat{R_i}^{CL} = \widehat{\mathsf{E}}\left[Y_{i,J} \mid \mathcal{D}_I\right] - Y_{i,I-i} = Y_{i,I-i}\left(\widehat{f}_{I-i} \cdots \widehat{f}_{I-1} - 1\right).$$

Under the assumption of CL method, several useful properties of derived estimates can be shown. One of the most important results is that $\widehat{Y_{i,J}}^{CL}$ is unbiased estimator for $\mathsf{E}\left[Y_{i,J} \mid \mathcal{D}_I\right]$ and in addition $\widehat{Y_{i,J}}^{CL}$ is (unconditionally) unbiased estimate of $\mathsf{E}\left[Y_{i,J}\right]$. Other properties of estimates derived using CL method along with their proves are stated in Wütrich and Merz (2008).

2.2.2 Generalized linear models

Generalized linear models (GLM) present the class of statistical models that is a natural generalization of classical linear models. When comparing GLM approach to linear models, one of the advantages consists in possibility of using other distributions of response variable than the normal. The response variable is permitted to have a distribution belonging to the *exponential family* of distributions.

A distribution is said to be of the exponential type, if its density function can be expressed as

$$f\left(y, \theta, \phi\right) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)}\right\} + \kappa(y, \phi),$$

where a, b and κ are real-valued functions, the parameter θ is called a *canonical parameter* and $\phi > 0$ is referred to as a *dispersion parameter*. The function b is assumed to be twice-differentiable and plays an important role in GLM, because it describes the relationship between the mean and the variance of given distribution.

Many of the widely-used distributions, such as normal, gamma, but also distributions of discrete random variables as Poisson or binomial, are members of exponential family. The specific forms of functions a, b and κ and other properties of several distributions are listed in Olsson (2002).

With this parametrization, it can be shown that the mean and the variance of the random variable Y with distribution from the exponential family are in the following form:

$$\mathsf{E}\left(Y\right) = \mu = b'\left(\theta\right) \tag{2.4}$$

and

$$\mathrm{var}\left(Y\right) = b''\left(\theta\right) a(\phi).$$

The expression $b''\left(\theta\right)$ is called a *variance function*. Since the dpendence on the canonical parameter implies through (2.4) the dependence on the mean μ, the variance function is typically expressed as a function of μ and denoted by $V\left(\mu\right)$.

GLM approach operates with the *response variable* (also called as *dependent variable*) $\boldsymbol{Y}$, of which we have a vector of observations

$$\boldsymbol{y} = \left(y_1, \ldots, y_n\right)^T.$$

The other way, in which GLM present a generalization of classical linear models, is that instead of modelling the mean of the response variable $\mathsf{E}\left(\boldsymbol{Y}\right) = \boldsymbol{\mu} = \left(\mu_1, \ldots, \mu_n\right)^T$ directly as a function of so-called *linear predictor* $\boldsymbol{X\beta}$, the function $g\left(\boldsymbol{\mu}\right)$ of $\boldsymbol{\mu}$ is modelled. Hence, the model becomes

$$g\left(\boldsymbol{\mu}\right) = \boldsymbol{\eta} = \boldsymbol{X\beta},$$

where $\boldsymbol{X}$ is a known matrix of dimension $n \times p$ referred to as a *design matrix*, $\boldsymbol{\beta}$ is a p-dimensional vector of parameters to be estimated and g is called a *link function*. The link function must be monotone and differentiable. The choice of the appropriate link function depends on the modelled data. Some of commonly used link functions, such as *logit* link, *probit* link or *complementary log-log* link, are stated in Olsson (2002). The important class of link functions, presented in McCullagh and Nelder (1989), are functions which transform the mean to a canonical parameter, i.e. $g(\boldsymbol{\mu}) = \boldsymbol{\theta}$. These functions are referred to as *canonical links*.

Once the appropriate distribution belonging to exponential family and link function are specified, the parameters of the model can be estimated. In GLM theory, the estimation is often provided via maximum likelihood. The unknown parameters are represented by the vector $\boldsymbol{\beta}$, which is in fact the function of θ. Thus, the log likelihood function is differentiated with respect to the elements of $\boldsymbol{\beta}$ and subsequently is set to equal to zero. This procedure leads to the system of likelihood equations which is often solved using numerical methods, for example using *Newton-Raphson* algorithm.

When the maximum likelihood estimators of the model are obtained, the fit of the model to data can be explored. There are several method of assessing the fit of the model. One of these methods is based on comparison of nested models using *deviance*. An alternative to the deviance is comparing models through *Pearson* χ^2 *statistic* or using the *Akaike's information criterion* which penalized the likelihood functions in order to get simpler model. For further details on the goodness-of-fit tests, see Olsson (2002).

The suitability of the chosen model can be also explored by analysing of residuals. Various types of residuals considered within the GLM concept are listed in McCullagh and Nelder (1989). For the purposes of this thesis, we will mention only one example and that are so-called *Pearson residuals* defined as

$$r_i^{(P)} = \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}},$$

for $i = 1, \ldots, n$, where $\widehat{\mu}_i$ is estimate of the mean μ_i.

Chapter 3

Claims reserving with copulae

In the Chapter 2, we described two methods used to determine claims reserves. In these methods, we assume that the observed data comes from a single line of business. In practice, almost every major insurance company has typically several lines of business in its portfolio. It is natural to suppose, that these lines of business are related to each other in some way, from which the problem of so-called *additivity* arises. Hence, there is a question whether an insurance company should consider the claims reserve from the sum of the lines of business or should it take into account the sum of claims reserves, each determined from a single line of business.

As we mentioned, there is an extensive actuarial literature dealing with claims reserving methods for the single line of business (run-off triangle respectively). In this chapter, we will deal with capturing the dependence structure among several lines of business using copula approach and we describe how this dependence structure is treated in order to set the claims reserves.

3.1 Modelling of claim amounts

The approach we will describe operates with incremental data introduced in the Chapter 2. We will assume that a portfolio of insurance company consists of N lines of business and for each line of business incremental payments forming associated run-off triangles are available. Under this assumptions we will denote the incremental payments by $X_{i,j}^{(n)},\ i \in \{0,\dots\ I\}$ and $j \in \{0,\dots\ J\}$, where n indicates payments related to nth portfolio (run-off triangle), $n \in \{1,\dots\ N\}$. Recall we suppose that $I = J$ and that all claims will be closed in J years. For the purposes of this model, it is more convenient to work with standardized data. Thus, we assume that $X_{i,j}^{(n)}$ are standardized incremental payments and we can write $X_{i,j}^{(n)} = \widetilde{X}_{i,j}^{(n)}/\omega_i^{(n)}$, where $\widetilde{X}_{i,j}^{(n)}$ are original observed incremental payments and $\omega_i^{(n)}$ denotes the exposure variable for ith accident year and nth line of business. With this notation, the claims reserves for accident year i at time I are given by

$$\sum_{n=1}^{N}\sum_{k=I-i+1}^{J} X_{i,k}^{(n)}\omega_i^{(n)} = \sum_{n=1}^{N}\sum_{k=I-i+1}^{J} \widetilde{X}_{i,k}^{(n)},$$

for all $i = 1,\dots,I$.

Our intention is to estimate the ultimate claim amounts and in the same time capture the dependencies among the N run-off triangles. We will assume the balanced model in terms of number of lines of business, which means that each portfolio is of the same size ($I = J$ has the same value for each n) and there are not missing values in run-off triangles.

The sample sizes of run-off triangles are typically small. Due to this fact, the model is limited on considering only parametric distributional families. We assume that $X_{i,j}^{(n)}$ comes from a parametric distribution $F_{i,j}^{(n)}$:

$$F_{i,j}^{(n)} = \mathsf{P}\left\{X_{i,j}^{(n)} \leq x_{i,j}^{(n)}\right\} = F^{(n)}\left(\eta_{i,j}^{(n)}; \boldsymbol{\gamma}^{(n)}\right),\ n = 1, \ldots, N, \tag{3.1}$$

where the vector $\boldsymbol{\gamma}^{(n)}$ of additional parameters represents the scale and shape of the distribution. The location is expressed by the term $\eta_{i,j}^{(n)}$ which is supposed to capture the systematic component of the distribution.

A regression model will be used in order to model the systematic component. Hence the systematic component is a linear function of explanatory variables (covariates) denoted by $\boldsymbol{\beta}^{(n)}$. We will assume two types of independent explanatory variables, namely accident year and development period. Suppose that $\alpha_i^{(n)},\ i \in \{0, \ldots, I\}$ and $\tau_j^{(n)},\ j \in \{0, \ldots, J\}$ represent the accident year effect and development lag effect respectively. With this notation, the systematic component for the nth run-off triangle can be written as:

$$\eta_{i,j}^{(n)} = \zeta^{(n)} + \alpha_i^{(n)} + \tau_j^{(n)},\ n \in \{1, \ldots, N\},$$

where $\zeta^{(n)}$ is intercept and constraints $\alpha_0^{(n)} = \tau_0^{(n)} = 0$ are used for estimation of parameters. Therefore, $\boldsymbol{\beta}^{(n)} = \left(\zeta^{(n)}, \alpha_1^{(n)}, \ldots, \alpha_I^{(n)}, \tau_1^{(n)}, \ldots, \tau_J^{(n)}\right)^T$.

For better illustration of the modelling of systematic component, we will describe approach which is typically used in context of claims reserving. This approach is used in Shi and Frees (2011) and adopted also by Abdallah et al. (2015), which both deal with using of couplae in estimating the ultimate claims. In these articles, authors assume gamma distribution for the incremental claims for one of the analysed line of business.

Gamma model

Assume that standardized incremental claims of nth line of business are independent and follow gamma distribution with shape parameter $\kappa^{(n)}$ and scale (location) parameter $\theta_{i,j}^{(n)}$, denoted $X_{i,j}^{(n)} \sim Gamma\left(\kappa^{(n)}, \theta_{i,j}^{(n)}\right)$. Then, from the properties of gamma distribution, we have

$$\mathsf{E}\left(X_{i,j}^{(n)}\right) = \mu_{i,j}^{(n)} = \kappa^{(n)}\theta_{i,j}^{(n)}.$$

In the case of gamma distribution, one could apply the canonical inverse link $g(\mu) = \eta = \mu^{-1}$, so we can write

$$\eta_{i,j}^{(n)} = \zeta^{(n)} + \alpha_i^{(n)} + \tau_j^{(n)} = \left(\mu_{i,j}^{(n)}\right)^{-1} = \left(\kappa^{(n)}\theta_{i,j}^{(n)}\right)^{-1}.$$

The unknown parameters to be estimated are $\zeta^{(n)}, \alpha_i^{(n)}, \tau_j^{(n)}$ and the dispersion parameter $\phi^{(n)}$ from the expression of gamma distribution in the associated

form of exponential family distribution. Hence in the case of gamma model, the vector of additional parameters $\boldsymbol{\gamma}^{(n)}$ is one-dimensional with the only element $\phi^{(n)}$. For the shape parameter of gamma distribution, it holds that $\kappa^{(n)} = 1/\phi^{(n)}$. Finally, the scale parameter is calculated using estimates of other parameters and the relationship among them. The estimated reserve in accident year i for nth line of business based on GLM model is $\sum_{k=I}^{J} \widehat{x}_{i,k}^{(n)} \omega_i^{(n)}$, where $\widehat{x}_{i,k}^{(n)}$ is the projected standardized incremental payment. For the gamma distribution we have

$$\widehat{x}_{i,k}^{(n)} = \left(\widehat{\eta}_{i,k}^{(n)}\right)^{-1} = \left(\widehat{\zeta}^{(n)} + \widehat{\alpha}_i^{(n)} + \widehat{\tau}_k^{(n)}\right)^{-1}.$$

Other possibility when dealing with gamma model is to use log-link function $g(\mu) = \eta = \log \mu$. According to Wütrich and Merz (2008), logarithmic link is a natural choice for the claims reserving purposes. Then

$$\eta_{i,j}^{(n)} = \zeta^{(n)} + \alpha_i^{(n)} + \tau_j^{(n)} = \log\left(\mu_{i,j}^{(n)}\right) = \log\left(\kappa^{(n)}\theta_{i,j}^{(n)}\right);$$

and for projected standardized incremental claims we have

$$\widehat{x}_{i,k}^{(n)} = \exp\left(\widehat{\eta}_{i,k}^{(n)}\right) = \exp\left(\widehat{\zeta}^{(n)} + \widehat{\alpha}_i^{(n)} + \widehat{\tau}_k^{(n)}\right).$$

For logarithmic link function, the mean of standardized incremental claims has the multiplicative structure, which is the function of accident year effect and development lag effect.

