ABSTRACT

It is crucial for the insurance business to create risk profiles for their customers to be able set a fair price for their insurances. This thesis will present an alternative method for measuring the dependence between the numbers of claims made by a customer holding two insurances. The group of insurance holders consisted of 74770 unique customers holding at least two insurances during one year. The method uses copulas to model the dependence between the numbers of claims made in each insurance. An advantage using copulas to model the dependence is that the margins and the dependence structure can be modeled separately. The copulas used in the analysis were four Archimedean copulas namely Clayton, Frank, Gumbel and Ali-Mikhail-Haq. These copulas were chosen for their simple explicit expressions and variety in dependence structure.

A flag has to be raised regarding the fact that the margins were discrete and the largest part of applied copula theory handles continuous margins. This will lead to complications in the modeling that were not expected. The marginal parameters were estimated using the ML-method and a 2-test decided that the negative binomial distribution respectively the zero-inflated negative binomial distribution were the best fits for the insurances. Regarding the copula, the dependence measure used was the rank correlation coefficient Kendall's Tau expressed in the the copula functions. The copula parameter was then estimated by inverting the formula for Kendall's tau. By performing a parametric bootstrap using Cramer von Mises method the Gumbel copula was shown to provide the best fit.

When a bivariate distribution was obtained from the model, it was compared to the empirical counterpart. Conditional distributions and conditional expected values were calculated from the bivariate model and they were compared to their empirical equivalent. It would appear

that the model provided an overall good fit, but the best fit was in the lower tail.

Keywords: Copula theory, Count data, Marginal distributions, Goodness of fit.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, my supervisor Nader Tajvidi for his guidance, patience and support throughout this research and the writing of this thesis. I would also like to thank Fredrik Thuring for an interesting problem and for providing the material and knowledge of the methods used in insurance business today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. OVERVIEW

1.1 Introduction

This thesis presents a copula approach to modeling dependence between the number of insurance claims made in two insurances. The two insurances are held by the same customer in the same insurance company, and the data consists of 74770 different customers. The dependence explored in this thesis has been modeled previously using a multivariate credibility method in Thuring (2011). A disadvantage with this method is that it uses assumptions of linear dependence which has its drawbacks, as will be explained in Section 2.1.2. Therefore an alternative method using copulas for modeling the dependence is considered in this thesis.

1.2 Background

In the last two decades copulas have become a popular tool for modeling dependence in different fields of applications, for example in actuarial and financial businesses (Kojadinovic and Yan, 2010b). Copulas provides a possibility to create a multivariate distribution by modeling the marginal distributions and the dependence beween the margins separately. Another great advantage with copulas is that they allow the margins to have a variety of distribution functions. A challenge using copulas to model count data is that the litterature in this area is fairly limited. This in comparison with the very popular multivariate normal distribution and its related distributions such as Student t's distribution. It is known to be easy to use the multvariate normal distribution because the margins are assumed to be normal and the association can be described by only the marginal distributions and the correlation coefficient. A problem using the assumption of normal distributions is that they are often not adequate to describe real life data.

1.3 Goal and purpose

The goal of this thesis is to make a copula model to model the dependence between the number of claims in two insurances held by one customer. From Thuring (2011) it is validated that a dependence exists, the objective in this thesis is to see if it is possible to model this dependence using copulas.

1.4 Structure of the paper

The first part of this thesis will present a theory chapter describing useful concepts used for the modeling. Since dependence measure is a central part of this thesis, different kinds of dependence measures will introduce Section 2.1. This will build up a base for explaining the copula concept in Section 2.3. The discrete distributions used for modeling the margins are found in Section 2.2. Complementary theory to Section 2 is found in Appendix A. Procedures and results are presented in Section 3. The final Section 4 consists of a summary and discussion of the results.

CHAPTER 2. THEORETICAL CONCEPTS

This chapter contains the necessary background to create a bivariate distribution for the number of claims made in the two insurances. Although a large part of the copula theory is assuming continuous margins, the theory is used in this thesis with smaller adjustments due to the discrete margins, such as using the dependence measure τ_β instead of τ_K. Some basic probability theory are found in Section A.1 in Appendix A.

2.1 Dependence measures

Desirable properties for dependence measures are listed by Embrechts *et al.* (2002), let $\delta(X, Y)$ stand for a scalar measure of dependence.

(I) $\delta(X, Y) = \delta(Y, X)$ (symmetry)

(II) $-1 \leq \delta(X, Y) \leq 1$ (normalization)

(III) $\delta(X, Y) = 1 \Leftrightarrow (X, Y)$ are comonotonic
$\delta(X, Y) = -1 \Leftrightarrow (X, Y)$ are countermonotonic

(IV) For a strictly monotonic transformation $T : \mathbb{R} \to \mathbb{R}$ of X:

$$\delta(T(X), Y) = \begin{cases} \delta(Y, X) & T \text{ increasing} \\ -\delta(Y, X) & T \text{ decreasing} \end{cases}$$

2.1.1 Comonotonicity and Countermonotonicity

The two random variables X_1, X_2 are comonotonic if there exists a random variable Z and the increasing functions f_1, f_2 such that,

$$(X_1, X_2) = (f_1(Z), f_2(Z)) \tag{2.1}$$

If one of f_1 or f_2 is decreasing the variables are countermonotonic. Comonotonicity and countermonotonicity can be considered the strongest concepts of dependence as the realization of X_i is entirely determined by $X_j, i, j = 1, 2, i \neq j$ (Yener, 2011, p.8).

2.1.2 Linear correlation

For definitions of linear correlation see Section A.2 in Appendix A. Linear correlation is a good measure of dependence for multivariate normal distributions which is a distribution used in many applications. When the margins have other distributions than normal, the weaknesses listed below will occur. (Trivedi and Zimmer , 2007, p.21)

- If X and Y are independent then $\rho = 0$, but the converse is not true, that is, even if $\rho = 0$, X and Y are not necessarily independent because ρ only detects linear dependence.
- The variances of X and Y must be finite, or the correlation coefficient can't be defined, this is a clear shortcoming for describing dependence for heavy-tailed distributions.
- The linear correlation is not invariant under strictly increasing nonlinear monotonic transform.

The linear correlation coefficient only fulfills requirements (I) and (II) in the list for desirable properties of dependence measures.

2.1.3 Rank correlation

An alternative measure of association to linear correlation is rank correlation, where the observations are ranked and compared regarding their concordance. Rank correlation measures monotone association and is invariant to all monotone increasing transformations of the original data. Rank correlation fulfills the requirements (I)-(IV) in the list for desirable properties for dependence measures. (Yener, 2011, p.17)

2.1.3.1 Concordance

A valid measure of dependence to examine is whether the number of claims of customers having one insurance tend to rise or fall together with his other insurance. Let (x_i, y_i) and

(x_j, y_j) denote pairs of random variables. Then (x_i, y_i) and (x_j, y_j) are said to be, (Yener, 2011, p.16)

$$\text{concordant if} \quad \begin{cases} x_i < x_j \text{ and } y_i < y_j \\ x_i > x_j \text{ and } y_i > y_j \end{cases}$$

$$\text{discordant if} \quad \begin{cases} x_i < x_j \text{ and } y_i > y_j \\ x_i > x_j \text{ and } y_i < y_j \end{cases}$$

2.1.3.2 Kendall's Tau

The rank correlation coefficient used in the modeling was Kendall's Tau τ_K.

Let $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ denote a random sample of n observations from a vector (X, Y) of random variables. Each pair (x_i, y_i) and (x_j, y_j) are concordant or discordant, and there are $\binom{n}{2}$ distinct pairs. Let c denote the number of concordant pairs and d number of discordant pairs. Then the sample version for Kendall's Tau τ_K when (X,Y) are continuous is,

$$\tau_K = \frac{c-d}{c+d} = \frac{c-d}{\binom{n}{2}} \tag{2.2}$$

$$-1 \leq \tau_K \leq 1 \tag{2.3}$$

(Nelsen , 2006, p.158, formula(5.1.1))

When (X, Y) are independent, then $\tau_K = 0$. For discrete data ties may occur,

$$\left.\begin{array}{c} x_i = x_j \\ \text{or} \\ y_i = y_j \end{array}\right\} \text{tie}$$

Kendall's Tau including ties will be referred to as τ_β and is defined as,

$$\tau_\beta = \frac{c-d}{\sqrt{(c+d+t_x)}\sqrt{(c+d+t_y)}} \tag{2.4}$$

$$\tag{2.5}$$

Where c, d is the number of concordant respectively discordant pairs, t_x, t_y the number of ties in (X, Y). According to Denuit (2005) in the discrete case τ_K is restricted to a narrower range

than $[-1, 1]$. The rank correlation measure τ_K can also be expressed by the copula function as, (Nelsen , 2006, p.161, Eq.(5.1.7))

$$\tau_{X,Y} = \tau_C = 4 \iint_{\mathbf{I^2}} C(u, v) dC(u, v) - 1 \tag{2.6}$$

$$\tau_C = 4E(C(U, V)) - 1 \tag{2.7}$$

2.2 Distribution for count data

To model count data that contains many zeros there are a few discrete alternatives such as the Poisson, zero-inflated Poisson, negative binomial and zero-inflated negative binomial distribution. The Poisson distribution is the standard distribution used for modeling the number of claims in insurance business today (Boucher, Denuit and Guillen , 2008, p.137), see definition in Section 2.2.1. A problem with the Poisson distribution is that mean and variance are equal, but in many applications the variance usually increases with the mean. The negative binomial distribution has larger variance than Poisson and it is sometimes referred to as overdispersed Poisson. The definition for the negative binomial distribution is presented in Section 2.2.2 and the relationship between the Poisson and negative binomial distribution is defined in Section 2.2.3.

2.2.1 Poisson distribution (Poi)

Let X denote the number of events in a unit interval of time, then X has the PMF,

$$P(X = k) = \frac{e^{-k}\lambda^k}{k!}, \qquad k = 0, 1, 2, ...; \lambda > 0 \tag{2.8}$$

(Krishnamoorthy , 2006, p.71) where λ is the mean numbers of events in a unit interval of time. The expected value and variance for the Poisson distribution is,

$$E(X) = \lambda \tag{2.9}$$

$$V(X) = \lambda \tag{2.10}$$

2.2.2 Negative Binomial distribution (NB)

Let X be a random variable and consider a sequence of independent Bernoulli trials with success probability p. Let X denote the number of failures until the nth success, then the PMF is,

$$P(X=k) = \binom{n+k-1}{k} p^n (1-p)^k, \qquad k = 0,1,2,...; 0 < p < 1 \tag{2.11}$$

$$E(X) = \mu_{NB} = \frac{n(1-p)}{p} \tag{2.12}$$

$$V(X) = \frac{n(1-p)}{p^2} \tag{2.13}$$

(Krishnamoorthy , 2006, p.97).

2.2.3 Relationship Poisson and negative binomial distributions

To make a comparison to the Poisson distribution with $E(X) = V(X) = \lambda$, the NB distribution can be written as a gamma-distributed mixture of Poisson distributions,

$$P(X=k) = \frac{\Gamma(\frac{1}{\alpha}+k)}{k!\Gamma(\frac{1}{\alpha})} \cdot \frac{\frac{1}{\alpha}^{\frac{1}{\alpha}} \lambda^k}{\left(\frac{1}{\alpha}+\lambda\right)^{\frac{1}{\alpha}+k}}$$

$$\lambda, \alpha > 0, y = 0,1,2,...$$

(Kattan, 2009, p.888) with $E(X) = \lambda$ and $V(X) = \lambda + \alpha\lambda^2$, where α is referred to as a dispersion parameter. When $\alpha \to 0$ NB $\to$ Poisson.

2.2.4 Zero-inflated distributions

A zero-inflated distribution is intuitively a distribution with extra zeros. This is accomplished by a mixture of the original distribution and a Bernoulli distribution.
Let X be a random variable and p the probability of success, then the PMF for the Bernoulli distribution is, (Krishnamoorthy , 2006, p.45)

$$P(X_i = 1) = p, \qquad i = 0,1,...n; 0 < p < 1 \tag{2.14}$$

$$P(X_i = 0) = 1 - p \tag{2.15}$$

A zero-inflated distribution is defined as, (Minkova, 2012, p.7)

$$P(X = 0) = p_z + (1 - p_z)P(0), \qquad k = 0 \tag{2.16}$$

$$P(X = k) = (1 - p_z)P(k), \qquad k = 1, 2... \tag{2.17}$$

where $P(X = k)$ is the PMF of the underlying distribution and p_z is the probability of zero-inflation. A result of the zero-inflation is that these distributions have a larger variance than their non-inflated counterparts.

2.2.5 Zero-inflated Poisson distribution (ZIP)

The PMF for the zero-inflated Poisson distribution is,

$$p(X = 0) = (1 - p_z) + p_z e^{-\lambda}, \tag{2.18}$$

$$p(X = k) = p_z \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 1, 2... \tag{2.19}$$

$$E(X) = (1 - p_z)\lambda \tag{2.20}$$

$$V(X) = \lambda(1 - p_z)(1 + \lambda p_z) \tag{2.21}$$

2.2.6 Zero-inflated Negative Binomial distribution (ZINB)

The PMF for the zero-inflated Negative Binomial is,

$$P(X = 0) = (1 - p_z) + p_z p^n \tag{2.22}$$

$$P(X = k) = p_z \binom{n + k - 1}{k} p^n (1 - p)^k, \tag{2.23}$$

$$k = 0, 1, 2, ...; 0 < p < 1; 0 < p_z < 1 \tag{2.24}$$

The PMF can also be expressed according to,

$$P(X=0) = (1-p_z) + p_z \left(\frac{\frac{1}{\alpha}}{\lambda + \frac{1}{\alpha}} \right)^{\frac{1}{\alpha}} \tag{2.25}$$

$$P(X=k) = p_z \frac{\Gamma(\frac{1}{\alpha}+k)}{k!\Gamma(\frac{1}{\alpha})} \cdot \frac{\frac{1}{\alpha}^{\frac{1}{\alpha}} \lambda^k}{\left(\frac{1}{\alpha}+\lambda\right)^{\frac{1}{\alpha}+k}} \tag{2.26}$$

$$k = 1, 2, ..., ; 0 < p_z < 1 \tag{2.27}$$

$$E(X) = (1-p_z)\lambda \tag{2.28}$$

$$V(X) = \lambda(1-p_z)(1+\lambda(p_z+\alpha)) \tag{2.29}$$

2.2.7 Goodness of fit - Margins

The parameters in the distributions were estimated using the maximum likelihood method defined in Appendix A.3.

2.2.7.1 AIC

The Akaike's Information Criterion (AIC) is a measure of the relative goodness of a statistical model. Choosing between models it decides which of them provides the best fit, but it does not give information of how well the model fits in its absolute sense. It is based on the log-likelihood function $\ell\ell h$ for the models and is defined as,

$$\text{AIC} = -2\ell\ell h + 2n_e \tag{2.30}$$

where n_e is number of estimated parameters in the model and the $\ell\ell h$ is the logarithm of $L(\theta)$ defined in Eq. A.12 in Appendix A. The best fit amongst the models will have the largest $\ell\ell h$ and the smallest AIC value. (Everitt and Skrondal, 2010, p.10)

2.2.7.2 χ^2-test

To answer the question of how well the model fits the data a p-value was calculated using a χ^2-test. The χ^2-test was calculated by the method described in Krishnamoorthy (2006), the method is found in Appendix A.

