碩 士 學 位 論 文

Copula 함수를 이용한 결합분포함수 추정 및 연관성 비교

高麗大學校 大學院

情報統計學科

鄭 準 永

2013年 12月

全 秀 榮 敎 授 指 導

碩 士 學 位 論 文

Copula 함수를 이용한 결합분포함수 추정 및 연관성 비교

이 論文을 理學 碩士學位 論文으로 提出함.

2013年 12月

高 麗 大 學 校 大 學 院

情 報 統 計 學 科

鄭 準 永 (印)

鄭 準 永의 理學 碩士學位論文
審査를 完了함.

2013年　12月

委員長　정두영　(印)

委　員　최종후　(印)

委　員　진서훈　(印)

요 약 문

Copula 함수는 다차원분포함수와 1차원 한계분포를 연결시키는 함수를 의미하며 데이터 사이의 종속성 구조와 개별 자료의 분포를 분리하여 모형화 함으로써 추정과 시뮬레이션을 용이하게 한다. 따라서 각 분포들의 개별 밀도함수만 추정하고 그 자료에 맞는 Copula 함수를 찾아내면 그 분포들의 결합밀도함수를 찾는 것보다 더 쉽고 simulation도 더 쉽게 할 수 있게 해준다.

본 연구에서는 Copula 함수에 관한 내용을 소개하고자 하며, 실증분석에서는 모 보험회사의 처리 기간과 보험 금액에 관한 상관성을 Copula를 이용하여 알아보았으며, 모 카드회사의 카드 사용 건수와 카드 사용액의 관계를 이용하여 두 분포의 결합분포함수를 추정하는데 Copula 함수를 사용하였다. 또한 본 논문에서는 fat tail 분포 추정하는 방법과 한 쌍의 데이터와 두 쌍의 데이터일 경우 더 효율적으로 결합분포함수를 찾는 방법 두 가지를 제시하였으며 실증분석을 통하여 제시한 두 방법이 효과적임을 보여주었다.

목 차

표 목 차

그 림 목 차

제 1 장 서 론

금융권 시장이 활발해지면서 금융권에서 수집되는 자료들은 점점 늘어나고 있다. 금융권에서 자료를 쌓아두기만 하고 제대로 분석하여 활용하지 않으면 그저 쓰레기에 불과하다. 금융권 시장에서 수집되는 자료들은 여러 종류의 자료가 있다. 보험회사의 손해액이나 카드회사의 거래내역, 증권회사의 주식과 같은 여러 가지의 자료들이 수집된다. 이렇게 쌓여가는 자료들은 중요 요인들 간의 상관성을 보고 분석하여 좋은 결과를 이끌어 내거나 미래를 예측하는 중요한 자료가 될 수 있다.

상관성을 분석하는 방법에는 여러 가지가 있다. 그 중에서 대부분 일반적으로 알 수 있는 방법은 상관계수를 통하여 분석하는 것이다. 하지만 상관계수만으로는 한계가 있기 때문에 보다 정확하고 유용한 방법을 이용하여 자료에 적용시켜야 한다.

금융권 시장과 같은 곳의 자료들은 fat tail의 형태를 가진 자료들이 많다. fat tail 분포를 상관계수로 판단하기에는 무리가 있다. 그래서 fat tail 분포와 같은 다양한 형태에 적합한 Copula 함수를 이용하여 좀 더 유용하고 정확한 상관성을 분석해야한다.

본 논문에서는 일반적인 Copula에 대한 설명을 통하여 기본적인 Archimedean Copula를 이용하여 분석을 실시하며 거기서 좀 더 나아가 상관성 분석을 좀 더 정확하게 알 수 있는 방법을 제안하며 모의실험 및 실증분석을 시행해 보았다.

본 논문은 2장에서 Copula에 대한 전반적인 설명을 하고(Copula 함수의 최근 연구동향에 대한 보다 이론적인 논의는 (Nelson. 1999)를 참조하기 바람.), 3장에서 본 논문에서 제안하는 방법을 제시하며 4장에서는 모의실험, 5장에서는 실증분석을 통하여 결과를 살펴보겠다. 마지막으로 6장에서 본 논문의 결론을 정리하였다.

제 2 장 Copula

2.1 Copula 함수의 정의

Copula 함수는 다변량 분포함수와 주변분포함수를 연결시키는 함수이며 개별 데이터의 분포와 데이터 간의 종속성 구조(dependence structure)를 분리하여 모형화 할 수 있게 함으로써 결합분포함수를 직접 모형화하는 것보다 추정과 시뮬레이션을 용이하게 할 수 있다(김명직, 신성환. 2003).

[정의 : Copula] : n-차원 Copula C는 다음과 같은 특성을 만족하는 함수이다

(1) $C: [0,1]^n \rightarrow [0,1]$

(2) C 는 “grounded” 함수이고 n-증가함수이다.

(3) C 는 모든 $u \in [0,1]$에 대하여 다음을 만족하는 한계분포함수 $C_i (i=1,2,...,n)$을 갖는다.

$$C_i(u) = C(1,\cdots,1,u,1,\cdots,1) = u$$

위의 세 가지 특성을 만족하는 Copula 함수가 결합분포와 같다는 것은 다음과 같은 Sklar 정리에 근거한다.

[정리 : Sklar's Theorem] H 가 결합분포함수를 나타내고 이의 연속인 한계분포함수가 $F_1,\cdots,F_n$이라고 하자. 이 경우 모든 실수 $x_{1,} \cdots,x_n$에 대하여 다음과 같은 관계를 만족하는 Copula 함수 C 가 유일하게 존재한다.

$$H(x_{1,} \cdots,x_n) = C(F_1(x_1),\cdots,F_n(x_n))$$

[보조정리] H 가 n-차원 결합분포함수를 나타내고 이의 연속인 한계분포함수

가 $F_1,\cdots,F_n$이며 Copula 함수 C 를 갖는다고 하자. 모든 $u=(u_1,\cdots,u_n)\in[0,1]^n$ 에 대하여 다음이 성립한다.

$$C(u_1,\cdots,u_n)=F(F_1^{-1}(u_1),\cdots,F_n^{-1}(u_n))$$

단, F_i^{-1} 은 F_i 의 일반역함수를 나타낸다.

기존의 상관성 구조를 파악하기 위한 방법으로 선형 상관계수(correlation coefficient)가 있으나, 보험이나 재무, 금융과 관련된 자료는 꼬리부분이 두터운 분포를 보이기 때문에 꼬리부분을 다루는 경우에는 선형 상관관계보다는 Copula를 이용한 방법이 더 타당하다고 할 수 있다.

2.2 Copula의 종류

Copula를 크게 2가지로 나누게 되면 Elliptical Copula와 Archimedean Copula로 나눌 수 있다. Elliptical Copula에는 Gaussian Copula와 Student's t-Copula가 있는데 이 두 Copula는 타원형 분포와 관련된 Copula이고, Archimedean Copula는 Clayton Copula, Frank Copula, Gumbel Copula로 나눌 수 있다. 축을 중심으로 대칭형 Copula는 Gaussian Copula와 Student's Copula, Frank Copula가 되며, 반대인 비대칭형 Copula는 나머지인 Clayton Copula와 Gumbel Copula가 있다.

2.2.1 Elliptical Copula

Elliptical Copula의 대표적인 종류는 Gaussian Copula와 Student's t-Copula이다. 우선 Gaussian Copula는 n개의 정규 확률변수와 선형 상관계수행렬 R을 통하여 만든 $C_R^{Ga}(u)=\Phi_R^n(\Phi^{-1}(u1),\cdots,\Phi^{-1}(u_n))$를 Gaussian Copula

라 한다. 이변량의 경우로 보게 된다면 다음과 같은 식이 된다.

$$C_R^{Ga}(u,v)=\int_{-\infty}^{\Phi^{-1}(u)}\int_{-\infty}^{\Phi^{-1}(v)}\frac{1}{2\pi(1-R_{12})^{\frac{1}{2}}}exp\left\{-\frac{s^2-2R_{12st}+t^2}{2(1-R_{12}^2)}\right\}dsdt.$$

그 다음은 Student's t-Copula로 다변량 Student's t-분포의 Copula 함수가 Student's t-Copula이다. 상관계수행렬 ρ와 자유도 ν인 표준 다변량 Student's t 분포함수이다. 마찬가지로 $C_{\nu,R}^t(u)=t_{\nu,R}^n(t_\nu^{-1}(u1),\cdots,t_\nu^{-1}(u_n))$를 Student's t-Copula라 하며 다음과 같이 표현 할 수 있다.

$$C_{\nu,R}^t(u,v)=\int_{-\infty}^{t_\nu^{-1}(u)}\int_{-\infty}^{t_\nu^{-1}(v)}\frac{1}{2\pi(1-R_{12})^{\frac{1}{2}}}exp\left\{1+\frac{s^2-2R_{12st}+t^2}{2(1-R_{12}^2)}\right\}^{-(\nu+2)/2}dsdt.$$

2.2.2 Archimedean Copula

Archimedean Copula는 다른 Copula 함수들에 비하여 모수를 하나만 가지며 계산하기 용이할 뿐만 아니라 여러 가지 다른 종속성 구조를 설명해 주어 Copula 함수 중에서도 많이 쓰여지고 있다.

[정의] 함수 $\varphi:[0,1]\rightarrow[0,\infty]$는 다음과 같은 속성을 가진다.

(1) φ는 연속함수이다.

(2) 모든 $u\in[0,1]$에 대하여, $\varphi'(u)<0$

(3) $\varphi(1)=0$

$\varphi:[0,1]\rightarrow[0,\infty]$과 같이 pseudo-inverse를 정의하면

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

Archimedean Copula 식은 $C(u,v)=\varphi^{[-1]}(\varphi(u)+\varphi(v))$와 같이 표현할 수 있으며, φ는 Copula의 생성인자(Generator)이다.

[정리] C 를 생성함수 φ에 의한 Archimedean Copula라 하면 다음의 성질을 만족한다.

(1) C 는 대칭이다. 즉 구간 $[0,1]$에서의 모든 u,v에 대해서 $C(u,v)=C(v,u)$ 이다.