Another distribution which is typically used to model claims payments in non-life insurance is normal distribution.

Normal model

The model assumes there exist parameters $\mu_{i,j}^{(n)}$ and $\left(\sigma^{(n)}\right)^2$ such that the standardized incremental claims $X_{i,j}^{(n)}$ are independent normally distributed with

$$X_{i,j}^{(n)} \sim \mathcal{N}\left(\mu_{i,j}^{(n)}, \left(\sigma^{(n)}\right)^2\right).$$

Let further assume the identity link $g(\mu) = \eta = \mu$ within the GLM approach, so we can write

$$\mathsf{E}\left(X_{i,j}^{(n)}\right) = \mu_{i,j}^{(n)} = \eta_{i,j}^{(n)} = \zeta^{(n)} + \alpha_i^{(n)} + \tau_j^{(n)}.$$

It can be seen that the normal model with identity link, which is canonical link function for normal distribution, is a classical linear model with normally distributed error terms.

In the concept of the normal model, one could apply logarithmic link function as defined below:

$$\eta_{i,j}^{(n)} = \zeta^{(n)} + \alpha_i^{(n)} + \tau_j^{(n)} = \log\left(\mu_{i,j}^{(n)}\right).$$

Then for the mean we have

$$\mu_{i,j}^{(n)} = \exp\left\{\eta_{i,j}^{(n)}\right\} = \exp\left\{\zeta^{(n)} + \alpha_i^{(n)} + \tau_j^{(n)}\right\}.$$

With this parametrization, the mean of incremental claims has a multiplicative structure, that is the product of the accident year effect and development year effect. For both link functions, the unknown parameters to be estimated are $\zeta^{(n)}, \alpha_i^{(n)}, \tau_j^{(n)}$ and $\sigma^{(n)}$ whose second power equals to the dispersion parameter $\phi^{(n)}$. The estimated reserve in accident year i for nth line of business derived using GLM model is $\sum_{k=I}^{J} \widehat{x}_{i,k}^{(n)} \omega_i^{(n)}$, where $\widehat{x}_{i,k}^{(n)}$ is the projected standardized incremental payment. For the normal distribution we have

$$\widehat{x}_{i,k}^{(n)} = \begin{cases} \widehat{\zeta}^{(n)} + \widehat{\alpha}_i^{(n)} + \widehat{\tau}_k^{(n)}, & \text{for identity link function,} \\ \exp\left\{\widehat{\zeta}^{(n)} + \widehat{\alpha}_i^{(n)} + \widehat{\tau}_k^{(n)}\right\}, & \text{for log-link function.} \end{cases}$$

3.2 Copula regression

In this section, we will describe the approach which relaxes two common assumptions used when dealing with multiple lines of business: the limitation of dependence structure to the concept of linear structure; and the assumption of common probability distribution for all run-off triangles. Our focus is on using copulae to analyse more general concept of dependence and that is association. In addition, combining copula approach with multivariate regression enables to consider different distribution for individual lines of business.

Following the Theorem 4, the joint distribution function of standardized incremental claims $\left(X_{i,j}^{(1)}, \ldots, X_{i,j}^{(N)}\right)$ can be represented by a unique copula function

$$\begin{aligned} F_{i,j}\left(x_{i,j}^{(1)}, \ldots, x_{i,j}^{(N)}\right) &= \mathsf{P}\left\{X_{i,j}^{(1)} \leq x_{i,j}^{(1)}, \ldots, X_{i,j}^{(N)} \leq x_{i,j}^{(N)}\right\} \\ &= C\left(F_{i,j}^{(1)}\left(x_{i,j}^{(1)}\right), \ldots, F_{i,j}^{(N)}\left(x_{i,j}^{(N)}\right); \boldsymbol{\theta}\right), \end{aligned} \tag{3.2}$$

where the form of $F_{i,j}^{(1)}, \ldots, F_{i,j}^{(N)}$ can be expressed by (3.1) and $C(\bullet; \boldsymbol{\theta})$ is the copula with the parameter vector $\boldsymbol{\theta} \in \Theta$; Θ is an open subset of $\mathbb{R}^p$ for some integer $p \geq 1$. In this specification, parameter $\boldsymbol{\theta}$ captures the association relationship among the lines of business. Definitions and some properties of non-linear measures of association, such as Spearman's rho, Kendall's tau and parameters of tail dependence, are stated in the Section 1.4.

The specification of the model includes the choice of particular copula function as well as the distribution functions of incremental payments for each run-off triangle. Once a particular model is selected, one needs to estimate its parameters. In the model of copula regression in the form (3.2), the unknown parameters of the model to be estimated are given by the vectors $\eta_{i,j}^{(n)}, \boldsymbol{\gamma}^{(n)}$ for each run-off triangle $n \in \{1, \ldots, N\}$ and by the vector $\boldsymbol{\theta}$ related to copula function. We will denote the vector of unknown parameter by $\boldsymbol{\phi}$, where $\boldsymbol{\phi} = \left(\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(N)}, \boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(N)}, \boldsymbol{\theta}\right)^T$. As the model is limited to parametric distributions and parametric copulae, the estimation of parameters is provided via *maximum likelihood method*.

In general, the maximum likelihood method selects the set of values of model parameters that maximizes the likelihood function, respectively the log-likelihood function. The likelihood function is given by the joint density of random variables $X_{i,j}^{(1)}, \ldots, X_{i,j}^{(N)}$. Hence, to use the maximum likelihood method, one first specifies

the joint density function for all observations. The joint distribution function of standardized incremental payments is given by copula in (3.2). Rewriting this expression for the case of absolutely continuous copula C, the joint density function of $X_{i,j}^{(1)}, \ldots, X_{i,j}^{(N)}$ is given by

$$f_{i,j}\left(x_{i,j}^{(1)}, \ldots, x_{i,j}^{(N)}\right) = c\left(F_{i,j}^{(1)}\left(x_{i,j}^{(1)}\right), \ldots, F_{i,j}^{(N)}\left(x_{i,j}^{(N)}\right); \boldsymbol{\theta}\right) \prod_{n=1}^{N} f_{i,j}^{(n)}.$$

In the formulation above, $f_{i,j}^{(n)}$ denotes the density of marginal distribution $F_{i,j}^{(n)}$, that is $f_{i,j}^{(n)} = f^{(n)}\left(x_{i,j}^{(n)}; \eta_{i,j}^{(n)}; \boldsymbol{\gamma}^{(n)}\right)$ for $n = 1, \ldots, N$.

Once the joint density function is specified, the likelihood and log-likelihood functions can be defined for given realizations $x_{i,j}^{(1)}, \ldots, x_{i,j}^{(N)}$ of a random sample $X_{i,j}^{(1)}, \ldots, X_{i,j}^{(N)}$ for $(i,j) \in \{(i,j) : i + j \leq I\}$:

$$L\left(\boldsymbol{\phi}, x_{i,j}^{(n)}\right) = \prod_{i=0}^{I} \prod_{j=0}^{I-i} c\left(F_{i,j}^{(1)}\left(x_{i,j}^{(1)}\right), \ldots, F_{i,j}^{(N)}\left(x_{i,j}^{(N)}\right); \boldsymbol{\theta}\right) \prod_{n=1}^{N} f_{i,j}^{(n)};$$

and

$$\begin{aligned} \ell\left(\boldsymbol{\phi}, x_{i,j}^{(n)}\right) = & \sum_{i=0}^{I} \sum_{j=0}^{I-i} \log c\left(F_{i,j}^{(1)}\left(x_{i,j}^{(1)}\right), \ldots, F_{i,j}^{(N)}\left(x_{i,j}^{(N)}\right); \boldsymbol{\theta}\right) \\ & + \sum_{i=0}^{I} \sum_{j=0}^{I-i} \sum_{n=1}^{N} \log f_{i,j}^{(n)}. \end{aligned} \tag{3.3}$$

The maximum likelihood estimator $\widehat{\boldsymbol{\phi}}$ can be found by solving the optimization problem

$$\widehat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argsup}}\, \ell\left(\boldsymbol{\phi}, x_{i,j}^{(n)}\right).$$

The number of the parameters to be estimated depends on the number of analysed lines of business N, the length of observed accident period I and further on the number of parameters p that specify the chosen copula. The final dimension of vector $\boldsymbol{\phi}$ results into $2N(I+1)+p$. As the dimension $2N(I+1)+p$ gets large, the number of parameters increases, and the optimization problem becomes harder and more complex. One issue arises from the optimization task itself and is caused by the form of the log-likelihood function which is supposed to be optimized. There is a question whether it is possible to compute the optimal solution which presents the global maximum of the log-likelihood function. Another limitation of this estimation consists of the potential computational difficulty due to relatively large number of parameters. This problem may get significant when methods such as bootstrap or Monte Carlo experiment are needed. We will refer the afore-described estimation as *joint model*.

Due to this limitations, an alternative approach, that is computationally more convenient, has been developed. This approach, known as *inference functions for margins* (IFM), consists of two steps. Firstly, the IFM method estimates the parameters of the margins under the assumption 3.1. The estimates

$\widehat{\boldsymbol{\phi}}_{IFM} = \left(\widehat{\boldsymbol{\beta}}^{(1)}, \ldots, \widehat{\boldsymbol{\beta}}^{(N)}, \widehat{\boldsymbol{\gamma}}^{(1)}, \ldots, \widehat{\boldsymbol{\gamma}}^{(N)}\right)^T$ are obtained by maximizing the following function:

$$\ell\left(\boldsymbol{\phi}_{IFM}, x_{i,j}^{(n)}\right) = \sum_{i=0}^{I} \sum_{j=0}^{I-i} \sum_{n=1}^{N} \log f_{i,j}^{(n)}.$$

Therefore

$$\widehat{\boldsymbol{\phi}}_{IFM} = \underset{\boldsymbol{\phi}_{IFM}}{\operatorname{argmax}} \sum_{i=0}^{I} \sum_{j=0}^{I-i} \sum_{n=1}^{N} \log f_{i,j}^{(n)}.$$

In the second step, the estimate $\widehat{\boldsymbol{\phi}}_{IFM}$ is plugged into the expression of the distribution function of incremental claims 3.1 and we get

$$\widetilde{F}_{i,j}^{(n)} = F^{(n)}\left(\widehat{\eta}_{i,j}^{(n)}; \widehat{\boldsymbol{\gamma}}^{(n)}\right),\ n = 1, \ldots, N,$$

Subsequently, these values are used to estimate the parameter of copula $\boldsymbol{\theta}$. The partial (pseudo) log-likelihood function is given by

$$\ell\left(\boldsymbol{\theta}, x_{i,j}^{(n)}\right) = \sum_{i=0}^{I} \sum_{j=0}^{I-i} \log c\left(\widetilde{F}_{i,j}^{(1)}\left(x_{i,j}^{(1)}\right), \ldots, \widetilde{F}_{i,j}^{(N)}\left(x_{i,j}^{(N)}\right); \boldsymbol{\theta}\right).$$

The above function is then maximized over the set Θ and the copula parameter is obtained as

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=0}^{I} \sum_{j=0}^{I-i} \log c\left(\widetilde{F}_{i,j}^{(1)}\left(x_{i,j}^{(1)}\right), \ldots, \widetilde{F}_{i,j}^{(N)}\left(x_{i,j}^{(N)}\right); \boldsymbol{\theta}\right).$$

It can be shown that under certain assumptions the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ is consistent and asymptotically normal. These properties and also the correctness of this approach were shown by Chen and Fan (2006). Further details about the maximum likelihood method in the copula concept can be found in Hofert et al. (2012). In this article, the derivation of closed form expressions for the multivariate density function of a few Archimedean copulae, notably the Gumbel, the Clayton and the Frank copula, are listed.