2.3 Copulas

An elementary description of a copula is that it is a multivariate distribution with uniform margins. A great advantage of copulas is that the margins and the dependence structure can be modeled separately and that the margins are allowed to take on different kinds of distributions. The dependence is modeled by rank correlation. The copula theory will be explained in the two-dimensional case although the theory can be translated into multiple dimensions if necessary.

2.3.1 Definition

Consider a pair of random variables X and Y with distribution functions $F(x) = P(X \leq x)$ and $G(y) = P(Y \leq y)$, respectively, and a joint distribution function $H(x, y) = P(X \leq x, Y \leq y)$. Each of these distribution functions are in the interval $[0, 1]$, this implies that each pair (x, y) of real numbers leads to a point $(F(x), G(y))$ in the unit square $[0, 1] \times [0, 1]$, this pair will correspond to a number $H(x, y)$ in $[0, 1]$. This linking function is called copula.(Nelsen , 2006, p.7).

2.3.1.1 Sklar's theorem

Let H be a joint distribution function with margins F and G. Then there exists a copula C such that for all $x, y \in \mathbf{R}$,

$$H(x, y) = C(F(x), G(y)) \tag{2.31}$$

If F and G are continuous then C is unique, otherwise C is uniquely determined on $Ran(F) \times Ran(G)$. Conversely if C is a copula and F and G are distribution functions, then the function H defined by 2.31 is a joint distribution function with margins F and G (Nelsen , 2006, Theorem 2.2.3, p.18).

2.3.1.2 Deriving the joint PMF

It is easy to derive the joint density $h(x, y)$ from $H(x, y)$ when the margins are continuous. The derivation is made using partial derivatives. In the discrete case the PMF $h(x, y)$ must be derived by using finite differences. Let $H(x_i, y_j) = C(F(x_i), G(y_j))$ be the same as in Eq. (2.31),

with number of observations $i, j = 1, ..n$ then the joint PMF $h(x_i, y_j) = c(F(x_i), G(y_j))$ is given by, (Nikoloulopoulos and Karlis , 2010, p.174)

$$h(x_i, y_i) = \begin{cases} C(F(x_i), G(y_j)), & i = j = 1 \\ C(F(x_1), G(y_j)) - C(F(x_1), G(y_{j-1})), & i = 1, j > 1 \\ C(F(x_i), G(y_1)) - C(F(x_{i-1}), G(y_1)), & i > 1, j = 1 \\ C(F(x), G(y)) - C(F(x-1), G(y)) & \\ -C(F(x), G(y-1)) + C(F(x-1), G(y-1)) & i, j > 1 \end{cases}$$

2.3.1.3 Properties of copulas

A copula is a function C defined from $\mathbf{I}^2$ to $\mathbf{I}$ having the following properties,

- For every u, v in $\mathbf{I}$,

  $C(u, 0) = 0 = C(0, v)$

  $C(u, 1) = u$ and $C(1, v) = v$

- For every u_1, u_2, v_1, v_2 in $\mathbf{I}$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$ the rectangular inequality holds according to,

  $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$

(Nelsen , 2006, Eq. 2.2.2-2.2.3, p.10)

The invariance theorem (Nelsen , 2006, Theorem 2.4.3, p) states that a copula with continuous random variables X, Y is invariant under strictly increasing transformations of X, Y.

2.3.1.4 Fréchet-Hoeffding bounds

Suppose that $C(u, v)$ is a copula, then for every u, v in $\mathbf{I}$ the Fréchet-Hoeffding bounds is,

$$W(u, v) \leq C(u, v) \leq M(u, v) \tag{2.32}$$

where $M(u, v)$ is called upper Fréchet-Hoeffding bound. $W(u, v)$ is called lower Fréchet-Hoeffding bound,

- $M(u, v) = min(u, v)$

- $W(u,v) = max(u+v-1, 0)$

(Nelsen , 2006, p.47)

Another important copula is the product copula $\prod(u,v)$. The applications of these copulas is that they represent the dependence measures: independence ($\prod(u,v)$), comonotonicity ($M(u,v)$) and countermonotonicity ($W(u,v)$) these can be seen in Figure 2.1.

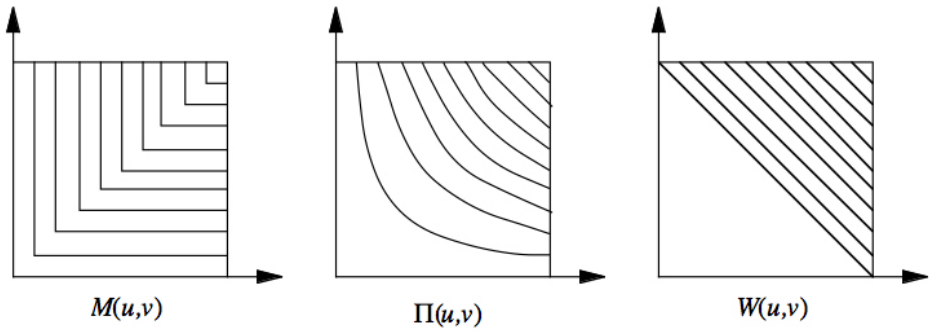


Figure 2.1 Contour plots of the copulas M, $\prod$ and W.

(Nelsen , 2006, Fig 2.2, p. 13)

2.3.2 Different copulas

There are basically three types of copulas: fundamental copulas, implicit copulas and explicit copulas. The focus in this thesis is a group of the explicit copulas called Archimedean copulas. They have simple closed form expressions and are easy to construct.

2.3.2.1 Archimedean copulas

Archimedian copulas have analytical expressions and they are popular because they allow modeling dependence using only one parameter. The two-dimensional Archimedean copula is defined as (Nelsen , 2006, p.110 Formula (4.1.3)),

$$C(u,v) = \phi^{-1}(\phi(u) + \phi(v)) \tag{2.33}$$

where ϕ is called the generator function that has the properties (Nelsen , 2006, Chap. 4.1 and 4.6),

- $\phi : [0,1] \to [0,\infty] \Longrightarrow \phi^{-1} : [0,\infty] \to [0,1]$

- $\phi(0) = \infty$ (Strict)
- $(-1)^k \frac{d^k}{dt^k}\phi(t)$, (Completely monotone)

  where d^k is the k_{th} derivative of ϕ, for all t in the interior of the interval J and $k = 0, 1, 2, ...$

The Archimedean copulas can be expressed using τ_C according to, (Nelsen , 2006, p.163 Formula (5.1.9)),

$$\tau_C = 1 + 4 \int_0^1 \frac{\phi(t)}{\phi'(t)} dt \tag{2.34}$$

Three copulas from the Archimedean family that are widely used are the Clayton, Frank and Gumbel copulas for their easy analytical expressions and different tail dependencies. Another Archimedean copula used in this thesis is the Ali-Mikhail-Haq (AMH) copula. The AMH copula can exhibit lower tail dependence and model both negtive and positive dependence. They are all members of the one-parameter θ families of copulas and will be denoted by $C_\theta(u, v)$. All copulas are strict for all values of θ, except for Clayton, that is only strict for $\theta \geq 0$.

Table 2.1 Archimedean copulas

	$C_\theta(u,v)$	$\phi_\theta(t)$	$\theta \in$	Limiting Case
[1]	$[\max(u^{-\theta}) + v^{-\theta} - 1, 0]^{\frac{-1}{\theta}}$	$\frac{1}{\theta}(t^{-\theta} - 1)$	$[-1, \infty)/\{0\}$	$C_{-1} = W, C_\infty = M$
[2]	$\frac{-1}{\theta} ln(1 + \frac{(e^{\theta u}-1)(e^{\theta v}-1)}{e^\theta - 1})$	$-ln\frac{e^{\theta t}-1}{\theta}(t^{-\theta} - 1)$	$(-\infty, \infty)/\{0\}$	$C_{-\infty} = W, C_\infty = M$
[3]	$\exp(-[(-\ln(u)^\theta) + (-\ln(v)^\theta)]^{\frac{1}{\theta}})$	$(-\ln(t))^\theta$	$[1, \infty)$	$C_1 = \prod, C_\infty = M$
[4]	$\frac{uv}{1-\theta(1-u)(1-v)}$	$\ln \frac{1-\theta(1-t)}{t}$	$[-1, 1)$	$C_0 = \prod, C_1 = \frac{\prod}{\Sigma - \prod}$

(Nelsen , 2006, p. 116 Table 4.1) where [1], [2], [3], [4] are the Clayton, Frank, Gumbel and AMH copula respectively.

Figures 2.2-2.3 show plots of 2000 random numbers drawn from copulas [1]-[4]. Using Eq.(2.34) the copula parameters are set so that $\tau_C = 0.3$. From the plots it is clear that Clayton exhibit lower tail dependence, Gumbel upper tail dependence, Frank no tail dependence and AMH lower tail dependence (with a parameter close to one).

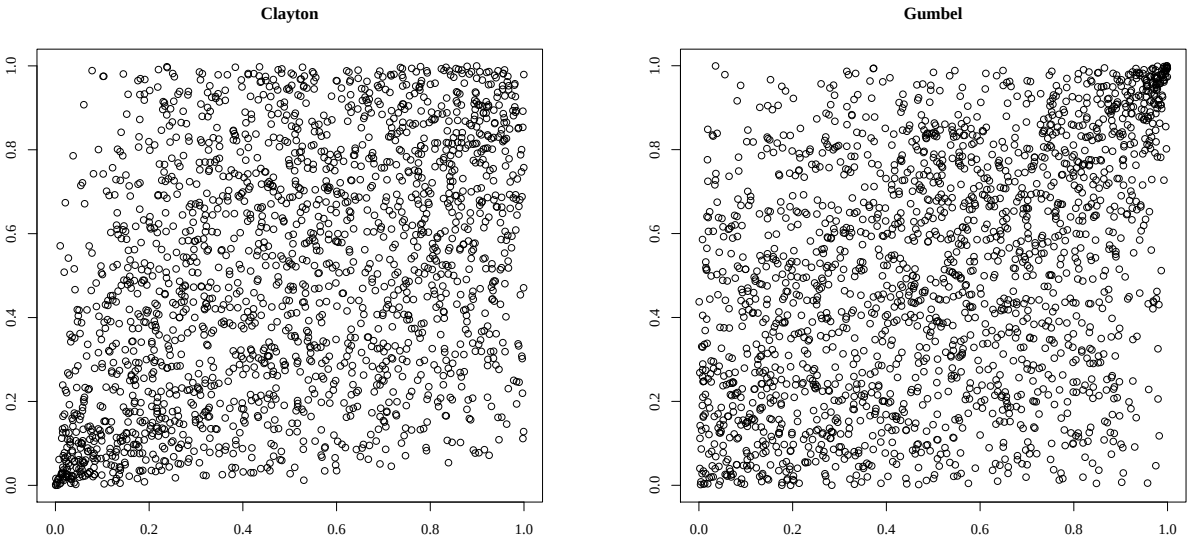


Figure 2.2 Random numbers from Clayton and Gumbel copula, $\tau_C = 0.3$

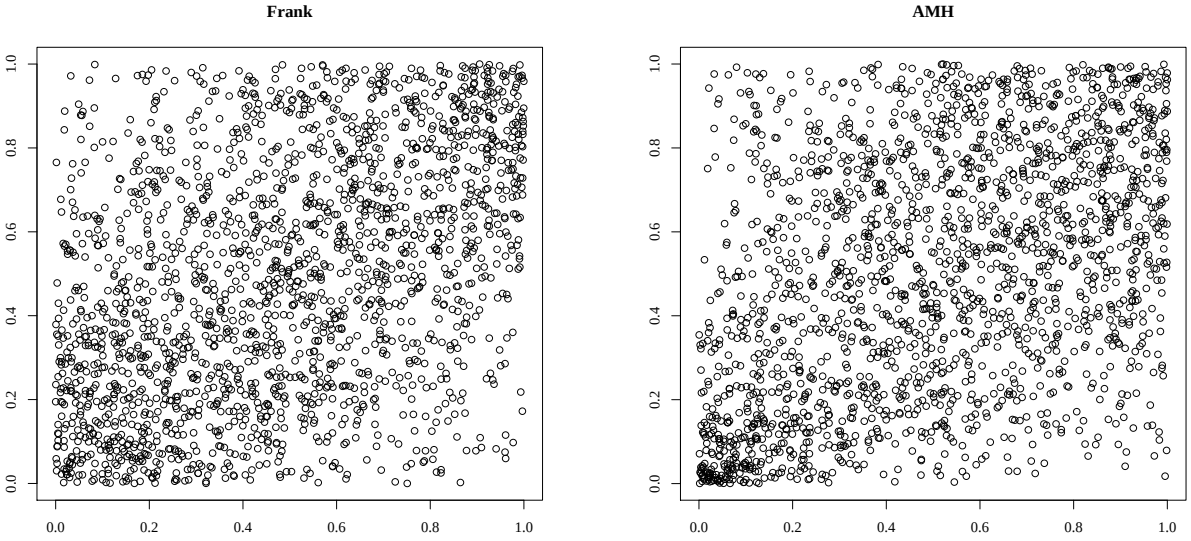


Figure 2.3 Random numbers from Frank and AMH copula, $\tau_C = 0.3$

For continuous margins the dependencies for upper tail dependence λ_u and lower tail λ_l respectively can be written as (Schmidt, 2007, Eq.15, p17),

$$\lambda_u = \lim_{q \to 0} P(X_2 > F_2^-(q) | X_1 > F_1^-(q)) = 2 + \lim_{q \to 0} \frac{C(1-q, 1-q) - 1}{q} \tag{2.35}$$

$$\lambda_l = \lim_{q \to 0} P(X_2 \leq F_2^-(q) | X_1 \leq F_1^-(q)) = \lim_{q \to 0} \frac{C(q, q)}{q} \tag{2.36}$$

For $\theta > 0$ Clayton has lower tail dependence, Gumbel has upper tail dependence for $\theta > 1$. AMH exhibits lower tail dependence for $\theta = 1$ and Frank has no tail dependence.

2.3.3 Copula parameter estimation

When using predefined copulas, there exists a collection of bivariate or multivariate distributions with desired marginal distribution, this is a consequence of Sklar's theorem (2.31). The Archimedean copulas chosen for modeling in this thesis have analytical expressions which makes it possible to estimate the copula parameters simply by inverting Eq. 2.34,

$$\hat{\theta} = \tau_C^{-1}$$

2.3.4 Goodness of fit

It is important to have a good measure of how well a copula model fits the data. The well-known and widely used χ^2 test does not work when replacing the marginal distributions with estimates of these. (Fermanian , 2005). When testing different copulas, the best fit among these can be decided using the log-likelihood value and Akaikes Information Criterion. However these methods do not give information of how well a model fit the original data. To find the answer to how well the copula fits the original data and give an answer to the hypothesis below, the Cramér von Mises method is used.

$$H_0 : C \in C_0$$

$$H_1 : C \notin C_0$$

where C_0 is a class of copula $C_\theta : \theta \in O$, O is an open subset in $\mathbb{R}^p$ for some integer $p \geq 1$, θ is the copula parameter. (Genest, Rémillard and Beaudoin , 2009, p.199)

2.3.4.1 Cramér von Mises method

The Cramér von Mises method measures the distance between the empirical and the estimated copula. The formula for Cramér von Mises test statistic is,

$$S_n = \sum_{i=1}^{n} \{C_n(\mathbf{u}_i) - C_{\theta_n}\}^2 \tag{2.37}$$

where C_n is the empirical copula, C_{θ_n} is the estimated copula and $\mathbf{u}$ is uniformly distributed variables. The empirical copula in d dimensions is calculated using Eq.(2.38).

$$C_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(U_{i1} \leq u_1, ..., U_{id} \leq u_d) \tag{2.38}$$

$$\mathbf{u} = (u_1, ..., u_d) \in [0,1]^d \tag{2.39}$$

(Genest, Rémillard and Beaudoin , 2009, p.201 Formula (1))

Fermanian (2005) comments that goodness of fit based on the empirical process "seem to be unpractical, except by bootstrapping". Therefore a bootstrap algorithm of the Cramér von Mises test is necessary. The following parametric bootstrap for calculating the an approximate p-value for the test based on S_n comes from Appendix A in Genest, Rémillard and Beaudoin (2009).