(2) C 는 결합적이다. 즉 구간 $[0,1]$에서의 모든 u,v,w에 대해서 다음이 성립한다.

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

앞서 언급하였듯이 Archimedean Copula의 대표적인 3가지 Copulas는 Clayton Copula, Frank Copula, Gumbel Copula이다.

먼저 Clayton Copula는 왼쪽 꼬리의 강한 종속성을 잘 설명하는 특징을 가진다. 예를 들자면 두 주식의 가격이 계속 하락할 때 강한 종속성을 보이는 경우에 효과적이다. 반대로 Gumbel Copula는 결합확률분포의 상단 꼬리에 강한 종속성을 가지는데 두 주식의 가격이 상승할 때 강한 종속성을 보이는 경우에 효과적이다. Clayton과 Gumbel은 반대의 경우로 생각할 수 있다. 마지막으로 Frank Copula는 Clayton과 Gumbel의 두 가지 경우 모두에 해당하는 왼쪽과 오른쪽 꼬리의 종속성을 똑같이 증대시키는 경우에 효과적이다.

다음은 세 가지 Archimedean Copula의 Copula 함수식과 생성인자(Generator)에 관한 표이다.

<표 2.1> Archimedean Copula와 Generator

Copula	$C_\alpha(u,v)$	$\varphi_\alpha(t)$
Clayton	$\max([u^{-\alpha}+v^{-\alpha}-1]^{-1/\alpha},0)$	$\frac{1}{\alpha}(t^{-\alpha}-1)$
Frank	$-\frac{1}{\alpha}ln(1+\frac{(e^{-\alpha u}-1)(e^{-\alpha v}-1)}{e^{-\alpha}-1})$	$-\ln(\frac{\exp(-\alpha t)-1}{\exp(-\alpha)-1})$
Gumbel	$\exp\{-[(-\ln u)^{\alpha}+(-\ln v)^{\alpha}]^{1/\alpha}\}$	$(-\ln t)^{\alpha}$

2.3 Archimedean Copula와 Kendall's tau의 관계

Copula의 모수를 추정하는 방법에는 The Maximum Likelihood Method(MLE), The Inference Function For Margins Method(IFM), The Canonical Maximum Likelihood Method(CML)와 같은 방법들이 있다. 이러한 여러 방법들로 모수를 추정하여 Copula 함수를 구하는 법과는 조금 다른 방법으로 Archimedean Copula를 구할 수 있다.

이 방법은 위의 여러 가지 방법들보다는 쉽지만 이변량 확률분포에만 적용된다는 단점을 가지고 있다. 이 방법은 Kendall's τ를 이용하여 Copula 모수를 추정한다.(Calibration method) Kendall's tau에 대한 간략적인 설명을 하자면 Kendall's τ는 순서통계량을 이용하여 일치성의 확률에서 불일치성의 확률을 차감한 것과 같다. 앞에서 언급하였듯이 Archimedean Copula는 Generator function을 이용하여 구하므로 Kendall's τ와 Generator function의 관계를 이용하여 모수를 추정한다.

[정리] 두 확률변수 X, Y에 대하여 τ_c는 다음과 같이 주어진다.

$$\tau_c = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}dt$$

먼저 표본을 통하여 Kendall's τ를 추정한 후에 Generator function을 이용하여 τ_c의 함수를 만든 후 Copula의 모수를 추정한다. 앞에서 대표적으로 언급한 Archimedean Copula 3가지들의 τ는 다음과 같다.

<표 2.2> Archimedean Copula와 τ

Copula	$\varphi_\alpha(t)$	τ_α
Clayton	$\frac{1}{\alpha}(t^{-\alpha}-1)$	$\frac{\alpha}{(\alpha+2)}$
Frank	$-\ln\left(\frac{\exp(-\alpha t)-1}{\exp(-\alpha)-1}\right)$	$1+\frac{4[D_1(\alpha)-1]}{\alpha}$
Gumbel	$(-\ln t)^\alpha$	$\frac{\alpha-1}{\alpha}$

이와 같이 τ를 추정한 후 Generator function을 이용하여 모수 값을 추정할 수 있다. 모수 값을 추정한 후에는 모수 값을 이용하여 쉽게 이변량 모의실험에 효과적으로 이용할 수 있다.

제 3 장 연구방법

3.1 자료소개

일반적으로 금융권에서 얻어진 데이터들은 대부분 fat tail을 가지는 분포들의 데이터이다. 예를 들면, 증권회사나 보험회사, 투자신탁회사, 종합금융회사, 상호저축은행 등 이러한 곳에서 발생하는 데이터들은 대부분이 fat tail 분포의 모양을 가진다.

fat tail을 가지는 분포들은 Weibull distribution(와이블 분포), Exponential distribution(지수 분포), Log-normal distribution(로그-노말 분포), Pareto distribution(파레토 분포) 등이 있다. 이 분포들은 모두 x>0 인 값들을 가지며 fat tail을 가지기 때문에 손해액 분포에 적합하며 Weibull < Exponential < Log-normal < Pareto 로 갈수록 fat tail을 가진다(최진석, 2010).

<표 3.1> 각 분포들의 pdf와 cdf

분포	$f(x)$	$F(x)$
Weibull	$\frac{k}{\lambda}(\frac{x}{\lambda})^{k-1}e^{-(x/\lambda)^k}$	$1-e^{-(x/\lambda)^k}$
Exponential	$\lambda e^{-\lambda x}$	$1-e^{-\lambda x}$
Log-normal	$\frac{1}{x\sqrt{2\pi\sigma}}e^{-\frac{(-\ln x-\mu)^2}{2\sigma^2}}$	$\frac{1}{2}+\frac{1}{2}erf[\frac{\ln x-\mu}{\sqrt{2}\sigma}]$, $\Phi(\frac{\ln x-\mu}{\sigma})$
Pareto	$\frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}$	$1-(\frac{x_m}{x})^{\alpha}$

3.2 연구방법론

본 논문에서는 여러 금융권 자료들의 특성인 fat tail 분포들 간의 상관성을 설명할 수 있는 Copula 함수를 이용하여 요인들 간의 결합분포함수를 추정해 보고 원 자료와 모의실험을 통하여 비교분석 하겠다. 본 논문에서 이용하는 자료들은 왼쪽부분에 치우쳐 Clayton Copula를 이용하여 상관성을 알아보도록 하겠다.

먼저 금융권 자료들의 특성이 fat tail 형태지만 어느 분포를 따르는지는 확인 해야한다. 자료의 분포를 추정하기 위하여 본 논문에서는 첫 번째 새로운 연구방안을 제시하였다. 기존의 Kolmogorov-Smirnov Test로만 판단하는 것이 아니라 QQ-plot을 이용하여 추가적인 확인 작업을 시행하였다. 분포 추정 과정을 설명하게 되면 fat tail 특성의 분포들 중 크게 4가지 (Weibull, Exponential, Log-normal, Pareto) 분포를 이용하였다. 어떠한 분포를 따르는지 모르기 때문에 4가지 모두 임의의 원 자료를 생성하였다. 이렇게 생성된 원 자료의 평균과 분산을 이용하여 각각의 분포의 모수 값들을 추정하였다.

<표 3.2> 각 분포들의 평균과 분산

분포	평균	분산
Weibull	$\lambda\Gamma(1+\frac{1}{k})$	$\lambda^2[\Gamma(1+\frac{2}{k})-(\Gamma(1+\frac{1}{k}))^2]$
Exponential	λ^{-1}	λ^{-2}
Log-normal	$e^{\mu+\sigma^2/2}$	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$
Pareto	$\frac{\alpha x_m}{\alpha-1}$	$\frac{x_m^2\alpha}{(\alpha-1)^2(\alpha-2)}$

이러한 방법으로 생성된 모수들을 가지고 새로운 임의의 4가지 분포를 생성한다. 그 다음 생성한 원 자료의 Kendall's τ를 측정한 후 Copula 함수의 모수 α를 추정한 후 모의실험을 통하여 X, Y 두 분포를 생성한다. 이렇게 생성된 X, Y와 기존의 평균과 분산을 이용하여 생성된 분포를 QQ-plot을 이용하여 비교하였고 Kolmogorov-Smirnov Test를 통하여 추가적으로 검정하였다. 비교하였을 때 가장 유사한 형태를 보이고 있는 분포로 추정하도록 하였다.

그 다음 본 논문에서 제시한 방안은 자료의 상관성 분석으로 자료가 한 쌍일 경우와 두 쌍인 경우로 나누어 실시하였다. 예를 들어, 한 쌍인 경우는 보험 처리기간과 보험 금액 같은 한 쌍인 경우는 Copula 함수를 이용하여 상관성 분석을 실시하였다. 그러나 카드 거래 지역 건수와 카드 거래 지역 금액을 한 쌍으로 보고 분석을 실시한 것과 카드 거래 지역을 거주 지역과 타 지역으로 나누어 보게 되면 거주 지역 건수와 타 지역 건수, 그리고 거주 지역 금액과 타 지역 금액으로 나누어 보게 되면 한 쌍이 두 개인 두 쌍으로 나눌 수 있다. 두 쌍으로 보게 되면 한 쌍을 먼저 Copula 함수를 이용하여 결합분포함수를 구하고 다른 쌍도 마찬가지로 Copula 함수를 이용하여 결합분포함수를 구한 후 두 결합분포함수를 한 쌍으로 보아 Copula 함수를 이용하여 최종 결합분포함수를 구하였다. 기본적으로 카드 거래 지역 건수와 카드 거래 지역 금액을 비교한 결과와 새로이 제시한 방법을 이용하여 나온 결과 값을 비교하여 어느 방법이 더 좋은지 평가하도록 하겠다.