3.3 Goodness-of-fit tests for copulae

There are several families of copulae which can be used to model the data. The most common examples of copulae are introduced in the Section 1.5. So, there is a natural question, which of the considered copulae fits the data in the most appropriate way.

The described copula regression model considers only parametric distributions of margins and parametric families of copulae. Formally we can write the following: the copula C capturing the association relationship between N lines of business is assumed to belong to a class of parametric copulae $\mathcal{C}_0 = \{C_\theta : \theta \in \Theta\}$; and further the distribution of margins $F^{(n)}$ comes from a parametric families $\mathcal{F}^{(n)} = \left\{F^{(n)}\left(\eta_{i,j}^{(n)}; \boldsymbol{\gamma}^{(n)}\right) : \eta_{i,j}^{(n)}(resp.\ \boldsymbol{\beta}^{(n)}) \in \Upsilon^{(n)}, \boldsymbol{\gamma}^{(n)} \in \Gamma^{(n)}\right\}$. Hence the parameters of the model are estimated under the assumptions:

1. $H_0' :\ F^{(1)} \in \mathcal{F}^{(1)}, \ldots, F^{(N)} \in \mathcal{F}^{(N)}$ in the first step;

2. $H_0 :\ C \in \mathcal{C}_0$ in the second step.

To review the suitability of the chosen model, the joint null hypothesis $H_0 \cap H_0'$ corresponding to a full parametric model is tested. In addition, one needs to distinguish between the joint and the IFM model.

3.3.1 Joint model

Since only one likelihood respectively log-likelihood function is maximized within the joint model, one of the approaches to review the fit of this model is likelihood-based measure represented by the concept of *Akaike's information criterion*. The Akaike's information criterion (AIC) is defined as follows:

$$AIC = -2\ell + 2k,$$

where $k = 2N(I+1) + p$ denotes the number of parameters of the model and $\ell = \ell\left(\widehat{\boldsymbol{\phi}}, x_{i,j}^{(n)}\right)$ is the maximal value of log-likelihood function for given observations. One computes the values of AIC for each candidate model and in the end the model with the lowest value of AIC should be chosen. The Akaike's information criterion is justified as a penalization for complexity and tends to chose simpler models. The penalization concept is not considered in the introduced case of copula regression, as the form of the model is set.

The AIC is widely used in practice due to computational simplicity. Further details on information criteria used to explore the suitability of candidate models are listed in Grønneberg and Hjort (2011). In this publication, the specific goodness-of-fit measure for copulae is introduced under the name the *Copula information criterion*.

3.3.2 IFM model

As it was introduced afore, the IFM model consists of two steps. In each of these steps, the log-likelihood function is maximized in order to obtain the estimates of the parameters of marginal distribution in the first step and using these estimates obtain the copula parameter in the second step. AIC is a measure based on the value of log-likelihood function, hence it is possible to express it in each individual step too. The problem is that in this case, information criteria allow to assess the suitability only of the partial task under the assumptions H_0 or H_0' and they do not provide the overall regard on the model with assumptions $H_0 \cap H_0'$. When testing the joint hypothesis $H_0 \cap H_0'$, the marginal distributions are considered as nuisance parameters.

Recently, many procedures have been developed for goodness-of-fit testing for copula-based models. They can be divided into several groups upon their properties and demands allowing their implementation. One of these groups is known as so-called *blanket tests* whose application requires no strategic choice of additional parameters, weight function, etc. The critical review of blanket tests and possible extensions of these procedures are introduced in Genest et al. (2009). The authors assess the level and the power of blanket tests using large Monte Carlo experiment.

Genest et al. (2009) describe in total seven different values which is possible to use to test the goodness-of-fit of copula regression model. For the purposes of the thesis and empirical analysis on real data, we will describe three of them, namely statistics stated by $S_n, S_n^{(B)}$ and $S_n^{(C)}$. In the notations $S_n, S_n^{(B)}$ and $S_n^{(C)}$, the value n stands for the number of independent observations available for individual random variable. In the copula regression model, we assume that the claims are observable for the pairs (i,j) for which $i+j \leq I$. Hence we have at our disposal $\frac{(I+1)(I+2)}{2}$ observations for every line of business (we recall that $i \in \{0,\ldots,I\}$ and $j \in \{0,\ldots,J\}$). To avoid the misleading discrepancies when mentioning the number of observation and individual business lines, we will denote the number of observations of one lines of business by m, i.e. $m = \frac{(I+1)(I+2)}{2}$, and the statistics will be denoted by $S_m, S_m^{(B)}$ and $S_m^{(C)}$.

Before defining afore-mentioned statistics, it is necessary to introduce next vectors and notation:

- $\boldsymbol{X}_{i,j} = \left(X_{i,j}^{(1)}, X_{i,j}^{(2)}, \ldots, X_{i,j}^{(N)}\right)$ for all pairs $(i,j) : i+j \leq I$;
- for $n \in \{1,\ldots,N\}$, $R_{i,j}^{(n)}$ is the rank of the $X_{i,j}^{(n)}$ among the claims of n-th line of business $\frac{(I+1)(I+2)}{2}$ variables $X_{0,0}^{(n)}, \ldots, X_{i,j}^{(n)}, \ldots, X_{I,J-I}^{(n)}$;
- $\boldsymbol{U}_{i,j} = \left(U_{i,j}^{(1)}, U_{i,j}^{(2)}, \ldots, U_{i,j}^{(N)}\right)$ for $(i,j) : i+j \leq I$ where $U_{i,j}^{(n)} = \dfrac{R_{i,j}^{(n)}}{\frac{(I+1)(I+2)}{2}+1}$.

The vectors $\boldsymbol{U}_{i,j}$ are considered as pseudo-observations and are interpreted as a sample from the underlying copula C. When providing the goodness-of-fit test based on these pseudo-information, one should take into account that they are not mutually independent and that their components are approximately uniform on the interval $(0,1)$. The blanket tests are based on ranks, since the ranks present the transformations with respect to which the considered statistics are invariant.

Using the pseudo-informations, the associated empirical distribution $C_m(\boldsymbol{u})$ is derived as

$$C_m(\boldsymbol{u}) = \frac{1}{m}\sum_{i=0}^{I}\sum_{j=0}^{I-i}\mathbb{I}\left(U_{i,j}^{(1)} \leq u_1, \ldots, U_{i,j}^{(N)} \leq u_N\right),\ \boldsymbol{u} = (u_1,\ldots,u_N) \in [0,1]^N.$$

C_m present the natural way how to test the hypothesis $H_0 :\ C \in \mathcal{C}_0$. This test is based on the distance between C_m and the estimation C_{θ_m} of the true underlying copula C under H_0. The parameter θ_m is assumed to be an estimate of θ based on the pseudo-observations. The goodness-of-fit test is based on the following measure:

$$\mathbb{C}_m = \sqrt{m}\,(C_m - C_{\theta_m}).$$

Genest et al. (2009) state the rank-based version of the Crameér-von Mises statistic derived from the $\mathbb{C}_m$ as follows

$$S_m = \int_{[0,1]^N} \mathbb{C}_m(\boldsymbol{u})^2\,\mathrm{d}C_m(\boldsymbol{u}).$$

The hypothesis H_0 is rejected for the big values of S_m. It is possible to derive approximate p-values from the limiting distribution of the statistic S_m which

depends on the asymptotic properties of $\mathbb{C}_m$. In addition, it was shown that tests based on S_m are consistent meaning that if $C \notin \mathcal{C}_0$, then as $m \to \infty$ the H_0 is rejected with probability 1.

$S_m^{(B)}$ and $S_m^{(C)}$ represent two new procedures proposed by Genest et al. (2009). They are both based on *Rosenblatt's transform*, the probability integral transformation. In contrary to the rank-based transformation, it provides the decomposition into elements that are mutually independent and uniformly distributed on the unit interval. Rosenblatt's probability integral transform of a copula C is the mapping $\mathcal{R} : (0,1)^N \to (0,1)^N$ that assigns to every vector $\boldsymbol{u} = (u_1, \ldots, u_N) \in (0,1)^N$, the vector $\mathcal{R}(\boldsymbol{u}) = (e_1, \ldots, e_N)$ where $e_1 = u_1$ and for each $n \in \{2, \ldots, N\}$

$$e_n = \frac{\partial^{n-1} C(u_1, \ldots, u_n, 1, \ldots, 1)}{\partial u_1 \cdots \partial u_{n-1}} \Bigg/ \frac{\partial^{n-1} C(u_1, \ldots, u_{n-1}, 1, \ldots, 1)}{\partial u_1 \cdots \partial u_{n-1}}$$

The most important property used for the purposes of goodness-of-fit testing of copulae is that under H_0 it is possible to interpret the Rosenblatt's transformations of pseudo-observations, $\boldsymbol{E}_{i,j} = \mathcal{R}_{\theta_m}(\boldsymbol{U}_{i,j})$ for $(i,j) : i + j \leq I$, as a sample from independent copula Π^N. This consequence is essential for deriving the tests based on statistics $S_m^{(B)}$ and $S_m^{(C)}$. First we will denote by D_m the empirical distribution function of transformed pseudo-observations $\boldsymbol{E}_{i,j}$:

$$D_m(\boldsymbol{u}) = \frac{1}{m} \sum_{i=0}^{I} \sum_{j=0}^{I-i} \mathbb{I}(\boldsymbol{E}_{i,j} \leq \boldsymbol{u}), \ \boldsymbol{u} \in [0,1]^N.$$

Hence, under the hypothesis H_0, the distance between D_m and Π^N is supposed to be "small". $S_m^{(B)}$ and $S_m^{(C)}$ can be again seen as the version of Crameér-von Mises statistic and are defined below:

$$S_m^{(C)} = m \int_{[0,1]^N} \left(D_m(\boldsymbol{u}) - \Pi^N(\boldsymbol{u})\right)^2 \mathrm{d}D_m(\boldsymbol{u});$$

and

$$S_m^{(B)} = m \int_{[0,1]^N} \left(D_m(\boldsymbol{u}) - \Pi^N(\boldsymbol{u})\right)^2 \mathrm{d}\boldsymbol{u}.$$

It is not clear which of the afore-introduced test is the most convenient. The goodness-of-fit tests based on these statistics depend on many other factors, which are not necessarily known before testing. Genest et al. (2009) suggest following preference ranking:

$$S_m^{(B)} \succ S_m \succ S_m^{(C)}.$$

All results presented in this publication are deducted from the comparative power study of the blanket tests based on the large Monte Carlo procedure. The approximate distribution of the tests statistics under H_0 is derived via bootstrap method. The overall recommendation of this study yields that S_m and $S_m^{(B)}$ provide the best results when assessing the fit of chosen copula model.

Chapter 4

Empirical analysis of the real data

In the previous chapters, the claims reserving method within a copula framework is introduced and its theoretical background is described. Hereafter, the copula regression model is applied to the claims triangles of a property-casualty insurer from USA. The empirical analysis is divided into two parts. First, it is important to find the appropriate GLM model for the margins. Subsequently, this model will be used to fit the copula and capture the association relationship between the analysed lines of business.

The computational part of the empirical analysis is performed in the environment of the R software for statistical calculations and graphics.