1. Compute C_n from Formula 2.38, estimate θ by first computing τ_K and then inverting it for the copula of interest.

2. Compute the value of S_n from Formula 2.37

3. For a large integer N, repeat the following steps for every $k \in \{1, ..., N\}$:

   (a) Generate a random sample $\mathbf{Y^*_{1,k}}, ..., \mathbf{Y^*_{n,k}}$ from C_θ and compute their associated rank vectors $\mathbf{R^*_{1,k}}, ..., \mathbf{R^*_{n,k}}$.

   (b) Compute $\mathbf{U^*_{1,k}} = \frac{\mathbf{R^*_{i,k}}}{n+1}$ for $i \in \{1, ..., n\}$ and let $C^*_{n,k}(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\mathbf{U^*_{i,k}} \leq \mathbf{u})$, $\mathbf{u} \in [0,1]^d$ and estimate $\theta^*_{n,k}$ as in step 1.

   (c) Compute $S^*_{n,k} = \sum_{i=1}^{n} \{C^*_{n,k}(\mathbf{U^*}_{i,k}) - C_{\theta^*_{n,k}}(\mathbf{U^*}_{i,k})\}^2$

4. An approximate p-value for the test is then given by $\frac{1}{N} \sum_{k=1}^{N} \mathbf{1}(S^*_{n,k} > S_n)$

CHAPTER 3. PROCEDURES and RESULTS

3.1 Introduction

According to the definition of copula, the margins could be fitted first and secondly the dependence structure can be modelled using Archimedean copulas. The Archimedean copulas used were the Clayton, Frank, Gumbel and AMH copulas. The properties of the copulas are presented in Chapter 2. When a bivariate distribution was obtained the expected values were compared to the empirical data. The significance level used throughout the modeling was $\alpha = 0.05$ and the number of simulations in the bootstrap algorithms was $N = 1000$.

3.2 Data

The data was received from a large Danish insurance company representing 74770 unique customers i having two products during $T_{i,j}$ number of years. The products represents building insurance and content insurance, and the time spanning over $1 \leq j \leq 5$, making the total number of records in the data 306196. The columns of the data consisted of the number of claims customer i had made during year T_{ij}. The modeling will be made for customers having the insurances year one $T_{i,1}$. The rest of the data will be used for validation of τ_K in Section 3.3.3. The number of claims made in the products year $T_{i,1}$ will be referred to as Build and Cont in this thesis. The histograms for Build and Cont can be seen in Figure 3.1.

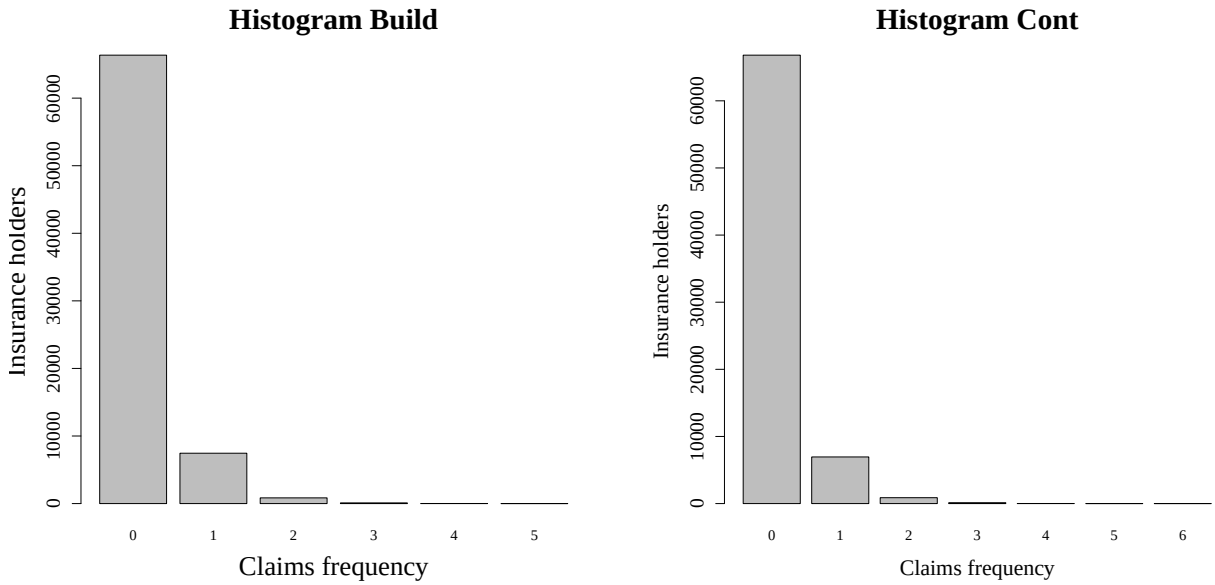


Figure 3.1 Histograms for Build and Cont data

3.3 Modeling Marginal Distributions

To understand which distribution would provide a suitable fit the histograms in Figure 3.1 were studied. The estimated mean and variance for the claims frequency for Build and Cont were calculated and are presented in Table 3.1.

Table 3.1 Mean and variance estimations

Variable	Mean	Var
Build	0.12699	0.14374
Cont	0.12240	0.14423

The variance is sligthly larger than the mean for both Build and Cont as seen in Table 3.1. This implies that an overdispersion is present and the NB or any of the zero-inflated distributions should provide a better fit than the Poisson distribution. Parameter estimations were made for the Poisson, Zero-inflated Poisson, Negative Binomial and Zero-inflated Negative Binomial distributions.

3.3.1 Goodness of fit

One might be tempted to consider the negative binomial and the zero-inflated negative binomial as nested models of each other, with the zero-inflated version having one more parameter than the non-inflated one. This will lead to using a likelihood ratio test to decide which model is the most appropriate one. Although using the definition of the zero-inflated negative binomial according to Eq. 2.22 and Eq. 2.23 the negative binomial distribution already consists of a Bernoulli variable. When adding another Bernoulli variable to create the zero-inflated version, the Bernoulli probabilities are non-identifiable, and is counted as one parameter. This results in that the negative binomial and the zero-inflated negative binomial will have equal amount of parameters and they will be compared using AIC. The ML-estimated distribution parameters and χ^2 p-value, for the parametric marginal distributions for Build and Cont are presented in Table 3.2. The p-values for the negative binomial and the zero-inflated negative binomial were larger than the significance level $\alpha = 0.05$ for both Build and Cont. The $\ell\ell h$ was largest and the AIC-value lowest for the negative binomial distribution for Build and the zero-inflated negative binomial distribution for Cont. Therefore these distributions were selected for modeling the Build respectively Cont. The differences between the AICs were very small and can be seen in (3.1) and (3.2) (they are not seen in Table 3.2).

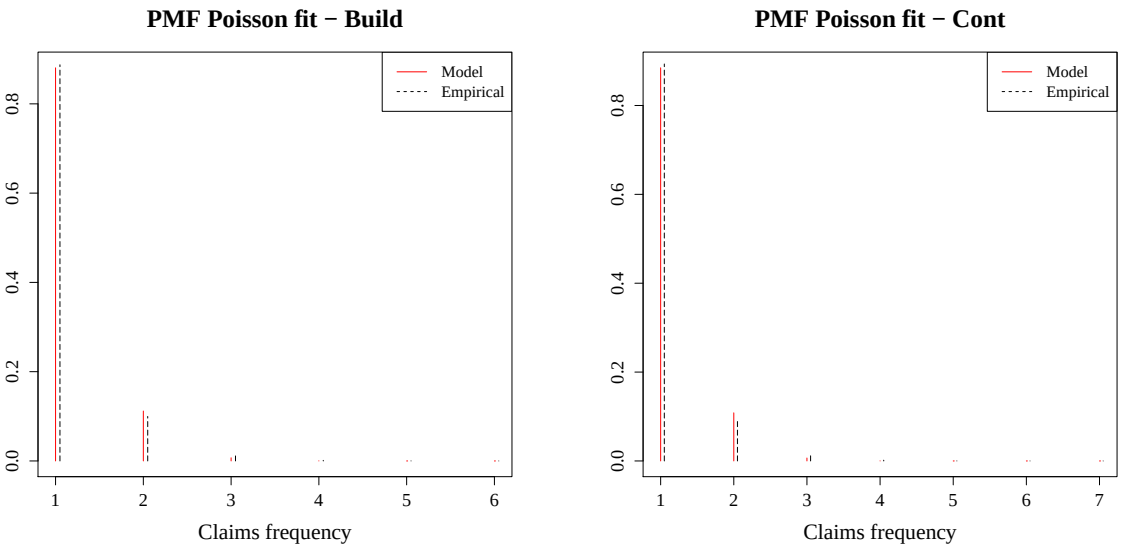
$$AIC^{NB}_{Build} - AIC^{ZINB}_{Build} = -2.8630 \cdot 10^{-6} \tag{3.1}$$

$$AIC^{NB}_{Cont} - AIC^{ZINB}_{Cont} = 4.0745 \cdot 10^{-4} \tag{3.2}$$

To get a graphical presentation of the goodness of fit of the distributions, plots of the PMFs and CDFs for the empirical data and the model are displayed in Figures 3.2:3.9.

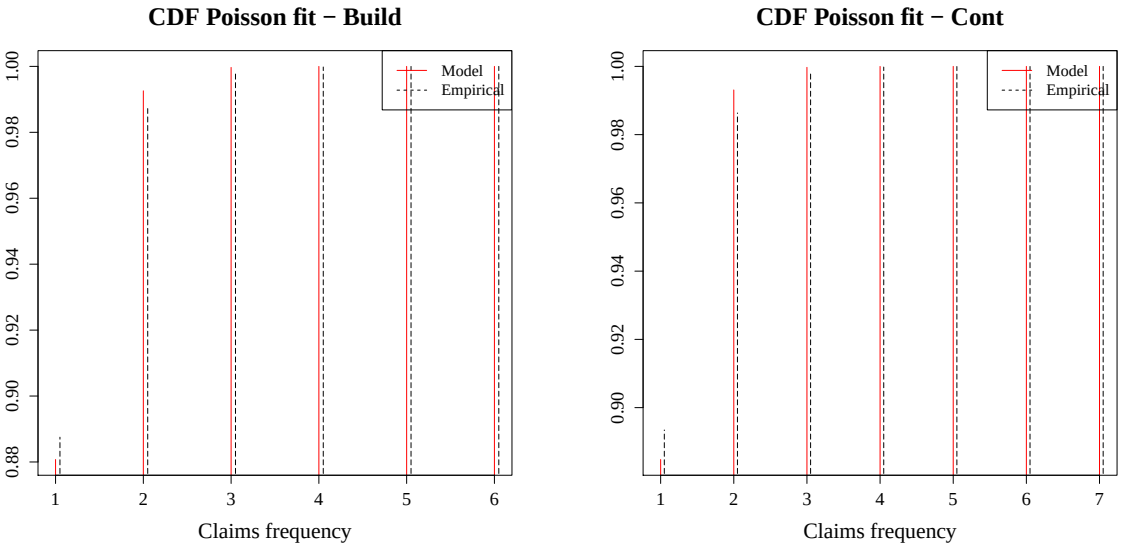
Table 3.2 Parametrical Marginal distributions

Distribution	Est. parameters	Std error	*llh*	AIC	χ^2-statistic	p-value
Poi						
Build	$\lambda = 0.1270$	0.001303	-29898	59798	634.8	0
Cont	$\lambda = 0.1224$	0.001279	-29270	58543	1191	0
ZIP						
Build	μ =0.2481	0.007388	-29680	59364	29.84	0
	σ =0.4880	0.01443				
Cont	μ =0.2842	0.008075	-28910	57825	34.47	0
	σ =0.5693	0.01152				
NB						
Build	$\mu_{NB} = 0.1270$	0.001386	-29668	59340	0.1147	0.7348
	$n = 0.9657$	0.05962				
Cont	$\mu_{NB} = 0.1224$	0.001388	-28892	57788	0.2441	0.6212
	$n = 0.6883$	0.03547				
ZINB						
Build	$\mu_{ZI} = 0.1270$	0.001386	-29668	59340	0.1994	0.9953
	$n = 0.9657$	0.05962				
Cont	$\mu_{ZI} = 0.1224$	0.001388	-28892	57788	1.798	0.7729
	$n = 0.6887$	0.03551				