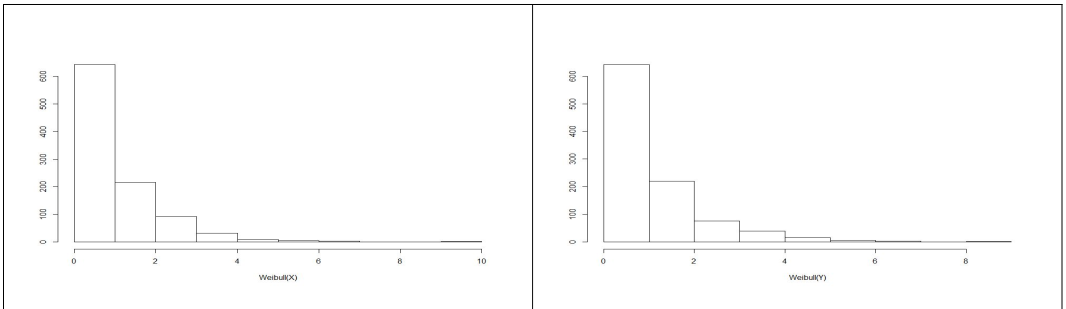
제 4 장 모의실험

4장에서는 2장과 3장에서 설명한 Copula와 연구방법이 맞는지에 대하여 모의실험을 진행하였다. 모의실험을 위한 통계 패키지로는 R-software(3.0.2)를 사용하였다.

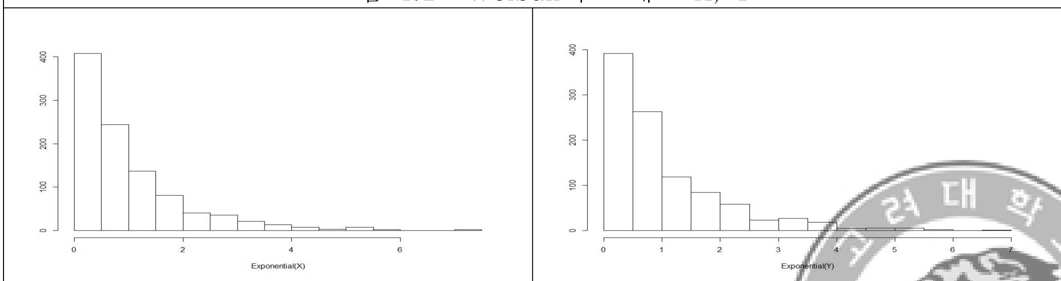
본 논문의 모의실험에서는 4가지 Weibull, Exponential, Log-normal, Pareto 분포를 따르는 경우만 고려하였다. 그래서 두 분포 사이의 상관성이 낮을 경우와 높을 경우로 나누어 각각 4가지 분포에 대한 누적확률분포(Cumulative Probability Distribution) 자료를 각각 생성하였다.

4.1 상관성이 낮은 경우

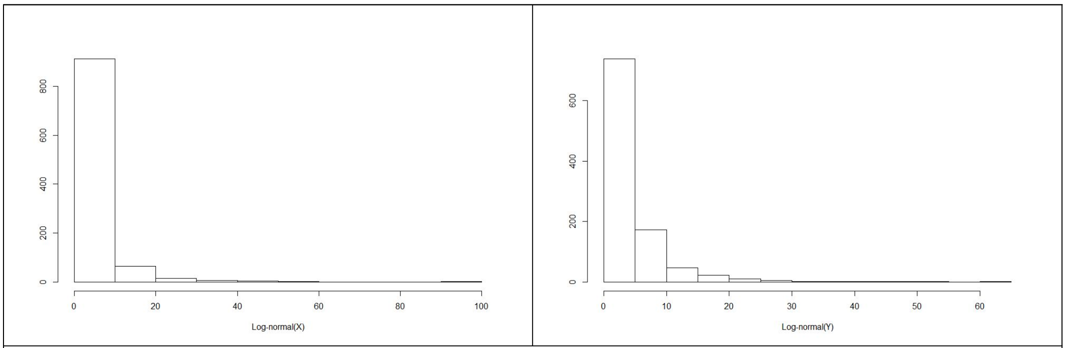
R-software(3.0.2)에 기존으로 설정되어 있는 모수 값들로 4가지 Weibull, Exponential, Log-normal, Pareto 분포로 각각 X, Y 값을 생성하였다.



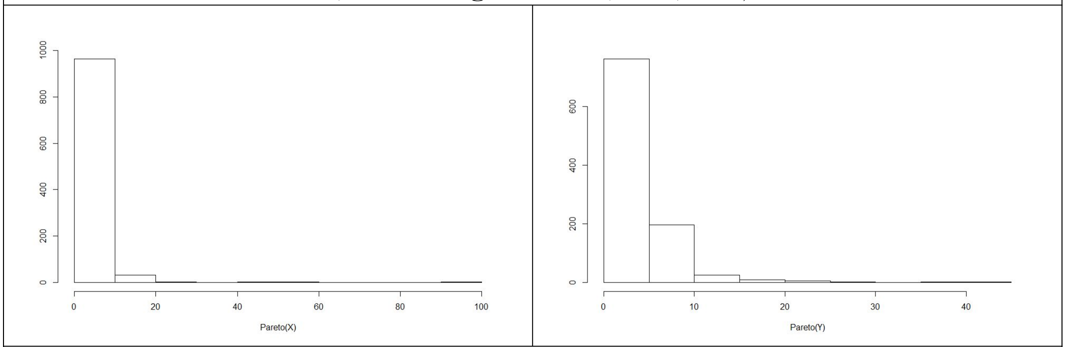
<그림 4.1> Weibull의 그래프 X, Y



<그림 4.2> Exponential의 그래프 X, Y



<그림 4.3> Log-normal의 그래프 X, Y



<그림 4.4> Pareto의 그래프 X, Y

위의 자료들은 두 분포의 상관성이 낮을 경우를 고려하여 생성한 자료이다. 각각 생성하였기 때문에 두 분포간의 상관성을 알아보기 위하여 Kendall's τ 값을 산출하였다.

<표 4.1> 두 분포간의 Kendall's τ

분포	τ
Weibull	0.00492
Exponential	0.00236
Log-normal	0.02612
Pareto	0.02465

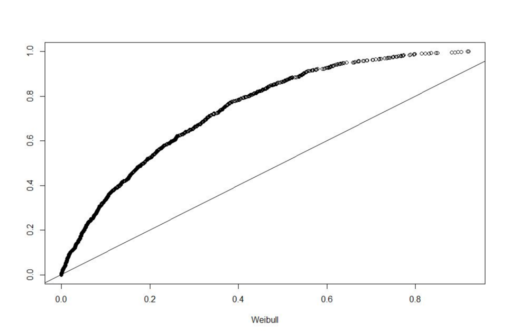
Kendall's τ 값을 살펴보면 앞에서 예상했듯이 4가지 분포의 대한 서로의 상관성이 낮은 수치로 나타났다. 얻어진 Kendall's τ 값으로 추정된 Copula의 모

수 값은 다음과 같다.

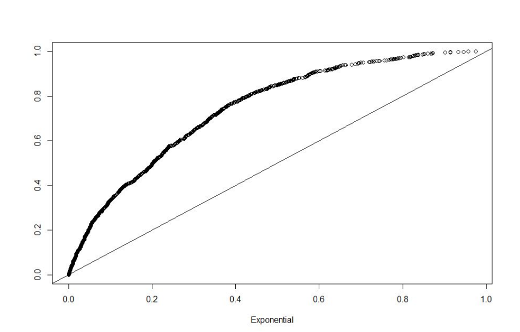
<표 4.2> Kendall's τ 에서 생성된 Copula의 모수

분포	모수
Weibull	0.00989
Exponential	0.00474
Log-normal	0.05365
Pareto	0.05055

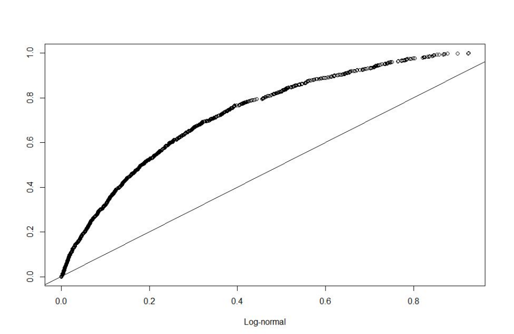
추정된 Copula의 모수로 앞서 언급하였듯이 좌측 하단 꼬리부분을 잘 반영하는 Clayton Copula의 경우가 더 적합하므로 Clayton Copula를 이용하여 값을 구하였다. 각각 분포 사이의 Clayton Copula 값과 생성된 4가지 분포와 QQ-Plot을 이용하여 비교하였다.