4.1 Datasets

The claims triangle data were published in the practical study of Meyers and Shi (2011) and come from the Analysis of Losses and Loss Expenses of NAIC (National Association of Insurance Commissioners, an organization of insurance regulators) database. The data were published in order to allow to provide claims reserving studies. The database includes both incurred and paid losses of major personal and commercial lines of business for property-casualty insurers.

The copula regression model is used in order to explore the mutual behaviour of several lines of business. In this analysis, the portfolio consisting of two lines of business will be considered, namely personal auto and commercial auto. The data describes ten years (from 1988 to 1997) of observed paid losses with ten development periods, hence according to the afore-introduce notation $I = J = 9$. Table 4.1 and Table 4.2 display the run-off triangles of incremental paid losses for personal and commercial auto, respectively. The earned premium and observations from the lower triangles are also given.

We observe, that according to the earned premium the portfolio is not equally distributed between the lines of business. In the beginning of the observed period, the personal auto line had the bigger portion (62 %) of the entire portfolio. However, after ten years, the portion of commercial auto line increased and yielded into 54 % of total portfolio, which makes the 46% share for personal auto. Hence the volume of the personal auto line decreased approximately by 26 %. Naturally, the total incremental claims amounts depends on the volume of the business,

Accident year i	Earned Premium	Development year j									
		0	1	2	3	4	5	6	7	8	9
0	51 228	15 318	12 422	7 671	4 793	2 148	1 338	491	60	123	31
1	52 526	15 031	15 101	7 814	4 425	1 504	643	220	351	5	52
2	55 816	16 994	14 620	7 985	5 344	1 399	1 311	213	219	12	144
3	61 549	17 717	16 050	8 974	4 140	2 236	1 302	222	146	155	38
4	64 970	17 842	13 275	8 319	5 435	1 939	611	788	515	91	318
5	66 559	20 266	17 200	8 255	4 920	1 603	997	553	299	375	3
6	67 046	18 778	14 438	8 814	5 665	1 557	750	544	253	88	3
7	64 535	19 900	16 542	7 143	5 592	2 875	1 098	270	68	161	10
8	66 791	20 395	15 402	8 019	3 871	2 781	617	513	156	2	158
9	69 057	20 622	15 844	8 123	5 950	2 321	1 026	724	186	252	32

Table 4.1: Development triangle of incremental claims and earned premium for personal auto in USD.

Accident year i	Earned Premium	Development year j									
		0	1	2	3	4	5	6	7	8	9
0	30 939	4 381	5 121	5 653	3 737	2 053	405	371	213	25	1
1	36 910	5 456	4 431	3 451	4 167	2 657	797	878	22	35	69
2	43 540	7 083	8 128	5 880	6 597	1 037	669	147	39	15	110
3	48 693	9 800	7 807	5 792	6 519	2 213	1 352	203	1 016	47	15
4	54 988	8 793	10 395	7 550	4 834	2 646	952	984	47	55	30
5	59 222	9 586	8 711	7 701	5 637	2 125	1 025	868	126	58	15
6	65 567	11 618	10 675	11 242	5 717	3 362	1 771	258	128	470	308
7	70 125	12 402	15 511	11 226	5 918	2 593	2 624	231	49	33	0
8	76 223	15 095	12 715	7 711	8 545	4 242	1 753	1 276	567	112	1 879
9	80 374	16 361	12 184	12 395	9 509	3 763	2 510	936	76	149	23

Table 4.2: Development triangle of incremental claims and earned premium for commercial auto in USD.

which is in this case measured by earned premium. Therefore, we normalize the payments dividing the amounts by this exposure variable of the corresponding accident. Further analysis will be provided on the standardized data. The normalization will be taken into account when expressing the estimated value of the reserves.

The graphical illustrations of standardized data are presented on the Figure 4.1. The plot captures the development pattern of the claims, where each line corresponds to single accident year. In both cases, it can be seen decreasing trend which signifies that all claims will be closed within ten development years as assumed. When comparing the patterns, we see higher volatility during first development years for commercial auto. The figure also shows that the relative volume of claims amounts is lower in case of commercial auto line.

Since the data provides the observations of claims also for accident years i and development lags j for which $i+j > I$, it is possible to express the real outstanding losses. We performed this calculation for each accident year separately and then get the total claims reserve, see Table 4.3. In this table, the values of claims reserves obtained by chain-ladder method are stated as well. Despite the fact, that chain-ladder method is distribution-free and does not require other inputs except the observations of upper triangle, it provides relatively similar results as real observed reserves. The CL estimates are higher by 3 % and 11 % in comparison with observed values for personal auto line and commercial auto respectively. These results can be considered as benchmark for assessing the reasonableness of

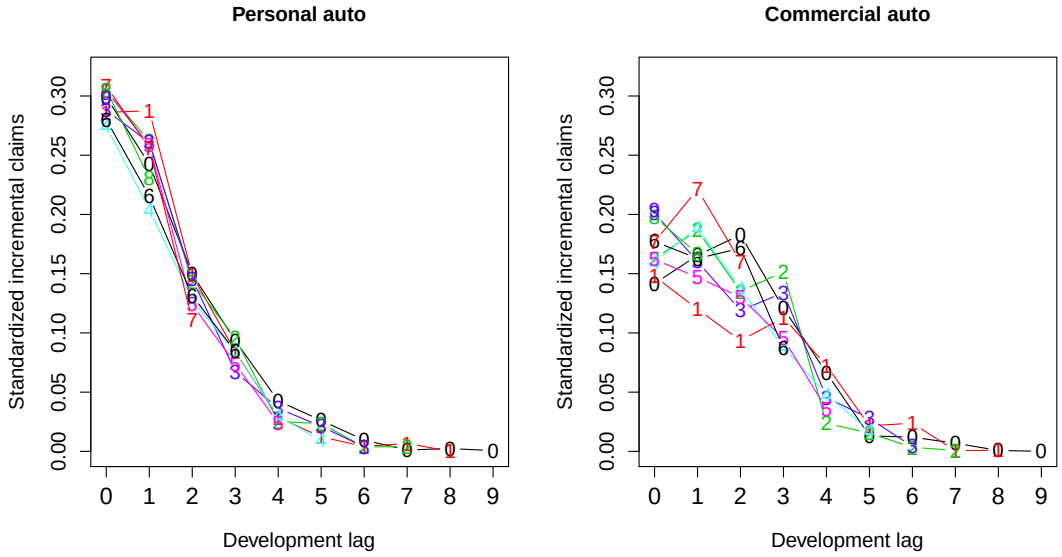


Figure 4.1: Development of standardized incremental claims amounts

Accident year i	1	2	3	4	5	6	7	8	9	Total
Observed claims resrves										
Personal auto	52	156	339	1 712	2 227	3 195	10 074	16 117	34 458	68 330
Commercial auto	69	125	1 078	1 116	2 092	6 297	11 448	26 085	41 545	89 855
Overall portfolio	121	281	1 417	2 828	4 319	9 492	21 522	42 202	76 003	158 185
Chain-ladder claims resrves										
Personal auto	32	103	342	614	1 881	3 707	9 153	18 280	36 462	70 571
Commercial auto	1	42	174	719	1 746	5 378	15 760	27 509	48 450	99 779
Overall portfolio	32	144	516	1 333	3 627	9 085	24 913	45 788	84 912	170 350

Table 4.3: Observed and chain-ladder values of claims reserves in USD

copula regression model. Moreover, the overall reserve of the portfolio is given by simple addition of individual reserves of lines of business which assumes no dependence structure. It can be seen that the overall reserve estimated by chain-ladder method is approximately higher by 10 % in comparison with observed reserve.

4.2 Marginal distributions fitting

Since the parametric copula fitting requires the specification of the marginal distributions, the preliminary analysis is needed. Our goal is to find the suitable GLM model which will be used in copula regression fitting. First, the initial analysis of the marginal distribution will be provided and subsequently the chosen distribution is plugged into the GLM model with different link functions. In both cases two possibilities are considered: normal and gamma distribution; and identity and logarithmic link function within GLM framework. The exactly same approach will be used in case of both analysed lines of business, so we will describe in detail the distribution and GLM fitting only for personal auto. All

the graphics for commercial auto line can be found in Appendix A.

4.2.1 Distribution fitting

In this part of fitting process, we consider two parametric distributions, namely gamma and normal distribution. The parameters of this distributions are obtain via maximum likelihood method. Subsequently the formal statistic tests are provided. The appropriate distribution is tested using Kolmogorov-Smirnov test which compares empirical distribution function of a random sample with the given reference probability distribution. The results of the test are reported in Table 4.4. The test was performed with significance level $\alpha = 5\%$. Hence the p-values in the table suggest that the normal distribution provide more suitable fit for both personal and commercial auto line.

	Personal auto	Commercial auto
Gamma distribution	≈ 0	≈ 0
Normal distribution	0.106	0.189

Table 4.4: p-values of Kolmogorov-Smirnov test

The other process to asses the suitable distribution is through the graphical comparison of empirical and theoretical measures. The Figure 4.2 illustrates qq-plots of gamma and normal distribution for personal auto line. In addition, so-called pp-plot, which compares the theoretical probabilities from given distribution with the series $1/(m+1), 2/(m+1), \ldots, m/(m+1)$, where m is the number of observations, $m = \frac{(I+1)(I+2)}{2} = 55$. In according to the displayed graphs, there is a place for potential improvement of the tails fitting. The provided analysis is restricted on gamma and normal distribution, which are not generally considered to capture well the heavier tail of the random sample. The typical representatives of heavy-tailed distributions are *Pareto* or *log-normal* distributions.

Based on the performed analysis, the normal distribution is considered to provide more appropriate fit for both personal and commercial auto lines.

4.2.2 GLM fitting

Once the distribution of standardized incremental claims belonging to exponential family is specified, the GLM model can be fitted. For each line of business, two GLM models are considered, the first one using the identity link function and the other one applying the logarithmic link function for normally distributed response variable.

The choosing of more appropriate link function is gained by the residual analysis of both performed models. In order to explore the properties of the Pearson residuals (further just residuals), the following plots are constructed:

- *residuals vs. fitted values*: this graphical illustration allows to observe how the residuals are located around the zero line to review their empirical mean and volatility;
- *observed vs. fitted values*: through this plot it can be seen how close are the fitted values to the observed ones;