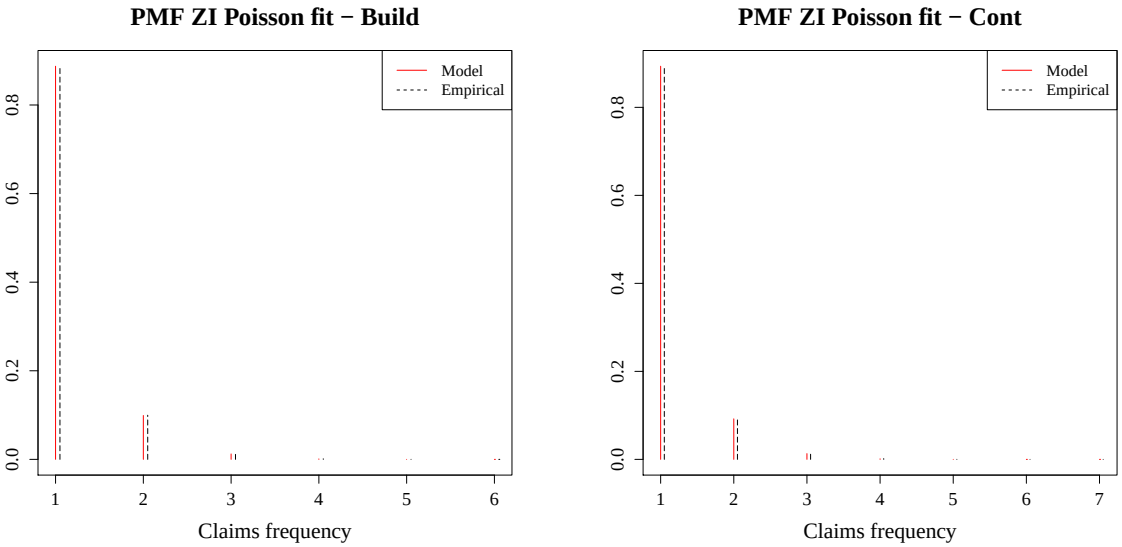
(a) Build (b) Cont

Figure 3.2 PMF for the Poisson fit



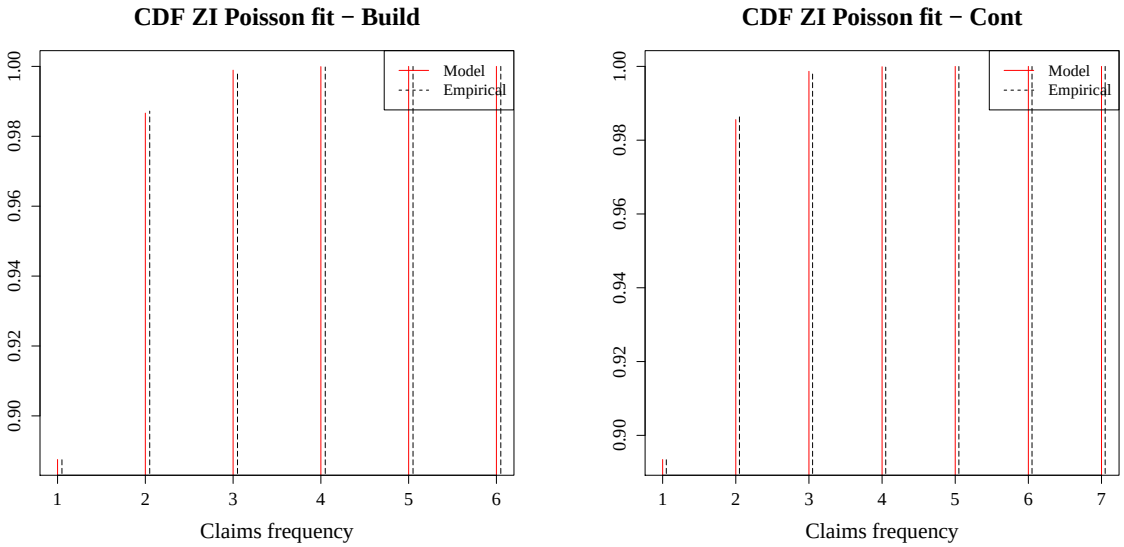
(a) Build (b) Cont

Figure 3.3 CDF for the Poisson fit



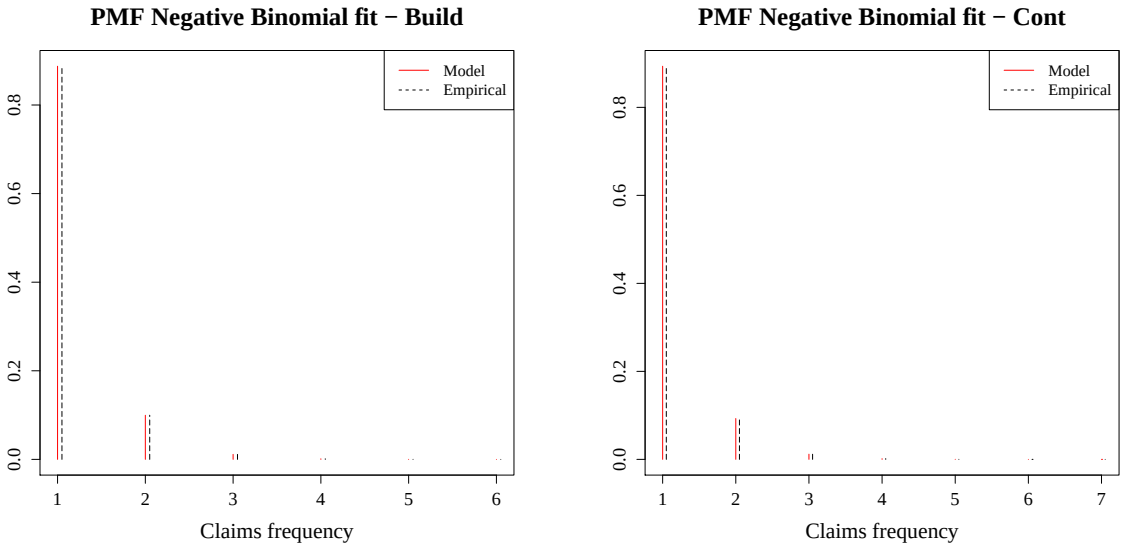
(a) Build (b) Cont

Figure 3.4 PMF for the zero-inflated Poisson fit



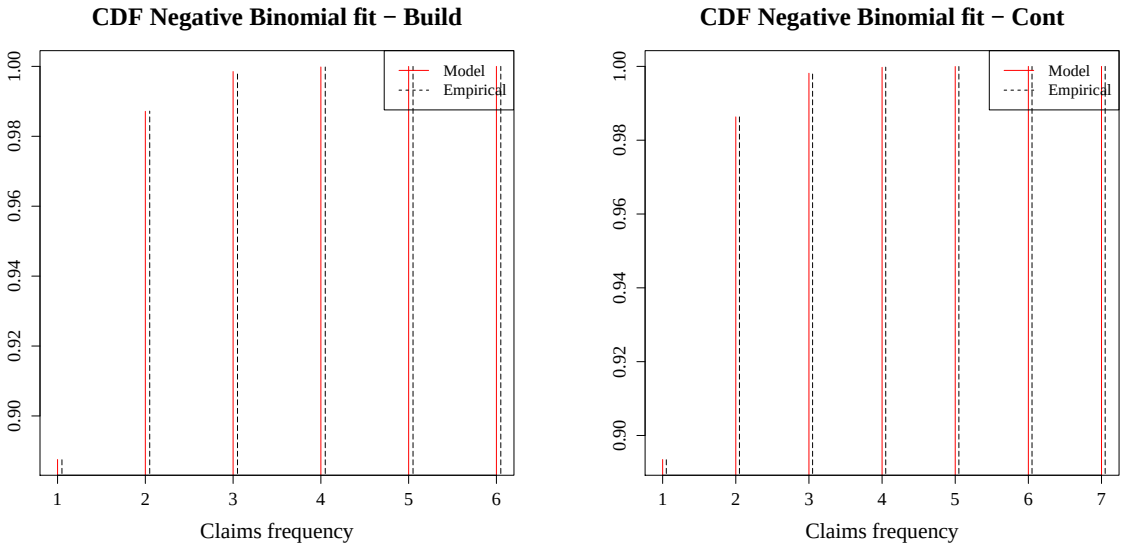
(a) Build (b) Cont

Figure 3.5 CDF for the zero-inflated Poisson fit



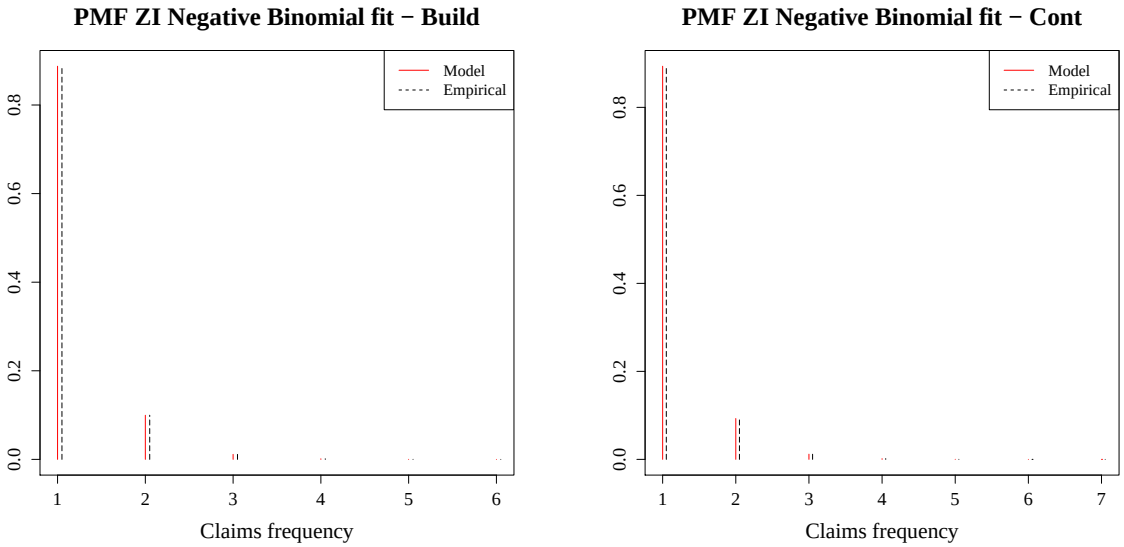
(a) Build (b) Cont

Figure 3.6 PMF for Negative Binomial fit



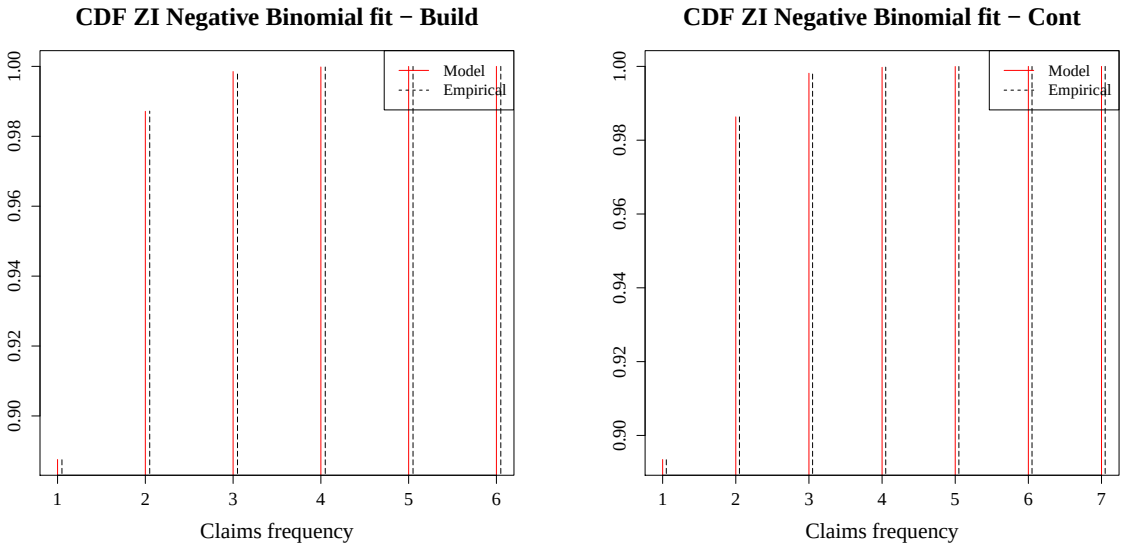
(a) Build (b) Cont

Figure 3.7 CDF for Negative Binomial fit



(a) Build (b) Cont

Figure 3.8 PMF for zero-inflated Negative Binomial fit



(a) Build

(b) Cont

Figure 3.9 CDF for zero-inflated Negative Binomial fit

The plots validates the conclusions made above, although it is difficult to distinguish which of the negative binomial and zero-inflated negative binomial provides the best fit for the margins.

3.3.2 Dependence

To see if there existed a dependence between the margins a few tests were performed. First a contingency table of the observed values can be seen in Figure 3.10. The lower triangular shape implies that a lower dependence exist.

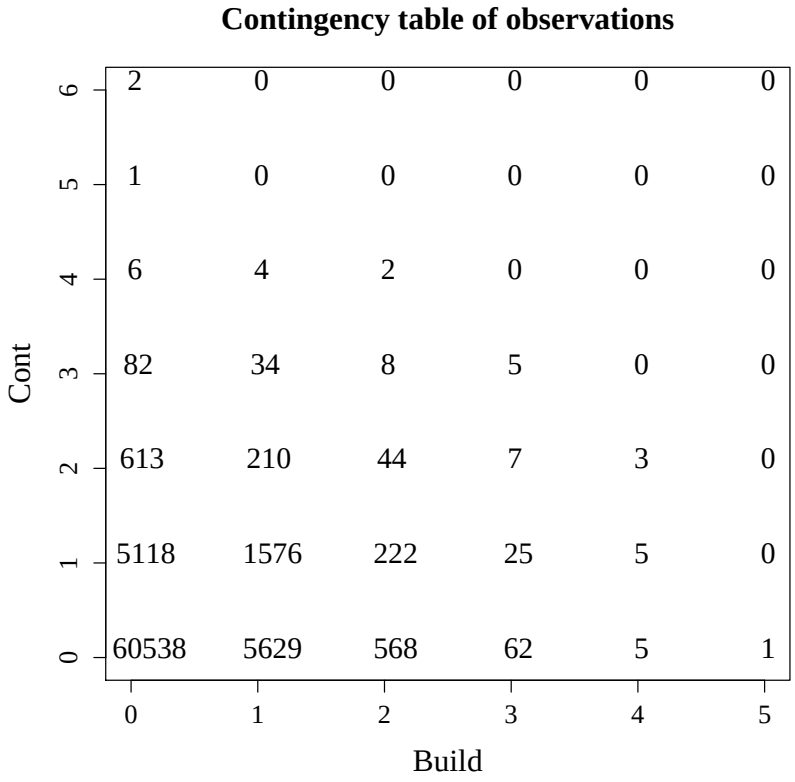


Figure 3.10 Observations for Build and Cont

A comparison between the empirical joint PMF and the independent joint PMF was made using the contingency tables in Figure 3.11. The empirical PMF in was calculated by dividing the observed frequencies seen in Figure 3.10 elementwise with the total number of customers $i = 74770$. The independent PMF was calculated by first calculating the margins PMFs using the number of observations in Build and Cont respectively and dividing with the total number of customers. Then the margins PMFs were multiplied using Eq. A.4 in Appendix A. If there exists a dependence the joint and the independent PMFs should not be equal which they weren't.

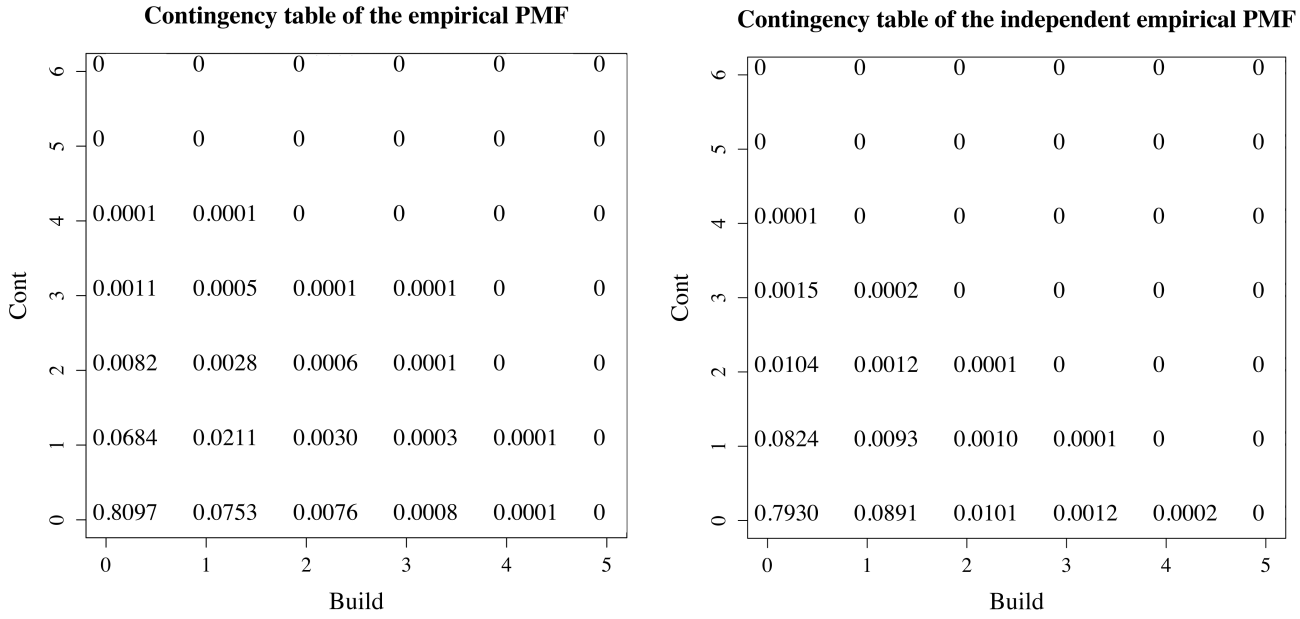


Figure 3.11 The empirical joint PMF fo Build and Cont. Joint on to the left and independent joint to the right.