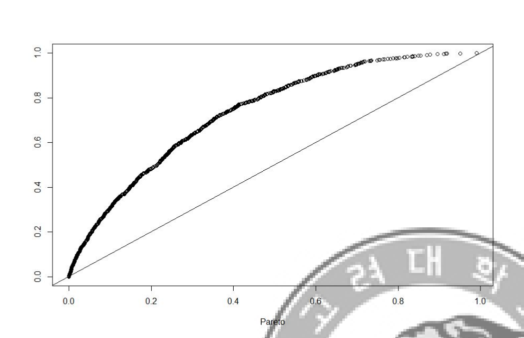
<그림 4.5> Weibull 분포



<그림 4.6> Exponential 분포



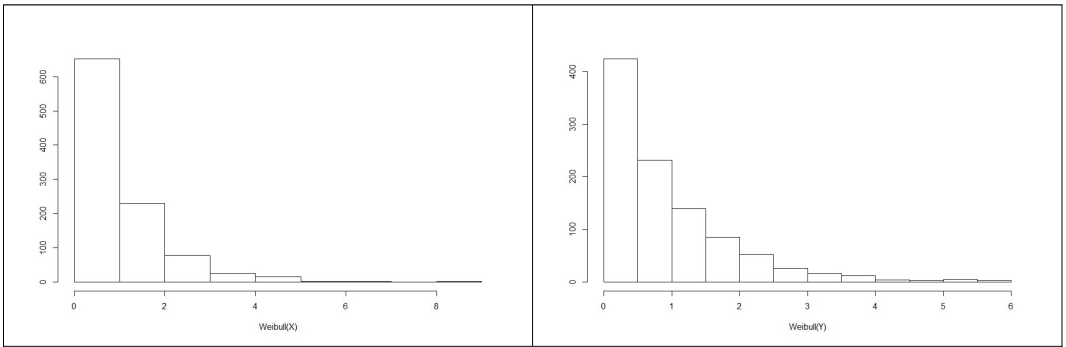
<그림 4.7> Log-normal 분포



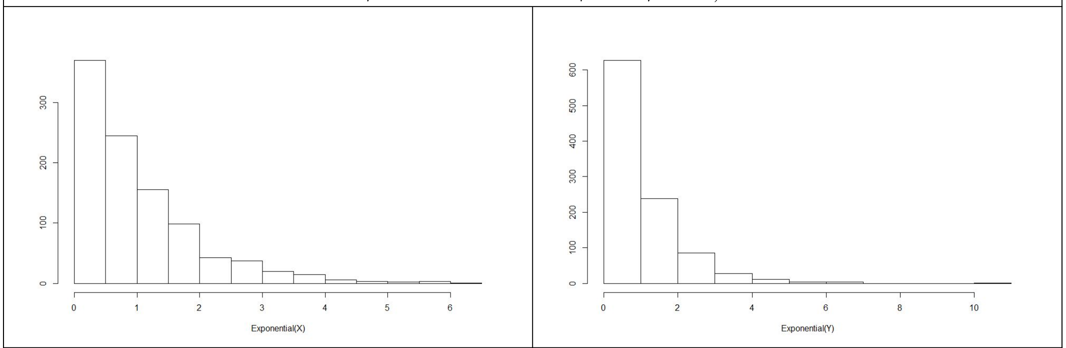
<그림 4.8> Pareto 분포

4.2 상관성이 높은 경우

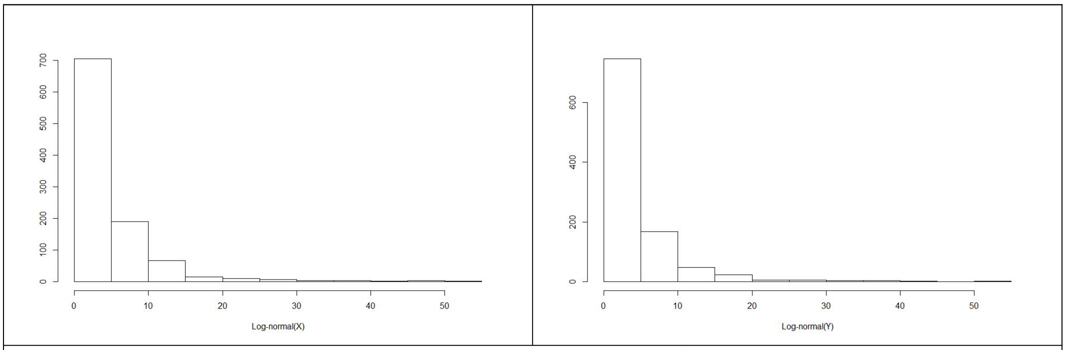
마찬가지로 R-software(3.0.2)를 이용하여 상관성이 높은 경우의 4가지 Weibull, Exponential, Log-normal, Pareto 분포들로 각각 X, Y를 생성하였다.



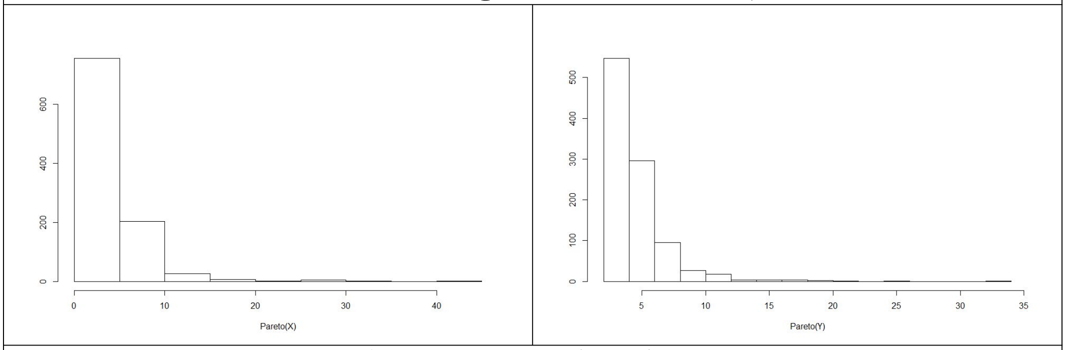
<그림 4.9> Weibull의 그래프 X, Y



<그림 4.10> Exponential의 그래프 X, Y



<그림 4.11> Log-normal의 그래프 X, Y



<그림 4.12> Pareto의 그래프 X, Y

위의 자료들은 두 분포의 상관성이 높을 경우를 고려하여 생성한 자료이다. 각각 생성하였기 때문에 두 분포간의 상관성을 알아보기 위하여 Kendall's τ 값을 산출하였다.

<표 4.3> 두 분포간의 Kendall's τ

분포	τ
Weibull	0.52992
Exponential	0.51401
Log-normal	0.50235
Pareto	0.50588

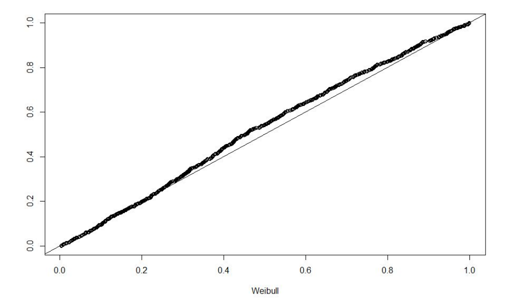
Kendall's τ 값을 살펴보면 앞에서 예상했듯이 4가지 분포의 대한 서로의 상관성이 이전의 생성한 자료보다는 상대적으로 높은 수치로 나타났다. 얻어진

Kendall's τ 값으로 추정된 Copula의 모수 값은 다음과 같다.

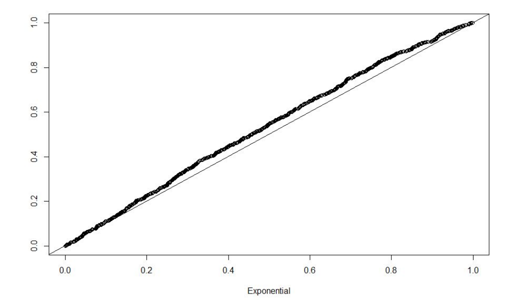
<표 4.4> Kendall's τ 에서 생성된 Copula의 모수

분포	모수
Weibull	2.25461
Exponential	2.11531
Log-normal	2.01889
Pareto	2.04768

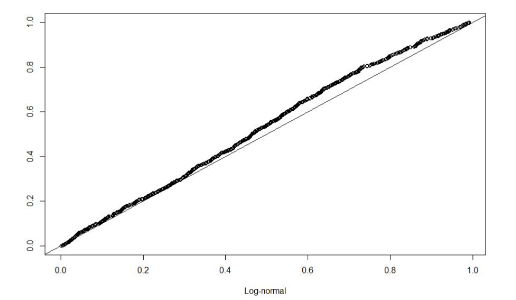
추정된 Copula의 모수로 앞서 언급하였듯이 좌측 하단 꼬리부분을 잘 반영하는 Clayton Copula의 경우가 더 적합하므로 Clayton Copula를 이용하여 값을 구하였다. 각각 분포 사이의 Clayton Copula 값과 생성된 4가지 분포와 QQ-Plot을 이용하여 비교하였다.