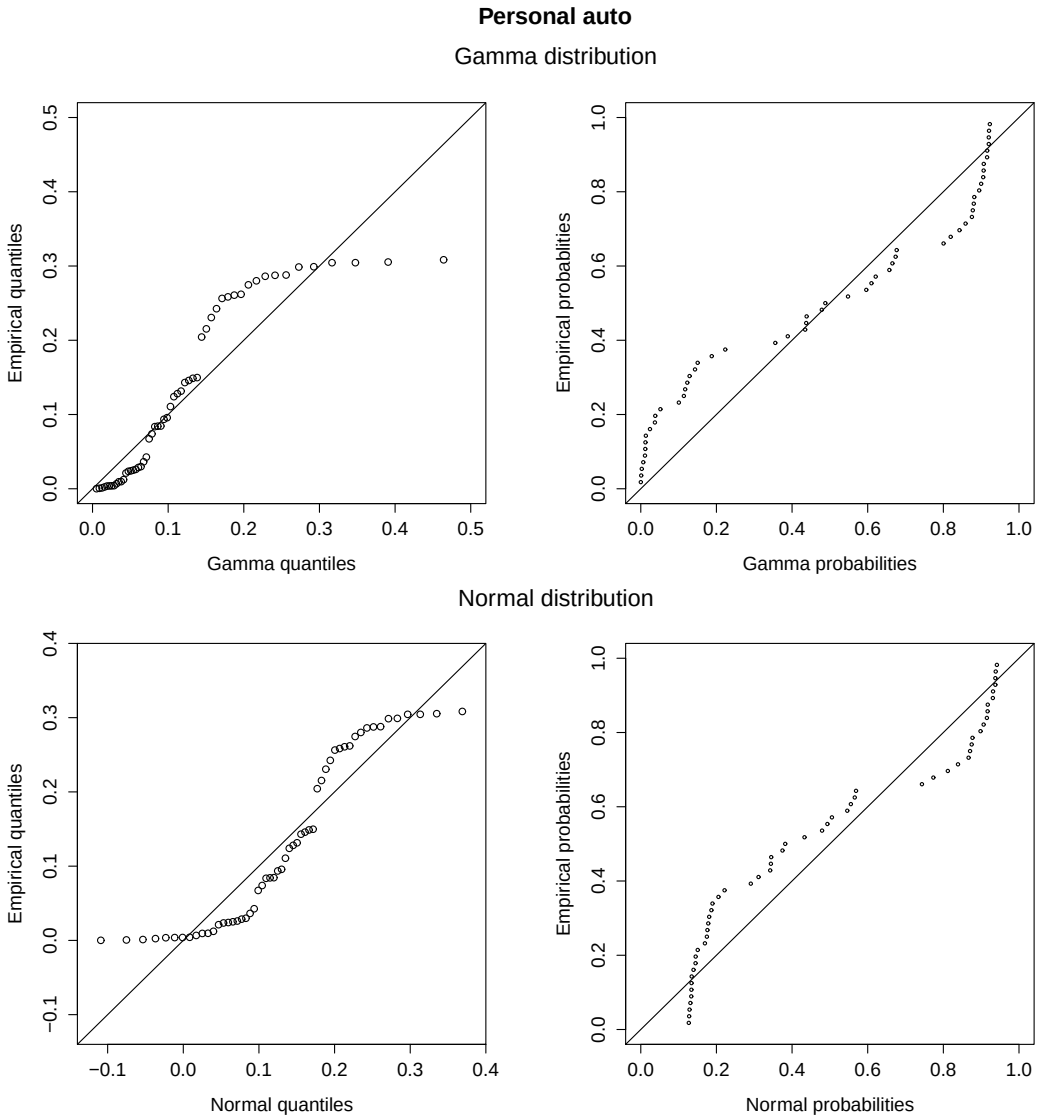


Figure 4.2: Empirical vs. theoretical measures (personal auto)

- *histograms of residuals*: histogram assess the approximate probability distribution of residuals, which are expected to be asymptotically normal;
- *scatter plot of residuals and residuals lagged in time by one period*: the scatter plot shows the possible autocorrelation among the residuals. If the residuals seem to be autocorrelated, there is a concern, that some dependence relationship might not be captured by the model.

All of afore-listed plots for personal auto line are reported in Figure 4.3, Figure 4.4, Figure 4.5 and Figure 4.6.

The first three plots exhibit quite similar features for both identity and logarithmic link function. The residuals seem to be uniformly spread through the range of fitted values with a few outlier values and are symmetrically located around the zero line. From the Figure 4.4 it can be seen that the fitted values are very closed to the observations. This "close" distance can be formally measured by *mean squared error* (MSE) which is given as arithmetic mean of the squared

distance between fitted and observed values. The MSE is approximately 0.00012 and 0.00010 for identity and logarithmic link function respectively.

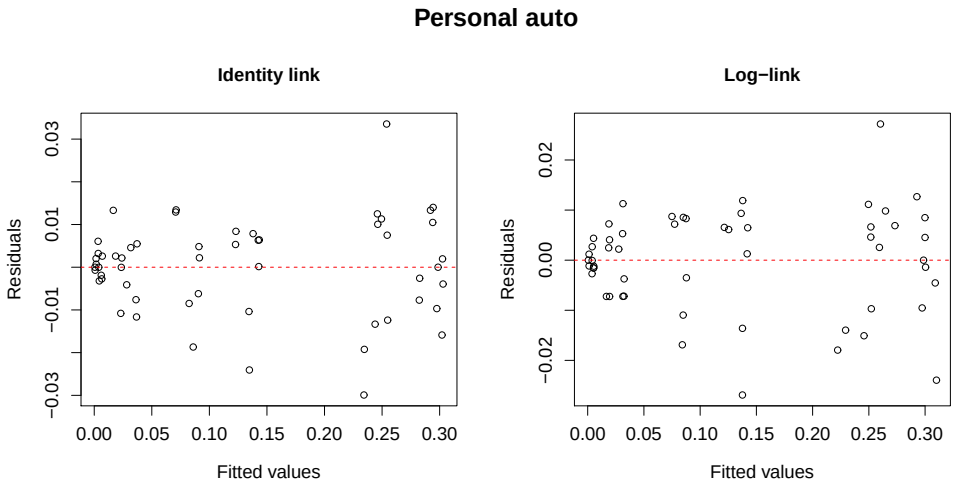


Figure 4.3: Pearson residuals of the fitted models (personal auto)

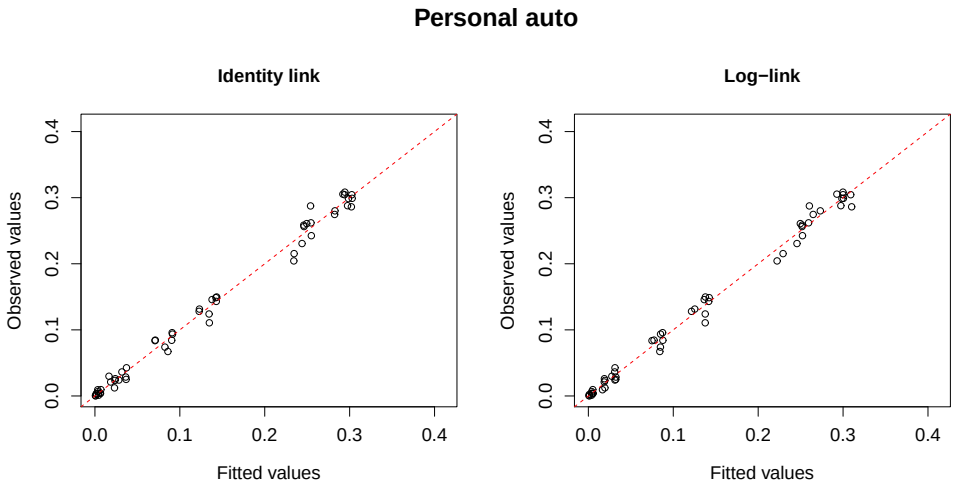


Figure 4.4: Observed vs. fitted values of the fitted models (personal auto)

In according to the histograms, the residuals of both models seem to be approximately normal. The formal statistical test can be performed using Shapiro-Wilk test for normality. The corresponding p-values fro identity and log link are 0.185 and 0.063 respectively, which suggest that the residuals admits the normal distribution.

The Figure 4.6 shows slightly higher rate of linear correlation among the residuals for logarithmic link function.

Considering all relevant graphs and measures, the final GLM model is specified by the normal distribution of response variables and identity link function in case of both lines of business.

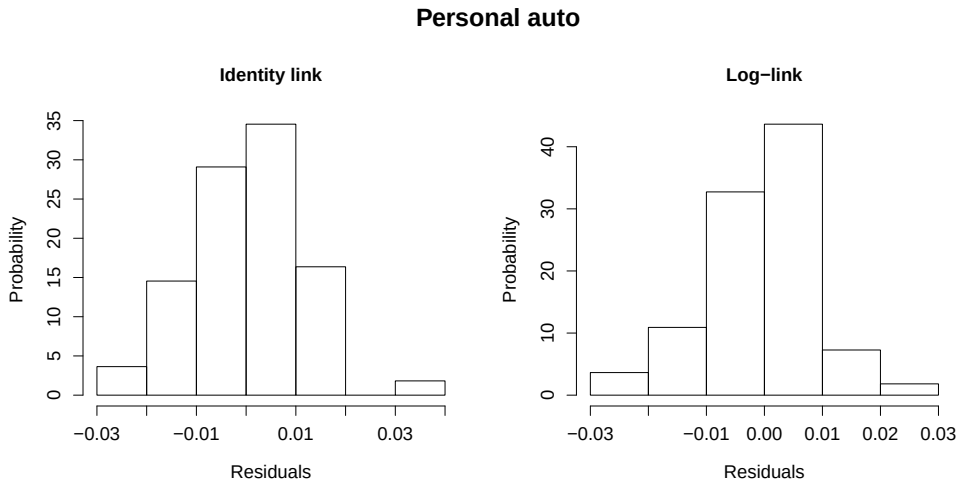


Figure 4.5: Histograms of the Pearson residuals (personal auto)

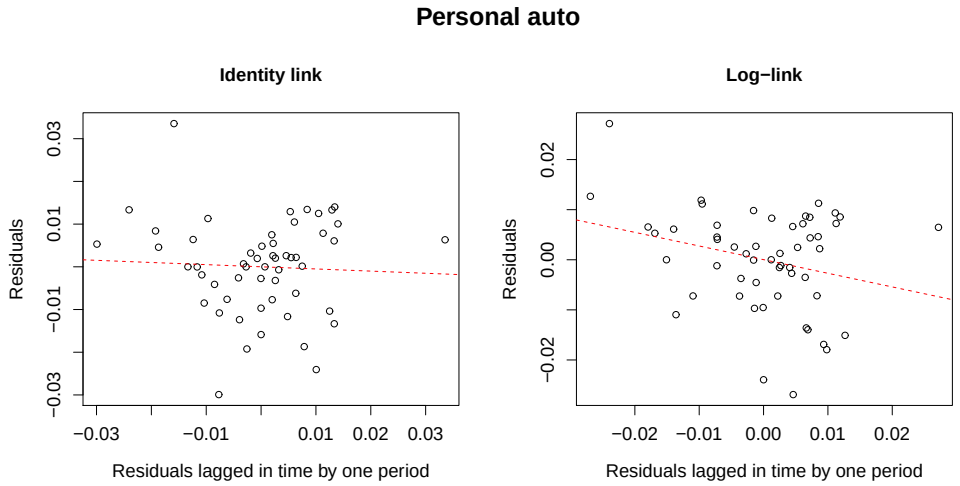


Figure 4.6: Autocorrelation of the Pearson residuals (personal auto)

4.3 Copula regression

In this part we focus on modelling of the association relationship between lines of business. The copula regression model will be fitted as described in Chapter 3. First the initial preliminary analysis of dependence structure is performed.

The scatter plot of standardized incremental claims is stated in Figure 4.7. The plot indicates quite strong positive relationship between personal and commercial auto line. The corresponding estimate of linear correlation reaches the value 0.880. Though, the observed dependence structure seems to be non-linear. Hence, the copula approach might be reasonable to model this association relationship.

To model the single lines of business, the normal model with identity link function is applied. By expressing the residuals of these models, the data poured off the effects of accident year and development lag are obtained. The Figure 4.8

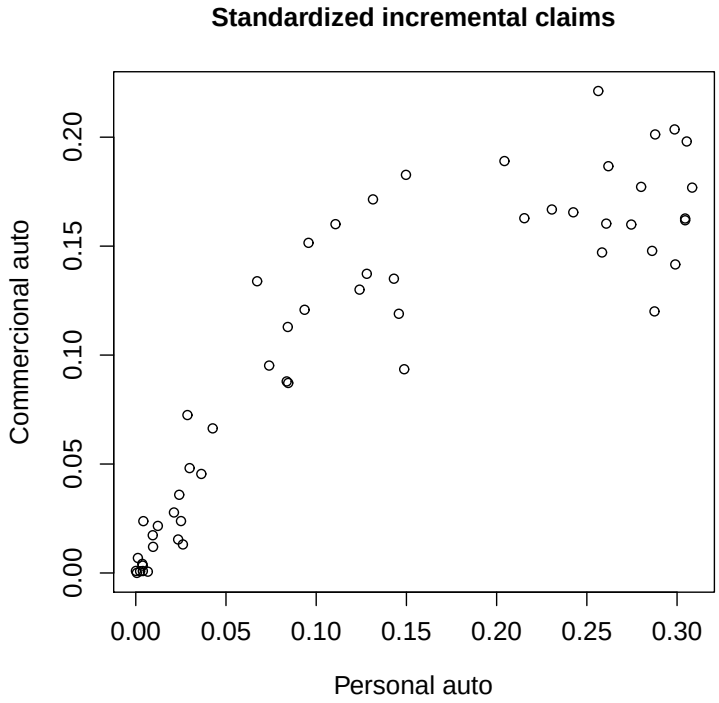


Figure 4.7: Standardized incremental claims of personal and commercial auto lines

displays the scatter plot of the residuals. The interesting result can be observed: the estimated correlation coefficient is -0.194, which signifies the negative relationship between residuals of lines of business.