3.3.3 Kendall's Tau

Because of the invariance property of rank correlation described in Section 2.1.3, the parametric and empirical margins will have the same τ_K. Therefore the parametric margins are used for modeling henceforth. The estimated $\hat{\tau}_\beta$ for discrete data was calculated from the parameteric margins to,

$$\hat{\tau}_\beta = 0.17099$$

$\hat{\tau}_\beta$ shows that the dependence between Build and Cont is positive. A non-parametric bootstrap was performed for the parametric margins. Assuming that the model was correct, it can be seen that $\hat{\tau}_\beta$ was centered in the plot of bootstrapped values and the variance was quite small as would be preferrable. The plot of the bootstapped values and the quantiles can be seen in Figure 3.12.

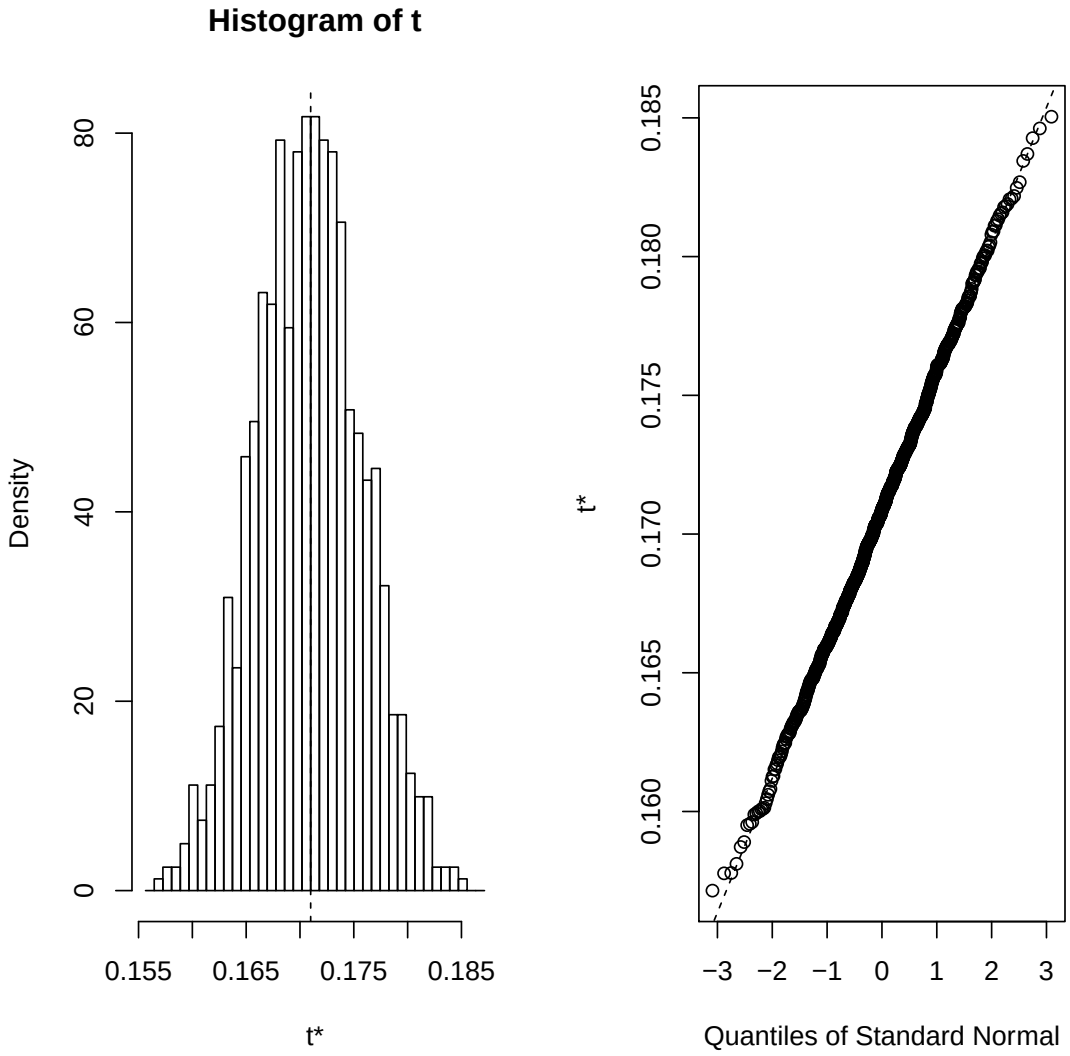


Figure 3.12 Non-parametric bootstrap for the parametric margins, $t = \tau_\beta$

The 95th and 5th percentile, $p_{0.05} \leq \hat{\tau}_\beta \leq p_{0.95}$, were calculated to,

$$0.16308 \leq \hat{\tau}_\beta \leq 0.17884$$

Again assuming that the model was correct, further validation could be attained by estimating τ_β for the empirical data all years. The dataset received from the insurance company included information regarding the number of claims made for customers up to $T_{i,j}, 1 \leq j \leq 5$ years. For customers where the data did not exist, the data from the previous year was included. The

estimated $\hat{\tau}_\beta^j$ is presented below,

$$\hat{\tau}_\beta^2 = 0.17674$$

$$\hat{\tau}_\beta^3 = 0.17484$$

$$\hat{\tau}_\beta^4 = 0.17475$$

$$\hat{\tau}_\beta^5 = 0.16439$$

The $\hat{\tau}_\beta^j$ estimates for the different years was inside the interval of the 5th and 95th percentile of $\hat{\tau}_\beta^1$. This suggests that under the assumption that the model is correct, $\hat{\tau}_\beta^1$ should be a good estimate for year one.

3.4 Modeling dependence - Copula

3.4.1 Copula Parameters

When $\hat{\tau}_\beta^1$ was estimated as described above the copula parameters were estimated using inverse τ_β described in Section 2.3.3. The estimates of the copula parameters $\hat{\theta}_C$, $\hat{\theta}_F$, $\hat{\theta}_G$ and $\hat{\theta}_{AMH}$, for Clayton, Frank, Gumbel and AMH and a non-parametric bootstrap was performed, see Figures 3.13 for the plots.

The 95th and 5th percentiles for the copula parameters $p_{0.05} \leq \hat{\theta} \leq p_{0.95}$ were calculated to,

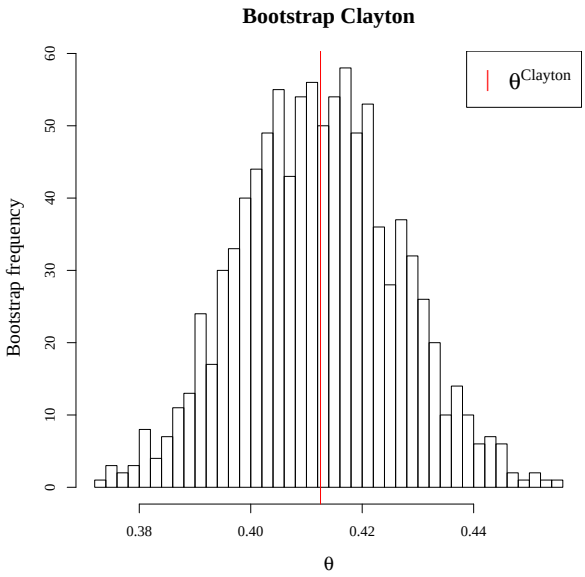
$$\hat{\theta}_C \quad 0.38971 \leq 0.41252 \leq 0.43557$$

$$\hat{\theta}_F \quad 1.50022 \leq 1.57653 \leq 1.65268$$

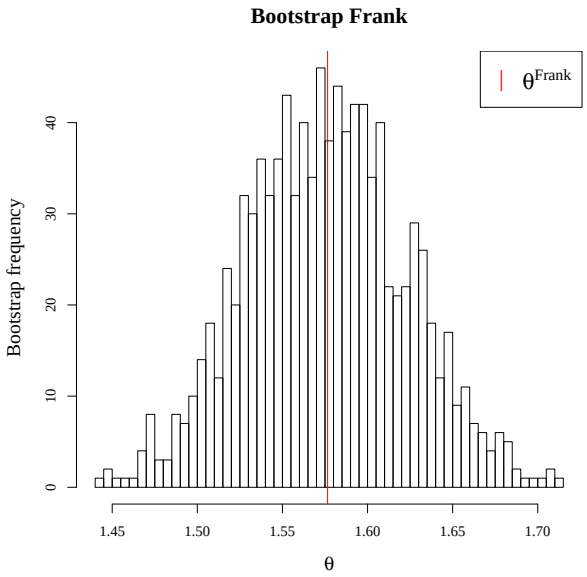
$$\hat{\theta}_G \quad 1.19485 \leq 1.20626 \leq 1.21779$$

$$\hat{\theta}_{AMH} \quad 0.60813 \leq 0.63166 \leq 0.65447$$

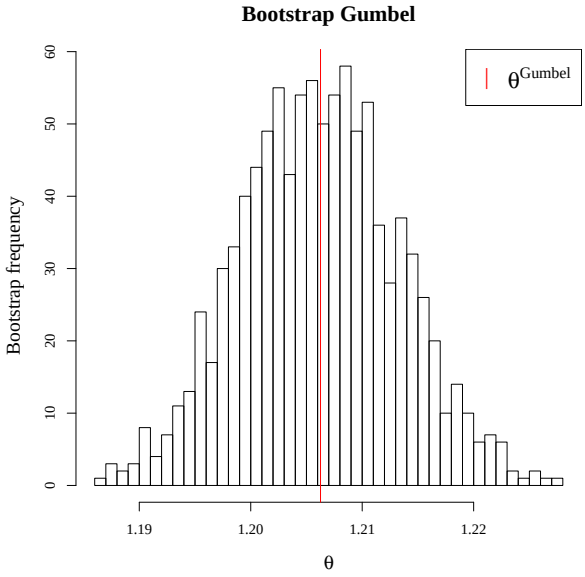
Assuming that the model was correct the parameter estimates were centered in the plot of bootstrapped values and the variance were quite small as for $\hat{\tau}_\beta^1$.



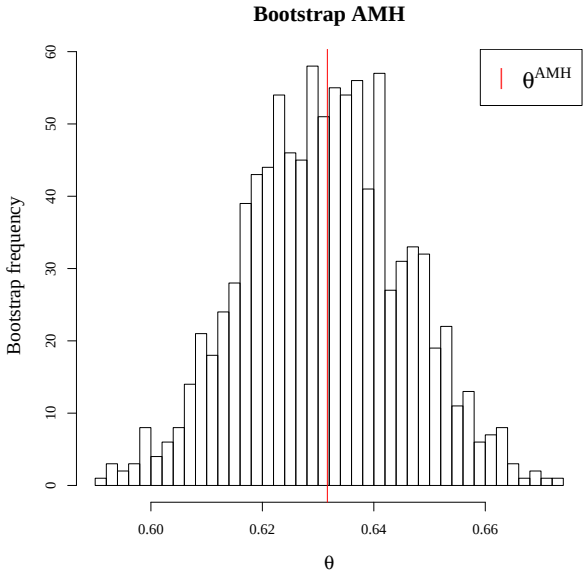
(a) Clayton



(b) Frank



(c) Gumbel



(d) AMH

Figure 3.13 Non-parametric bootstrap for the copula parameters

3.4.2 Goodness of Fit - Copula

The Cramér von Mises bootstrap described in Section 2.3.4.1, was perfomed to calculate the p-value for the copula parameters. The estimated copula parameter $\hat{\theta}$, $\ell\ell h$ value, AIC-value and p-value from the Cramér von Mises test are presented in Table 3.3.

Table 3.3 Parameter estimates and summary statistics for Build and Cont

Copula	$\hat{\theta}$	$\ell\ell h$	AIC	p-value (CvM)
Clayton	0.41252	19970	-39939	0
Frank	1.57653	32047	-64091	0.011
Gumbel	1.20626	39539	-79077	0.774
AFM	0.63166	26099	-52196	0

The copula with the largest $\ell\ell h$ and the lowest AIC was Gumbel, therefore this copula would be the best choice among the copulas. The only copula with p-value larger than the significance level was Gumbel, this fact together with the results form the AIC proposes that the Gumbel copula provided the best dependence structure for the data among the considered models. Plots of the Cramér von Mises bootstraps are shown in Figures 3.14:3.15.

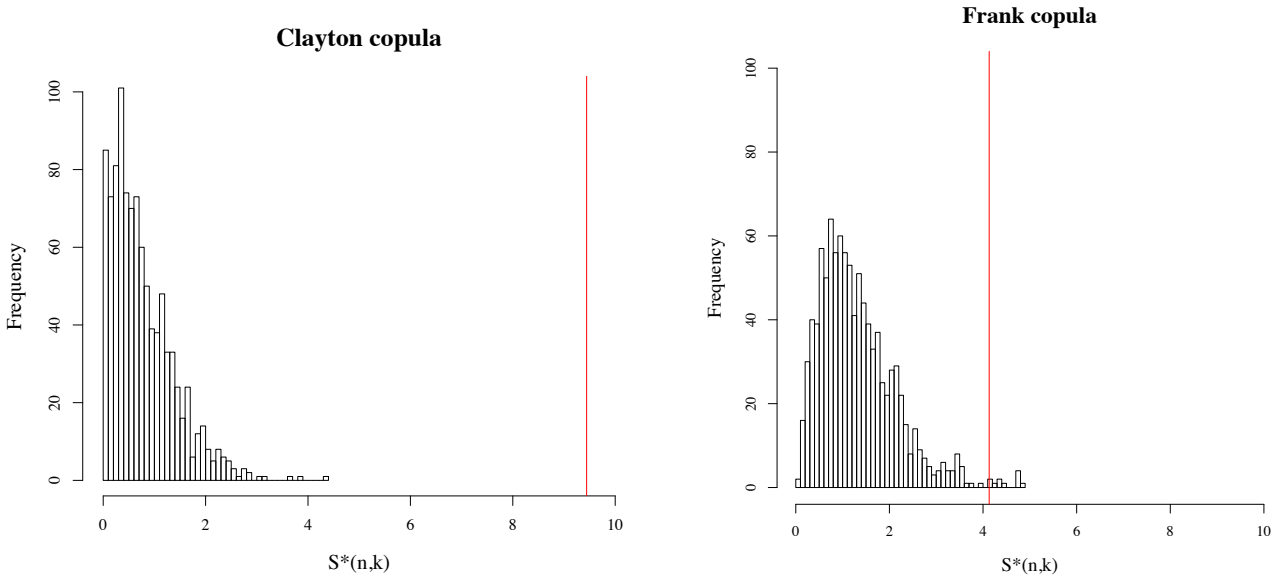


Figure 3.14 Cramér von Mises parametric bootstrap for the Clayton and Frank copulas, the red line is the original S_n

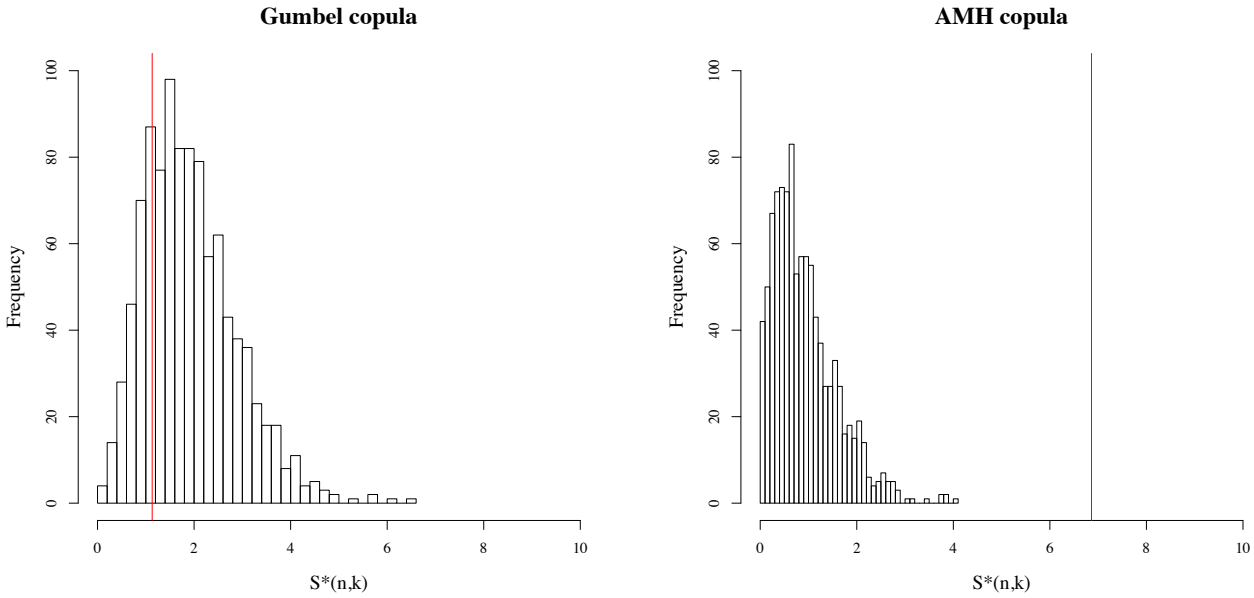


Figure 3.15 Cramér von Mises parametric bootstrap for the Gumbel and AMH copulas, the red line is the original S_n

3.4.3 Fitted bivariate distribution

For further validation of the model a bivariate CDF was created using the function **pMvdc()** in the Copula package in R. The new joint distribution consisted of one negative binomial, one zero-inflated negative binomial margin and a Gumbel copula with their respective parameters presented above. The PMF for the model was calculated using the method in Section 2.3.1.2. The PMF for the model respective the empirical data are found in Figure 3.16. The PMFs for the model and the empirical data are rather alike. This suggests that the model provided a good fit.