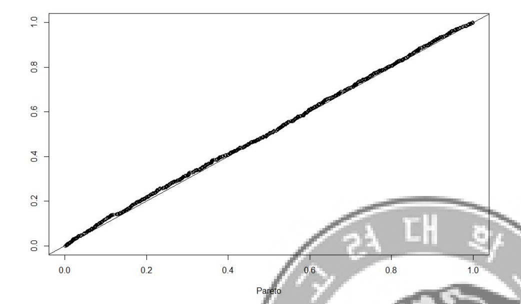
<그림 4.13> Weibull 분포



<그림 4.14> Exponential 분포



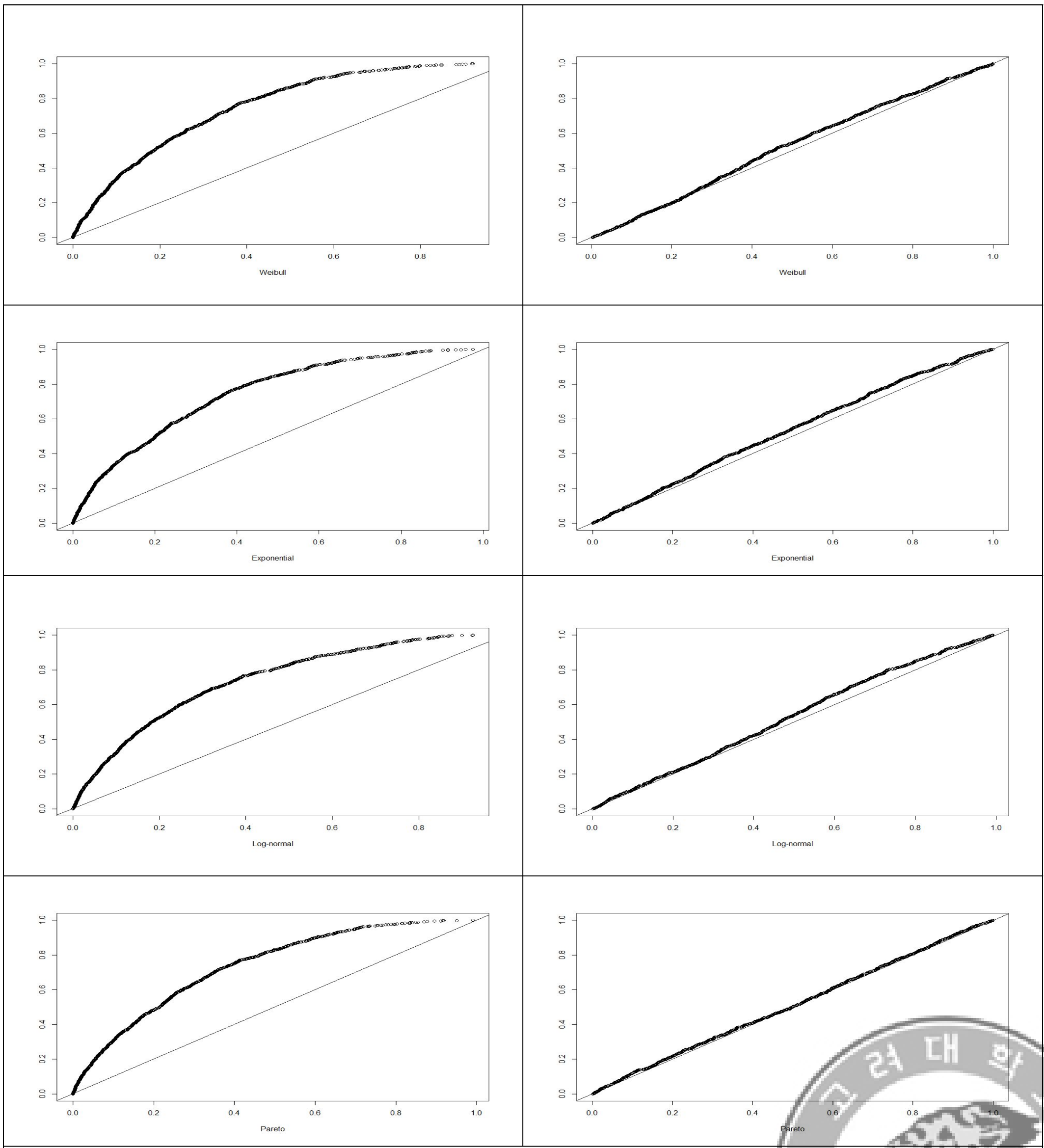
<그림 4.15> Log-normal 분포



<그림 4.16> Pareto 분포

4.3 모의실험 결론

상관성이 낮을 경우와 상관성이 높을 경우로 나누어 모의실험을 진행하였다. fat tail을 가지는 4가지 Weibull, Exponential, Log-normal, Pareto 분포를 이용하여 비교하였다.



<그림 4.17> 상관성에 따른 QQ-Plot

Copula 함수를 이용해 상관성이 낮을 경우와 상관성이 높을 경우로 나누어 두 분포의 결합함수를 도출하였다. 서로 같은 형태의 두 분포가 결합함수를 통하여 하나의 분포로 합쳐지게 되면 결합함수의 분포 형태도 같은 분포를 따른다는 가정으로 두 경우의 Copula 함수 값을 비교해 보았다. <그림 4.13>에서 왼쪽 부분인 상관성이 낮을 경우 분석되어 나열된 그래프들을 보게 되면 Copula 함수 값들과 각각 해당하는 분포와의 QQ-Plot은 기울기가 1인 대각선과 거리가 좀 떨어져 있는 것을 볼 수 있다. 일반적으로 상관성을 보기 위하여 기초적인 상관분석을 실시한다. 앞에서 상관분석을 실시한 결과 0에 가까워 서로 상관성을 거의 찾지 못하는 결과를 보였지만 실제로 Copula 함수를 이용하여 상관성을 보았더니 <그림 4.13>과 같은 결과를 볼 수 있었다. 단순한 상관계수를 통하여 상관성을 보는 것보다는 Copula 함수를 이용하여 상관성을 보는 것이 더 정확하고 효과적이다. 반면에 상관성이 높은 경우의 Copula 함수 값은 대각선과 가깝게 겹쳐져 있는 것을 볼 수 있다. 모의실험을 통하여 제시한 분포를 추정하는 과정은 상관성이 높을 경우 더 좋은 결과 값을 가지는 것을 볼 수 있다.

결과적으로는 상관성이 낮을 경우와 높을 경우를 비교해 보면 어느 정도의 차이는 있지만 기본적인 상관계수로 판단하는 것보다 Copula 함수를 이용하여 구해진 값들이 좀 더 세부적으로 분석된다고 할 수 있다.

제 5 장 실증분석

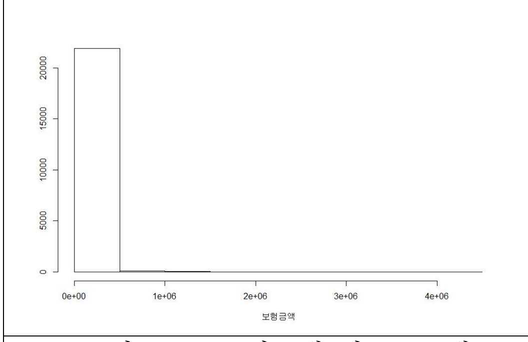
5.1 배 경

금융권 시장의 데이터들이 fat tail의 형태를 가지고 있기 때문에 금융권에서 보험금액 데이터와 카드내역 데이터를 이용하여 실제로 적용하여 얼마만큼의 상관성이 존재하는지를 알아보고자 한다.

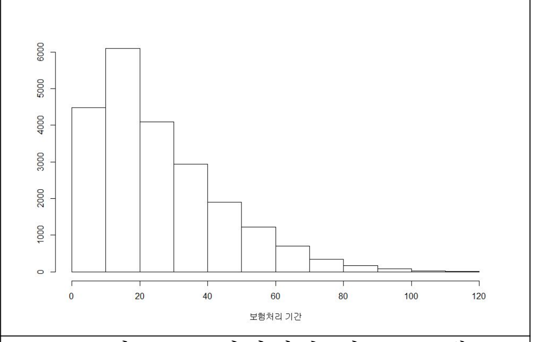
5.2 자료 설명

본 논문의 실증분석에 사용되는 자료는 모 보험회사의 10년간 보험금액과 사고가 일어난 후 보험금을 탈 때까지의 기간을 기록한 자료와 모 카드회사의 일정기간 동안의 카드거래 건수와 카드거래 금액을 기록한 자료이다.

먼저 모 보험회사의 10년간 기록된 보험금액과 사고가 일어난 후 보험금을 탈 때까지의 기간을 월 단위로 끊어서 기록한 자료의 일부분이다. 자료의 수는 약 22000건 정도이며 10년간 기록된 자료이기 때문에 최대 120개월까지이며 짧게는 1달로 구분할 수 있다.

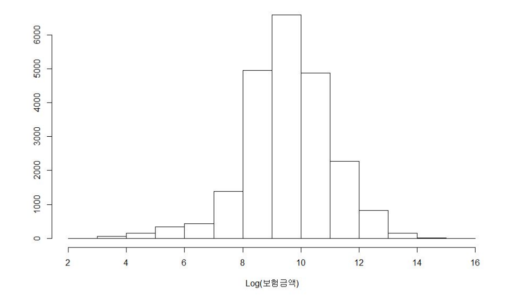
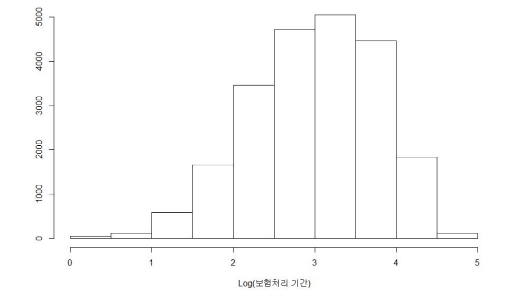


<그림 5.1> 보험금액 히스토그램



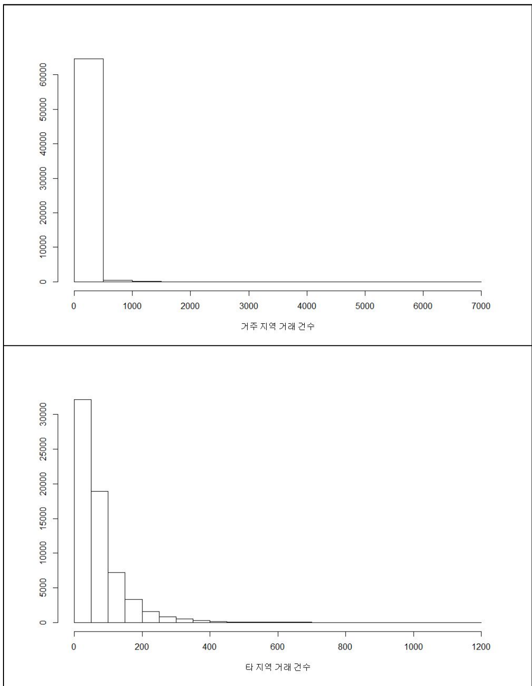
<그림 5.2> 처리기간 히스토그램

보험금액과 처리기간의 히스토그램을 보게 되면 둘 다 왼쪽으로 많이 치우친 분포 형태를 볼 수 있다. 전체 히스토그램만으로는 세부적으로 볼 수 없기 때문에 log 변환을 실시해 보았다.

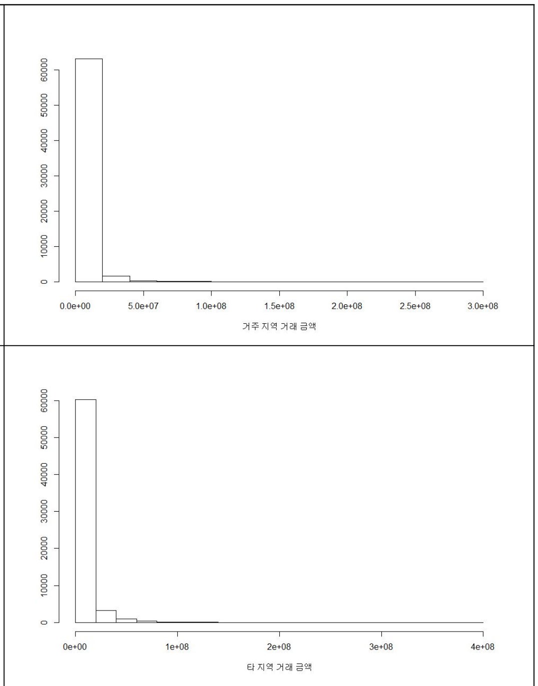
	
<그림 5.3> log(보험금액) 히스토그램	<그림 5.4> log(처리기간) 히스토그램

log 변환을 한 보험금액 히스토그램은 대체적으로 정규분포 형태와 유사하였으며, log 변환을 한 처리기간의 히스토그램은 오른쪽 경사가 좀 더 가파르지만 정규분포의 모양과 비슷하게 나왔다.