4.3.1 IFM model fitting

The first copula model that is applied is two-steps method of inference function of margins. In fact, the first step of this method is already performed, since the single GLM model has been already fitted when exploring the appropriate link function. The estimates of the parameters of normal model with identity link for personal and commercial auto are stated in Table 4.5 (the abbreviation AY stands for accident year and DL denotes the development lag).

Once the GLM model is estimated, the copula can be fitted. The GLM model allows to express the corresponding form of normal distribution of standardized incremental claims. The mean is given by the estimates of the parameters and the variance is in the case of normal model equalled to the dispersion parameter of the GLM model. Hence the distribution function of marginals is exactly specified and its values in observed standardized incremental claims can be expressed. This way, the sets of observations which are expected to be uniformly distributed on unit interval are obtained. Using these derived sets of observations, the copula is fitted. The fitting procedure can be summarized as follows:

- the estimates of means $\widehat{\mu}_{i,j}^{(n)}$ and of dispersion parameter $\widehat{\phi}^{(n)}$ are obtained by fitting the GLM model;

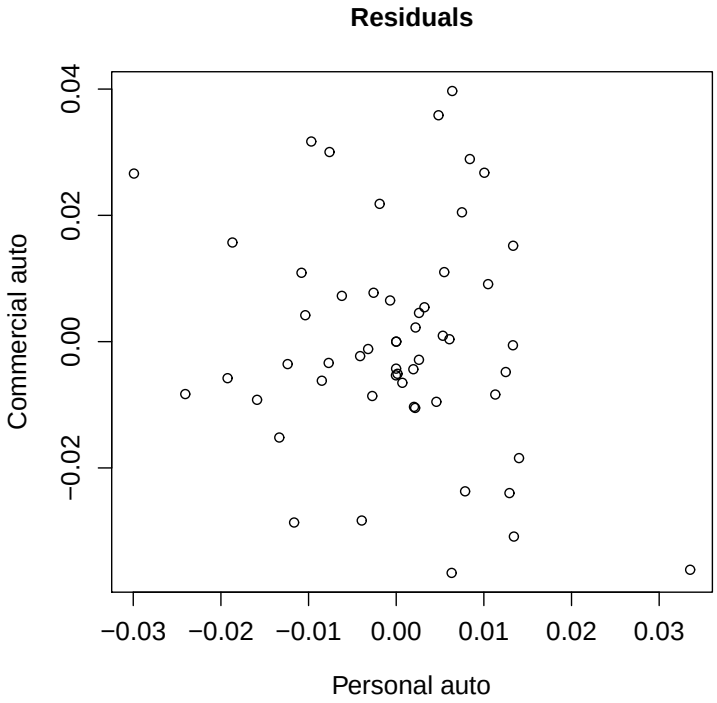


Figure 4.8: Residuals of GLM models for personal and commercial auto lines

- for the normal model, it holds that $\left(\sigma^{(n)}\right)^2 = \phi^{(n)}$, i.e. the estimate of the variance is given by $\left(\widehat{\sigma}^{(n)}\right)^2 = \widehat{\phi}^{(n)}$;
- the estimated distribution function $\widetilde{F}_{i,j}^{(n)}$ of $X_{i,j}^{(n)}$ is $\mathcal{N}\left(\widehat{\mu}_{i,j}^{(n)}, \left(\widehat{\sigma}^{(n)}\right)^2\right)$;
- the set of uniformly distributed observations $u_{i,j}^{(n)}$ is obtained by expressing the value of $\widetilde{F}_{i,j}^{(n)}$ in the observed standardized incremental claims;
- the parameter $\boldsymbol{\theta}$ of the copula C can be estimated via maximum likelihood function by maximizing the corresponding density function $c\left(u_{i,j}^{(1)}, \ldots, u_{i,j}^{(N)}\right)$ over the set Θ.

After constructing the observations $u_{i,j}^{(n)}$, the copula function is fitted. In the performed analysis, the following copulae will be considered: Gaussian, Clayton, Gumbel and Frank copula. Results of the fitting process are summarized in the Table 4.6. The table reports the estimated parameter of the copula and the associated measures as the Spearman's rho, Kendall's tau and coefficients of tail dependence. As it can be seen, the results for Gumbel copula are missed. This is caused by the fitting algorithm inside R, which did not converge and hence did not provide any estimate. However, the divergence of the algorithm can be also considered as a kind of result as it suggests that Gumbel copula is not able to appropriately capture the dependence of analysed data.

Once, the copulae are fitted, there is the question which of them express the dependence structure of the lines of business in the most suitable way. To review this ability, the goodness-of-fit tests for copulae are performed. The blanket

	Estimates	
	Personal auto	Commercial auto
Intercept	0.303	0.170
AY1	-0.001	-0.013
AY2	0.000	-0.003
AY3	-0.005	0.000
AY4	-0.021	-0.007
AY5	-0.009	-0.017
AY6	-0.020	0.000
AY7	-0.009	0.025
AY8	-0.011	0.013
AY9	-0.004	0.034
DL1	-0.048	-0.001
DL2	-0.160	-0.027
DL3	-0.212	-0.051
DL4	-0.266	-0.115
DL5	-0.279	-0.146
DL6	-0.296	-0.155
DL7	-0.299	-0.162
DL8	-0.301	-0.163
DL9	-0.302	-0.170

Table 4.5: Estimated parameters of normal model

	θ	ρ_C	τ_C	λ_L	λ_U
Gaussian	-0.335	-0.332	-0.218	0	0
Clayton	-0.274	-0.234	-0.159	0	0
Frank	-1.447	-0.235	-0.157	0	0

Table 4.6: Estimated parameters of fitted copulae

tests described in Chapter 3 are exhibited and the associated p-values are shown in the Table 4.7. We observe, that on the significance level $\alpha = 5\%$, according to the stated p-values the Gaussian copula does not provide the suitable fit. Both, Clayton and Frank copula seem to be appropriate upon the test based on statistic S_m. In addition, the Frank copula might be suitable choice also in consonance with tests based on Rosenblatt's transformation $S_m^{(B)}$ and $S_m^{(C)}$. In line with the preference ranking of Genest et al. (2009), the decision will be based on the test statistic $S_m^{(B)}$ which suggests the Frank copula as the most appropriate fit of the analysed data.

	S_m	$S_m^{(B)}$	$S_m^{(C)}$
Gaussian	0.003	0.025	0.020
Clayton	0.059	0.027	0.029
Frank	0.065	0.306	0.239

Table 4.7: p-values of selected blanket tests

The final copula regression model fitted by two-step IFM approach is given by Frank copula with marginals characterized by normal GLM model with identity link. After the model is specified and parameters are estimated, the future outstanding claims reserves can be predicted. One of the advantage of parametric

model is that it is able to predict the reserves by simulating. The predictive distribution of the reserves can be constructed and will be used to derive the final value of the reserves. Beside that, the total distributional specification of the reserves allows to provide other distributional quantities. The prediction of reserves using the copula regression IFM model is exhibited according to the following procedure:

- the realizations $\left(u_{i,j}^{(1)}, \ldots, u_{i,j}^{(N)}\right)$ are generated from the estimated Frank copula for the accident years i and development lags j for which $i + j > I$;
- the standardized incremental claims $x_{i,j}^{(n)}$ are simulated as the value of inverse function of $\widetilde{F}_{i,j}^{(n)}$ given by parameters estimated by GLM model;
- the unpaid losses for accident year $i \in \{1, \ldots, I\}$ are obtained by

$$\sum_{n=1}^{N} \sum_{k=I-i+1}^{J} x_{i,k}^{(n)} \omega_i^{(n)};$$

- the total reserve of the overall portfolio is

$$\sum_{i=1}^{I} \sum_{n=1}^{N} \sum_{k=I-i+1}^{J} x_{i,k}^{(n)} \omega_i^{(n)}.$$

These steps are repeated B times, hence we get B simulations of the reserve value either for every accident year i separately or aggregated for all $i \in \{1, \ldots, I\}$. Based on these simulation, the empirical distribution function can be determined. Then the final value of reserves is given as mean of this distribution. For the purposes of our analysis, we set $B = 10000$.

Following the afore-described procedure, the estimated reserve for overall portfolio is 187 186 USD. The detailed results for separate lines of business and accident years i are shown in Table 4.8. We observe that there are negative values of claims for some accident years. This can occur as a consequence of the normal distribution we consider as the normal distribution does not ensure non-negative values of reserves. In practice, the negative values of incremental claims can occur when there are some returns of already paid claims because of some later findings in claims settlements or lawsuits.

4.3.2 Joint model fitting

In this section, we will describe the fitting procedure of the joint model which estimates the parameters of GLM models and of copula jointly in one step. There is no package and pre-defined function in R dealing with this problem. Thus it was needed to construct the log-likelihood function given by 3.3. The formulation of this function is stated in Appendix B. Then this function is optimize using the same optimization method as classical maximum likelihood estimation procedure designed in R. As the initial values for parameters, the estimations from IFM model are used.

It is notable, that in general this optimizing task might not converge and reach the results. In our case, with the claims data of two lines of business, the

Accident year i	Personal auto	Commercial auto	Total
1	-16	-475	-491
2	80	69	149
3	-587	698	110
4	-4 472	202	-4 269
5	-475	-1 895	-2 369
6	-3 148	6 973	3 825
7	6 837	28 424	35 261
8	14 829	36 142	50 971
9	36 284	67 716	104 000
Total	49 332	137 854	**187 186**

Table 4.8: Reserving results obtained by IFM model (in USD)

calculation is not so crucial as the sample size is quite low. There are situations, where more lines of business or generally more margins are considered with high sample size. In those instances, it might happen that the optimizing assignment becomes really challenging and it can be hard or even unreal to obtain results.

Beside the optimizing task itself, there is also problem of computational difficulty and efficiency. When considering the bootstrap method or simulating within this approach, the calculation can yield into undesirable duration and the overall model becomes inefficient from the practical point of view.

As it is mentioned, due to our quite low sample size, the optimizing task converged and reached the results in finite number of iterations. The estimates of the parameters are summarized in the Table 4.9. The values maximized log-likelihood and AIC are stated as well.

In the case of joint model, the selection of the most suitable model is simply based on the AIC. As it is seen in the table, the lowest value of AIC belongs to the Clayton copula. Thus the Clayton copula regression model is chosen as the most suitable.

After the specification and estimation of the model, the values of incremental claims, based on which reserves are calculated, are predicted. The estimation of the reserves is exhibited by the same procedure as in the case of IFM model. The Table 4.10 displays the reserving results for separate accident years and lines of business. The total reserve of the portfolio consisting of personal and commercial auto lines is 189 549 USD. Again, we observe the negative values for some accident years. The comparison of the models will be described in the following section.