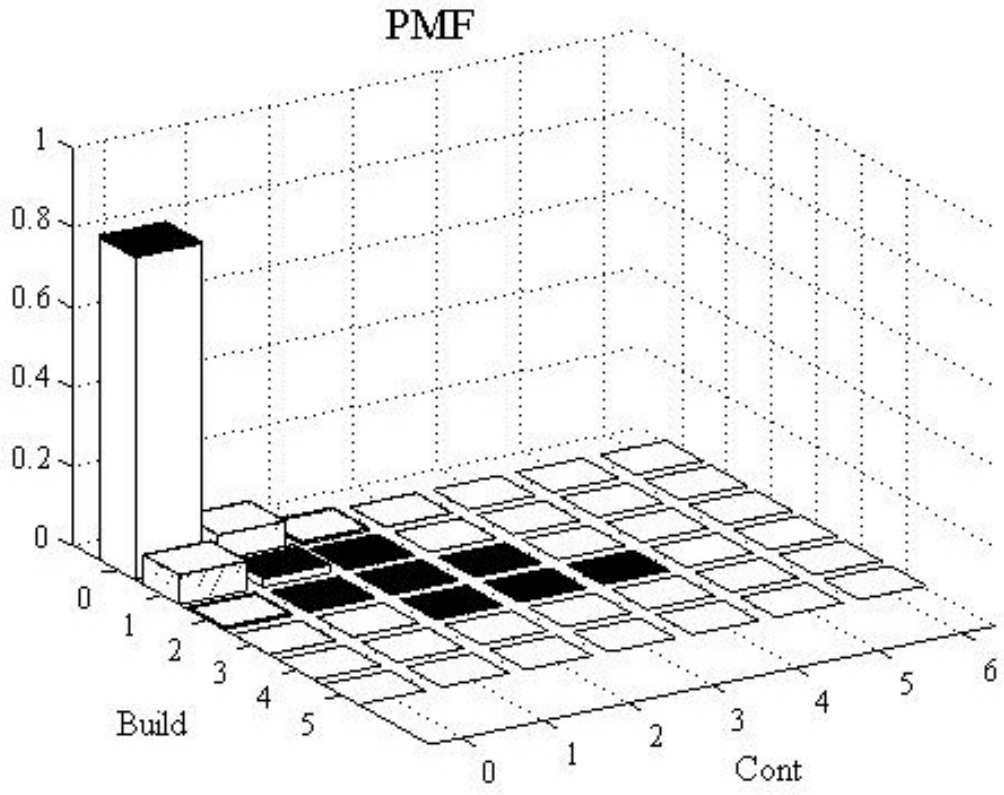


Figure 3.16 PMF for Build and Cont. Empirical PMF white, model PMF black.

3.4.4 Conditional distributions

To validate the model further the conditional distributions for Build conditional on Cont and vice versa were calculated. The conditional distributions for the empirical data and the model were calculated using Eq. (3.3)-(3.4). Let X be a random variable for Build taking values $j = 0, ..., 5$ and Y a random variable for Cont taking values $k = 0, ..., 6$ where i, j is the empirical number of claims made in Build and Cont respectively. Then Build conditional on Cont is presented in Eq.(3.3) and Cont conditional on Build in Eq.(3.4).

$$p_{X|Y}(j;k) = P(X = j|Y = k) = \frac{p_{X;Y}(j;k)}{p_Y(k)}, \qquad j = 0, ..., 5; k = 0, ..., 6 \tag{3.3}$$

$$p_{Y|X}(k;j) = P(Y = k|X = j) = \frac{p_{Y;X}(k;j)}{p_X(j)}, \qquad k = 0, ..., 6; j = 0, ..., 5 \tag{3.4}$$

The empirical PMFs for the joint distribution and the margins were the ones defined in Section 3.3.2. The joint PMF for the model was calculated according to Section 3.4.3. The

marginal PMFs for the model were calculated using Eq. A.2.

The models conditional PMFs compared to their empirical counterparts for Build are displayed in Figures 3.17-3.20 and in Figures 3.21-3.23 for Cont.

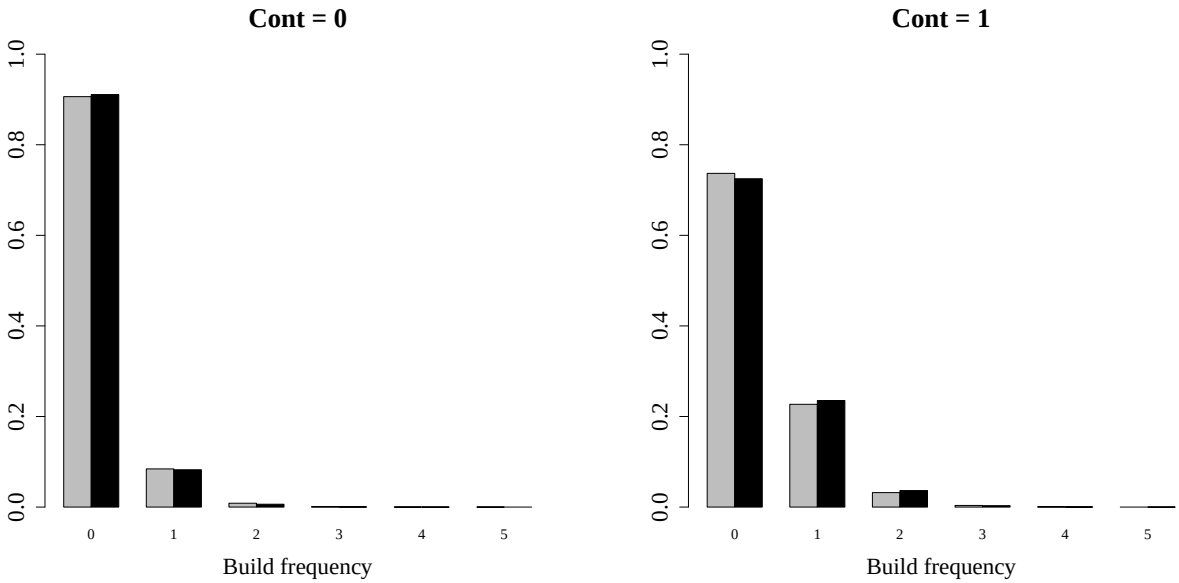


Figure 3.17 Build conditional on Cont = 0, 1. Grey is the empirical value and black is the models value

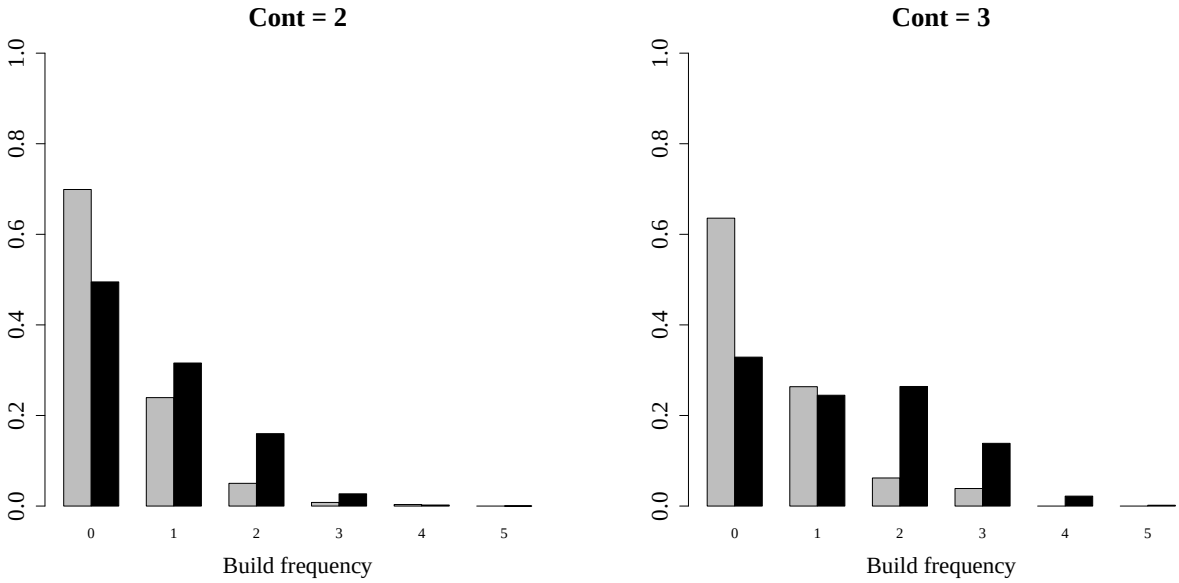


Figure 3.18 Build conditional on Cont = 2, 3. Grey is the empirical value and black is the models value

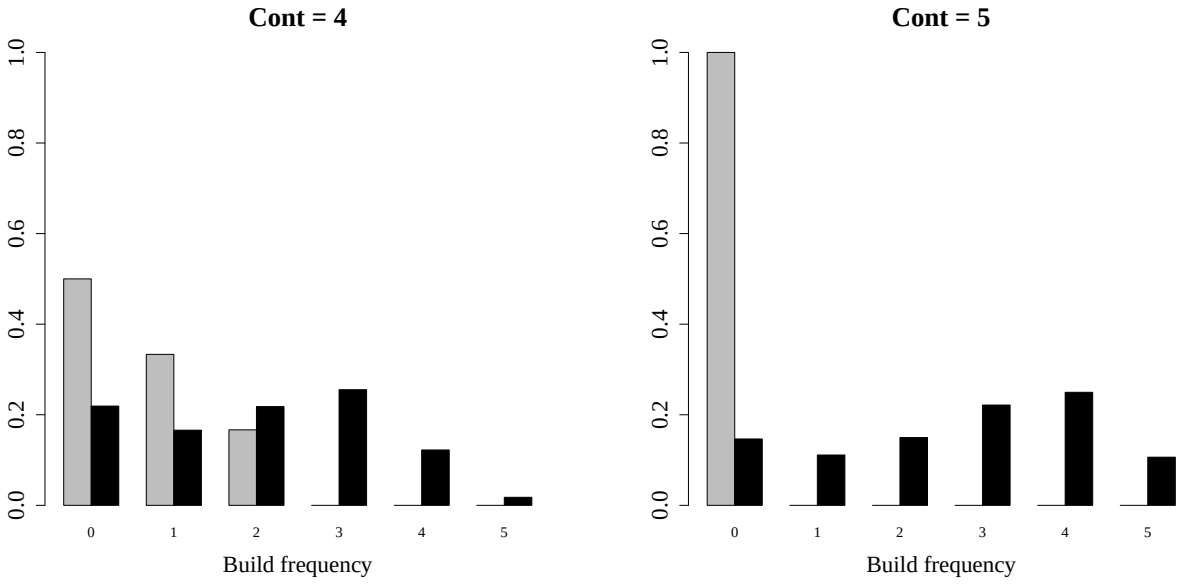


Figure 3.19 Build conditional on Cont = 4, 5. Grey is the empirical value and black is the models value

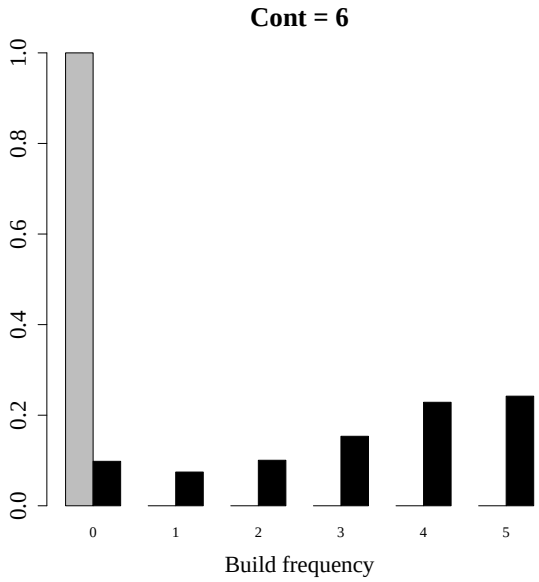


Figure 3.20 Build conditional on Cont = 6. Grey is the empirical value and black is the models value

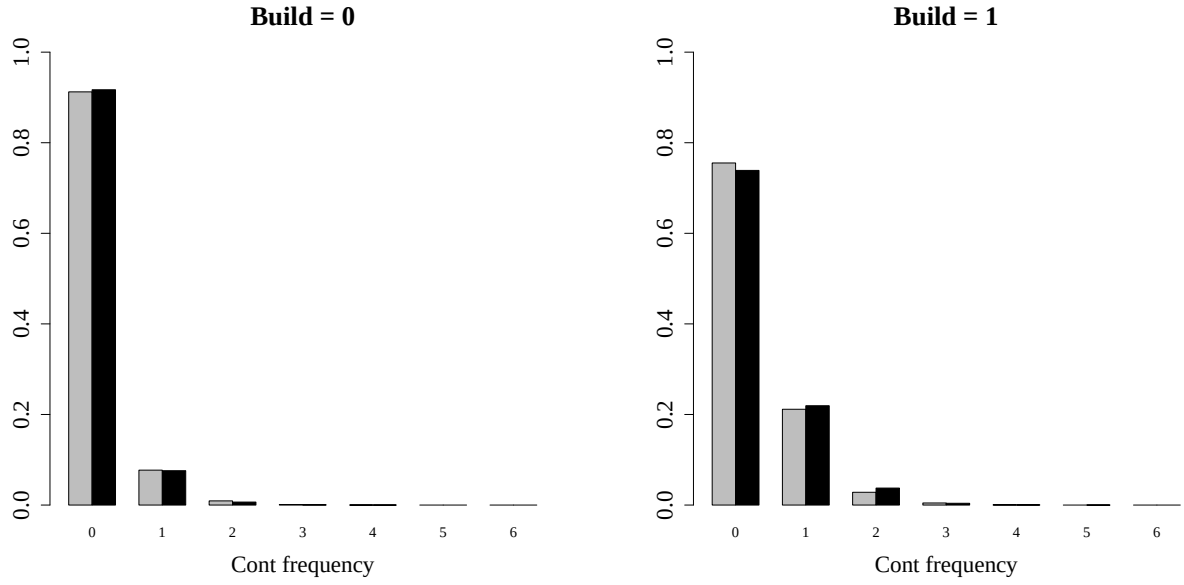


Figure 3.21 Cont conditional on Build = 0, 1. Grey is the empirical value and black is the models value