그 다음 분석을 하게 될 자료 역시 모 카드회사의 일정기간 동안의 카드거래 건수와 카드거래 금액이다. 카드거래 건수와 카드거래 금액은 각각 실제 거주 지역과 실제 거주 지역을 벗어난 타 지역으로 구분하여 기록되어있는 자료이다. 자료의 수는 약 65000건 정도이며 거주 지역 카드거래 건수와 타 지역 카드거래 건수, 거주 지역 카드거래 금액과 타 지역 카드거래 금액 4가지로 기록되어 있다.

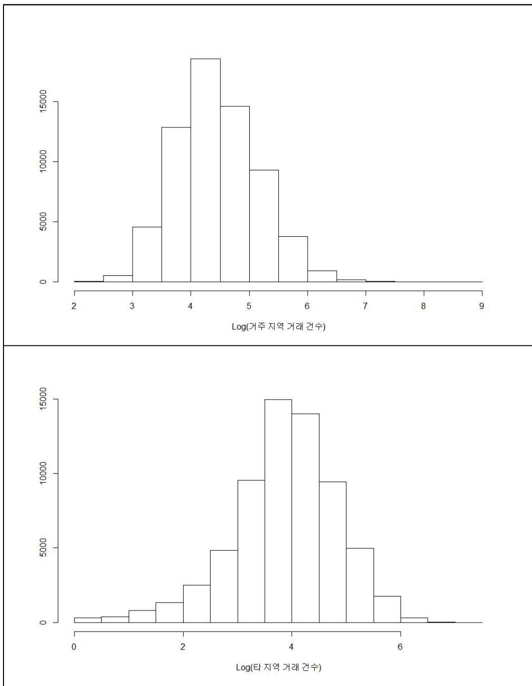


<그림 5.5> 카드거래 건수 히스토그램

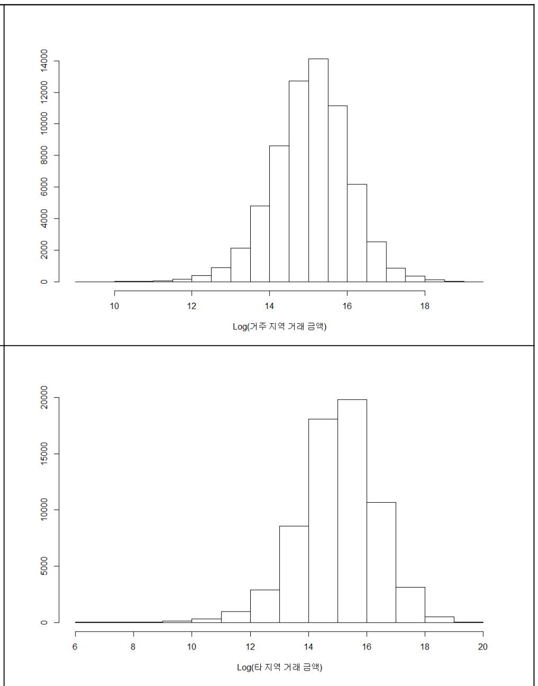


<그림 5.6> 카드거래 금액 히스토그램

카드회사의 자료도 보험회사의 자료와 마찬가지로 왼쪽으로 치우친 분포 형태를 보이고 있다. 마찬가지로 히스토그램만으로는 자세히 보기 힘들어 log 변환을 실시해 보았다.



<그림 5.7> log(카드거래 건수) 히스토그램



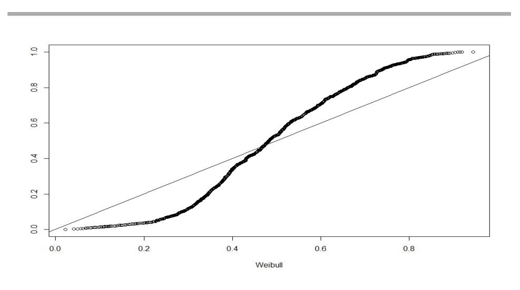
<그림 5.8> log(카드거래 금액) 히스토그램

log 변환을 한 4가지 히스토그램을 살펴보게 되면 대부분이 정규분포 형태를 보이고 있다. 거주 지역 카드거래 건수는 왼쪽 부분이 경사가 더 가파른 것을 확인 할 수 있고, 상대적으로 타 지역 카드거래 건수는 오른쪽 부분이 약간 더 가파른 형태를 보이고 있다. 거주 지역 카드거래 금액과 타 지역 카드거래 금액은 좌우 차이가 있긴 하지만 정규분포 형태를 잘 보이고 있다.

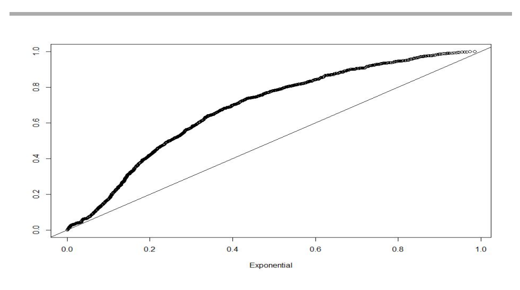
5.3 분석 결과

5.3.1 보험회사 자료 분석 결과

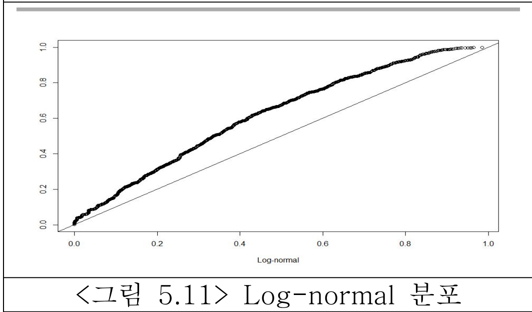
보험 처리기간과 보험금액에 관한 상관성을 알아보기 위하여 Copula 함수를 이용하여 분석을 실시하였다. 우선 보험 처리기간 자료와 보험금액 자료에 대한 분포를 추정하기 위하여 모의실험을 통한 방법으로 분포를 추정하였다. 4가지 Weibull, Exponential, Log-normal, Pareto분포 중에 한 가지 분포를 따른다고 가정하여 각각에 대하여 어느 분포와 가장 유사한지 QQ-plot을 그려보았다.



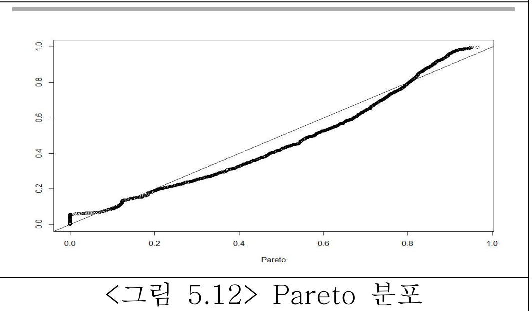
<그림 5.9> Weibull 분포



<그림 5.10> Exponential 분포



<그림 5.11> Log-normal 분포



<그림 5.12> Pareto 분포

QQ-plot의 결과를 보게 되면 Weibull, Exponential, Log-normal 분포에 비하여 Exponential 분포만 멀리 떨어져 있으며 나머지 3개의 분포의 QQ-plot은 비슷한 형태를 보이고 있다. 육안으로는 QQ-plot만으로 구분하기 힘들어 적합도 검정인 Kolmogorov-Smirnov Test를 실시하여 좀 더 유사한 형태의 분포를 추정하였다.

<표 5.1> 보험처리 기간 Kolmogorov-Smirnov Test 결과

분 포	Kolmogorov-Smirnov statistic
Weibull	0.1662
Exponential	0.2962
Log-normal	0.0586
Pareto	0.2843

<표 5.2> 보험금액 Kolmogorov-Smirnov Test 결과

분 포	Kolmogorov-Smirnov statistic
Weibull	0.2958
Exponential	0.1585
Log-normal	0.075
Pareto	0.29

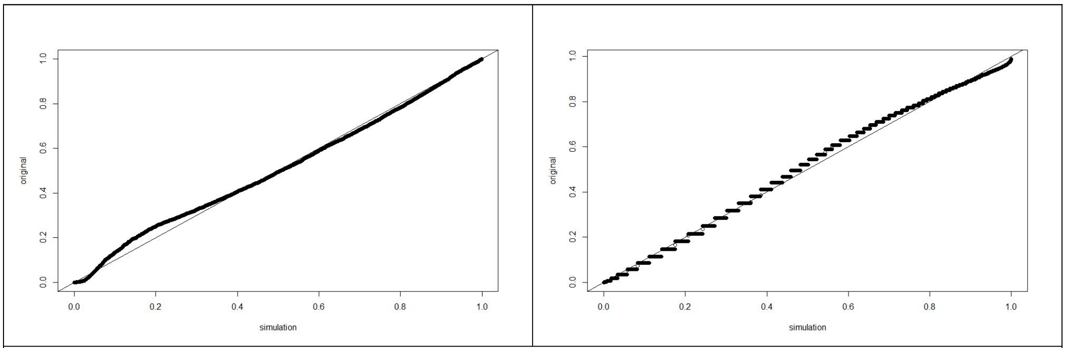
Kolmogorov-Smirnov Test 결과를 보면 통계량의 수치가 적을수록 그 분포와의 형태가 유사한 것으로 볼 수 있는데 보험처리 기간 자료와 보험금액 자료 둘 다 Weibull, Exponential, Pareto 분포에 비하여 Log-normal 분포와 더 유사한 결과로 나온 것을 볼 수 있다. QQ-plot에서는 Pareto 분포가 전체적으로 유사한 형태로 보였으나 실제로는 많이 어긋난 부분이 많아서 적합도 검정 통계량도 높은 수치로 나타났다.

보험처리 기간 자료와 보험금액 자료를 Log-normal 분포로 두고 모의실험을 통한 Copula 모수를 추정한 후 실제 자료와 비교하였다.

<표 5.3> Kendall's τ 에서 생성된 Copula의 모수

Copula	모수
Clayton Copula	1.679041

Kendall's τ 에서 생성된 모수를 이용하여 모의실험을 시행하였다. 원 자료와 비교해보기 쉽게 동일한 자료의 수로 시행하였다.