	Gaussian copula		Clayton copula		Frank copula	
	Personal auto	Commercial auto	Personal auto	Commercial auto	Personal auto	Commercial auto
Intercept	0.303	0.170	0.301	0.169	0.303	0.170
AY1	-0.001	-0.013	0.000	-0.012	-0.000	-0.013
AY2	-0.000	-0.003	-0.001	-0.004	-0.000	-0.003
AY3	-0.005	-0.000	-0.003	0.001	-0.006	-0.000
AY4	-0.021	-0.007	-0.018	-0.005	-0.021	-0.006
AY5	-0.009	-0.017	-0.006	-0.016	-0.009	-0.017
AY6	-0.020	-0.000	-0.019	-0.001	-0.020	-0.001
AY7	-0.009	0.025	-0.008	0.024	-0.008	0.025
AY8	-0.011	0.013	-0.012	0.011	-0.011	0.013
AY9	-0.004	0.034	-0.001	0.035	-0.004	0.034
DL1	-0.048	-0.001	-0.047	-0.000	-0.048	-0.001
DL2	-0.160	-0.027	-0.160	-0.027	-0.160	-0.027
DL3	-0.212	-0.051	-0.210	-0.050	-0.211	-0.052
DL4	-0.266	-0.115	-0.268	-0.116	-0.266	-0.114
DL5	-0.279	-0.146	-0.277	-0.145	-0.279	-0.146
DL6	-0.296	-0.155	-0.294	-0.154	-0.296	-0.155
DL7	-0.299	-0.162	-0.296	-0.160	-0.299	-0.162
DL8	-0.301	-0.163	-0.299	-0.161	-0.301	-0.162
DL9	-0.302	-0.170	-0.299	-0.168	-0.302	-0.170
Copula parameter	-0.121		-0.188		-0.588	
Log-likelihood	-290.743		-299.441		-297.932	
AIC	-499.486		-516.882		-513.864	

Table 4.9: Estimated parameters of fitted joint models

Accident year i	Personal auto	Commercial auto	Total
1	69	-426	-357
2	23	55	78
3	-77	1 034	958
4	-3 654	701	-2 952
5	470	-1 510	-1 039
6	-3 067	6 901	3 833
7	6 622	27 573	34 194
8	13 705	34 813	48 518
9	37 601	68 715	106 316
Total	51 692	137 857	**189 549**

Table 4.10: Reserving results obtained by joint model (in USD)

4.3.3 Comparison of the models

We have already introduced and fitted the copula regression model for both approaches, the IFM and joint model. Hereafter, the results of these methods along with the observed values and chain-ladder estimation will be summarized and compared. All obtained result are displayed in Table 4.11. When comparing the results of copula regression models with observed reserves, we observe that the reserves of IFM model are higher by 18 % and in the case of joint model, they are greater by 20 %. The chain-ladder method can be viewed as industry's benchmark as it is one of the most commonly used methods in practice. The IFM model and joint model provides the results that are 10 % and 11 % respectively higher than results exhibited by chain-ladder method.

Accident	Observed			Chain-ladder model		
year i	Personal auto	Commercial auto	Total	Personal auto	Commercial auto	Total
1	52	69	121	32	1	32
2	156	125	281	103	42	144
3	339	1 078	1 417	342	174	516
4	1 712	1 116	2 828	614	719	1 333
5	2 227	2 092	4 319	1 881	1 746	3 627
6	3 195	6 297	9 492	3 707	5 378	9 085
7	10 074	11 448	21 522	9 153	15 760	24 913
8	16 117	26 085	42 202	18 280	27 509	45 788
9	34 458	41 545	76 003	36 462	48 450	84 912
Total	68 330	89 855	**158 185**	70 571	99 779	**170 350**

Accident	IFM model			Joint model		
year i	Personal auto	Commercial auto	Total	Personal auto	Commercial auto	Total
1	-16	-475	-491	69	-426	-357
2	80	69	149	23	55	78
3	-587	698	110	-77	1 034	958
4	-4 472	202	-4 269	-3 654	701	-2 952
5	-475	-1 895	-2 369	470	-1 510	-1 039
6	-3 148	6 973	3 825	-3 067	6 901	3 833
7	6 837	28 424	35 261	6 622	27 573	34 194
8	14 829	36 142	50 971	13 705	34 813	48 518
9	36 284	67 716	104 000	37 601	68 715	106 316
Total	49 332	137 854	**187 186**	51 692	137 857	**189 549**

Table 4.11: Reserving results (in USD)

Hereafter, we focus on copula regression approaches. These two approaches

can be compared upon two aspects: from the statistical and numerical point of view. In the terms of statistic measures, this comparison is based on mean squared error (MSE).

MSE measures the average of the squared differences of the "errors". We fitted the models using the observations for which $i+j \leq I$, this set of observations forms so-called *in-sample data*. On the contrary data satisfying $i + j > I$ are denoted as *out-of-sample* data. In this case, we will use MSE to assess the reasonableness of the model by measuring the difference between out-of-sample observations and the predictions given by the model estimated by using in-sample data. Thus MSE is given by

$$MSE = \frac{1}{l} \sum_{i=1}^{I} \sum_{k=I-i+1}^{J} \left(x_{i,j}^{(n)} - \widehat{x}_{i,j}^{(n)} \right)^2,$$

where l is the number of considered observations, in our case $l = \frac{(I+1)(I+1)}{2} - (I+1)$, and $\widehat{x}_{i,j}^{(n)}$ denotes the predicted values of standardized incremental claims.

The results of MSE for both lines of business and both approaches are shown in Table 4.12. As it can be seen, from the point of view based on MSE measure, IFM as well as joint model provide comparable results.

	Personal auto	Commercial auto
IFM model	0.00261	0.02080
Joint model	0.00250	0.02048

Table 4.12: Mean squarred errors

As it was already mentioned, there is a concern that joint model might be challenging in terms of the optimizing the joint log-likelihood function. To assess this aspect, the duration of estimation and prediction of the models was measured and reached the results summarized in the Table 4.13. We observe that the joint model is time demanding in both calculations.

We recall that the joint model is fitted using only 110 observations and two lines of business. In situations which necessitate to model more than two margins with relatively higher sample sizes, the calculations can become really time-consuming and inefficient.

	Estimation	Prediction
IFM model	0.03	8.64
Joint model	6.44	24.49

Table 4.13: Duration od selected calculations (in seconds)

Conclusion

The primary goal of the presented thesis was the study of different approaches to model the dependence among loss triangles using multivariate copulae. If losses of different lines of business are somehow related, this relationship needs to be reflected by aggregate reserves. To describe the dependence structure of multiple lines of business, the copula approach was proposed.

Our intention was to describe the model that is able to reflect the dependence structure of lines of business into a reserve of overall portfolio. The common practice when determining the value of aggregate reserve is the assumption of independent lines of business and then the simple addition of individual reserves. However, with the implementation of new regulatory standards, it is very important to understand the reserves as a whole, not just by lines of business. Hence the parametric approach of copula regression model was theoretically described and subsequently practically implemented in order to estimate the overall reserve of portfolio consisting of two lines of business.

Both variations of copula regression, the joint and IFM model, have been applied on real claims data. At first, the estimation of the considered models were obtained and using goodness-of-fit test, the most appropriate IFM and joint model was chosen. In the case of IFM model, the Frank copula seemed to be the most suitable while the Clayton copula provided the best fit for the joint model. Both approaches provided comparable results in terms of the final value of the reserves and MSE measure. When dealing with computational efficiency, the joint model seemed to be quite time demanding, especially the prediction of outstanding claims reserves.

The described copula regression model has both strengths and limitations. The copula model allows to consider different parametric regression for different lines of business, although the analysis of chosen losses of US insurer suggested normal regression for both lines of business. In addition, beside the point estimate of reserves, the parametric approach enables to construct also prediction errors and prediction distribution. Moreover, a strength of the parametric approach is that it has been widely used for small data sets which are typical in claims reserving framework.

The limitation of the model is the fact that we focus only on two representatives of the exponential family in the implementation of GLM. It would be interesting to explore whether the incremental payments might not support another distribution, possibly the distribution with heavier tails. The set of analysed copulae could be also extended. As it was mentioned, the dependence structure may appear on many levels, hence the model allowing the dependence relation among all the observations belonging to the same calendar year could be considered.

One possible improvement is described by Abdallah et al. (2015), which pro-

posed the use of new models based on hierarchial Archimedean copulae providing more flexibility and more intuitive interpretation of dependence relationship. Another method describes Regis (2011) which combines a Bayesian approach for the estimation of marginal distribution of single line of business and a Bayesian copula procedure for the estimation of aggregate reserves. It would be interesting to examine both these proposals, investigate the change in claims predictions and compare the obtained results.

Bibliography

A. Abdallah, J.P. Boucher, and H. Cossette. Modeling dependence between loss triangles with hierarchical archimedean copulas. *ASTIN Bulletin*, 2015.

R. Ballerini. Archimedean copulas, exchangeability, and max-stability. *Journal of Applied Probability*, 31:383–390, 1994.

S. Cambanis, S. Huang, and G. Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11:368–385, 1981.

X. Chen and Y. Fan. Estimation of copula-based semiparametric time series models. *J. Econometrics*, 36(2):307–335, 2006.

P. D. England and R. J. Verrall. Stochastic claims reserving in general insurance. *Brit. Actuar. J.*, 102(8):443–544, 2002.

Ch. Genest, B. Rémillard, and D. Beaudoin. Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44(2): 199–213, 2009.

S. Grønneberg and N.L. Hjort. The copula information criteria. `http://www.bi.edu/InstitutterFiles/Samfunns%C3%B8konomi/Papers/Spring%202012/Gr%C3%B8nneberg.pdf`, 2011. [Online; accessed 1-July-2015].

M. Hofert, M. Mächler, and A.J. McNeil. Likelihood inference for archimedean copulas in high dimensions under known margins. *Journal of Multivariate Analysis*, 110:133–150, 2012.

H. Joe, H. Li, and A.K. Nikoloulopoulos. Tail dependence functions and vine copulas. *Journal of Multivariate Analysis*, 101:252–270, 2010.

A. Juri and M.V. Wütrich. Tail dependence from a distributional point of view. *Extremes*, 6:213–246, 2003.

H.O. Lancaster. Measures and indices of dependence. In S. Kotz and N.L. Johnson, editors, *Encyclopedia of Statistical Sciences*. Wiley, New York, 1982.

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

P. McCullagh and P. Nelder. *Generalized Linear Models*. Second edition. Chapman & Hall, London, 1989. ISBN 978-0412317606.

A.J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management*. Princeton series in finance. Princeton University Press, Princeton, 2005. ISBN 0-691-12255-5.

G. G. Meyers and P. Shi. Loss reserving data pulled from NAIC Schedule P. `http://www.casact.org/research/index.cfm?fa=loss_reserves_data`, 2011. [Online; accessed 1-July-2015].

A. Müller and M. Scarsini. Archimedean copulae and positive dependence. *Journal of Multivariate Analysis*, 93:434–445, 2005.

R.B. Nelsen. *An Introduction to Copulas*. Second edition. Springer, New York, 2006. ISBN 978-0387-28659-4.

U. Olsson. *Generalized Linear Models - An Applied Approach*. Studentlitteratur, Lund, 2002. ISBN 91-44-03141-6.

L. Regis. A bayesian copula model for stochastic claims reserving. Technical report, Collegio Carlo Alberto, 2011.

B. Schweizer and A. Sklar. *Probabilistic Metric Spaces*. North-Holland, New York, 1983.

B. Schweizer and E.F. Wolff. On nonparametric measures od dependence for random variables. *Ann. Statist.*, 9(4):879–885, 1981.

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

A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ Inst Statist Univ Paris*, 3(8):229–231, 1959.

A. Sklar. Random variables, distribution functions, and copulas - a personal look backward and forward. In L. Rüschendorf, B. Schweizer, and M.D. Taylor, editors, *Distributions with Fixed Marginals and Realted Topics*. Institute of Mathematical Statistics, Hayward, CA, 1996.

E.F. Wolff. N-dimensional measures of dependence. *Stochastica*, 4(3):175–188, 1980.