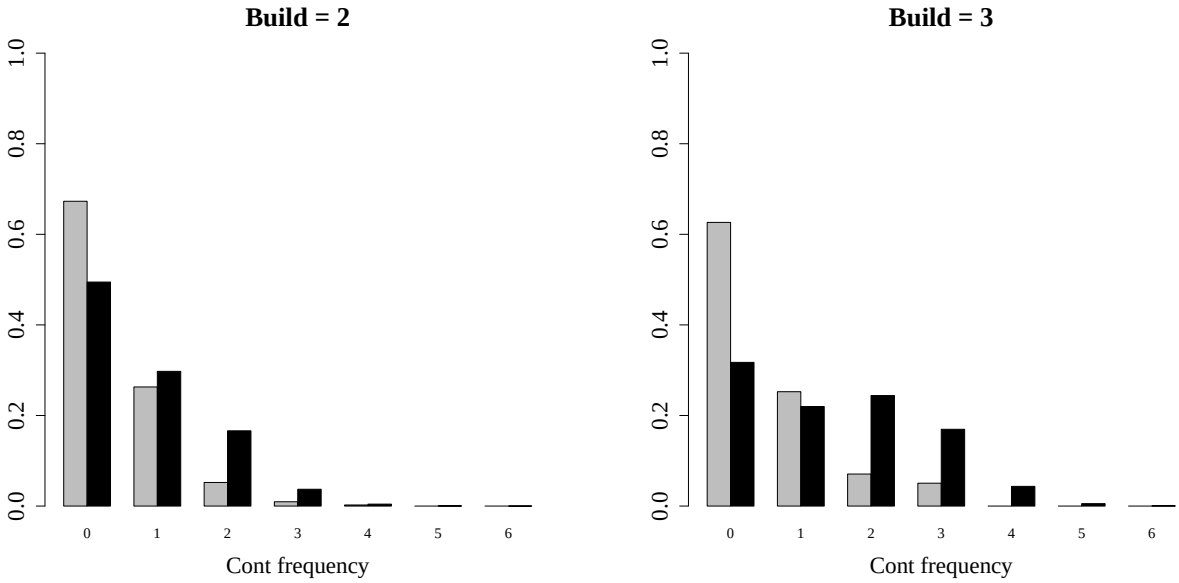


Figure 3.22 Cont conditional on Build = 2, 3. Grey is the empirical value and black is the models value

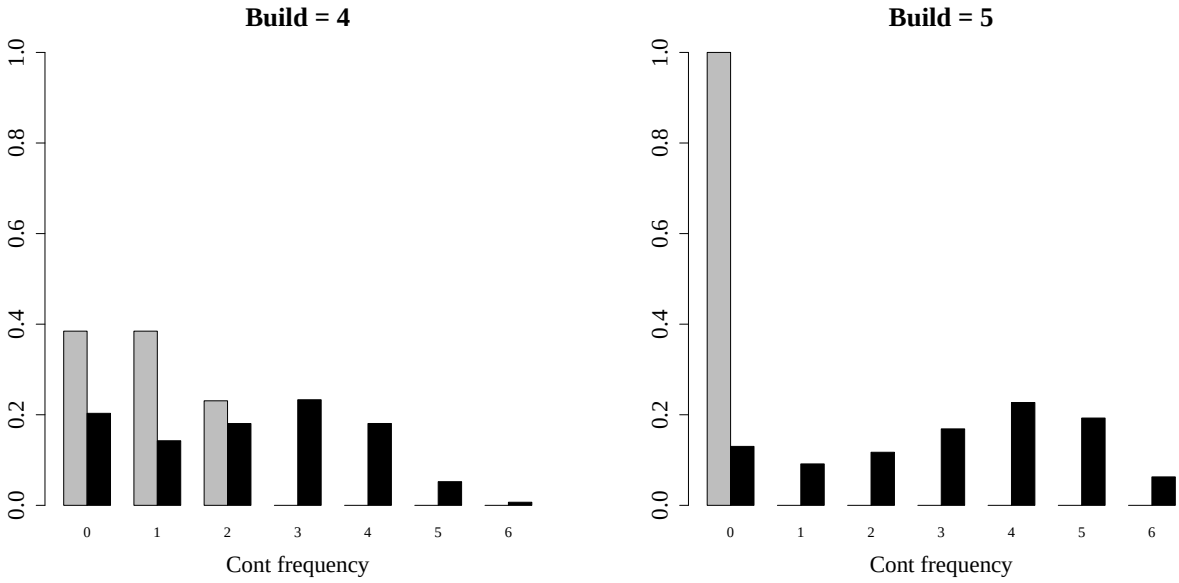


Figure 3.23 Cont conditional on Build = 4, 5. Grey is the empirical value and black is the models value

It is obvious that the empirical and fitted values are most alike when conditioning on the other distributions zero and one see Figures 3.17 and 3.21. The largest difference between the empirical and fitted PMFs are when for Build conditioning on Cont = 6 and for Cont conditioning on Build = 5, see Figures 3.20 and 3.23. Looking at the definition of the conditional PMF in Eq.(A.5) in Appendix A, there is an elementwise division between the joint distribution and the marginal distribution. The small differences in the joint PMF are enhanced when dividing with the margins PMFs (which also contains small differences). This is explained by the fact that there were fewer observations for the large number of claims and therefore it demands a greater exactness when modelling these numbers.

3.4.5 Expected number of claims

The conditional expected number of claims for Build and Cont are presented in Figure 3.24. The conditional expected values indicate that, as the conditional distributions did above, that the model has the best fit in 0 and 1. Although it looks like the empirical values and the model have very similar results in Cont conditioned on Build = 4 in Figure 3.24, this is not the case. Actually this is only a coincidence since multiplying the frequencies in the conditional distributions in Figure 3.23 with the numbers 0-6 evens out the total expected values, although the distributions were quite different.

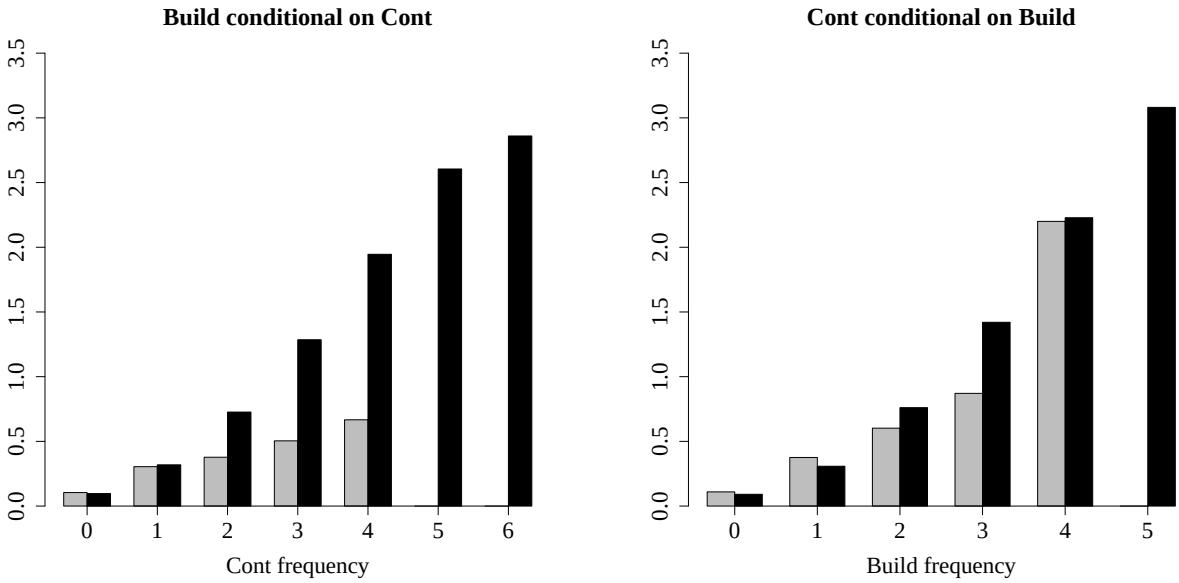


Figure 3.24 Conditional expected number of claims. Grey is the empirical value and black is the models value. Build conditional on Cont to the left, Cont conditional on Build to the right.

The values for Figures 3.16-3.24 are presented in table form in Appendix B.

CHAPTER 4. SUMMARY AND DISCUSSION

In this thesis a copula method was considered to model a bivariate distribution for the number of insurance claims. The goal was to see if information regarding the number of claims in a customers insurance had impact on the number of claims made in the customers other insurance.

The margins were modeled using a set of discrete distributions that were able to capture specific characteristics of data with huge number of zeros. The marginal parameters were estimated using the ML method. To decide the best fit amongst the marginal models a comparison of the AIC-values was made and a χ^2 goodness of fit test was performed. The results were that the negative binomial distribution provided a good fit for Build and the zero-inflated negative binomial distribution provided a good fit for Cont. The copula parameters were estimated using inverse τ_K. Then Cramér von Mises test with a parametric bootstrap showed that the best copula fit was the Gumbel copula.

It is not certain that the Gumbel copula actually provided the best dependence structure for all points in the discrete data. The Gumbel copula provided the best fit in (0,0) and therefore Cramér von Mises test found that Gumbel was the best fit. The reason for this is that Cramér von Mises method measures the distance between the estimated and empirical copula in a point. For example there is a large number of observations in point (0,0) and adding all of them will create a large quantity for Cramér von Mises in this point. The absolute differences between the empirical copula and the estimated copula for the parametric margins are displayed in Figure 4.1.

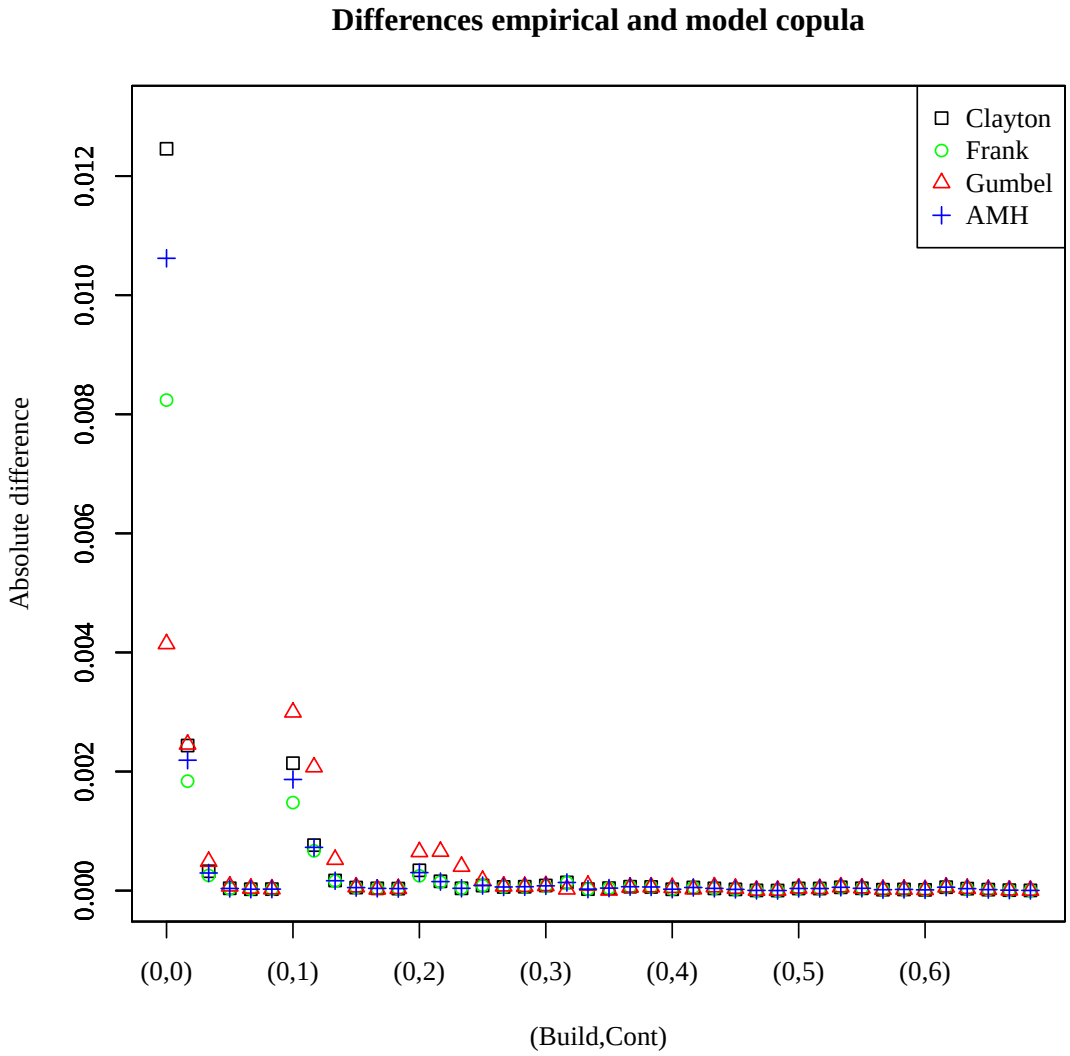


Figure 4.1 Absolute differences between the empirical and estimated copula. Between the points (0,0) and (0,1) lies the points (1,0), (2,0).. etc.

In Figure 4.1 for example the number of observations in $(0,0) = 60538$, $(0,1) = 5629$, $(0,2) = 568$. The rest of the observations are found in Figure 3.10 in Section 3.3.2. In Figure 4.1 it is obvious that the Gumbel copula provided the best fit amongst the copulas tested in point (0,0), that is it had the smallest difference between the empirical and esimated copula. When looking at the other points where the other copulas had reasonable large differences, the Gumbel copula actually had the greatest differences. Because there were extremely many observations

in this point this value becomes much smaller than say for the Clayton copula, where the distance was a bit larger. This number is then squared in Cramér von Mises test which makes it even larger. Even if the other copulas had smaller differences in the other points this will not be noticed in the total test statistic S_n in Cramér von Mises test. This is because there are not as many observations in the other points. This is an important point that has to be considered when working with discrete data.

When having this model an application in insurance business can be to predict the number of claims by using the conditional expectation in Section 3.4.5. This can be made because a risk profile is created by this modeling of dependence. It tells how many claims is it probable that a customer makes in one insurance, having made a certain numbers of claims in the second insurance.