<그림 5.13> 원 자료와 모의실험 결과 QQ-plot

왼쪽의 QQ-plot은 보험금액에 대한 원 자료와 모의실험 결과의 QQ-plot이다. 전체적으로 큰 오차 없이 거의 비슷한 형태를 보이고 있다고 할 수 있다. 오른쪽의 QQ-plot은 보험처리 기간에 대한 원 자료와 모의실험 결과의 QQ-plot으로 마찬가지로 크게 벗어난 곳 없이 원 자료와 모의실험 결과가 비슷하게 나온 것을 보여주고 있다. 수치상 확인을 위하여 분위수를 통해 알아보았다.

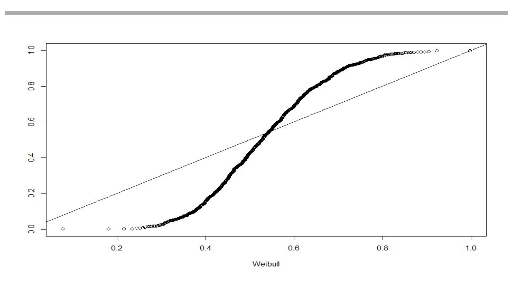
<표 5.4> 보험 자료의 원 자료와 모의실험 결과 분위수 비교

Quantile	보험처리 기간		보험 금액	
	원 자료	모의실험	원 자료	모의실험
1%	0.0067257	0.0106450	0.0011775	0.0096370
25%	0.2496038	0.2543484	0.2901758	0.2531438
50%	0.5435856	0.5033536	0.4945298	0.5005278
75%	0.7726014	0.7540846	0.7331648	0.7513830
99%	0.9674855	0.9903458	0.9885736	0.9901038

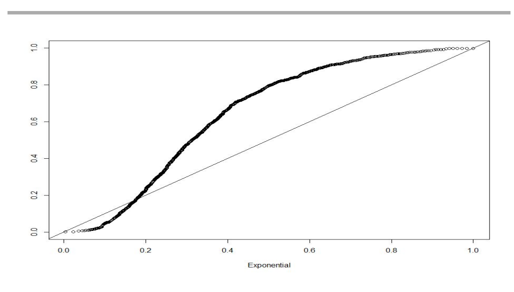
분위수 결과를 보게 되면 역시나 크게 차이나는 부분이 없이 비슷한 수치를 나타내고 있다.

5.3.2 카드회사 자료 분석 결과

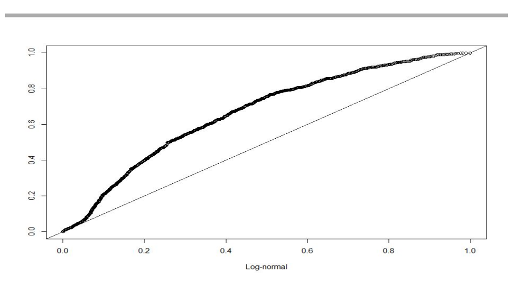
카드회사 자료는 65141개의 자료로 거주 지역 거래건수와 거래금액 그리고 타 지역 거래건수와 거래금액이 기록된 자료이다. 우선 거래건수와 거래금액간의 상관성을 보기 위하여 거주 지역과 타 지역의 자료를 합친 후 분석을 실시하였다. 마찬가지로 4가지 Weibull, Exponential, Log-normal, Pareto 분포로 각각의 거래건수와 거래금액의 분포를 추정하였다. 마찬가지로 QQ-plot을 이용하여 비교하였다.



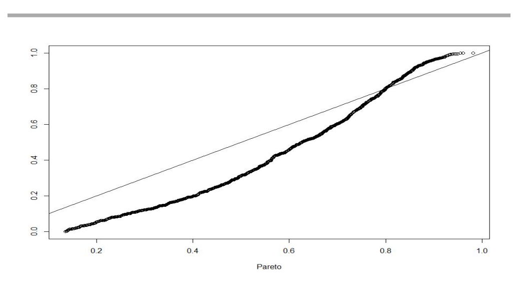
<그림 5.14> Weibull 분포



<그림 5.15> Exponential 분포



<그림 5.16> Log-normal 분포



<그림 5.17> Pareto 분포

위의 그래프를 보게 되면 Log-normal 분포와 Exponential 분포가 4가지 분포 중에서 다른 2가지 분포보다 상대적으로 더 유사한 형태를 보이고 있다. 얼핏 보기에는 Pareto 분포가 가장 가까운 것처럼 보이지만 실제 시작부분은 멀리 떨어져 있으며 다른 분포에 비해 유사하지 않은 형태를 보이고 있다. Exponential 분포 보다는 Log-normal이 좀 더 유사한 형태인지 확인하기 위하여 적합도 검정인 Kolmogorov-Smirnov Test를 시행하였다.

<표 5.5> 거래 건수 Kolmogorov-Smirnov Test 결과

분 포	Kolmogorov-Smirnov statistic
Weibull	0.5620
Exponential	0.2771
Log-normal	0.0707
Pareto	0.2321

<표 5.6> 거래 금액 Kolmogorov-Smirnov Test 결과

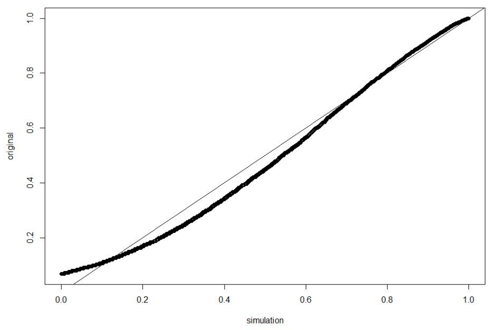
분 포	Kolmogorov-Smirnov statistic
Weibull	0.3833
Exponential	0.0951
Log-normal	0.0215
Pareto	0.619

Kolmogorov-Smirnov Test 통계량을 보면 역시나 Log-normal 분포가 다른 분포에 비하여 수치가 적은 것을 확인 할 수 있다. 거래 건수와 거래 금액 모두 Exponential 보다 Log-normal이 더 유사한 형태를 보인다. 거래 건수와 거래 금액을 Log-normal 분포로 가정하여 Kendall's τ에서 생성된 모수로 모의실험을 하여 원 자료와 얼마나 유사한지 알아보겠다.

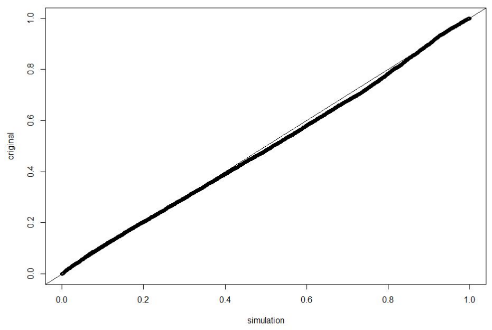
<표 5.7> Kendall's τ 에서 생성된 Copula의 모수

Copula	모수
Clayton Copula	1.118153

Kendall's τ 에서 생성된 Copula의 모수 값을 이용하여 원 자료의 같은 개수로 모의실험을 시행한 후 비교하였다.



<그림 5.18> 거래 건수 QQ-plot



<그림 5.19> 거래 금액 QQ-plot

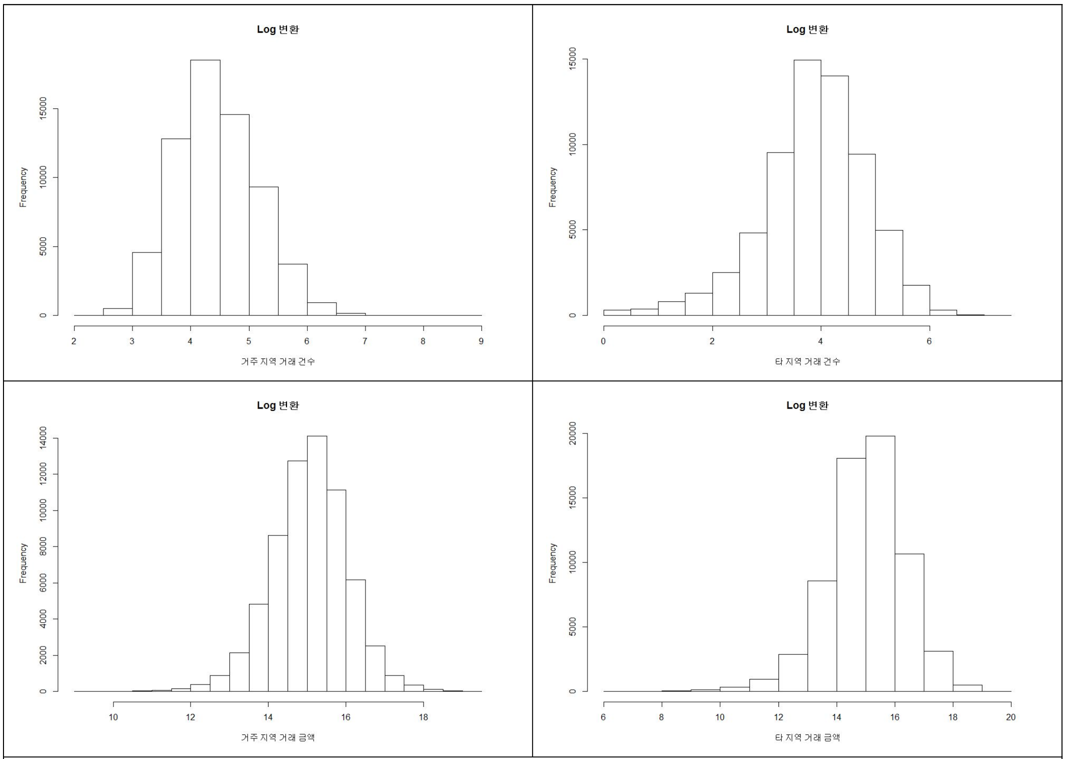
원 자료와 모의실험을 이용한 거래 건수와 거래 금액 QQ-plot을 보게 되면 두 자료가 거의 비슷한 형태를 보이고 있다. Copula 함수를 이용하여 원 자료와 모의실험 결과가 비슷하다는 것은 거래 건수와 거래 금액간의 결합함수를 잘 구한 것이라 할 수 있으며 Copula 함수의 값들이 거래 건수와 거래 금액의 상관성을 잘 나타낸다고 할 수 있다. 수치상으로 확인하기 위하여 분위수에 따른 값들을 비교하였다.