M.V. Wütrich and M. Merz. *Stochastic Claims Reserving Methods in Insurance*. Wiley Finance Series. John Wiley & Sons, West Sussex, 2008. ISBN 978-0-470-72346-3.

List of Figures

List of Tables

Appendix A

Commercial auto

A.1 Distribution fitting

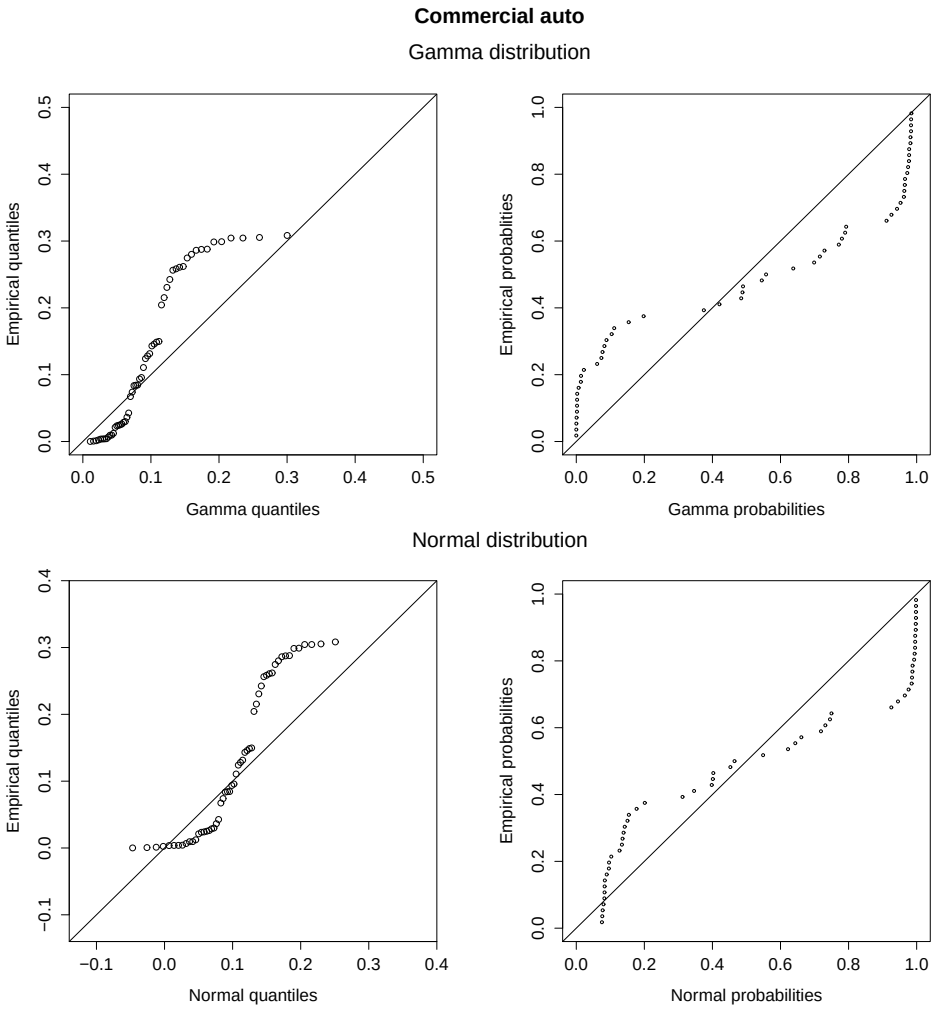


Figure A.1: Empirical vs. theoretical measures (commercial auto)

A.2 GLM fitting

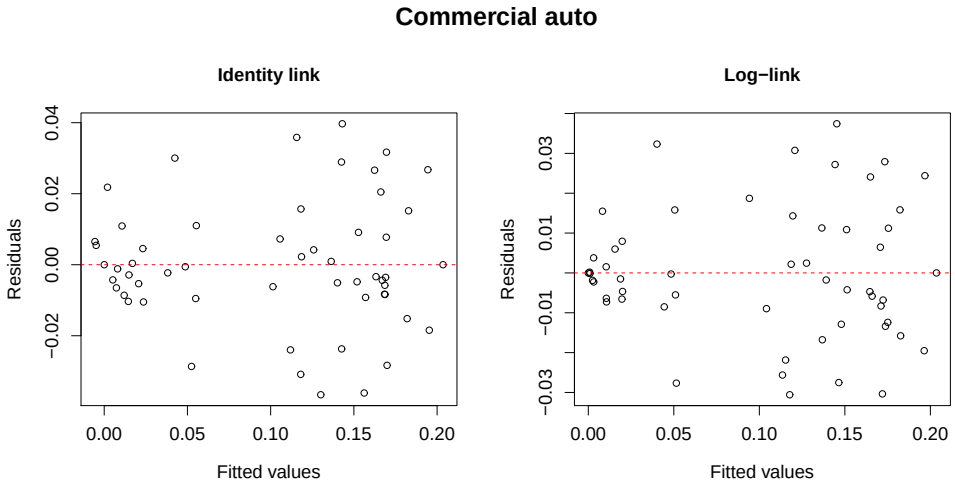


Figure A.2: Pearson residuals of the fitted models (commercial auto)

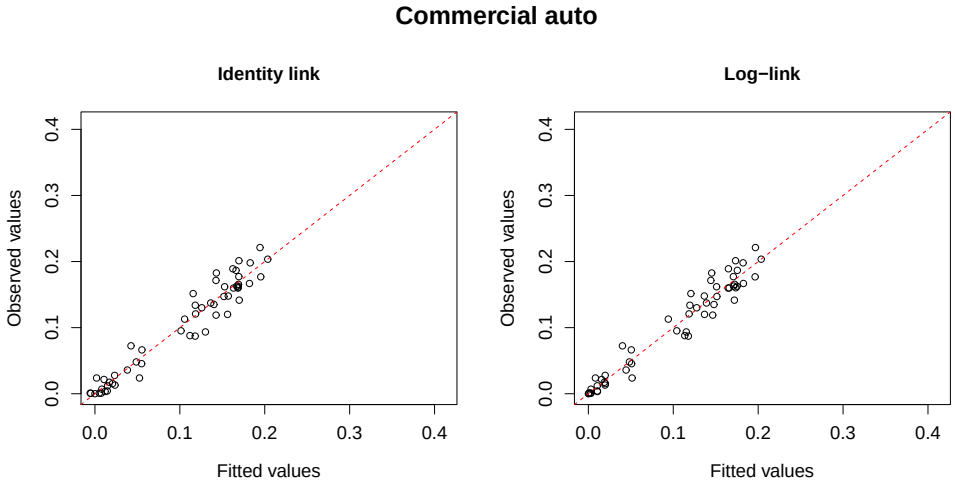


Figure A.3: Observed vs. fitted values of the fitted models (commercial auto)

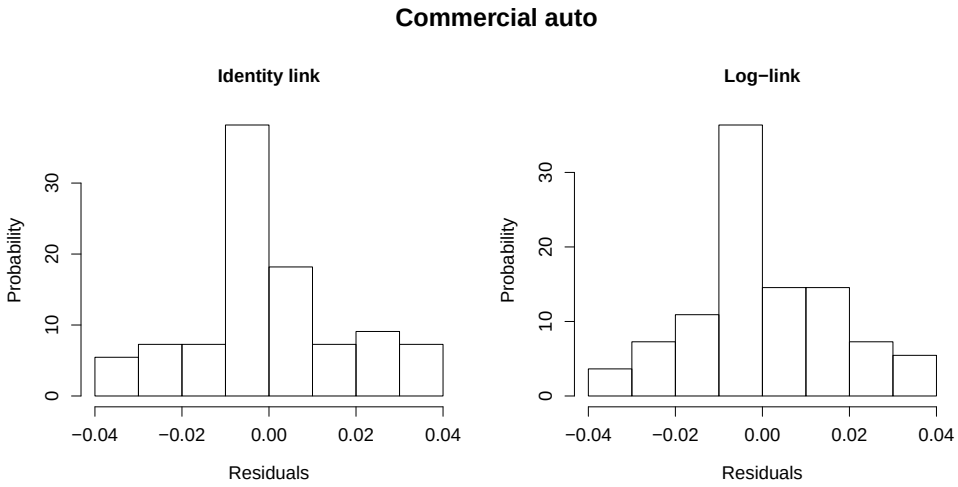


Figure A.4: Histograms of the Pearson residuals (commercial auto)

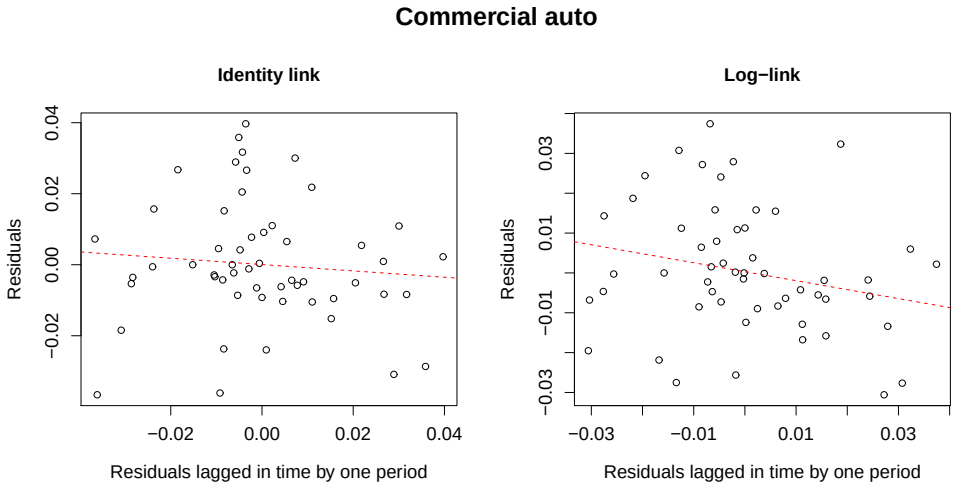


Figure A.5: Autocorrelation of the Pearson residuals (commercial auto)

	MSE	p-value of Shapiro-Wilk test
Identity link	0.00030	0.170
Logarithmic link	0.00026	0.338

Table A.1: Selected measures for commercial auto line

Appendix B

Source code

```
### Log-likelihood for margins

loglikMarginal_norm <- function(beta,disp,yyy,Xmat) {
  #### yyy is the response, Xmat is the model matrix
  xmat <- as.matrix(Xmat)
  xmat <- cbind(rep(1, NROW(xmat)),xmat)

  mu = xmat%*%beta
  std= sqrt(disp)

  llk <- sum(dnorm(yyy, mean=mu, sd=std ,log=TRUE))
  u <- pnorm(yyy,mean=mu, sd=std)
  u<-ifelse (u > 0.999999999999999, 0.999999999999999, u)

  Results <- list(llk, u)
  return(Results)
}

### The value of log-likelihood

loglikMarginal <- function(beta,disp,yyy,Xmat) {
  llk <- loglikMarginal_norm(beta,disp,yyy,Xmat)[[1]]
  return(llk)
}

### Log-likelihood for joint model

loglikCopula <-function(theta, y1, Xmat1, y2, Xmat2, copula) {
  copulaparam = theta[1]
  beta1 = theta[2:(NCOL(Xmat1)+2)]
  beta2 = theta[(NCOL(Xmat1)+3):NROW(theta)]

  temp1 <- loglikMarginal_norm(beta1,disp=sigma,yyy=y1,Xmat=Xmat1)
  temp2 <- loglikMarginal_norm(beta2,disp=sigma,yyy=y2,Xmat=Xmat2)
```

```
  uu = cbind(temp1[[2]],temp2[[2]])
  personal_loglik = temp1[[1]]
  com_loglik = temp2[[1]]
  copula@parameters <- copulaparam

  Loglik <- personal_loglik + com_loglik + sum(log(dCopula(uu,copula)))
  return(-Loglik)
}
```