4.1 Further work

To get better parameter estimates of the copula parameter a ML-estimation can be made. A warning when using ML-estimation is that the functions in R for calculating the PMF for the copulas have problem evaluating the PMF expressions at the boundary points. (Yan, 2007, p.12). In the discrete case it was even more difficult since the calculations of PMF in the copula package uses the continuous definition and there is no alternative in the discrete case. This problem was experienced during the modeling in this thesis and therefore no ML-estimations of the copula parameters could be performed. A suggestion to find better models of the margins is to include covariates in them. An easy implementation can be made using the example code in Appendix A in Yan (2007). A proposition to be able to avoid the downsides using discrete margins is to make a continuous extension of the discrete data. An example of this can be found in Denuit (2005).

Now that it has been shown that a dependence structure was attained by using a copula approach, it would be interesting to compare this model to the multivariate credibility model made by Thuring (2011). Another area to look deeper into is the multivariate modeling for customers holding several insurances. This is possible since copulas can be used to describe dependence in several dimensions.

APPENDIX A. ADDITIONAL MATERIAL

A.1 Probability theory

Some general relationships for random variables are presented below.

The joint probability mass function (PMF) for two discrete random variables X and Y and marginal distribution are defined in Eq. (A.1)-(A.2). (Blom *et al.*, 2010, p.84,99)

$$p_{X;Y}(j;k) = P(X = j \cap Y = k), \qquad j = 0, 1, 2..., k = 0, 1, 2... \tag{A.1}$$

$$p_X(j) = \sum_{k=0}^{\infty} p_{X|Y=k}(j) p_Y(k) = \sum_{k=0}^{\infty} p_{X;Y}(j;k) \tag{A.2}$$

The expected value for a random variable X is defined in Eq.(A.3). (Blom *et al.*, 2010, p.108)

$$E(X) = \sum_{k} k p_X(k) \tag{A.3}$$

The definition for independence between the discrete random variables X and Y is stated in Eq. (A.4), (Blom *et al.*, 2010, p.90).

$$p_{X;Y}(j;k) = p_X(j) p_Y(k), \qquad j = 0, 1, 2..., k = 0, 1, 2... \tag{A.4}$$

The definition for conditional PMF for the random variable X given $Y = k$ is stated in Eq. (A.5), (Blom *et al.*, 2010, p.98).

$$p_{X|Y}(j;k) = P(X = j|Y = k) = \frac{p_{X;Y}(j;k)}{p_Y(k)}, \qquad j = 0, 1, 2... \tag{A.5}$$

A.2 Linear correlation

The linear correlation coefficient is defined in Eq. (A.6), (Blom *et al.*, 2010, p.123).

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{(Var(X) Var(Y))}} \tag{A.6}$$

where

$$Cov(X,Y) = E(XY) - E(X)E(Y) \tag{A.7}$$

$$E(X) = \sum_k k p_X(k) \tag{A.8}$$

$$Var(X) = E(X^2) - [E(X)]^2 \tag{A.9}$$

From reduction and rewriting of Eq. (A.11) one finds that $-1 \leq \rho \leq 1$ only if X and Y are linearly dependent, that is if $Y = aX + b, a, b \in \mathbb{R}$. If $a > 0$ then $\rho = 1$, if $a < 0$ then $\rho = -1$ (Blom *et al.*, 2010, p.124).

$$E\left(\left[\frac{X - E(X)}{\sqrt{(Var(X))}} \pm \frac{Y - E(Y)}{\sqrt{(Var(Y))}}\right]^2\right) \geq 0 \tag{A.10}$$

$$\frac{Var(X)}{Var(X)} + \frac{Var(Y)}{Var(Y)} \pm 2\rho \geq 0 \tag{A.11}$$

A.3 Maximum likelihood

One of the most widely used methods for estimating parameters for a data set given a parametric statistical model is the maximum-likelihood (ML) method. Let $\{x_1, ..., x_n\}$ be outcomes of the random variables $\{X_1, ..., X_n\}$, that has a distribution depending on an unknown parameter θ, then the likelihood function is in the discrete case defined as,

$$L(\theta) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n; \theta) \tag{A.12}$$

(Blom *et al.*, 2010, p.254 Definition 11.6)

If $\{X_1, ..., X_n\}$ are independent variables $L(\theta)$ becomes the product of the probability mass functions. The ML-estimation of θ is the value θ^* where $L(\theta)$ takes its maximum value in its parameter space Ω_Θ. (Blom *et al.*, 2010, p.255 Definition 11.7)

When maximizing $L(\theta)$ it is often beneficial to maximize $l(\theta) = \ln L(\theta)$. Since the logarithm is a monotonically increasing function $L(\theta)$ and $\ln L(\theta)$ will take on the same maximum.

A.4 $\chi^2 - test$

Let X be a discrete random variable with the support $\{x_1, ..., x_m\}$. Make a hypothesis that $x_1 \leq ... \leq x_m$. Let $X_1, ..., X_n$ be a sample of n observations on X. Make an assumption

that the sample is from a particular discrete distribution with the PMF $f(k|\theta)$, where θ is an unknown parameter. Then the hypothesis can be tested as;

1. Find the number O_j of data points that are equal to x_j, $j = 0, 1, 2, ..., m$, the O_j's are called observed frequencies.
2. Compute an estimator $\hat{\theta}$ of θ based on the sample.
3. Compute the probabilities $p_j = f(x_j|\hat{\theta})$ for $j = 1, 2, ..., m-1$ and $p_m = 1 - \sum_{j=1}^{m-1} p_j$.
4. Compute the expected frequencies $E_j = p_j \times n$, $j = 1, .., m$
5. Calculate the χ^2-statistic
$$\chi^2 = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j} \tag{A.13}$$
6. Compare to the $(1-\alpha)$th quantile $q_{1-\alpha}$ of a χ^2 distribution with $m - d - 1$ degrees of freedom.

If $\chi^2 > q_{1-\alpha}$ the hypothesis is rejected at a significance level α. An important rule of thumb for the χ^2-test is that the value of expected frequencies > 5. If it is not, it is possible to summarize the frequencies so that the total number adds up to five, beware to consider this when deciding the degrees of freedom.

Packages used in R

- fitdistrplus
- gamlss
- VGAM
- copula
- boot
- graphics
- xtable

Adjusted functions for the extensive data

Several functions were adjusted by dividing the discrete data into the discrete points, and then multiplying with the frequencies in each point. This procedure was made when calculating τ_K, log-likelihood values for the copulas and the Cramér von Mises test statistic.

APPENDIX B. STATISTICAL RESULTS

B.1 PMFs in Table form

Cont/Build	0	1	2	3	4	5	6
0	0.8138	0.0673	0.0059	0.0005	0.0000	0.0000	0.0000
1	0.0736	0.0218	0.0037	0.0004	0.0000	0.0000	0.0000
2	0.0056	0.0034	0.0019	0.0004	0.0000	0.0000	0.0000
3	0.0004	0.0003	0.0003	0.0002	0.0001	0.0000	0.0000
4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Empirical							
0	0.8097	0.0684	0.0082	0.0011	0.0001	0.0000	0.0000
1	0.0753	0.0211	0.0028	0.0005	0.0001	0.0000	0.0000
2	0.0076	0.0030	0.0006	0.0001	0.0000	0.0000	0.0000
3	0.0008	0.0003	0.0001	0.0001	0.0000	0.0000	0.0000
4	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table B.1 Joint PMF for Build and Cont. Fitted model on top, empirical below.

Freq	$Build^{Emp}$	$Build^{Fit}$	$Cont^{Emp}$	$Cont^{Fit}$
0	0.8875	0.8875	0.8935	0.8934
1	0.0996	0.0997	0.0928	0.0929
2	0.0114	0.0113	0.0118	0.0117
3	0.0013	0.0013	0.0016	0.0017
4	0.0002	0.0002	0.0002	0.0002
5	0	0	0	0
6			0	0

Table B.2 Marginal PMF, empirical and fitted

Build model	0	1	2	3	4	5
Conditional on						
0	0.9108	0.0824	0.0063	0.0005	0.0000	0.0000
1	0.7249	0.2353	0.0365	0.0031	0.0002	0.0000
2	0.4949	0.3158	0.1598	0.0270	0.0023	0.0002
3	0.3289	0.2447	0.2641	0.1385	0.0219	0.0018
4	0.2195	0.1662	0.2184	0.2557	0.1224	0.0178
5	0.1488	0.1129	0.1520	0.2249	0.2535	0.1080
6	0.1094	0.0830	0.1120	0.1710	0.2547	0.2698
Empirical						
Conditional on						
0	0.9062	0.0843	0.0085	0.0009	0.0001	0.0000
1	0.7368	0.2269	0.0320	0.0036	0.0007	0.0000
2	0.6990	0.2395	0.0502	0.0080	0.0034	0.0000
3	0.6357	0.2636	0.0620	0.0388	0.0000	0.0000
4	0.5000	0.3333	0.1667	0.0000	0.0000	0.0000
5	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table B.3 Conditional distributions, Build conditional on Cont. Model on top, empirical below.

Cont Model	0	1	2	3	4	5	6
Conditional on							
0	0.9169	0.0758	0.0066	0.0006	0.0001	0.0000	0.0000
1	0.7389	0.2193	0.0375	0.0039	0.0004	0.0000	0.0000
2	0.4946	0.2974	0.1661	0.0371	0.0043	0.0004	0.0000
3	0.3173	0.2198	0.2440	0.1695	0.0435	0.0053	0.0005
4	0.2033	0.1427	0.1808	0.2333	0.1809	0.0523	0.0068
5	0.1314	0.0923	0.1184	0.1705	0.2293	0.1947	0.0634
Empirical							
Conditional on							
0	0.9123	0.0771	0.0092	0.0012	0.0001	0.0000	0.0000
1	0.7553	0.2115	0.0282	0.0046	0.0005	0.0000	0.0000
2	0.6730	0.2630	0.0521	0.0095	0.0024	0.0000	0.0000
3	0.6263	0.2525	0.0707	0.0505	0.0000	0.0000	0.0000
4	0.3846	0.3846	0.2308	0.0000	0.0000	0.0000	0.0000
5	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table B.4 Conditional distributions. Cont conditional on Build. Model on top, empirical below.

Conditional on	Build cond. Cont	Empirical data	Cont cond. Build	Empirical data
0	0.0965	0.1044	0.0910	0.1095
1	0.3185	0.3045	0.3078	0.3756
2	0.7264	0.3774	0.7604	0.6021
3	1.2855	0.5039	1.4203	0.8710
4	1.9486	0.6667	2.2298	2.2000
5	2.6454	0.0000	3.1115	0.0000
6	3.1879	0.0000		

Table B.5 Conditional expected values. Cont conditional on Build to the left and Build conditional on Cont to the right.

REFERENCES

Blom, G., Enger, J., Englund, G., Grandell, J. and Holst, L. (2005). *Sannolikhetsteori och statistikteori med tillmpningar*. 5:7 ed. Studentlitteratur Lund.

Boucher, J.P., Denuit, M., Guillen, M. (2008). Models of Insurance Claim Counts with Time Dependence based on Generalization of Poisson and Negative Binomial Distributions. *Variance Journal*, **34(9)**, 135–162

Bühlmann, H., Gisler, A. (2005). *A Course in Credibility Theory and its Applications*. Springer-Verlag Berlin Heidelberg.

Cohen, A. (2005). Assymetric Information and Learning: Evidence from the Automobile Insurance Market. *The Review of Economics and Statistics*, **87(2)**, 197–207

Coles, S. (2001). *An introduction to Statistical Modeling of Extreme Values*. Springer-Verlag London Limited.

Denuit, M., Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, **93**, 40–57.

Dodge, Y. (2010). *The Concise Encyclopedia of Statistics*. Springer Science Business Media, LLC 2010

Embrechts, P., McNeil, A. and Straumann, D. (2002). Correlation and dependence in risk management: Properties and pitfalls. *M. A. H. Dempster (ed.): Risk Management: Value at Risk and Beyond.* Cambridge University Press, pp.176-223.

Englund, M., Guilln, M., Gustafsson J., Nielsen, L.H and Nielsen, J.P. (2008). Multivariate latent risk: A credibility Approach. *Astin Bulletin*, **38(1)**, 137–146.

Englund, M., Gustafsson J., Nielsen, J.P. and Thuring, F. (2009). Multidimensional Credibility with Time Effects - An application to commercial Business Lines. *The Journal of Risk and Insurance*, **76(2)**, 443–453.

Everitt, B.S., Skrondal, A. (2010). *The Cambridge Dictionary of Statistics*. Cambridge University Press, 4th Edition

Fermanian, J. (2005). Goodness-of-fit tests for copulas. *Journal Of Multivariate Analysis*, **95(1)**, 119–152.

Frees, E.W, Valdez, E.A. (1998). Understanding Relationships Using Copulas. *North American Actuarial Journal*, **2(1)**, 1–25.

Genest, C., Ghoudi, K., Rivest, L.,P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, **82(3)**, 543–552.

Genest, C., Rémillard, B. and Beaudoin, D. (2007). Goodness-of-fit tests for copulas: A review and a power study. *Insurance Mathematics and Economics*, **44(2009)**, 199–213.

Genest, C., Werker, B.J.M. (2002). Conditions on the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models. *Distributions with Given Marginals and Statistical Modelling* Kluwer, Dordrecht, The Netherlands. Cuadras, C.M, Fortiana, J., Rodriguez-Lallela, J.a. pp. 102-112.

Gräler, B. (2011). Introduction to Copulas. *Spatio-temporal dependence* University of Muenster, Institute for Geoinformatics, unpublished.

Joe, H., Xu, J. (1996).The Estimation Method of Inference Fnctions for Margins for Multivariate Models. *JTechnical Report 166*, Departmant of Statistics, University of British Columbia.

Kamakura, W.A., Wedel, M., de Rosa, F. and Mazzon, J.A. (2003). Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in Marketing*, **20(1)**, 45–65.

Kattan, W.M. (2009). *Encyclopedia of Medical Decision Making*. Sage Publications, Inc.

Kim, G., Silvapulle M.J., and Silvapulle, P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics and Data Analysis*, **51(6)**, 2836–2850.

Kojadinovic, I., Yan, J. (2010). Comparison of three semiparametric methods for estimating dependence parameters in copula models. *Insurance Mathematics and Economics*, **47(1)**, 52–63.

Kojadinovic, I., Yan, J. (2010). Modeling Multivariate Distributions with Continuous Margins Using the copula R Package. *Journal of Statistical Software*, **34(9)**, 1–20.

Krishnamoorthy, K. (2006). *Handbook of statistical distributions with applications*. Chapman and Hall/CRC. Boca Raton.

Lindsey, J.K. (1997). *Applying Generalized Linear Models*. Springer-Verlag New York Inc.

Minkova, L.D. (2012). *Distributions in Insurance RIsk Models* Ph.D. Sofia University "St.Kl.Ohridski".

Nelsen, R.B. (2006). *An introduction to Copulas*. 2nd ed. Springer Science+Business Media, NY.

Nikoloulopoulos, A.K., Karlis, D. (2010). Modeling Multivariate Count Data Using Copulas. *Communications in Statistics - Simulation and Computation*, **39(1)**, 172–187

Schmidt, T.(2007). Coping with Copulas. *Risk Books, J. Rank*

Thuring, Fredrik. (2011). A credibility method for profitable cross-selling of insurance products. *Annuals of Actuarial Science*, **6(1)**, 65–75.

Trivedi, P.K. and Zimmer, D.M. (2007). Copula Modeling: An Introduction for Practitioners. *Foundations and Trens in Econometrics*, **1(1)**, 1–111.

Yan, J. (2007). Enjoy the Joy of Copulas: With a Package copula. *Journal of Statistical Software*, **21(4)**, 1–21.

Yener, T. (2011). *Risk Management Beyond Correlation* Ph.D. Fakultät für Mathematik, Informatik und Statistik der Ludwig-Maximilians-Universität München.