<표 5.8> 원 자료와 모의실험 결과 분위수 비교

Quantile	거래 건수		거래 금액	
	원 자료	모의실험	원 자료	모의실험
1%	0.0711224	0.0085938	0.0112815	0.0085557
25%	0.2038041	0.2545938	0.2516567	0.2525938
50%	0.4500901	0.4809077	0.4918400	0.5191397
75%	0.7511720	0.7452415	0.7389255	0.7602445
99%	0.9948271	0.9901630	0.9932364	0.9892150

분위수 수치를 비교해 보면 1%에서만 약간의 차이가 있고 대체적으로 거의 근사한 수치를 보인다.

다음은 Copula를 좀 더 많이 이용하여 자료들 간의 상관성을 보는 분석을 실시하였다. 이번 분석에서는 우선 지역별로 나누어 거주 지역 거래 건수와 거주 지역 거래금액간의 상관성을 분석하고 타 지역 거래 건수와 타 지역 거래금액의 상관성을 분석 한 후 마지막으로 두 상관성 분석 수치로 분석을 실시하였다. 거래 건수와 거래금액이 Log-normal 분포로 가정하고 분석하였기 때문에 거주 지역 거래 건수와 거래금액, 타 지역 거래 건수와 거래금액 역시 Log-normal 분포 형태를 가지는지 분석하였다.



<그림 5.20> Log 변환 히스토그램

거주 지역 거래 건수와 거래금액, 타 지역 거래 건수와 거래금액 모두 Log-normal 분포 형태의 모습을 보인다. Kolmogorov-Smirnov Test를 통하여 좀 더 유사한 형태를 보이는지 분석하였다.

<표 5.9> Kolmogorov-Smirnov Test 결과

분 포	Kolmogorov-Smirnov statistic
거주 지역 거래 건수	0.0419
거주 지역 거래 금액	0.0570
타 지역 거래 건수	0.0231
타 지역 거래 금액	0.0340

Kolmogorov-Smirnov Test 결과 역시 통계량이 아주 작은 값으로 나오는 것을 볼 수 있다. 이전에 실시한 분석과 마찬가지로 4가지 거주 지역 거래 건수와 거래 금액, 타 지역 거래 건수와 거래 금액은 Log-normal을 따르는 분포로 가정하고 Copula 함수를 이용하여 상관성을 알아보았다. Copula 함수를 이용하여 상관성을 알아보기 위하여 Kendall's τ에서 생성된 모수를 구하였다.

<표 5.10> Kendall's τ 에서 생성된 Copula의 모수

Copula	모수
거주 지역과 타 지역의 거래 건수	0.4421815
거주 지역과 타 지역의 거래 금액	0.4765997

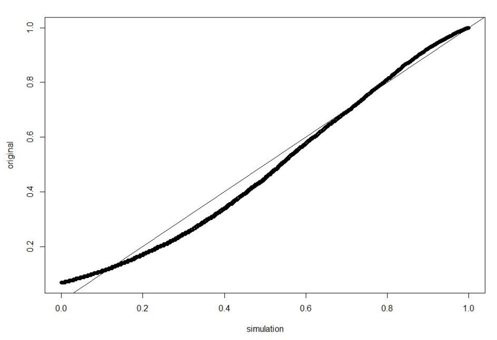
거주 지역 거래 건수와 거래 금액에서 생성된 모수와 타 지역 거래 건수와 거래 금액에서 생성된 모수를 이용하여 각각의 Copula 함수를 이용하여 상관성을 분석하였다.

그 다음 이 모수들로부터 생성된 결합함수의 값들로 다시 한번 Copula 함수에 적용하여 상관성을 분석하기 위해 두 자료를 이용하여 Kendall's τ에서 생성되는 모수를 구하였다.

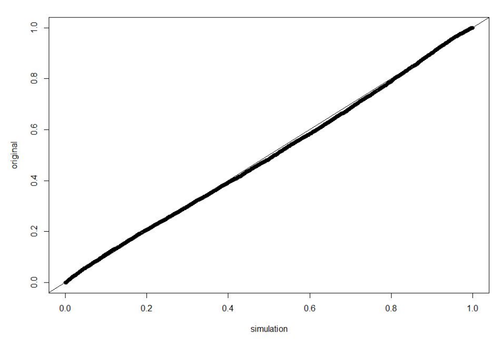
<표 5.11> Kendall's τ 에서 생성된 Copula의 모수

Copula	모수
전체 거래 건수와 거래 금액	1.510882

거주 지역과 타 지역의 거래 건수와 거주 지역과 타 지역의 거래 금액 간의 상관성을 보기 위하여 3번의 Copula 함수를 이용하였다. 3번의 과정을 거치면서 Kendall's τ에서 생성된 모수를 이용하여 모의실험을 시행하였다. 모의실험 결과와 원 자료를 비교하였다.



<그림 5.21> 거래 건수 QQ-plot



<그림 5.22> 거래 금액 QQ-plot

거래 건수의 QQ-plot을 보게 되면 약간의 차이는 있지만 크게 벗어나는 곳이 없어서 유사한 형태로 보이며, 거래 금액의 QQ-plot은 거의 일치하는 형태를 보이고 있어서 Copula 함수를 이용한 상관성 분석이 잘 되었다고 할 수 있다. 분위수를 이용하여 수치상으로 확인을 실시하였다.

<표 5.12> 원 자료와 모의실험 결과 분위수 비교

Quantile	거래 건수		거래 금액	
	원 자료	모의실험	원 자료	모의실험
1%	0.0711224	0.0094266	0.0112815	0.0088158
25%	0.2038041	0.2518094	0.2516567	0.2490169
50%	0.4500901	0.4795450	0.4918400	0.5057381
75%	0.7511720	0.7491048	0.7389255	0.7515942
99%	0.9948271	0.9898170	0.9932364	0.9900621

마지막으로 원 자료와 Copula 함수를 1번 이용했을 경우와 3번 이용했을 경우를 비교하기 위하여 분위수 비교를 실시하였다. 간편하게 Copula 함수를 1번 이용한 경우는 모의실험 1이라 하고 Copula 함수를 3번 이용한 경우는 모의실험 2로 표기하였다. 오차는 절대값으로 표기하였다.

<표 5.13> 원 자료와 모의실험 1,2 결과 분위수 비교

Quantile	거래 건수				
	원 자료	모의실험 1	오 차	모의실험 2	오 차
1%	0.0711224	0.0085938	0.0625286	0.0094266	0.0616958
25%	0.2038041	0.2518051	0.0480010	0.2518094	0.0480053
50%	0.4500901	0.4809077	0.0308176	0.4795450	0.0294549
75%	0.7511720	0.7452415	0.0059305	0.7491048	0.0020672
99%	0.9948271	0.9901630	0.0046641	0.9898170	0.0050101
Quantile	거래 금액				
	원 자료	모의실험 1	오 차	모의실험 2	오 차
1%	0.0112815	0.0085557	0.0027258	0.0088158	0.0024657
25%	0.2516567	0.2545938	0.0029371	0.2490169	0.0026398
50%	0.4918400	0.5191397	0.0272997	0.5057381	0.0138981
75%	0.7389255	0.7602445	0.0213190	0.7515942	0.0126687
99%	0.9932364	0.9892150	0.0040214	0.9900621	0.0031743

거래 건수와 거래 금액 모두 원 자료와 모의실험 1과 모의실험 2의 오차는 큰 차이가 없었다. 거래 건수는 자료의 수치가 작기 때문에 큰 차이가 보이지 않지만 거래 금액에서는 모의실험 2가 모의실험 1보다 미세하게나마 오차가 더 작은 것을 볼 수 있다.

제 6 장 결 론

본 논문에서는 보험회사 자료와 카드회사 자료를 이용하여 각 자료 내에 있는 요인들의 분포를 추정하며 요인들 간의 상관성을 분석하였다. 금융권과 관련된 자료들은 크게 fat tail의 형태를 가지고 있는 Weibull, Exponential, Log-normal, Pareto 분포를 따른다는 가정 하에 분포를 유추하여 Copula 함수를 이용해 상관성을 알아보았다.

상관계수만으로 요인끼리의 상관성을 분석하기에는 많은 한계들이 있기 때문에 Copula 함수를 이용하여 결합분포함수를 추정하는 것이 더 효과적이다. 또한, 결합분포함수를 추정하는 과정에서 모수를 이용하여 모의실험을 시행하여 결과를 비교해 봤을 때 좀 더 세부적으로 분석되었다고 판단되어질 수 있다.

fat tail의 형태를 가지는 분포들이 Weibull, Exponential, Log-normal, Pareto의 형태만을 가지고 있지 않기 때문에 다른 fat tail의 형태도 본 논문에서 제시한 모의실험 방법으로 좀 더 다양한 분포를 유추하여 시행한다면 좀 더 정확한 분포를 유추 할 수 있을 것이라고 판단되어진다.

실증분석에서 비슷한 성격의 요인들을 일반적으로 합쳐서 Copula 함수를 이용하여 상관성을 구하는 것은 일반적인 상관계수로 상관성을 분석하는 것보다는 좀 더 정확하게 알 수 있는 효과적인 방법이다. 하지만 본 논문에서 실증분석을 실시하며 제안한 방법은 비슷한 성격의 요인들끼리도 Copula 함수를 이용하여 결합분포함수를 추정하고 그 결합분포함수들을 이용하여 Copula 함수를 한 번 더 이용하게 된다면 상관성을 좀 더 명확하게 분석하는 방법이라는 것을 잘 보여주고 있다.

참 고 문 헌

[1] Roger B.Nelsen. (1999). An Introduction to Copula. Springer.

[2] 김명직, 신성환. (2003). Copula 함수의 추정과 시뮬레이션. 한국선물학회/선물연구 제 11권 제 2호.

[3] 최진석. (2010). Copula를 이용한 대물배상과 대인배상의 결합분포 함수 추정. 성균관대학교 일반대학원 석사학위논문.